Endogenous Attention and the Spread of False News[*]

Tuval Danenberg[†] and Drew Fudenberg[‡]

June 3, 2024

Abstract

We study the impact of endogenous attention in a dynamic model of social media sharing. Each period, a distinct user randomly draws a story from the pool of stories on the platform and decides whether or not to share it. Users want to share stories that are true and interesting, but differentiating true stories from false ones requires attention. Before deciding whether to share a story, users choose their level of attention based on how interesting the story is and the platform's current proportions of true and false stories. We characterize the limit behavior of the share of true stories using stochastic approximation techniques. For some parameter specifications, the system has a unique limit. For others, the limit is random—starting from the same initial conditions, the platform may end up with very different proportions of true and false stories and different user sharing behavior. We present various comparative statics for the limit. Endogenous attention leads to a counterbalancing force to changes in the credibility of false stories but can intensify the effects of changes in false stories' production rate.

[*] We are grateful to Michel Benaim, Krishna Dasaratha, Kevin He, David Rand and Alexander Wolitzky for helpful comments and conversations.

[†] Department of Economics, MIT, tuvaldan@mit.edu

[‡] Department of Economics, MIT, drew.fudenberg@gmail.com

1 Introduction

Misinformation on social media has become an issue of growing public concern so it is important to better understand the mechanisms by which it spreads. Vosoughi, Roy, and Aral (2018) shows that falsehoods on social media spread farther, faster and deeper than truth, and that this is mostly due to humans rather than bots. This underscores the need to understand users' sharing decisions and specifically their motivations for sharing false news. Various explanations have been proposed, including politically motivated reasoning and ideological alignment (e.g., Van Bavel and Pereira (2018), Allcott and Gentzkow (2017)) and digital illiteracy (e.g., Guess et al. (2020)). Recently, Pennycook, Epstein, et al. (2021) have suggested an inattention-based account, according to which users care about sharing accurate content but nevertheless share false news because the social media context focuses their attention on factors other than accuracy.

This paper starts from the premise that social media users want to share accurate content and that more attentive users are better at filtering false content. We combine this with three basic ideas: (i) users have some control over their attention levels; (ii) attention is costly; and (iii) users' choice of attention level depends on their beliefs regarding the credibility of false content and the relative share of false content. If a social media user believes that the relative share of false stories in their feed is negligible, they have little incentive to invest attention towards spotting false stories. Additionally, if false stories are blatantly false, users will not need to pay much attention to spot them. However, if the share of false stories is significant and false stories seem plausible, users are more likely to mistake them for true stories if they are inattentive, so they might be willing to pay a significant cost of attention to distinguish between true and false content.

We incorporate these ideas into an infinite-horizon dynamic model of social media consumption. In every period, a distinct user randomly draws a story out of the current set of stories on a social media platform and decides whether or not to share it. Users consider two factors when evaluating a story: its *veracity*, or truthfulness, and its *evocativeness*, or how interesting and stimulating it is. We ignore boring stories, which we assume users do not share, and consider two levels of evocativeness: mildly interesting (M) and very interesting (I). Each story's veracity is fixed throughout time; conditional on veracity, evocativeness is drawn i.i.d. for each user, capturing the idea the different users will find different stories very interesting. We also assume that false stories are more likely to be very interesting.

Before drawing the story, the user chooses their attention level and pays the cost of attention. Upon drawing the story, they receive a binary signal regarding the story's veracity. False stories are characterized by a credibility measure that captures how true they appear—when false stories are highly credible, signals about their veracity are less precise. On the other hand, the precision of the signal is increasing in the user's chosen attention level. We further assume that the signal's precision is supermodular in credibility and attention so that users' attention is increasing in credibility.[1] If the user decides to share the story, a fixed number of identical copies of the story are added to the platform. Regardless of the sharing decision, fixed numbers of true and false stories are exogenously added to the platform, which corresponds to original content creation.

Our main object of interest is the share of true stories in the system for each period $n \in \mathbb{N}$, which we denote by y_n. Users' optimal behavior depends on the value of y_n, and follows one of three sharing rules. When y_n is sufficiently high, the system is in the *sharing* region, where users share all stories for which they receive the signal suggesting the story is true. When y_n is low, the system is in the *no sharing* region, where users do not share any stories and do not pay attention. In between, there is an intermediate region, where users share either only mildly interesting stories or only very interesting stories, depending on the model parameters.

We analyze the evolution of y_n and its long run behavior. We find that y_n converges almost surely and provide a complete characterization of its limit. For some parameter values the limit is unique. For others it is random, so that starting from the same initial conditions the platform may end up with significantly different limit shares of true stories and different user behavior in the limit. This effect is most pronounced when the platform is new and the total number of stories is small, but it is still present in any finite-sized platform.

If users were to always follow one of the three sharing rules mentioned above, our system would be a generalized Polya urn model (henceforth GPU) and limit behavior could be analyzed using existing results. This is not the case, as users follow the sharing rule that maximizes their payoffs and the optimal sharing rule depends on the current share of true stories. Hence, our system is a concatenation of a finite

[1] We assume a specific functional form for the signal function for now but hope to generalize it.

number of GPUs. To analyze its limit, we extend results on stochastic approximation for GPUs to cover such concatenations.

We find that each of the GPUs corresponding to the different sharing rules has a unique limit share of true stories, which we call a *quasi steady state*. A quasi steady state is a limit point for our system, i.e., a point to which y_n converges with positive probability, if and only if it is within the region where its associated sharing rule is optimal. The only other possible limit points for our system are the *thresholds* between regions—points where users are indifferent between two sharing rules.

After characterizing the set of limit points, we consider comparative statics of the limit points with respect to the model parameters. For the quasi steady states, the share of true stories is decreasing in false story credibility for low credibility levels, but an opposite effect may arise when credibility is high. The intuition is that while false stories of high credibility are harder to identify, users also pay more attention to them. When credibility is high, user responses to an increase in credibility may more than compensate for the direct effect of this increase, thereby leading to an increase in the limit share of true stories. The comparative statics imply that producers of false stories may choose low credibility levels even when credibility is free. They also imply that platforms who aim to counter the spread of false news by fact-checking false stories might be better off not fact-checking at all than fact checking only a small share of stories. This is because increasing the share of stories flagged as false leads users to put more trust in stories that were not flagged.

We find that the limit share of true stories in the quasi steady states may be either increasing or decreasing in a measure of the *reach* on the platform—the number of friends who will see a shared story—and in the probability that false stories are very interesting. We also find that when the production of rate of false stories is sufficiently high, the system has a unique limit in which users do not share any stories, while when this production rate is sufficiently low the system has a unique limit in which users share all stories for which they receive the signal suggesting the story is true. This implies that when moving from high to low false story production rates, users' reactions will further increase the limit share of true stories. Thus, while user responses lead to a counterbalancing force to changes in the credibility of false stories, they may intensify the effect of changes in false stories' production rate.

When the system converges to a point where users are indifferent between two sharing rules, the comparative statics can be different than for the limit points where the users strictly prefer one rule.[2] For example, the limit share of true stories may be increasing in the cost of attention, because the cost of attention enters negatively into users' payoffs while the share of true stories enters positively. So when the cost of attention increases, the share of true stories required for indifference increases as well. In contrast, increasing the cost of attention lowers the share of true stories at the other limit points.

2 Related Literature

Empirical Evidence A basic assumption of our model is that users care about the accuracy of the stories they share. Pennycook, Epstein, et al. (2021) reports a representative survey of Americans who rated accuracy as the most important factor affecting their sharing decisions, over other factors like humor, interest, and political alignment. Chen, Pennycook, and Rand (2023) conducts a factor analysis of the content dimensions affecting sharing decisions in a series of experiments, and finds that the main factors are perceived accuracy, evocativeness, and familiarity. The evocativeness factor captures characteristics such as the extent to which content is surprising, amusing, or provokes anxiety and other negative feelings. The association between these characteristics and sharing intentions is supported by Berger and Milkman (2012), and motivates our assumption that evocativeness influences sharing decisions.[3]

Chen, Pennycook, and Rand (2023) finds that all three content dimensions are significantly positively correlated with sharing intentions. In line with Pennycook, Epstein, et al. (2021), the accuracy focused factor has the highest coefficients in a regression of sharing intentions on the factors, while the evocativeness factor had the lowest coefficients. Consistent with this, we assume that users will not share stories that they know are false even if they are very interesting. Chen, Pennycook, and Rand (2023) also finds that users ratings on the evocativeness dimension are negatively correlated with stories' objective veracity. This supports our assumption that false stories are more likely to be very interesting.

[2] This is analogous to the difference in comparative statics between pure-strategy and mixed-strategy Nash equilibrium in games.

[3] Our model does not track the number of times an individual story has been shared, so it does not capture the positive effect of familiarity found in e.g., Chen, Pennycook, and Rand (2023), or the related "illusory truth" effect in Pennycook, Cannon, and Rand (2018).

In our model, inattention plays a central role in the sharing of false content. Pennycook, Epstein, et al. (2021) claims that inattention to veracity is one of the key mechanisms leading users to share false political stories, and that shifting users' attention to accuracy significantly increases the accuracy of the content they share. Pennycook, McPhetres, Zhang, Lu, and Rand (2020) finds similar results in the context of information about COVID-19.

Theory of Online Misinformation Our paper contributes to the theoretical literature on online misinformation. Papanastasiou (2020) and Acemoglu, Ozdaglar, and Siderius (2023) analyze sequential models of the spread of a single story across a network. In Papanastasiou (2020), users only care about veracity. They are arranged in a line; if a user decides not to share the story it "dies" and otherwise it is shared with the next user in line. Before sharing, users can inspect the story at a cost, and inspection fully reveals the story's veracity. In Acemoglu, Ozdaglar, and Siderius (2023), the network structure is determined by a social media platform that aims to maximize engagement. Users care about veracity, but also want to share a story that will be liked by many subsequent users and do not want to share a story that will be disliked. Unlike in our model, beliefs and sharing decisions do not depend on the actions of previous users. The paper finds that a regulator who cares about the accuracy of users' beliefs may be better off censoring less content than is technologically feasible. Like our finding that increasing the share of flagged stories may lead to a decrease in the limit share of true stories, this relates to the "implied truth effect" studied in Pennycook, Bear, Collins, and Rand (2020). However, while we find that no flagging can be better than poor flagging, in their framework some censorship is always better than none. Mostagir and Siderius (2022) also considers various policies to to curb misinformation, focusing on the difference in the responses of rational and naive users. Other papers that study misinformation by tracking the spread of a single story include Bloch, Demange, and Kranton (2018) and Hsu, Ajorlou, and Jadbabaie (2021). Merlino, Pin, and Tabasso (2023) studies the diffusion of one true message and one false message in a network. Kranton and McAdams (2024) analyzes the interaction between information suppliers' quality choices and social media users' sharing decisions in a three period model.

Dasaratha and He (2023), like our paper, uses stochastic approximation to determine the evolution of the shares of true and false stories rather than the spread of a single story. Users do not know the state of the platform, beyond what they observe in their feed. The paper focuses on the weight the platform places on stories' virality when choosing what stories to display to users, and does not feature endogenous attention.[4] In contrast, our paper focuses on the interaction between user's endogenous attention and platform evolution and develops a richer model of user's sharing decisions. In Dasaratha and He (2023) users only care about veracity, while in our model they also care about evocativeness, and trade off their ability to filter false content with their cost of attention.

Stochastic Approximation Our model is very close to a generalized Polya urn with two colors, where stories are "balls" and colors are veracity levels.[5] Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) use stochastic approximation arguments to show that for such urn models, when the number of balls grows sufficiently fast, the urn's long-run behavior can be determined by studying the attractors of a deterministic differential equation. For a fixed decision rule, our model is covered by the assumptions of these papers, but the overall system is not, because of the discontinuous changes at the boundaries between regions. To extend Schreiber (2001)'s results to our setting, we employ results from Benaim, Hofbauer, and Sorin (2005) (henceforth BHS), which generalizes the dynamical systems approach to stochastic approximation to the case where the differential equation is replaced by a differential inclusion.

3 Model

We consider an infinite horizon model of a social media platform. The platform contains stories with two characteristics (v, e). A story's *veracity* is $v \in \{T, F\}$, with the story being true if $t = T$ and false otherwise. A story's *evocativeness* is $e \in \{M, I\}$, with the story being mildly interesting if $e = M$ and very interesting if $e = I$. While

[4] In their model sharing increases the "popularity score" of a story and this popularity score affects the probability that a story appears in a user's feed. A similar interpretation can be applied to our model.

[5] In the Polya urn model, an urn consists of balls of various colors. In each period one ball is drawn randomly from the urn. The ball is then returned together with one additional ball of the same color. The generalized urn model allows for the number of balls added in each period to be random, with probabilities that depend on the state of the system. See, e.g., Schreiber (2001) and Mahmoud (2008).

a story's veracity is fixed (the story is either always true or always false), a story might be mildly interesting to one user and very interesting to another.[6] When a user draws a story, the probabilities of each evocativeness level are:

$$Pr(e = I|t = T) = \frac{1}{2}; Pr(e = I|t = F) = \delta.$$

We assume that $\delta > \frac{1}{2}$, so false stories are more likely to seem very interesting, as in Chen, Pennycook, and Rand (2023), and that $\delta < 1$ as otherwise mildly interesting stories are always true.

The false stories are of *credibility* $\theta \in (0, 1)$. The credibility of a false story determines how difficult it is to distinguish from a true story, in a manner that will be described below. To keep the model simple we assume that all false stories have the same credibility.

The platform begins operating at time $t = 0$ with an exogenous stock of true and false stories (T_0, F_0). In each subsequent period $n \in \mathbb{N}$, 1 true story and κ false stories are exogenously added to the platform, and T_n and F_n respectively denote the numbers of true and false stories on the platform at the beginning of period n.[7] The vector $z_n := (T_n, F_n)$ summarizes the current state of the platform; we use the notation $|z_n| := T_n + F_n$ for the total number of stories in period n, and let $y_n := \frac{T_n}{|z_n|}$ denote the share of true stories.

Each period, a distinct user randomly draws a story among those currently on the platform and decides whether or not to share it. Before making the sharing decision, the user sees the story's evocativeness level and a noisy signal of its veracity. The precision of this signal depends on the user's *attention* as will be explained below. The parameter ρ describes the *reach* of shared stories on the platform—if the user decides to share the story, ρ copies of the story are added to the platform.

In summary, each period the current user:

1. Draws a story, and observes its realized evocativeness.

2. Chooses an attention level $a \in [0, 1]$.

3. Draws a signal whose distribution depends on a.

[6] In reality there are also boring stories that are rarely or never shared, we omit these.

[7] The analysis would be the same in a continuous-time model where the time the next user arrives is a random variable. If multiple users could arrive simultaneously the analysis would be slightly different.

4. Decides whether to share the story.

5. Receives payoffs.

Finally, 1 new true story and κ new false stories are posted, and ρ copies of the current story are added if it was shared.

3.1 User's choice and payoffs

In each period n, the current user knows the current share of true stories y_n.[8] After drawing a story and observing its evocativeness level e, the user chooses a level of attention a, which will determine the precision of the signals they get regarding the story's veracity. The cost $c(a)$ of attention level a is $\beta \cdot a^2$,where $\beta > 0$. The signal is $s \in \{T', F'\}$, with probabilities given by

$$\mathbb{P}(T'|T) = 1; \mathbb{P}(T'|F) = \theta(1-a). \tag{1}$$

The idea behind Equation 1 is that a false story of credibility θ is *clearly false* with probability $1-\theta$, where a clearly false story is one that users will recognize as false even when they do not pay attention. With probability θ, users will notice the story is false only if they pay attention. That is, an attentive user can perfectly detect false stories. A user's attention level a is the probability with which they pay attention. Thus, when a user's attention level is a and the credibility of false stories is θ, they will identify a false story as false with probability $Pr(F'|F) = 1-\theta+\theta a = 1-\theta(1-a)$. Additionally, regardless of their attention level, if the story is true the user receives the signal T' with certainty. This allows us to identify the signal F' with the situation in which the user realizes the story is false and the signal T' with the situation in which the user is uncertain about the story's veracity.

Users' payoffs when they do not share the story are normalized to 0. The payoff to sharing a (v, e)-story, ignoring the cost of attention, is

$$\begin{aligned} u(v, M) &= \lambda\left(\mathbb{1}(v=T) - \mu\mathbb{1}(v=F)\right) \\ u(v, I) &= \lambda\left(\mathbb{1}(v=T) - \mu\mathbb{1}(v=F)\right) + (1-\lambda) \end{aligned}$$

[8] This approximates the situation where users have seen a number of recent stories and the mix between true and false stories is not changing too quickly.

Hence, as in the Chen, Pennycook, and Rand (2023) experiments, users want to share stories that are true and very interesting. The parameter $\lambda > 0$ is the weight on veracity and $1 - \lambda$ is the weight on evocativeness, i.e., the weight on stories being very interesting. The parameter $\mu > 0$ captures the loss to sharing a false story relative to the gain from sharing a true story, which is normalized to 1.

We make two parametric assumptions:

Assumption 1. $\mu > \frac{1-\lambda}{\lambda}$.

Assumption 2. $\mu\theta < 2\beta$.

Assumption 1 implies users will not share very interesting stories they know are false, and therefore will not share any stories for which they received the signal F'.[9] It remains to analyze, for each evocativeness level, when they will share stories with signal T', which we do in the beginning of the next section. Assumption 2 implies that users attention levels conditional on sharing stories with signal T' are always given by solutions to first order conditions as in Lemma 1 below.

In summary, the model parameters are $(\rho, \kappa, \theta, \mu, \beta, \delta, \lambda)$. We assume throughout that all parameters are strictly positive, satisfy Assumptions 1 and 2, and that $\theta, \lambda < 1$ and $\delta \in (\frac{1}{2}, 1)$.

4 Analysis

We are interested in characterizing the composition of stories on the platform over time, i.e, analyzing the stochastic process $\{z_n\}$, and in particular the share of true stories $\{y_n\}$. To begin, we solve for the user-optimal attention level as a function of the current state.

4.1 Optimal Attention and the Sharing Decision

Users choose their attention level after seeing the story's evocativeness e. They will never share stories for which they received the signal F', so they either share stories with signal T' or do not share at all. Thus their payoff to a story of evocativeness e

[9] This assumption, which bounds λ from below, is consistent with the finding in Chen, Pennycook, and Rand (2023) that the content factor with the strongest positive correlation with users' sharing intentions is perceived accuracy.

is the maximum of 0 and their expected payoff from sharing a story with signal T', which is

$$U(a, y, e) := \mathbb{P}_{a,y}(T'|e)\mathbb{E}[u(v, e)|T', e] - c(a). \tag{2}$$

Let $a(y, e) := \operatorname{argmax}_a U(a, y, e)$ denote the optimal attention level conditional on sharing stories with signal T'.

Lemma 1. *The functions $U(a, y, M)$ and $U(a, y, I)$ are strictly concave, and the optimal attention levels (conditional on sharing T' stories) are*[10]

$$\begin{cases} 0 \leqslant a(y, M) = \dfrac{\lambda\mu(1-y)(1-\delta)\theta}{\beta(y + 2(1-y)(1-\delta))} \leqslant 1, \\[2ex] 0 \leqslant a(y, I) = \dfrac{(1-y)\delta\theta(\lambda\mu - (1-\lambda))}{\beta(y + 2(1-y)\delta)} \leqslant 1. \end{cases}$$

The proof of this and all other results stated in the text are in Appendix A. Intuitively, when $y = 1$ there is no need to pay attention, so $a(1, I) = a(1, I) = 0$. As y decreases the marginal gain from paying attention increases, and since the U's are strictly concave, $da/dy < 0$. However, when y is close enough to 0 the payoff from the $a(y, e)$ is so low that users prefer not to pay any attention at all.[11]

As we show in Appendix D, both of the conditionally optimal attention levels are decreasing in β and increasing in θ, μ, λ, and $a(y, I)$ is increasing in δ while $a(y, M)$ is decreasing in δ. That is, users pay more attention when false stories are highly credible, when the cost to sharing false stories is high, and when the weight on veracity is high, and pay less attention when the share of true stories is high and when the cost of attention is high. Users pay more attention to the veracity of very interesting stories (and less attention to mildly interesting stories) when false stories are more likely to be very interesting. These observations will be relevant for our discussion of comparative statics in Section 4.3.

The next lemma shows that there are interior thresholds $\hat{y}_M, \hat{y}_I$ for each evocativeness level such that if the share of true stories is below the corresponding threshold then users choose $a = 0$ and do not share the story, and if the share is above this

[10] In practice both of these attention levels are always strictly between 0 and 1. It straightforward to verify that $a(y, e) < 1$ for all y and $a(y, e) > 0$ if $y < 1$, and that the system can never reach a state where $y = 1$.

[11] We allow users to randomize when indifferent between $a = 0$ and $a = a(y, e)$.

threshold users choose the attention level given in Lemma 1 and share if and only if they received the signal T'.

Lemma 2. *Let $V(y,e) := U(a(y,e),y,e)$. $V(y,M)$ and $V(y,I)$ are strictly increasing in y, and there are (unique) $\hat{y}_M, \hat{y}_I \in (0,1)$ s.t $V(\hat{y}_M, M) = V(\hat{y}_I, I) = 0$.*

4.2 Dynamics

When $0 < y_n < \min\{\hat{y}_M, \hat{y}_I\}$, we say that the system is in the "no sharing" region (N), where users do not share any stories and do not pay attention. When $\max\{\hat{y}_M, \hat{y}_I\} < y_n < 1$, the system is in the "sharing" region (S) where users share all stories with signal T' and choose attention levels given by (1). When $\min\{\hat{y}_M, \hat{y}_I\} < y_n < \max\{\hat{y}_M, \hat{y}_I\}$, the system is in an intermediate region where users share only one type of story. The intermediate region is a "sharing very interesting" region (I), where users share T' stories only if they are very interesting, if $\hat{y}_I < \hat{y}_M$. It is a "sharing mildly interesting" region (M) if the sign is reversed. Thus, the system always has three regions: the extreme regions N to the left and S to the right, and an intermediate region which is either I or M depending on the ordering of $\hat{y}_I$ and $\hat{y}_M$. Numerical computations show that both $\hat{y}_M < \hat{y}_I$ and $\hat{y}_M > \hat{y}_I$ are possible so the intermediate region can be either of the two.

Let $p_R^T(y), p_R^F(y)$ be the probabilities that the agent shares a true or false story, respectively, when the current share of true stories is y under the sharing rule of region $R \in \{N, I, M, S\}$. These are given by,

$$p_R^T(y),\, p_R^F(y) = \begin{cases} y,\ (1-y)\theta\left(1-\delta a(y,I)-(1-\delta)a(y,M)\right), & R=S \\ \frac{y}{2},\ (1-y)\delta\theta\left(1-a(y,I)\right), & R=I \\ \frac{y}{2},\ (1-y)(1-\delta)\theta\left(1-a(y,M)\right), & R=M \\ 0,\ 0, & R=N \end{cases} \tag{3}$$

For example, $p_I^F(y) = (1-y)\delta\theta\left(1-a(y,I)\right)$ because in region I users share a false story if and only if all of the following occur: They drew a false story, the story is very interesting, and they observed the signal T'.

For each R, define the Markov process $\{z_{n;R}\}$ by:

$$z_{n+1;R} = z_{n;R} + \begin{cases} \begin{pmatrix} 1+\rho \\ \kappa \end{pmatrix}, & \text{with probability} \quad p_R^T(y_n) \\ \begin{pmatrix} 1 \\ \kappa+\rho \end{pmatrix}, & \text{with probability} \quad p_R^F(y_n) \\ \begin{pmatrix} 1 \\ \kappa \end{pmatrix}, & \text{w.p} \quad 1 - p_R^T(y_n) - p_R^F(y_n). \end{cases} \tag{4}$$

These Markov processes describe the evolution of counterfactual platforms where users follow the sharing rule of region R regardless of the current share of true stories. As shown in Appendix B.3, they are *generalized Polya urns* (GPUs), which lets us apply results from Schreiber (2001) and Benaim, Schreiber, and Tarres (2004).

The law of motion for y_n in region R, implied by the law of motion for $z_{n;R}$ in (4), is

$$y_{n+1} - y_n = \begin{cases} \dfrac{(1-y_n)(1+\rho)-\kappa}{|z_n|+1+\kappa+\rho}, & \text{with probability} \quad p_R^T(y_n) \\ \dfrac{(1-y_n)-(\kappa+\rho)}{|z_n|+1+\kappa+\rho}, & \text{with probability} \quad p_R^F(y_n) \\ \dfrac{(1-y_n)-\kappa}{|z_n|+1+\kappa}, & \text{w.p} \quad 1 - p_R^T(y_n) - p_R^F(y_n). \end{cases} \tag{5}$$

4.2.1 Limit differential inclusion

We will use tools from stochastic approximation theory to approximate the behavior of the discrete stochastic system $\{y_n\}_{n\geqslant 0}$ by a continuous and deterministic system. If our system was a single GPU, we could apply results in Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) to relate its limit behavior to that of an appropriately chosen *limit differential equation.* Since our system is a concatenation of the GPUs $\{z_{n;R}\}$, one for each region, we instead relate its limit behavior to that of a differential inclusion, an equation of the form $\frac{dx}{dt} \in F(x)$ for a set valued function F. We construct this inclusion, which we will refer to as the *limit differential inclusion* or LDI, by pasting together the limit ODEs associated with the GPUs $\{z_{n;R}\}$. In our model these ODEs are[12]

$$g_R(y) = 1 + p_R^T(y)\rho - y(1 + \kappa + \rho(p_R^T(y) + p_R^F(y)) \tag{6}$$

For an intuition for the limit ODEs, note that for the process $z_{n;R}$, the expected number of incoming true stories in the next period is $1+p_R^T(y)\rho$, and the total expected number of incoming stories in the next period is $1 + \kappa + \rho\left(p_R^T(y) + p_R^F(y)\right)$. So,

$$\begin{aligned} g_R(y) = \; & \mathbb{E}_R[\#\text{incoming true stories in period n+1}|y_n = y] \\ & - y\mathbb{E}_R[\#\text{total incoming stories in period n+1}|y_n = y]. \end{aligned}$$

Thus, according to the limit ODE $\frac{dy}{dt} = g_R(y)$, the share of true stories increases if and only if $\frac{\mathbb{E}_R[\#\text{incoming true stories in period n+1}|y_n=y]}{\mathbb{E}_R\#\text{total incoming stories in period n+1}|y_n=y]} > y$, i.e., if and only if the ratio of expected incoming true stories to total expected incoming stories is greater than the current share of true stories.

We now define the LDI as,

$$\frac{dy}{dt} \in F(y), \tag{7}$$

where in the interior of each region F takes the (singleton) value of the relevant limit ODE:

$$F(y) = \{g_R(y)\} \quad \text{for} \quad y \in R,$$

and at the thresholds, F takes on all values in the interval between the limit ODEs. If $\hat{y}$ is the threshold between regions R and W then:

$$F(\hat{y}) = [\min\{g_R(\hat{y}), g_W(\hat{y})\}, \max\{g_R(\hat{y}), g_W(\hat{y})\}]$$

We say that a point $y^* \in (0,1)$ is a *steady state* for the LDI if $0 \in F(y)$. We say that y^* is a *stable steady state* for the LDI if it is a steady state and there exists $\epsilon > 0$ such that for all $y \in (y^* - \epsilon, y^* + \epsilon)$ we have $\text{sign}(x) = \text{sign}(y^* - y)$ for all $x \in F(y)$. A steady state is *unstable* if it is not stable. In our model any unstable steady state must also be repelling, i.e., there exists $\epsilon > 0$ s.t for all $y \in (y^* - \epsilon, y^* + \epsilon)$ we have $\text{sign}(x) = -\text{sign}(y^* - y)$ for all $x \in F(y)$.

[12] See Appendix B.3 for the derivation of this equation.

Our main result, Theorem 2 below, is that y_n converges almost surely to a stable steady state for the LDI and, except for a special case where the system begins in the no sharing region and never leaves, converges to any stable steady state with positive probability. Before stating this result, we characterize the set of stable steady states of the LDI, which we denote by $\mathcal{S}_F$.

As we show in Lemma 4 in Appendix A, the ODEs $\frac{dy}{dt} = g_S(y)$, $\frac{dy}{dt} = g_I(y)$, $\frac{dy}{dt} = g_M(y)$, and $\frac{dy}{dt} = g_N(y)$, defined over $[0, 1]$, each have a globally stable steady state. We denote the steady state of $\frac{dy}{dt} = g_R(y)$ by $g^*_R(y)$, and refer to these as *quasi steady states* of the system and to $\hat{y}_I, \hat{y}_M$ as *thresholds*. We reserve the term *limit points* for values to which y_n converges with positive probability.

Recall the definitions of the regions: $N = (0, \min\{\hat{y}_I, \hat{y}_M\}), I = (\hat{y}_I, \hat{y}_M), M = (\hat{y}_M, \hat{y}_I), S = (\max\{\hat{y}_M, \hat{y}_I\}, 1)$. To draw phase diagrams for the LDI it suffices to know the positions of $\{y^*_S, y^*_I, y^*_M, y^*_N, \hat{y}_I, \hat{y}_M\}$. The positions of the thresholds $\hat{y}_I, \hat{y}_M$ determine the system's regions, and within each region R the flow is towards the corresponding steady state y^*_R. Thus, it is important to understand the possible orderings of these variables.

Lemma 3. $\min\{y^*_S, y^*_M\} > \max\{y^*_I, y^*_N\}$

An intuition for Lemma 3 is that, since users care more about filtering M content than I content, when users share an M story the expected inflow of true stories is greater than when they share an I story. This explains why $y^*_S, y^*_M > y^*_I$. Additionally, when users share M stories they are successfully filtering false content, so that the expected inflow of true stories is greater than the inflow without any sharing, implying $y^*_M > y^*_N$. When users share I stories the inflow of true stories may be greater than or less than the inflow without any sharing, so we cannot sign the relationship between y^*_S, y^*_M nor the relationship between y^*_I, y^*_N but can verify that $y^*_S > y^*_N$. Numerical calculations described in Appendix D show that both $y^*_S < y^*_M$ and $y^*_S > y^*_M$ are possible and similarly that y^*_N can be either greater or less than y^*_I. Moreover, the relationship between any threshold and any quasi steady state is also undetermined, i.e., both $\max\{y^*_N, y^*_I, y^*_M, y^*_S\} < \min\{\hat{y}_I, \hat{y}_M\}$ and $\min\{y^*_N, y^*_I, y^*_M, y^*_S\} > \max\{\hat{y}_I, \hat{y}_M\}$ are possible.

This means that Lemma 3 is the only restriction on the ordering of the quasi steady states and thresholds (for simplicity, we rule out the knife edge case of equality between any of these variables). Because regions M and I don't occur at same time, for given parameters only one of y_I^* and y_M^* matters. This means there are 40 possible strict configurations for the five variables that pin down the phase diagram: the two thresholds, and the quasi steady states for the system's three regions, i.e., y_S^*, y_N^* and one of y_I^*, y_M^*.

To see why there are 40 configurations, consider the case $\hat{y}_I < \hat{y}_M$. In this case, the five variables are $\{\hat{y}_I, \hat{y}_M, y_N^*, y_I^*, y_S^*\}$. We can now count the number of orderings of these variables that satisfy our restrictions. First, we can choose the relative positions of the two thresholds, giving $\binom{5}{2} = 10$ options. Now, since we assumed $\hat{y}_I < \hat{y}_M$, and by Lemma 3 we know $y_S^* > \max\{y_N^*, y_I^*\}$, the only degree of freedom is the order between y_N^*, y_I^*, for a total of 20 configurations in which $\hat{y}_I < \hat{y}_M$. Similarly, there are 20 configurations with $\hat{y}_I > \hat{y}_M$.

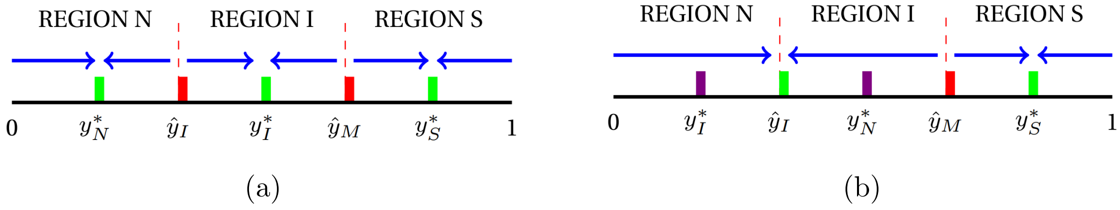

Figure 1: Examples of phase diagrams: (a) all quasi steady states are within their region, (b) only y_S^* is within its region.

Figure 1 presents two examples of phase diagrams. Stable steady states of the LDI are marked in green,unstable steady states are in red, quasi steady states that are not steady states are in purple, and thresholds are marked by dashed lines. An immediate observation is that all quasi steady states that are within their regions are stable steady states for the LDI. Additionally, since every limit ODE has a unique steady state, the only other candidate steady states for the inclusion are the thresholds. Thus, defining $\mathcal{Q} = \{y_R^* | y_R^* \in \text{region } R\}$ as the set of quasi steady states that are within their respective regions, we have $\mathcal{Q} \subset \mathcal{S}_F \subset \mathcal{Q} \cup \{\hat{y}_I, \hat{y}_M\}$. Note that $\mathcal{Q}$ depends on the model's parameters and contains at most three quasi steady states.

For a threshold $\hat{y}$ to be a stable steady state, the flow above it needs to point down and the flow below it needs to point up. This requires a "flip" of quasi steady states: Let W be the region to the left of $\hat{y}$, and Z the region to the right, a flip is: $y_Z^* < \hat{y} < y_W^*$. Flips around $\hat{y}_I$ occur when $\hat{y}_I < \hat{y}_M$ and $y_I^* < \hat{y}_I < y_N^*$ or $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$ (we find that both are possible). Using Lemma 3, we find that flips cannot occur around $\hat{y}_M$, which implies the following characterization of $\mathcal{S}_F$:

Theorem 1. *Either (a) $\mathcal{S}_F = \mathcal{Q} \cup \{\hat{y}_I\}$, or (b) $\mathcal{S}_F = \mathcal{Q}$. Case (a) obtains if and only if $\hat{y}_I < \hat{y}_M$ and $y_I^* < \hat{y}_I < y_N^*$ or $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$.*

Figure 2 presents phase diagrams for all possible configurations in which $\hat{y}_I < \hat{y}_M; y_I^* < y_N^*$. As above, stable steady states are in green, unstable steady states are in red, and quasi steady states that are not steady states are in purple. The numbers on the bottom left of each phase diagram are the indices of the positions of the two thresholds among the five variables that pin down the phase diagram (so in the diagram on the top left the thresholds are in the first and second positions since the order is $\hat{y}_I < \hat{y}_M < y_I^* < y_N^* < y_S^*$). Phase diagrams $(2,3), (2,4), (2,5)$ correspond to case (a) of Theorem 1, in which $\hat{y}_I$ is a stable steady state.

For each of the points $\{y_N^*, y_M^*, y_I^*, y_S^*, \hat{y}_I\}$, there exists a configuration in which this point is the only member of $\mathcal{S}_F$. Some of these appear in Figure 2. For example, in phase diagram $(4,5)$, $\mathcal{S}_F = \{y_N^*\}$. Figures 3, 4, and 5 in Online Appendix D.3 present the phase diagrams for the remaining possible configurations.

4.2.2 Limit behavior

Recall that y_0 is the initial share of true stories. Since behavior in the no sharing region (N) is deterministic—exactly 1 true story and κ false stories are added every period—if the system starts in region N and $y_N^* \in N$ then $y_n \to y_N^* = \frac{1}{1+\kappa}$ deterministically. Otherwise, y_n converges to any stable steady state with positive probability.

> **Theorem 2.** *y_n converges almost surely to a point in $\mathcal{S}_F$. If $y_N^* \in N$ and $y_0 \in N$ then y_n converges to y_N^*. Otherwise, for all $y^* \in \mathcal{S}_F$ there is positive probability that y_n converges to y^*.*

We prove Theorem 2 for the cases where y_n does not converge deterministically to y_N^* in three steps: First, we show that y_n converges to a steady state of the LDI almost surely, then that y_n converges to every stable steady state with positive probability, and finally that y_n almost surely does not converge to an unstable steady state.

For the first step, we prove a more general result, Theorem 3 in Appendix B, which applies to any system that is a concatenation of a finite number of GPUs (with two colors). We provide a formal definition of such systems in Appendix B, and refer to them as piecewise GPUs or PGPUs. Theorem 3 relates the limit behavior of a PGPU to an associated limit differential inclusion, defined as a concatenation

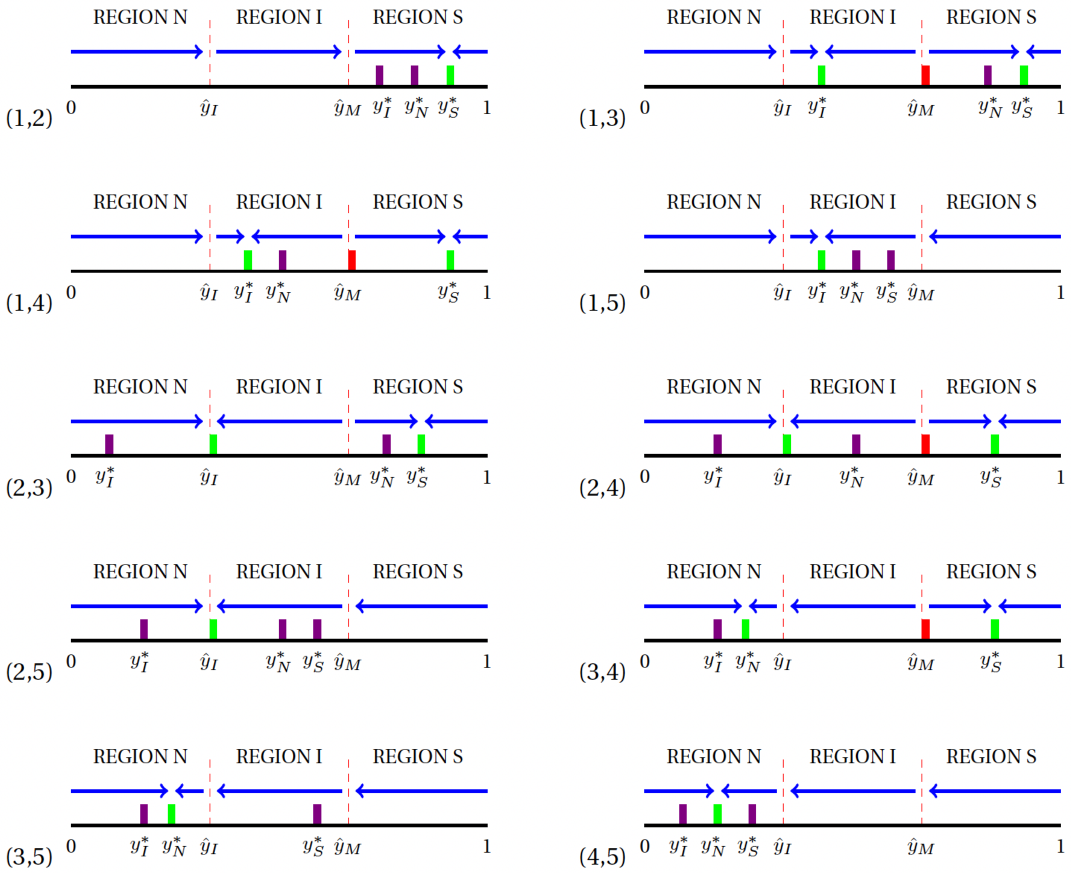

Figure 2: Phase diagrams for the case $\hat{y}_I < \hat{y}_M; y_I^* < y_N^*$.

of the limit ODEs of the GPUs. Applied to our system, the theorem implies that the limit set of y_n, $L(y_n) = \bigcap_{m>0} \overline{\{y_n : n > m\}}$, is almost surely a singleton that contains a steady state of the LDI. To prove this result, we follow standard stochastic approximation techniques, and define a continuous time version $\mathbf{Y}$ of y_n (or, in the general case, of the share of balls of color 1) by rescaling time and using a piecewise affine interpolation. We then extend a result in Schreiber (2001) to prove that $\mathbf{Y}$ is almost surely a *perturbed solution* to the associated inclusion.[13] We complete this step by applying a result in BHS that characterizes limits of perturbed solutions.

To prove the second step—positive probability of convergence to every stable steady state—consider first the quasi steady states y_R^*. We first show that y_n has positive probability of converging to y_R^* conditional on starting from states z_m with $|z_m|$ sufficiently large and y_m sufficiently close to y_R^*. This claim is true for the counterfactual process that follows the sharing rule of region R everywhere, because this process converges almost surely to y_R^*. This implies that the claim is also true for y_n, because: i) when y_n is in region R it follows the same law of motion as the counterfactual process; and ii) we show that starting from a state z_m with $|z_m|$ sufficiently large and y_m sufficiently close to y_R^* the counterfactual process (and therefore also y_n) has positive probability of never leaving region R. We complete the proof for this case by showing that the system has positive probability of arriving at a state z_m from which convergence occurs with positive probability. The proof for the case where the stable steady state is $\hat{y}_I$ is similar but uses a different counterfactual process.

Finally, to prove that y_n almost surely does not converge to an unstable steady state we use Theorem 4 in Appendix B, which shows that a sufficient condition for nonconvergence to an unstable steady state is that there is a positive uniform lower bound on the noise in the stochastic process. Intuitively, noise jiggles y_n away from the steady state, and because the steady state is unstable, the drift of the process will tend to move it further away. Theorem 4 follows from Theorem 2.9 in Pemantle (2007). Pemantle's result is for stochastic approximations with one dimensional limit ODEs, but it can be applied to PGPUs as it does not require that the RHS of the ODE is continuous.

[13] A continuous function $\mathbf{X} : [0, \infty) \to \mathbb{R}$ is a perturbed solution to a differential inclusion if it is absolutely continuous and satisfies conditions that ensure that it is eventually arbitrarily close to a solution to the inclusion. See appendix B for the formal definition, adapted from BHS.

4.2.3 Discussion

Our simplified representation of platform dynamics allows for rich limit behavior. Our finding that the limit share of true stories is random, though not mathematically surprising within the context of generalized urns, has notable implications for the evolution of platform composition. It implies that starting from the same initial platform composition and parameters, the system can end up at very different limits in terms of both the share of true stories and users' limit actions. For instance, in some cases the system has positive probability of converging to any of three limits: One in which the share of true stories is low and users do not share at all (since the probability of sharing a false story is high), one in which the share of true stories is intermediate and users share only stories with one evocativeness level (very interesting/mildly interesting), and one in which the share of true stories is high and users share both very interesting and mildly interesting stories. This path-dependence suggests sensitivity to shocks—when the system is close to a steady state, exogenously adding a large number of false (or true) stories may change the trajectory. Exogenous shocks will be more likely to affect limit behavior if they occur in the "early days" of the platform, when the overall number of stories is small. On the other hand, for each of the candidate limit points (the four quasi steady states and the threshold $\hat{y}_I$), there exist parameter configurations such that this point is the unique stable steady state of the LDI, so the system converges to this point almost surely starting from any initial platform composition.

4.3 Comparative statics

We will refer to points to which the system converges with positive probability as *limit points*. The previous section showed that the set of limit points is $\mathcal{S}_F$, and characterized $\mathcal{S}_F$ for every parameter specification. We now ask how the positions of the limit points and composition of $\mathcal{S}_F$ change with the parameters.

It is straightforward to verify that $y_N^* = \frac{1}{1+\kappa}$, so this candidate limit point is decreasing in κ and constant in all other parameters. Theorems 5-8 in Appendix C present comparative statics for each of the remaining candidate limit points with respect to all parameters. We now discuss the main takeaways from these theorems.

All candidate limit points are increasing in λ and μ. This is intuitive: Increasing the weight on veracity or the penalty for sharing false stories increases the limit share of true stories. Additionally, any limit point that is a quasi steady state is decreasing in κ and, with the exception of y_N^*, is decreasing in β. This is also intuitive—increasing the cost of attention or the exogenous inflow of false stories decreases the share of true stories on the platform.

Less intuitive is the possibility of a limit point increasing in β or being constant in κ. However, both of these arise when the limit point is $\hat{y}_I$. Recall that $\hat{y}_I$ is the point where users are exactly indifferent between sharing and not sharing very interesting stories. This point is increasing in β because users' payoffs are decreasing in the cost of attention and increasing in the share of true stories. Hence, when β goes up, the share of true stories required for indifference needs to go up as well to compensate for the utility loss. $\hat{y}_I$ does not depend on κ, since the exogenous inflow of false stories is is not an argument in users' utility functions. However, as we show below, when κ is sufficiently large $\hat{y}_I$ will not be a limit point.

Table 1: Comparative Statics for κ, β

Variable	**Comparative Statics w.r.t** κ
y_M^*, y_S^*, y_I^*	Everywhere decreasing in κ
$\hat{y}_I$	Constant in κ.
Variable	**Comparative Statics w.r.t** β
y_M^*, y_S^*, y_I^*	Everywhere decreasing in β
$\hat{y}_I$	Everywhere increasing in β.

Comparative statics with respect to the remaining parameters are more nuanced. We discuss each of them in turn:

The role of θ

Table 2: Comparative Statics for θ

Variable	**Comparative Statics w.r.t** θ
y_M^*	Decreasing in θ for $\theta < \theta_M$ and increasing for $\theta > \theta_M$, where $\theta_M \in (0, 1]$.
y_S^*	Decreasing in θ for $\theta < \theta_S$ and increasing for $\theta > \theta_S$, where $\theta_S \in (0, 1]$.
y_I^*	Decreasing in θ for $\theta < \theta_I$ and increasing for $\theta > \theta_I$, where $\theta_I \in (0, 1]$.
$\hat{y}_I$	Everywhere increasing in θ.

Recall that θ, the "credibility" of false stories, determines how hard it is to distinguish between a true story and a false one. When θ increases it is harder to identify false stories but users are aware of this and also pay more attention (both $a(y, I)$ and $a(y, M)$ are increasing in θ). This leads to two opposing forces on the limit share of true stories, and our model predicts that either one can prevail: The candidate limit points y_S^*, y_M^* and y_I^* are decreasing in θ up to a point and then increasing in θ, so for sufficiently large values of θ the increase in attention more than compensates for the increase in credibility.[14] The candidate limit point $\hat{y}_I$ behaves differently, as it is always increasing in θ: Users' payoffs from sharing are decreasing in θ so $\hat{y}_I$ needs to increase to maintain indifference.

Our model highlights the complex relationship between the credibility of false stories and their prevalence when attention is endogenous. The observation that increasing the credibility of false stories might actually lead to a decrease in their prevalence might explain why it sometimes seems as if the producers of false stories do not make much of an effort for these stories to appear true. This is especially relevant if we acknowledge that in reality different stories have different credibility levels and users do not know the credibility of a given story and may also not know the distribution of θ. Under these conditions, we might expect producers of false stories or other interested parties to put a spotlight on false stories with low credibility in order to bias users' estimations of θ downward. Relaxing our assumptions on θ and modeling the interests of fake news producers are interesting directions for future work.

Another interpretation of θ is that the social media platform employs some fact-checking scheme, and θ is the probability that a false story is *not* flagged as false (so the flagging rate is $1 - \theta$). This fits with the formulation of the signal function in (1) since users will recognize flagged stories as false regardless of their attention level. Under this interpretation, the comparative statics of the quasi steady states with respect to θ imply that if flagging rates are low, marginally improving them may have unintended consequences. Again, the intuition relates to a counterbalancing force driven by attention choices. When more stories are flagged, users pay less attention. This means they are more likely to share stories that have not been flagged, which can lead to an overall increase in the limit share of false stories. The comparative statics

[14] Note that the comparative statics in Table 2 allow for the case that a quasi steady state y_R^* is everywhere decreasing in θ (this is the case if $\theta_R = 1$). However, we show in Appendix D that for each quasi steady state y^* there are examples where y^* is both decreasing and increasing in θ when it is a limit point.

for the quasi steady states $\{y_S^*, y_I^*, y_M^*\}$ are a manifestation of the "implied truth effect" introduced and empirically demonstrated in Pennycook, Bear, Collins, and Rand (2020), where false content that is not flagged as false is considered validated and seen as more accurate. We contribute to the analysis of this effect by showing that it might imply a non-monotonic relationship between flagging rates and the share of true stories. Finally, the comparative statics with respect to $\hat{y}_I$ imply that the limit share of true stories may be everywhere decreasing in the flagging rate, through the constraint that users are indifferent, a mechanism distinct from the implied truth effect.

The role of δ

Table 3: Comparative Statics for δ

Variable	**Comparative Statics w.r.t δ**
y_M^*	Everywhere increasing in δ.
y_S^*	Decreasing in δ for δ sufficiently close to $\frac{1}{2}$, and increasing in δ for δ sufficiently close to 1.
y_I^*	Everywhere decreasing in δ.
$\hat{y}_I$	Everywhere increasing in δ.

Increasing δ means false stories are more likely to be very interesting, so the comparative statics for y_I^*, y_M^* are intuitive—the limit share of true stories decreases (increases) in δ when users share only very interesting (mildly interesting) stories. The quasi steady state y_S^*, where users share both types of stories, decreases in δ when δ is close to $\frac{1}{2}$, and increases in δ when δ is close to 1. Appendix D presents numerical examples where y_S^* is both decreasing and increasing in δ when it is a limit point. Intuitively, the non-monotonicity arises because when δ is close to $\frac{1}{2}$ users are sharing more very interesting stories than mildly interesting stories, since both types of stories are almost equally likely to be false and very interesting stories have additional value. Thus, in this case, the comparative statics with respect to δ are similar to the comparative statics for y_I^*, i.e., to those in the region where users are sharing only very interesting stories. As δ moves closer to 1, the stories that users share are more likely to be mildly interesting and comparative statics with respect to δ eventually become similar to those for y_M^*. Finally, $\hat{y}_I$ is increasing in δ because for a fixed y_n, increasing δ leads to a decrease in the value from sharing very interesting stories.[15]

The role of ρ

Table 4: Comparative Statics for ρ

Variable	**Comparative Statics w.r.t ρ**
y_M^*	Everywhere increasing in ρ.
y_S^*	Everywhere increasing in ρ.
y_I^*	Increasing in ρ if $\frac{1}{2} > \delta\theta\,(1 - a(y, I))$ and decreasing in ρ when the sign is reversed.
$\hat{y}_I$	Constant in ρ.

Candidate limit points are increasing in the reach parameter ρ when users are successfully filtering false content, i.e., when the share of true stories shared (out of all true stories) is greater than the share of false stories shared (out of all false stories). The only case where this may not happen is if the system is in region I. In this case, users are sharing $\frac{1}{2}$ of all true stories and $\delta\theta(1 - a(y, I))$ of all false stories. We find that both $\delta\theta(1 - a(y, I)) > \frac{1}{2}$ and $\delta\theta(1 - a(y, I)) < \frac{1}{2}$ are possible, and that both can occur when $y_I^* \in I$, so that y_I^* can be either increasing or decreasing in ρ when it is a limit point. Thus, in the model, increasing the reach of shared stories may contribute to the spread of false news only when users put high value on sharing very interesting stories, and there are enough false stories in the system so that users are better off not sharing mildly interesting stories.

The composition of $\mathcal{S}_F$

We now turn to comparative statics for the composition of $\mathcal{S}_F$. Making general statements here is challenging given the large number of possible configurations. One clear example is the effect of κ, the production rate of false stories. We find that the thresholds are constant in κ and all quasi steady states y_R^* are decreasing in κ. Additionally, fixing values for the other parameters, for any quasi steady state y_R^* we have $\lim_{\kappa\to 0} y_R^*(\kappa) = 1$ and $\lim_{\kappa\to 0} y_R^*(\kappa) = 0$, because when the number of false stories produced is sufficiently large the sharing decisions become inconsequential. Thus, for sufficiently large values of κ all quasi steady states fall in the no sharing

[15] This can lead to a counter-intuitive situation where asymptotically users only share very interesting stories, but when very interesting stories become more likely to be false the limit share of true stories increases. This happens when $\hat{y}_I$ is a limit point and it is between regions N and I (as in Figure 1b) so users are mixing between sharing very interesting stories and not sharing.

region and the unique limit point is y_N^*, and for sufficiently small values of κ all quasi steady states fall in the sharing region and the unique limit point is y_S^*. In other words, there are values $0 < \kappa_1 < \kappa_2$ such that when $\kappa \leqslant \kappa_1$, $\mathcal{S}_{\mathcal{F}} = \{y_S^*\}$ and when $\kappa \geqslant \kappa_2$, $\mathcal{S}_{\mathcal{F}} = \{y_N^*\}$. Thus, increasing the production rate of false stories from $\kappa \leqslant \kappa_1$ to $\kappa \geqslant \kappa_2$ will change users limit behavior from sharing both very interesting and mildly interesting stories to not sharing at all. Since we saw above that when users are sharing stories of both evocativeness levels they are successfully filtering false content, the exogenous decrease in the share of incoming stories that are true is amplified by user behavior.[16]

5 Conclusion

This paper analyzes a model of the sharing of stories on a social media platform when users' attention levels are endogenous and depend on the mix of true and false stories. The share of true stories converges almost surely, but the realized limit point is stochastic, and different possible limits have very different user sharing behavior. This randomness of the limit implies that the type of stories users happened to be exposed to in the early days of the platform and their subsequent sharing decisions can have long-term implications.

We show that the limit share of true stories may be either increasing or decreasing in each of the following parameters: the cost of attention, the credibility of false stories, the probability that false stories are very interesting, and the reach of shared stories.

Although endogenous attention creates a counterbalancing force to changes in the credibility/flagging of false stories, it can intensify the effect of producing more false stories. This suggests that interventions that target producers of false news might be more efficient than attempts to stop the spread of false news already on the platform.

Our model captures many important features in a tractable framework, and parts with most of the literature by tracking the evolution of the entire platform rather than the spread of a single story. Its key simplifying feature is that it has a one-dimensional state space. We maintain this feature while considering two-dimensional story char-

[16] Relatedly, some changes in κ will lead to discontinuous jumps in the distribution of $\lim_{n\to\infty} y_n$. This happens when a quasi steady state crosses a threshold so that it (or the threshold) are no longer a limit point.

acteristics by assuming that only a story's veracity is fixed while its evocativeness is drawn every period. It would be straightforward to analyze variations that preserve this structure. For instance, Allcott and Gentzkow (2017) emphasizes the importance of user heterogeneity and show that education, age, and total media consumption are strongly associated with discernment between true and false content. Such characteristics can be incorporated into our model by having the user's type drawn randomly every period. Allcott and Gentzkow (2017) also finds that in the run-up to the 2016 election, both Democrats and Republicans were more likely to believe ideologically aligned articles than nonaligned ones. Such partisan considerations can be incorporated into our model by having both the user's and story's partisanship drawn every period, and including the relation between them in the users' payoffs.

Other important features of social media behavior could in principle be handled with similar techniques but a larger state space. For example, if users conditioned their choices on the number of times a story has been shared or if some stories were always more interesting than others. The larger state space makes both the stochastic approximation arguments and the analysis of the associated deterministic continuous-time dynamics more complicated, and we leave this for future work.

References

Acemoglu, D., A. Ozdaglar, and J. Siderius (Dec. 2023). "A Model of Online Misinformation". *The Review of Economic Studies*, rdad111.

Allcott, H. and M. Gentzkow (2017). "Social media and fake news in the 2016 election". *Journal of economic perspectives* 31, pp. 211–236.

Benaim, M., J. Hofbauer, and S. Sorin (2005). "Stochastic approximations and differential inclusions". *SIAM Journal on Control and Optimization* 44, pp. 328–348.

Benaim, M., S. J. Schreiber, and P. Tarres (2004). "Generalized urn models of evolutionary processes".

Berger, J. and K. L. Milkman (2012). "What makes online content viral?" *Journal of marketing research* 49, pp. 192–205.

Bloch, F., G. Demange, and R. Kranton (2018). "Rumors and social networks". *International Economic Review* 59, pp. 421–448.

Chen, X., G. Pennycook, and D. Rand (2023). "What makes news sharable on social media?" *Journal of Quantitative Description: Digital Media* 3.

Dasaratha, K. and K. He (2023). "Learning from Viral Content". *arXiv preprint arXiv:2210.01267.*

Guess, A. M. et al. (2020). "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India". *Proceedings of the National Academy of Sciences* 117, pp. 15536–15545.

Hsu, C.-C., A. Ajorlou, and A. Jadbabaie (2021). "Persuasion, news sharing, and cascades on social networks". *2021 60th IEEE Conference on Decision and Control (CDC).* IEEE, pp. 4970–4975.

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

Mahmoud, H. (2008). *Pólya urn models.* CRC press.

Merlino, L. P., P. Pin, and N. Tabasso (2023). "Debunking rumors in networks". *American Economic Journal: Microeconomics* 15, pp. 467–496.

Mostagir, M. and J. Siderius (2022). *Naive and bayesian learning with misinformation policies.* Tech. rep.

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

Pemantle, R. (2007). “A survey of random processes with reinforcement”.

Pennycook, G., A. Bear, E. T. Collins, and D. G. Rand (2020). “The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings”. *Management Science* 66, pp. 4944–4957.

Pennycook, G., T. D. Cannon, and D. G. Rand (2018). “Prior exposure increases perceived accuracy of fake news.” *Journal of experimental psychology: general* 147, p. 1865.

Pennycook, G., Z. Epstein, et al. (2021). “Shifting attention to accuracy can reduce misinformation online”. *Nature* 592, pp. 590–595.

Pennycook, G., J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand (2020). “Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention”. *Psychological Science* 31, pp. 770–780.

Schreiber, S. J. (2001). “Urn models, replicator processes, and random genetic drift”. *SIAM Journal on Applied Mathematics* 61, pp. 2148–2167.

Van Bavel, J. J. and A. Pereira (2018). “The partisan brain: An identity-based model of political belief”. *Trends in cognitive Sciences* 22, pp. 213–224.

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

Appendix A: Proofs and Omitted Results

Proof of Lemma 1.

When $v = T$, $s = T'$ with probability 1 and $e = I$ with probability $\frac{1}{2}$. When $v = F$, $e = I$ with probability δ. Thus,

$$\mathbb{P}_{a,y}(T', T|I) = \frac{\mathbb{P}_{a,y}(T', T, I)}{\mathbb{P}_{a,y}(I)} = \frac{\frac{y}{2}}{\frac{y}{2} + (1-y)\delta} = \frac{y}{y + 2(1-y)\delta}.$$

Similarly,

$$\begin{aligned}
\mathbb{P}_{a,y}(T', T|M) &= \frac{y}{y + 2(1-y)(1-\delta)}, \\
\mathbb{P}_{a,y}(T', F|I) &= \frac{2(1-y)\delta\theta(1-a)}{y + 2(1-y)\delta}, \\
\mathbb{P}_{a,y}(T', F|M) &= \frac{2(1-y)(1-\delta)\theta(1-a)}{y + 2(1-y)(1-\delta)}.
\end{aligned}$$

By (2), the expected payoff when attention is a, evocativeness is M and the user will share the story if and only if they receive the signal T' is,

$$U(a, y, M) = \mathbb{P}_{a,y}(T', T|M)u(T, M) + \mathbb{P}_{a,y}(T', F|M)u(F, M) - \beta a^2.$$

Since $u(T, M) = \lambda$, and $u(F, M) = -\lambda\mu$, we have

$$U(a, y, M) = \lambda \left(\mathbb{P}_{a,y}(T', T|M) - \mu\mathbb{P}_{a,y}(T', F|M)\right) - \beta a^2,$$

so

$$U(a, y, M) = \frac{\lambda \left(y - 2\mu(1-y)(1-\delta)\theta\right)}{y + 2(1-y)(1-\delta)} + \frac{2\lambda\mu(1-y)(1-\delta)\theta}{y + 2(1-y)(1-\delta)}a - \beta a^2.$$

Similarly,

$$U(a, y, I) = \lambda \left(\mathbb{P}_{a,y}(T', T|I) - \mu\mathbb{P}_{a,y}(T', F|I)\right) + (1-\lambda)(\mathbb{P}_{a,y}(T', T|I) + \mathbb{P}_{a,y}(T', F|I)) - \beta a^2,$$

so

$$U(a, y, I) = \frac{\lambda(y - 2\mu(1-y)\delta\theta(1-a))}{y + 2(1-y)\delta} + \frac{(1-\lambda)(y + 2(1-y)\delta\theta(1-a))}{y + 2(1-y)\delta} - \beta a^2,$$

and after rearranging,

$$U(a,y,I)=\frac{y+2(1-y)\delta\theta((1-\lambda)-\lambda\mu)}{y+2(1-y)\delta}+\frac{2(1-y)\delta\theta(\lambda\mu-(1-\lambda))}{y+2(1-y)\delta}a-\beta a^2.$$

The functions $U(a,y,I), U(a,y,M)$ are strictly concave in a. Taking first order conditions we find that they are maximized at $a(y,I), a(y,M)$ respectively as defined in Lemma (1). Finally, using Assumptions 1 and 2 it straightforward to verify that $a(y,I), a(y,M) \in [0,1]$.

■

Proof of Lemma 2. Plugging the optimal attention levels from (1) back into $U(a,y,M), U(a,y,I)$ respectively we get,

$$\begin{aligned} V(y,M) &= \frac{\lambda(y-2\mu(1-y)(1-\delta)\theta)}{y+2(1-y)(1-\delta)}+\frac{1}{\beta}\left(\frac{\lambda\mu(1-y)(1-\delta)\theta}{y+2(1-y)(1-\delta)}\right)^2, \\ V(y,I) &= \frac{y+2(1-y)\delta\theta((1-\lambda)-\lambda\mu)}{y+2(1-y)\delta}+\frac{1}{\beta}\left(\frac{(1-y)\delta\theta(\lambda\mu-(1-\lambda))}{y+2(1-y)\delta}\right)^2. \end{aligned} \tag{8}$$

To prove that these value functions are strictly increasing in y, it suffices to show that $U(a,y,M), U(a,y,I)$ are strictly increasing in y for all a, as then for $y_2 > y_1$ we have $V(y_1)=U(a(y_1),y_1) < U(a(y_1),y_2) \leqslant U(a(y_2),y_2)=V(y_2)$. And

$$\begin{aligned} \frac{\partial U(a,y,M)}{\partial y} &= \frac{2(1-\delta)\lambda(1+(1-a)\mu\theta)}{(y+2(1-y)(1-\delta))^2} > 0 \\ \frac{\partial U(a,y,I)}{\partial y} &= \frac{2\delta\left[(\lambda\mu-(1-\lambda))\theta(1-a)+1\right]}{(y+2(1-y)\delta)^2} > 0, \end{aligned}$$

where the second inequality follows from Assumption 1. To show that both $\hat{y}_I, \hat{y}_M$ are interior, note that $V(1,I)=1>0$, and $V(1,M)=\lambda>0$. Additionally, from Assumptions 1 and 2,

$$\begin{aligned} V(0,M) &= \frac{\lambda\mu\theta(\lambda\mu\theta-4\beta)}{4\beta} < 0, \\ V(0,I) &= \theta(\lambda\mu-(1-\lambda))\left[\frac{\theta(\lambda\mu-(1-\lambda))}{4\beta}-1\right] < 0. \end{aligned}$$

■

Lemma 4. *For all $R \in \{S, I, M, N\}$, the ODE $\frac{dy}{dt} = g_R(y)$ defined over $[0,1]$ has a globally stable steady state $y_R^* \in (0,1)$.*

Proof of Lemma 4. First, note that by the definition of $g_R(y)$ in (6), for all $R \in \{S, I, M, N\}$ we have $g_R(0) = 1$ and $g_R(1) = -\kappa$. This follows from $g_R(0) = 1 + p_R^T(0)\rho$ and $p_R^T(0) = 0$ for all R, and $g_R(1) = -\kappa - p_R^F(1)\rho$ and $p_R^F(1) = 0$ for all R. For $R = N$ the ODE takes the simple form $g_N(y) = 1 - (1+\kappa)y$ and the conclusion follows immediately with $y_N^* = \frac{1}{1+k}$. For the other regions, it suffices to prove that $g_R'''(y) > 0$ for all $y \in [0,1]$. Indeed, for $g_R(y)$ to have more than one root in $[0,1]$ it must have a local minimum that is greater than the first root, followed by a local maximum (between the second root and $y = 1$). So, there need to be $0 < w < z < 1$ such that $g_R''(w) \geqslant 0$ while $g_R''(z) \leqslant 0$ which cannot be the case if $g_R'''(y) > 0$ for all $y \in [0,1]$. The derivatives are

$$
\begin{aligned}
g_S'''(y) &= \frac{12\theta^2\rho}{\beta}\left(\frac{(1-\delta)^3\lambda\mu}{(y+2(1-\delta)(1-y))^4} + \frac{\delta^3(\lambda\mu-(1-\lambda))}{(y+2\delta(1-y))^4}\right),\\
g_I'''(y) &= \frac{12\delta^3\theta^2 v(\lambda\mu-(1-\lambda))}{\beta\,(y+2\delta(1-y))^4},\\
g_M'''(y) &= \frac{12(1-\delta)^3\theta^2\rho\lambda\mu}{\beta\,(y+2(1-y)(1-\delta))^4}.
\end{aligned}
$$

All are strictly positive for $y \in [0,1]$. Stability follows from the existence of a unique root together with $g_R(0) = 1 > 0, g_R(1) = -\kappa < 0$ for all regions R. ■

Proof of Lemma 3. By Lemma 4, to prove $y_R^* > y_W^*$ for $R, W \in \{S, I, M, N\}$, it suffices to prove that $g_R(y) > g_W(y)$ for all $y \in (0,1)$. By (6), we have for any $R, W \in \{S, I, M, N\}$:

$$
g_R(y) - g_W(y) = \rho\left[(1-y)\left(p_R^T(y) - p_W^T(y)\right) - y\left(p_R^F(y) - p_W^F(y)\right)\right].
$$

So $g_R(y) > g_W(y)$ if and only if $(1-y)\left(p_R^T(y) - p_W^T(y)\right) > y\left(p_R^F(y) - p_W^F(y)\right)$. Hence,

$g_S(y) > g_I(y)$ for all $y \in (0,1)$ because, by (3),

$$(1-y)\left(p_S^T(y) - p_I^T(y)\right) = (1-y)\frac{y}{2},$$

and,

$$\begin{aligned} y\left(p_S^F(y) - p_I^F(y)\right) &= y(1-y)\theta\left(1 - \delta a(y,I) - (1-\delta)a(y,M) - \delta(1-a(y,I))\right) \\ &= y(1-y)\theta(1-\delta)\left(1-a(y,M)\right), \end{aligned}$$

and, for $y \in (0,1)$,

$$(1-y)\frac{y}{2} > y(1-y)\theta(1-\delta)\left(1-a(y,M)\right) \iff \frac{1}{2} > \theta(1-\delta)(1-a(y,M))$$

which always holds since $(1-\delta) < \frac{1}{2}$, $\theta < 1$ and $a(y,M) \leqslant 1$.

To see that $g_M(y) > g_I(y)$ for all $y \in (0,1)$ note that,

$$(1-y)\left(p_M^T(y) - p_I^T(y)\right) = 0,$$

and,

$$y\left(p_M^F(y) - p_I^F(y)\right) = y(1-y)\theta\left((1-\delta)\left(1-a(y,M)\right) - \delta\left(1-a(y,I)\right)\right).$$

So $g_M(y) > g_I(y)$ if and only if

$$(1-\delta)\left(1-a(y,M)\right) < \delta\left(1-a(y,I)\right).$$

Fix $y \in (0,1)$ and let $L(\delta) = (1-\delta)\left(1-a(y,M)\right); R(\delta) = \delta\left(1-a(y,I)\right)$. We will prove $L(\delta) < R(\delta)$ for all $\delta \in \left[\frac{1}{2}, 1\right)$ by showing that $L(\frac{1}{2}) < R(\frac{1}{2})$ and $L(\delta)$ is decreasing in δ while $R(\delta)$ is increasing in δ. Indeed,

$$R\left(\frac{1}{2}\right) = \frac{1}{4}\left(2 - \frac{\theta(1-y)\left(\lambda\mu - (1-\lambda)\right)}{\beta}\right) > \frac{1}{4}\left(2 - \frac{\theta(1-y)\lambda\mu}{\beta}\right) = L\left(\frac{1}{2}\right),$$

and

$$\frac{\partial L(\delta)}{\partial \delta} = \frac{2\lambda\mu\theta(1-\delta)(1-y)(1-\delta(1-y))}{\beta\left(y+2(1-y)(1-\delta)\right)^2} - 1.$$

Assumption 2 and $\lambda < 1$ imply that $\lambda\mu\theta < 2\beta$. Therefore, it suffices to prove

$4(1-\delta)(1-y)(1-\delta(1-y)) < (y+2(1-y)(1-\delta))^2$, which simplifies to $y^2 > 0$. Hence, $\frac{\partial L(\delta)}{\partial \delta} < 0$. To see that $\frac{\partial R(\delta)}{\partial \delta} > 0$, note that,

$$\frac{\partial R(\delta)}{\partial \delta} = 1 - \frac{2(\lambda\mu - (1-\lambda))\theta\delta(1-y)(y+\delta(1-y))}{\beta(y+2\delta(1-y))^2}.$$

Since $(\lambda\mu-(1-\lambda))\theta < \lambda\mu\theta < 2\beta$ it suffices to prove $4(\delta(1-y)(y+\delta(1-y)) < (y+2\delta(1-y))^2$, which simplifies to $y^2 > 0$. This completes the proof that $\min\{y_S^*, y_M^*\} > y_I^*$.

To see that $\min\{y_S^*, y_M^*\} > y_N^*$, note that $g_S(y) > g_N(y)$ if and only if

$$(1-y)y > y(1-y)\theta\left(1-\delta a(y,I) - (1-\delta)a(y,M)\right),$$

which always holds. Finally, $g_M(y) > g_N(y)$ if and only if

$$(1-y)\frac{y}{2} > y(1-y)(1-\delta)\theta\left(1-a(y,M)\right),$$

which follows from $\delta > \frac{1}{2}, \theta < 1$. ■

Proof of Theorem 1. That $\mathcal{Q} \subset \mathcal{S}_F$ follows immediately from the definitions of these sets and of F. Since each limit ODE has unique steady state, the only other possible members of $\mathcal{S}_F$ are the thresholds between the regions, so $\mathcal{S}_F \subset \mathcal{Q} \bigcup \{\hat{y}_I, \hat{y}_M\}$. A threshold $\hat{y}$ is a stable steady state if for all $y \in (\hat{y}-\epsilon, \hat{y}+\epsilon)$ we have $\text{sign}(x) = \text{sign}(\hat{y}-y)$ for all $x \in F(y)$. This holds only if there is a "flip" of quasi steady states: Let W be the region to the left of $\hat{y}$, and Z the region to the right, a flip is: $y_Z^* < \hat{y} < y_W^*$.

Flips around $\hat{y}_I$ occur if and only if one the following holds: $\hat{y}_I < \hat{y}_M$ and $y_I^* < \hat{y}_I < y_N^*$; or $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$. In Appendix D we show that both are possible. We now show that flips cannot occur around $\hat{y}_M$ so $\hat{y}_M \notin \mathcal{S}_F$. There are two possible cases:

1. $\hat{y}_I < \hat{y}_M$, so the region to the right of $\hat{y}_M$ is S and the region to left is I.

2. $\hat{y}_I > \hat{y}_M$, so the region to the right of $\hat{y}_M$ is M and the region to the left is N.

In Case 1 a flip cannot occur because by Lemma 3, $y_S^* > y_I^*$. In Case 2 a flip cannot occur because by Lemma 3, $y_M^* > y_N^*$. ■

Proof of Theorem 2. When $y^*_N \in N$ and $y_0 \in N$, the system follows the law of motion $z_{n+1} = z_n + \begin{pmatrix} 1 \\ k \end{pmatrix}$, so it never leaves the region N and converges deterministically to $y^*_N = \frac{1}{1+\kappa}$. We henceforth assume that $y^*_N \notin N$ or $y_0 \notin N$. By Theorem 3 in Appendix B, the limit set of y_n is almost surely chain transitive for the LDI in (7). Since the LDI is a one-dimensional autonomous inclusion, its chain transitive sets are simply its steady states, so y_n converges almost surely to a steady state of the LDI. We now show that there is positive probability of convergence to any stable steady state (Claim 1) and zero probability of convergence to an unstable steady state (Claim 2) (see Section 4.2.1 for definitions of stable and unstable steady states).

Claim 1. *If ψ is a stable steady state, there is positive probability that $y_n \to \psi$.*

Proof of Claim 1. Let ψ be stable steady state and $\epsilon > 0$.
Step 1: Defining five auxiliary processes.

The first four auxiliary processes are $\{z_{n;R}\}$ for $R \in \{S, I, M, N\}$ as defined in (4). Recall that $z_{n;R}$ is what the state would be in period n in the hypothetical case where users always follow the sharing rule of region R even when it does not maximize their utility. For example, $\{z_{n;S}\}$ is the process in which users always share all stories for which they received the signal T'. Let $y_{n;R}$ be the share of true stories in period n for the process $\{z_{n;R}\}$. The differential inclusion associated with $\{z_{n;R}\}$ is $\frac{dy}{dt} \in \{g_R(y)\}$. By Lemma 4, this inclusion has a unique steady state y^*_R, so by Theorem 3, $y_{n;R}$ converges almost surely to y^*_R.[17] In particular, for any $\epsilon > 0$ there exists $m_R \in \mathbb{N}$ such that starting from any y in the open ball $B_\epsilon(y^*_R)$, if the total number of stories is greater than m_R, then $y_{n;R}$ has positive probability of remaining in $B_\epsilon(y^*_R)$ for ever, i.e., $\mathbb{P}\left(y_{m;R} \in B_\epsilon(y^*_R) \forall m > n \,|\, y_{n;R} \in B_\epsilon(y^*_R), |z_{n;R}| > m_R\right) > 0$.

The fifth auxiliary process will be used to prove convergence to $\hat{y}_I$ when it is a stable steady state so we define it only for that case. Let L be the region to the left of $\hat{y}_I$ and R the region to the right of $\hat{y}_I$. Since $\hat{y}_I$ is a stable steady state, we have $y^*_R < \hat{y}_I < y^*_L$. Let O be the third region of the system (O is located either to the right of R or to the left of L). Define an alternative stochastic process $\{z_{n;H}\}$, with share of true stories $y_{n;H}$, where the law of motion in regions R, L is unchanged but the law of motion in region O is changed to be deterministic and such that $y_{n;H}$ moves

[17] This also follows directly from the results of Schreiber (2001) because these auxiliary processes are GPUs.

monotonically towards the other regions.[18] Let $\frac{dy}{dt} \in F_H(y)$ be the limit differential inclusion for this alternative process, as defined in Definition 5 in Appendix B. By construction, $\hat{y}_I$ is the unique steady state for this inclusion, so Theorem 3 implies that $y_{n;H}$ converges to $\hat{y}_I$ almost surely. In particular, there exists $m_H \in \mathbb{N}$ such that $\mathbb{P}\left(y_{m;H} \in B_\epsilon(\hat{y}_I) \forall m > n \,|\, y_{n;H} \in B_\epsilon(\hat{y}_I), |z_{n;H}| > m_H\right) > 0$.

Step 2: Positive probability of converging to ψ conditional on arriving at an open ball around it when $|z_n|$ is sufficiently large.

Assume w.l.o.g. that ϵ is small enough that $B_\epsilon(y^*_R) \subset R$ if $\psi = y^*_R$ for some region R and that $B_\epsilon(\hat{y}_I) \subset [0,1]\backslash O$ if $\psi = \hat{y}_I$ (recall from the previous step that O is the only region not adjacent to $\hat{y}_I$). For the case where $\psi = y^*_R$ we have $\mathbb{P}\left(y_m \in B_\epsilon(y^*_R) \forall m > n \,|\, y_n \in B_\epsilon(y^*_R), |z_n| > m_R\right) > 0$, since conditional on y_n remaining in $B_\epsilon(y^*_R)$ we have $y_n = y_{n;R}$ (i.e., they follow the same law of motion). The fact that $y_n = y_{n;R}$ conditional on y_n remaining in region R also implies $\mathbb{P}(y_n \to y^*_R | y_n \in B_\epsilon(y^*_R) \, \forall n > m) = 1$. So, if the system arrives at a state z_n such that $y_n \in B_\epsilon(y^*_R)$ and $|z_n| > m_R$ then y_n converges to y^*_R with positive probability.

In the case where $\psi = \hat{y}_I$, an analogous argument (replacing $y_{n;R}$ with $y_{n;H}$), implies that if the system arrives at state z_n such that $y_n \in B_\epsilon(\hat{y}_I)$ and $|z_n| > m_H$ then y_n converges to $\hat{y}_I$ with positive probability.

Step 3: Positive probability of arriving at such a ball.

We now prove that there is positive probability of arriving at z_n such that $y_n \in B_\epsilon(\psi)$ and $|z_n| > m$ where m is as defined above. By (6), for any region R,

$$y^*_R = \frac{1 + p^T_R(y^*_R)\rho}{1 + \kappa + \rho\left(p^T_R(y^*_R) + p^F_R(y^*_R)\right)}.$$

This implies that $\frac{1}{1+\kappa+\rho} < y^*_R < \frac{1+\rho}{1+\kappa+\rho}$: the first inequality is immediate and the second is equivalent to

$$\rho\left(\kappa(1 - p^T_R(y^*_R)) + p^F_R(y^*_R)(1+\rho)\right) > 0,$$

which always holds. Since any stable steady state is either is a quasi steady state or

[18] So if O is to the right of R then $y_{n;H}$ is decreasing and in region O and if O is to the left of L then $y_{n;H}$ is increasing in region O. It is easy to construct such a law of motion—for $y_{n;H}$ to increase we can require that every period one true story is added and no false stories and vice versa for $y_{n;H}$ to decrease.

a threshold bounded above and below by quasi steady states, the above implies that

$$\frac{1}{1+\kappa+\rho} < \psi < \frac{1+\rho}{1+\kappa+\rho} \quad \forall \psi \in \mathcal{S}_F. \tag{9}$$

Recall that we assumed that either $y_N^* \notin N$ or $y_0 \notin N$ (or both). If $y_0 \notin N$ and $\psi \notin N$ the claim follows immediately from (9) and Lemma 5 below, together with $|z_n| \to \infty$ surely. If $y_0 \notin N$ and $\psi \in N$ then it must be the case that $\psi = y_N^*$ and y_N^* is a stable steady state. In this case, Lemma 5 implies that there is positive probability of arriving at $\sup(N)$ (which is $\min\{\hat{y}_I, \hat{y}_M\}$). Additionally $y_N^* \in N$ implies that there is positive probability of arriving from $\sup(N)$ into $B_\epsilon(y_N^*)$. Finally, if $y_0 \in N$ and $y_N^* \notin N$, because the system converges deterministically towards y_N^* when the system is in region N, the system surely arrives at $y_n \notin N$ with $|z_n| > m$ after finite time and Lemma 5 implies that there is positive probability of arriving from this y_n to $B_\epsilon(\psi)$. This completes the proof of Claim 1. ■

Claim 2. *The system almost surely does not converge to an unstable steady state.*

Proof of Claim 2.

Since by Lemma 4 all quasi steady states are stable for their associated ODEs, the only possible unstable steady states for the LDI are the thresholds $\hat{y}_I, \hat{y}_M$. Let $\hat{y}$ be a unstable steady state. Let A denote the event "$y_n \in N$ infinitely often" and let A^C denote its complement. We will prove that $\mathbb{P}(y_n \to \hat{y}) = 0$ by proving that if $\mathbb{P}(A) > 0$ then $\mathbb{P}(y_n \to \hat{y}|A) = 0$, and if $\mathbb{P}(A^C) > 0$ then $\mathbb{P}(y_n \to \hat{y}|A^C) = 0$.

Assume $\mathbb{P}(A) > 0$ and consider a realization where $y_n \in N$ infinitely often. If $\hat{y}$ is not adjacent to region N then $y_n \in N$ i.o. rules out convergence to $\hat{y}$. If $\hat{y}$ is adjacent to region N, then by the instability of $\hat{y}$ it must be the case that $y_N^* \in N$. But then, if $y_n \in N$ for some n then y_n converges (deterministically) to $y_N^* \neq \hat{y}$. Thus, if $\mathbb{P}(A) = \mathbb{P}(y_n \in N \quad i.o) > 0$, then $\mathbb{P}(y_n \to \hat{y}|A) = 0$.

We now apply Theorem 4 in Appendix B to prove that if $\mathbb{P}(A^C) > 0$ then $\mathbb{P}(y_n \to \hat{y} = 0|A^C) = 0$. Assume $\mathbb{P}(A^C) > 0$ and consider a realization where $y_n \in N$ at most finitely often, so there exists $m \in \mathbb{N}$ such that $y_n \notin N$ for all $n > m$. To apply Theorem 4 we need to verify that $\mathbb{E}[\xi_n^+|\mathcal{F}_n]$ are uniformly bounded below by a positive number, where $\xi_{n+1} := (y_{n+1} - y_n - \mathbb{E}[y_{n+1} - y_n|z_n])|z_n|$, $\xi_n^+ := \max\{0, \xi_n\}$ and $\mathcal{F}_n$ is the σ-algebra generated by $(z_1, ..., z_n)$.

Consider the law of motion for y_n in Equation 5. Denoting $\Delta_T = \frac{(1-y_n)(1+\rho)-\kappa}{|z_n|+1+\kappa+\rho}$, $\Delta_F = \frac{(1-y_n)-(\kappa+\rho)}{|z_n|+1+\kappa+\rho}$, $\Delta_O = \frac{(1-y_n)-\kappa}{|z_n|+1+\kappa}$, we have $\Delta_T > \Delta_O > \Delta_F$, so that when y_n is in region R,

$$\mathbb{E}[\xi^+_{n+1}|\mathcal{F}_n] \geqslant p^T_R(y_n)\left(\Delta_T - \sum_{i\in\{T,F,0\}} p^i_R(y_n)\Delta_i\right)|z_n| \geqslant p^T_R(y_n)(1-p^T_R(y_n))(\Delta_T-\Delta_O)|z_n|.$$

Now,

$$\begin{aligned}(\Delta_T - \Delta_O) &\geqslant \frac{(1-y_n)(1+\rho)}{|z_n|+1+\kappa+\rho} - \frac{\kappa}{|z_n|+1+\kappa+\rho} - \frac{(1-y_n)}{|z_n|+1+\kappa+\rho} + \frac{\kappa}{|z_n|+1+\kappa+\rho} \\ &= \frac{(1-y_n)\rho}{|z_n|+1+\kappa+\rho},\end{aligned}$$

so,

$$(\Delta_T - \Delta_O)|z_n| \geqslant \frac{(1-y_n)\rho}{2+\kappa+\rho}.$$

Since $y_n \notin N$ from some point onward, by (3), $p^T_R(y_n) \in \{y_n, \frac{y_n}{2}\}$ for both of the adjacent regions R. Thus, for small $\epsilon > 0$, there exists $c > 0$ such that for any $y_n \in (\hat{y}-\epsilon, \hat{y}+\epsilon)$: $p^T_R(y_n)(1-p^T_R(y_n)) \geqslant c$. So, for any $y_n \in (\hat{y}-\epsilon, \hat{y}+e)$ we have $\mathbb{E}[\xi^+_{n+1}|\mathcal{F}_n] \geqslant \frac{c(1-\hat{y}-\epsilon)\rho}{2+\kappa+\rho} > 0$, which completes the proof of Claim 2. ■

Together, Claims 1 and 2 prove Theorem 2. ■

Lemma 5. *If $y^*_N \notin N$ then for any $\epsilon > 0$ and $y \notin N$ such that $y \in (\frac{1}{1+\kappa+\rho}, \frac{1+\rho}{1+\kappa+\rho})$ the system has a positive probability of arriving at some $y_m \in B_\epsilon(y)$ starting from any initial state z_n.*

Proof. Since the number of stories added each period is bounded, there exists some $n_\epsilon \in \mathbb{N}$ such that $|y_{n+1} - y_n| < \epsilon$ whenever $|z_n| > n_\epsilon$. Since $|z_n| \to \infty$ we can assume w.l.o.g. that the initial state z_n satisfies $|z_n| > n_\epsilon$. For such z_n, we consider two possible cases: $y_n < y$ and $y_n > y$.

If $y_n < y$ then $y < \frac{1+\rho}{1+\kappa+\rho}$ implies that if the user shares a true story in period n then $y_n < y_{n+1} < \frac{1+\rho}{1+\rho+\kappa}$.[19] Thus, there exists some $T > 0$ such that if users share a true story every period for T periods then $y_{n+T} \in B_\epsilon(y)$.

[19] Because if a true story story is shared in period n then $z_{n+1} = z_n + \begin{pmatrix} 1+\rho \\ \kappa \end{pmatrix}$.

If $y_n > y$ then, by a similar argument, $y > \frac{1}{1+\kappa+\rho}$ implies that there exists some $T' > 0$ such that if users share false stories for T' periods then $y_{n+T'} \in B_\epsilon(y)$.

We can assume w.l.o.g. that $y_n \notin N$, because if $y_n \in N$, the assumption $y^*_N \notin N$ and the fact that behavior in region N is deterministic imply that surely $y_m \notin N$ for some $m > n$. Since $y_n \notin N$ there is positive probability of sharing a false story and positive probability of sharing a true story. Also, since region N is always the leftmost region and $y \notin N$ then starting from $y_n > y$ and drawing T' false stories or starting from $y_n < y$ and drawing T true stories will not lead the system to enter region N. Thus there is positive probability of drawing T (T') true (false) stories consecutively so there is positive probability of $y_m \in B_\epsilon(y)$ for some $m > n$. ■

Appendix B: Stochastic Approximation and GPUs

This appendix extends and applies results from the theory of stochastic approximation. Specifically, we build on results from Schreiber (2001) and Benaim, Schreiber, and Tarres (2004), who apply stochastic approximation methods to analyze Generalized Polya Urns (GPUs). A key feature of these urn models is that the number of balls added each period is bounded so that as the overall number of balls grows the change in the system's composition between any two consecutive periods becomes arbitrarily small. Within each of the regions $\{N, I, M, S\}$, our system $\{z_n\}$ behaves like a GPU. To analyze the entire system, we define *Piecewise Generalized Polya Urns* (PGPUs), which we analyze by combining results on GPUs with results from BHS that extend the theory of stochastic approximation to cases where the continuous system is given by a solution to a differential inclusion rather than a differential equation. Theorem 3 relates the limit behavior of a PGPU to the limit behavior of the associated differential inclusion; we use it in the proof of Theorem 2. Section B.3 explains why the processes $\{z_{n;R}\}$ defined in (4) are GPUs and derives the corresponding limit ODEs. We conclude this appendix with a discussion of unstable steady states for limit inclusions and a result that we use in the proof of Theorem 2.

B.1 Definitions and Notation

We begin by introducing some notation and definitions from the literatures on stochastic approximation and generalized urns. Given a vector $w \in \mathbb{R}^2$ define $|w| = |w^1|+|w^2|$. Let $\{z_n\} = \{(z_n^1, z_n^2)\}$ be a homogeneous Markov chain with state space $\mathbb{Z}_+^2$ ($\mathbb{Z}_+$ are all the non-negative integers). Let $\Pi : \mathbb{Z}_+^2 \times \mathbb{Z}_+^2 \to [0,1]$ denote its transition kernel, $\Pi(z,z') = \mathbb{P}(z_{n+1} = z'|z_n = z)$. We interpret the process as an urn model, with z_n^i the number of balls of color i at time step n. We now define two types of stochastic processes. Definitions 1 and 3 are taken from Benaim, Schreiber, and Tarres (2004) (similar definitions appear in Schreiber (2001)).[20]

Definition 1. A Markov process $\{z_n\}$ as above is a *generalized Polya urn* (GPU) if:

i. Balls cannot be removed and there is a maximal number of balls that can be added, that is: For all n: $z_{n+1}^1 \geqslant z_n^1, z_{n+1}^2 \geqslant z_n^2$ and there is a positive integer m such that $|z_{n+1} - z_n| \leqslant m$.

ii. For each $w \in \mathbb{Z}_+^2$ with $|w| \leqslant m$ there exist Lipschitz-continuous maps $p^w : [0,1] \to [0,1]$ and a real number $a > 0$ such that

$$\left| p^w\left(\frac{z^1}{|z|}\right) - \Pi(z, z+w) \right| \leqslant \frac{a}{|z|}$$

for all nonzero $z \in \mathbb{Z}_+^2$.

We focus on the dynamics of the distribution of colors in the urn, and in particular the asymptotic distribution. In the two-color case considered here this means tracking the share of balls of color 1 (i.e., true stories), which we denote by $y_n = \frac{z_n^1}{|z_n|}$. A key observation is that $\{y_n\}$ is a stochastic approximation process, as defined below.

Definition 2. Let $\{x_n\}$ be a stochastic process in $[0,1]$ adapted to a filtration $\{\mathcal{F}_n\}$. We say that $\{x_n\}$ is a (one dimensional) stochastic approximation if for all $n \in \mathbb{N}$:

$$x_{n+1} - x_n = \gamma_n \left(g(x_n) + \xi_{n+1} + R_n \right), \tag{10}$$

where γ_n are non-negative with $\gamma_n \to 0, \sum_n \gamma_n = \infty$, g is a Lipschitz function on $\mathbb{R}$,

[20] Their definitions allow for more than two colors and for the possibility of balls being removed, but we focus on the two color case without removal of balls.

$\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$ and the remainder terms $R_n \in \mathcal{F}_n$ go to zero and satisfy $\sum_{n=1}^{\infty} \frac{|R_n|}{n} < \infty$ almost surely.

The function g is often called the *limit ODE*, because the limit of a stochastic approximation process $\{x_n\}$ can be related to the limits of solutions to the continuous deterministic system $\frac{dx}{dt} = g(x)$.[21]Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) provide a formula for the limit ODE for a GPU and prove that with this limit ODE the sequence $\{y_n\}$ of the share of balls of color 1 is a stochastic approximation process. Since we will later consider a system that includes several GPUs we introduce the notation $\{z_{n;k}\}$ to refer to a general GPU.

Definition 3. For a GPU $\{z_{n;k}\}$ with corresponding maps p_k^w, the corresponding *limit ODE* is $\frac{dy}{dt} = g_k(y)$ where $g_k : [0,1] \to [0,1]$ is given by

$$g_k(y) = \sum_{w \in \mathbb{Z}^2} p_k^w(y) \left(w^1 - y|w| \right). \tag{11}$$

B.2 Stochastic Approximation of PGPU's

We extend the literature on GPUs to concatenations of a finite number of GPUs.

Definition 4. A Markov process $\{z_n\}$ with transition kernel Π is a *piecewise generalized Polya urn* (PGPU) if there exists a finite number of GPUs $\{\{z_{n;k}\}\}_{k=1}^K$ (each with kernel Π_k), a finite integer K, and an interval partition $\{I_k\}_{k=1}^K$ of $[0,1]$, such that for all z', if $\frac{z^1}{|z|} \in int(I_k)$ then $\Pi(z,z') = \Pi_k(z,z')$.[22]

The next definition presents the analog of a limit ODE for a PGPU, which is no longer a differential equation but a differential inclusion, i.e., a set valued function.

Definition 5. For a PGPU $\{z_n\}$ we define the *limit differential inclusion* to be $\frac{dy}{dt} \in$

[21] The random shocks average out from the martingale LLN and the limit of the deterministic difference equation $x_{n+1} - x_n = \gamma_n g(x_n)$ can be approximated by the ODE.

[22] Note that we allow for an arbitrary law of motion $\Pi(z,z')$ for z such that $\frac{z^1}{|z|} = max(I_k) = min(I_{k+1})$, i.e, when the share of balls of color 1 is the boundary of an interval. The systems we consider will arrive at such states z with probability zero.

$F(y)$ where

$$F(y) = \begin{cases} \{g_k(y)\}, & y \in int(I_k) \\ \{g_1(0)\}, & y = 0 \\ \{g_K(1)\} & y = 1 \\ [min\{g_k(y), g_{k+1}(y)\}, max\{g_k(y), g_{k+1}(y)\}], & y = max(I_k), 1 \leqslant k < K \end{cases}$$

Henceforth, we fix a PGPU $\{z_n\}$ comprised of GPUs $\{\{z_{n;k}\}\}_{k=1}^{K}$, with share of balls of color 1 denoted $y_n = \frac{z_n^1}{|z_n|}$ and let

$$\frac{dy}{dt} \in F(y) \tag{12}$$

be the associated differential inclusion. In order to apply results from BHS, we need to verify that the paper's standing assumptions on the inclusion hold. These are:

BHS Standing Assumptions. 1. F has a closed graph.

2. $F(y)$ is non empty, compact, and convex for all $y \in [0,1]$.

3. There exists $c > 0$ such that for all $y \in [0,1]$, $\sup_{x \in F(y)} |x| \leqslant c(1 + |y|)$.

Lemma 6. *The inclusion* (12) *satisfies the standing assumptions in BHS.*

Proof. Assumptions 1 and 2 follow immediately from Definition 5. Assumption 3 follows from the fact that the $g_k(y)$ are continuous functions defined over compact sets. Since K is finite there exists some $c > 0$ such that $|g_k(y)| \leqslant c$ for all $y \in [0,1]$ and all $k \in \{1, ..., K\}$, and so for any $y \in [0,1]$: $sup_{x \in F(y)}|x| \leqslant c \leqslant c(1 + |y|)$. ■

To relate the limiting behaviors of y_n to the solutions to the differential inclusion 12, define the *piecewise affine interpolation* of y_n by

$$\mathbf{Y}(t) = y_n + \frac{t - \tau_n}{\gamma_{n+1}}(y_{n+1} - y_n), \quad t \in [\tau_n, \tau_{n+1}], \tag{13}$$

where $\tau_0 = 0$, $\tau_{n+1} = \tau_n + \frac{1}{|z_n|}$, and $\gamma_{n+1} = \frac{1}{|z_n|}$.

Definition 6. A continuous function $\mathbf{Y} : [0, \infty) \to \mathbb{R}$ is a *perturbed solution* to 12 (we also say a perturbed solution to F) if it is absolutely continuous, and there is a locally integrable function $t \mapsto U(t)$ such that

- $\lim_{t\to\infty} \sup_{0\leqslant h\leqslant T} |\int_t^{t+h} U(s)ds| = 0$ for all $T > 0$
- $\frac{d\mathbf{Y}(t)}{dt} - U(t) \in F(\mathbf{Y}(t))$ for almost every $t > 0$.

We now show that the continuous time version of y_n is a bounded perturbed solution to (12) and then complete our characterization of the limit of y_n by applying a result in BHS that characterizes the limit sets of perturbed solutions. The next lemma, which relates PGPUs to their corresponding limit inclusions, is an analog to results in Benaim, Schreiber, and Tarres (2004) and Schreiber (2001) that relate GPUs to their limit ODEs.

Lemma 7. *Let $\{z_n\}$ be a PGPU and (12) its limit differential inclusion, and let $\boldsymbol{Y}$ be the associated interpolated process given by (13). Then $\boldsymbol{Y}$ is a bounded perturbed solution to F.*

Proof. Since $\mathbf{Y}$ is piecewise affine, it is continuous and differentiable almost everywhere and hence absolutely continuous. Define $t \mapsto U(t)$ by

$$U(t) = \frac{y_{n+1} - y_n}{\gamma_{n+1}} - \tilde{F}(\mathbf{Y}(t)) \quad t \in [\tau_n, \tau_{n+1}],$$

where the function $\tilde{F} : [0,1] \to \mathbb{R}$ is such that for every $y \in [0,1]$: $\tilde{F}(y) \in F(y)$. Note that $\frac{d\mathbf{Y(t)}}{dt} = \frac{y_{n+1}-y_n}{\gamma_{n+1}}$ for $t \in [\tau_n, \tau_{n+1}]$, so $\frac{d\mathbf{Y(t)}}{dt} - U(t) = \tilde{F}(\mathbf{Y}(t)) \in F(\mathbf{Y}(t))$. It remains to show $\lim_{t\to\infty} \sup_{0\leqslant h\leqslant T} |\int_t^{t+h} U(s)ds| = 0$ for all $T > 0$. For this, we invoke the following theorem:[23]

Theorem 2.2 (Schreiber (2001)). Let $\{z_{n;k}\}$ be a GPU. Let ϕ^k be the flow of the limit ODE, and $\mathbf{Y}^k(t)$ the piecewise affine interpolation. On the event $\{\liminf_{n\to\infty} \frac{|z_{n;k}|}{n} > 0\}$, $\mathbf{Y}^k(t)$ is almost surely an asymptotic pseudotrajectory for ϕ^k. In other words for any $T > 0$

$$\lim_{t\to\infty} \sup_{0\leqslant h\leqslant T} |\mathbf{Y}^k(t+h) - \phi^k(\mathbf{Y}^k(t), h)| = 0.$$

Fix $T > 0$ and $0 \leqslant h \leqslant T$. Consider $\int_t^{t+h} U(s)ds$. On the event $\mathbf{Y}(s) \in I_k$ for all

[23] Schreiber (2001) states the theorem for piecewise constant interpolations, but it also applies to piecewise affine interpolations.

$s \in [t, t+h]$ we have

$$\begin{aligned}\int_t^{t+h} U(s)ds = \int_t^{t+h} \left(\frac{d\mathbf{Y}(s)}{ds} - \tilde{F}(x)\right) ds &= \int_t^{t+h} \left(\frac{d\mathbf{Y}^k(s)}{ds} - \frac{d\phi^k(\mathbf{Y}(s), s)}{ds}\right) ds \\ &= \mathbf{Y}^k(t+h) - \mathbf{Y}^k(t) - \left(\phi^k(\mathbf{Y}(t), h) - \phi^k(\mathbf{Y}(t), 0)\right) \\ &= \mathbf{Y}^k(t+h) - \phi^k(\mathbf{Y}(t), h).\end{aligned} \tag{14}$$

Since by Definition 4 a PGPU has a finite number of partition intervals I_k, in the interval $[t, t+h]$ the interpolation $\mathbf{Y(t)}$ transitions between intervals I_k a finite a number of times. Thus,

$$\int_t^{t+h} U(s)ds = \sum_{j=1}^{M} \left[\mathbf{Y}^{k_j}(t_j) - \phi^{k_j}(\mathbf{Y}(t_{j-1}), h_j)\right],$$

where $M > 0$ is some integer; $t = t_0 < t_1 < ... < t_M = t + h$; $h_j = t_j - t_{j-1}$, and $k_j \in 1, ..., K$ for all $1 \leqslant j \leqslant M$.[24] Thus, from Schreiber (2001)'s Theorem 2.2 above,

$$\lim_{t\to\infty} \sup_{0\leqslant h\leqslant T} |\int_t^{t+h} U(s)ds| \leqslant \sum_{j=1}^{M} \left(lim_{t\to\infty} \sup_{0\leqslant h\leqslant T} |\mathbf{Y}^{k_j}(t_j) - \phi^{k_j}(\mathbf{Y}(t_{j-1}), h)| \right) = 0.$$

■

We are now ready to prove Theorem 3. The proof combines the previous results with a direct application of the following theorem:

Theorem 3.6 (BHS). Let $\mathbf{x}$ be a bounded perturbed solution to F. Then, the limit set of $\mathbf{x}$, $L(\mathbf{x}) = \bigcap_{t\geqslant 0} \overline{\{\mathbf{x}(s) : s > t\}}$ is internally chain transitive in the sense of BHS.[25]

Theorem 3. *Let $\{z_n\}$ be a PGPU, $\{y_n\}$ the share of balls of color 1 and F the associated limit differential inclusion. Then the limit set of $\{y_n\}$, $L(y_n) = \bigcap_{m>0} \overline{\{y_n : n > m\}}$, is almost surely chain transitive for F.*

Proof. By Lemma 7, the interpolation $\mathbf{Y}$ is a perturbed solution to F. Note that it is also bounded since $\mathbf{Y}(t) \in [0, 1]$ for all $t \geqslant 0$. Thus, Theorem 3.6 in BHS implies that the limit set of $\mathbf{Y}$ is internally chain transitive for F. Note that the

[24] Note that $(M, (t_j)_{j=0}^M, (h_j)_{j=1}^M, (k_j)_{j=1}^M)$ is a random vector.

[25] BHS extend the standard definition to differential inclusions.

asymptotic behaviors of $\mathbf{Y}(t)$ and y_n are the same by the definition of interpolation, i.e., $L(y_n) = L(\mathbf{Y})$, which completes the proof.

■

B.3 The GPUs $\{z_{n;R}\}$

We now explain why the processes $\{z_{n;R}\}$ as defined in (4) are GPUs and derive the formula for their limit ODEs.

Lemma 8. *For each region $R \in \{N, I, M, S\}$, $\{z_{n;R}\}$ is a GPU with limit ODE given by (6).*

Proof. Let R be one of the four possible regions. To show that $\{z_{n;R}\}$ is a GPU we need to verify the conditions of Definition 1. Condition i) follows directly from (4), with the upper bound $m = 1 + \kappa + \rho$. For condition ii), let $w_1 = \begin{pmatrix} 1+\rho \\ \kappa \end{pmatrix}, w_2 = \begin{pmatrix} 1 \\ \kappa+\rho \end{pmatrix}, w_3 = \begin{pmatrix} 1 \\ \kappa \end{pmatrix}$, and let $p_R^T(y), p_R^F(y), 1 - p_R^T(y) - p_R^F(y)$ respectively be the maps p^w corresponding to these vectors. By (3) all three maps are Lipschitz-continuous. Let Π_R denote the transition kernel for $\{z_{n;R}\}$. By the law of motion (4), for any $w \in \{w_1, w_2, w_3\}$ and for any $z \in \mathbb{Z}_+^2$: $\Pi_R(z, z+w) = p^w\left(\frac{z^1}{|z|}\right)$. Since $\Pi_R(z, z+w) = 0$ for any $w \notin \{w_1, w_2, w_3\}$, condition ii) is satisfied.[26]

Next, (11) together with (3) and (4), implies that the ODE associated with the GPU $\{z_{n;R}\}$ is

$$\begin{aligned} g_R(y) = {} & p_R^T(y)(1 + \rho - y(1 + \rho + \kappa)) + p_R^F(y)(1 - y(1 + \rho + \kappa)) \\ & + (1 - p_R^T(y) - p_R^F(y))(1 - y(1 + \kappa)). \end{aligned}$$

Rearranging the above equation gives $g_R(y) = 1 + p_R^T(y)\rho - y\left(1 + \kappa) + \rho\left(p_R^T(y) + p_R^F(y)\right)\right)$, as in (6).

■

[26] Note that we do not use the additional flexibility in Definition 1, which does not require that $\Pi_R(z, z+w)$ be equal to $p^w\left(\frac{z^1}{|z|}\right)$ but only that it converges to it sufficiently quickly as $|z| \to \infty$.

B.4 Unstable Steady States

Consider a PGPU $\{z_n\}$, comprised of GPUs $\{z_{n;k}\}_{k=1}^K$ with associated intervals I_k, where g_k is the RHS of the limit ODE for GPU $\{z_{n;k}\}$. Let $y_{n;k} = \frac{z^1_{n;k}}{|z_{n;k}|}$. Recall that $y_n = \frac{z^1_n}{|z_n|}$ and that the LDI for this PGPU is given by (12). In this subsection, we apply a result in Pemantle (2007) to prove that if ψ is an unstable steady state for the LDI, then under a condition on the noise in the stochastic system, $\mathbb{P}(y_n \to \psi) = 0$.[27] We now add the following assumption, which is satisfied by the PGPU in our model:

Assumption 3. Each of the limit ODEs $\frac{dy}{dt} = g_k(y)$ has a globally stable steady state $y^*_k \in [0, 1]$.

Assumption 3 implies that the only possible unstable steady states for the LDI are the thresholds between the intervals I_k. Define these these as $\hat{y}_k = \max\{I_k\}$ for $k = 1, \ldots, K$. Finally, let $\mathcal{F}_n$ be the σ-algebra generated by $(z_1, ..., z_n)$, let $\xi_{n+1} = (y_{n+1} - y_n - \mathbb{E}[y_{n+1} - y_n|z_n])|z_n|$ and denote $\xi^+_n = \max\{0, \xi_n\}, \xi^-_n = -\min\{0, \xi_n\}$.

Theorem 4. *Let $\hat{y}_k$ be the threshold between intervals I_k, I_{k+1} and assume that $\hat{y}_k$ is an unstable steady state for the LDI. If there exist $\epsilon, r > 0$ such that for all $n \in \mathbb{N}$: $\mathbb{E}[\xi^+_n|\mathcal{F}_n] > r$ if $y_n \in (\hat{y}_k - \epsilon, \hat{y}_k + \epsilon)$, then $\mathbb{P}(y_n \to \hat{y}_k) = 0$.*

The proof applies the following result:

Theorem 2.9 (Pemantle (2007)). Suppose $\{x_n\}$ is a stochastic approximation process as defined in Definition 2 except that g need not be continuous. Assume that for some $p \in (0, 1)$ and $\epsilon > 0$: $\text{sign}(g(x)) = -\text{sign}(p - x)$ for all $x \in (p - \epsilon, p + \epsilon)$. Suppose further that the martingale terms ξ_n in the stochastic approximation equation (10) are such that $\mathbb{E}[\xi^+_{n+1}|\mathcal{F}_n], \mathbb{E}[\xi^-_{n+1}|\mathcal{F}_n]$ are bounded above and below by positive numbers when $x_n \in (p - \epsilon, p + \epsilon)$. Then $\mathbb{P}(x_n \to p) = 0$.

Proof of Theorem 4.

[27] See Section 4.2.1 for definitions of stable and unstable steady states for the LDI

Define the function $g : [0, 1] \to \mathbb{R}$. By

$$g(y) = \begin{cases} g_k(y), & y \in int(I_k) \\ g_1(0), & y = 0 \\ g_K(1) & y = 1 \\ g_k(y) & y = max(I_k), 1 \leqslant k < K \end{cases}$$

Recall that $\xi_{n+1} = (y_{n+1} - y_n - \mathbb{E}[y_{n+1} - y_n|z_n])|z_n|$, and let

$$R_n = |z_n|\mathbb{E}[y_{n+1} - y_n|z_n] - g(y_n).$$

Then ξ_n, R_n are adapted to $\mathcal{F}_n$, $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$ and

$$y_{n+1} - y_n = \frac{1}{|z_n|} (f(y_n) + \xi_{n+1} + R_n) \tag{15}$$

By Lemma 1 in Benaim, Schreiber, and Tarres (2004), and the fact that y_n follows the same law of motion as $y_{n;k}$ when $y_n \in \text{int}(I_k)$, there exists a real number $K > 0$ such that $|R_n| \leqslant \frac{K}{|z_n|}$. Thus, $\sum_{n=1}^{\infty} \frac{|R_n|}{n} < \infty$, so $\{y_n\}$ is a stochastic approximation. By the same Lemma, $|\xi_n| \leqslant 4m$ where m is the maximal number of balls added in each period. This implies that $\mathbb{E}[\xi_n^+|\mathcal{F}_n], \mathbb{E}[\xi_n^-|\mathcal{F}_n]$ are bounded from above by $4m$. To apply Theorem 2.9, it remains to prove that $\mathbb{E}[\xi_n^+|\mathcal{F}_n], \mathbb{E}[\xi_n^-|\mathcal{F}_n]$ are bounded from below by a positive number when $y_n \in (\hat{y} - \epsilon, \hat{y} + \epsilon)$. From $\xi_n = \xi_n^+ - \xi_n^-$ and $\mathbb{E}[\xi_n|\mathcal{F}_n] = 0$, it follows that $\mathbb{E}[\xi_n^+|\mathcal{F}_n] = \mathbb{E}[\xi_n^-|\mathcal{F}_n]$ so it suffices to find a positive lower bound for $\mathbb{E}[\xi_n^+|\mathcal{F}_n]$ when $y_n \in (\hat{y} - \epsilon, \hat{y} + \epsilon)$ and, by assumption, $r > 0$ is such a lower bound.

■

Appendix C: Comparative Statics

The following three theorems summarize comparative statics with respect to all parameters for the quasi steady states y_S^*, y_I^*, y_M^*.

Theorem 5. *The quasi steady state y_S^* is increasing in ρ, μ, and λ and decreasing in κ and β. There exists $\theta_S \in (0, 1]$ (whose value depends on the other parameters) such* *that y_S^* is decreasing in θ for $\theta < \theta_S$ and increasing in θ for $\theta > \theta_S$. y_S^* is decreasing in δ for δ sufficiently close to $\frac{1}{2}$ and increasing in δ for δ sufficiently close to 1.*

Proof of Theorem 5. Let $r_0 = (\rho_0, \kappa_0, \theta_0, \mu_0, \beta_0, \delta_0, \lambda_0)$ be a vector of parameters and consider $g_S(y)$ as a function $G(y,r) : \mathbb{R}^8 \to \mathbb{R}$. Let $y_0^* \in (0,1)$ be the unique $y \in [0,1]$ that solves

$$G(y_0^*, r_0) = 0. \tag{16}$$

Lemma 4 implies that $G(y, r_0) > 0$ for $y < y_0^*$ and $G(y, r_0) < 0$ for $y > y_0^*$ so it must be the case that $G_y(y_0^*, r_0) \leqslant 0$. Moreover, it cannot be the case that $G_y(y_0^*, r_0) = 0$ because that would imply that y_0^* is a local maximum for $G_y(\cdot, r_0)$ while the proof of Lemma 4 shows that the second derivative of this function (the third derivative w.r.t y of $G(y, r_0)$) is strictly positive over $[0,1]$, so $G_y(y_0^*, r_0) < 0$.

Since $G(y_0^*, r_0) = 0$ and $G_y(y_0^*, r_0) \neq 0$, by the implicit function theorem equation 16 defines a function $y_S^*(r) : \mathbb{R}^7 \to \mathbb{R}$ in some neighborhood of r_0, such that $y_S^*(r)$ is the unique steady state of the ODE $\frac{dy}{dt} = g_S(y)$ in $[0,1]$, and

$$\nabla y_S^*(r_0) = -\frac{1}{G_y(y_0^*, r_0)} \nabla_r G(y_0^*, r_0) \tag{17}$$

Furthermore, since $G_y(y_0^*, r_0) < 0$, for all $x \in (\rho, \kappa, \theta, \mu, \beta, \delta, \lambda)$: $\text{sign}(\frac{dy^*(r_0)}{dx}) = \text{sign}(G_x(y_0^*, r_0))$. Plugging the probabilities p_S^T, p_S^F from (3) into (6) and rearranging we have,

$$\begin{aligned} G(y,r) = 1 &+ (\rho(1-\theta) - 1 - \kappa)y - \rho(1-\rho)y^2 \\ &+ \frac{\rho y(\theta(1-y))^2}{\beta}\left(\frac{(1-\delta)^2\lambda\mu}{y + 2(1-y)(1-\delta)} + \frac{\delta^2(\lambda\mu - (1-\lambda))}{y + 2\delta(1-y)}\right). \end{aligned}$$

We now solve for the sign of each of the partial derivatives of G.

$\boldsymbol{\rho}$: $G_\rho(y,r) = (1-\theta)y(1-y) + \frac{y(\theta(1-y))^2}{\beta}\left(\frac{\lambda\mu(1-\delta)^2}{y+2(1-y)(1-\delta)} + \frac{\delta^2(\lambda\mu-(1-\lambda))}{y+2\delta(1-y)}\right) > 0.$

$\boldsymbol{\theta}$: $G_\theta(y,r) = \rho y(1-y)\left(\frac{2(1-y)\theta}{\beta}\left(\frac{(1-\delta)^2\lambda\mu}{y+2(1-y)(1-\delta)} + \frac{\delta^2(\lambda\mu-(1-\lambda))}{y+2\delta(1-y)}\right) - 1\right).$

So $G_\theta(y,r) > 0$ if and only if,

$$\theta > \frac{\beta}{2(1-y)}\left(\frac{1}{\frac{(1-\delta)^2\lambda\mu}{y+2(1-y)(1-\delta)} + \frac{\delta^2(\lambda\mu-(1-\lambda))}{y+2\delta(1-y)}}\right).$$

Note that the RHS is always positive, so that for sufficiently small θ, y_S^* is decreasing in θ. However, it is possible that the RHS is below 1 so that for large values of θ the relationship reverses. See Appendix D for an example.

$\boldsymbol{\kappa}$: $G_\kappa(y,r) = -y < 0$.

$\boldsymbol{\mu}$: $G_\mu(y,r) = \frac{\rho y(\theta(1-y))^2}{\beta}\left(\frac{(1-\delta)^2\lambda}{y+2(1-y)(1-\delta)} + \frac{\delta^2\lambda}{y+2\delta(1-y)}\right) > 0$.

$\boldsymbol{\beta}$: $G_\beta(y,r) = -\frac{\rho y(\theta(1-y))^2}{\beta^2}\left(\frac{(1-\delta)^2\lambda\mu}{y+2(1-y)(1-\delta)} + \frac{\delta^2(\lambda\mu-(1-\lambda))}{y+2\delta(1-y)}\right) < 0$.

$\boldsymbol{\delta}$: $G_\delta(y,r) = \frac{2\rho y(\theta(1-y))^2}{\beta(y+2\delta(1-y))^2}\left(\frac{(2\delta-1)\lambda\mu y^2}{(y+2(1-y)(1-\delta))^2} - \delta(1-\lambda)(y+\delta(1-y))\right)$.

So fixing all parameters except δ we have $\operatorname{sign} G_\delta(y,r) = \operatorname{sign}(s(y,\delta))$ where

$$s(y,\delta) := (2\delta-1)\lambda\mu y^2 - (y+2(1-y)(1-\delta))^2\,\delta(1-\lambda)(y+\delta(1-y)).$$

Note that $s(y,1/2) = -\frac{(1-\lambda)(1+y)}{4} < 0$, and $s(y,1) = y^2\left(\lambda\mu-(1-\lambda)\right) > 0$, so y_S^* is decreasing in δ for small values of δ and increasing in δ for large values of δ (recall that we assume $\delta \geqslant \frac{1}{2}$).

$\boldsymbol{\lambda}$: $G_\lambda(y,r) = \frac{vy(q(1-y))^2}{\beta}\left(\frac{(1-\delta)^2 h}{y+2(1-y)(1-\delta)} + \frac{\delta^2(1+h)}{y+2\delta(1-y)}\right) > 0$. ■

Theorem 6. *The quasi steady state y_I^* is increasing in μ and λ and decreasing in κ, β, and δ. y_I^* is increasing in ρ if $\frac{1}{2} > \delta\theta\left(1-a(y,I)\right)$ and decreasing in ρ when the sign is reversed, and both cases can arise in region I. There exists $\theta_I \in (0,1]$ (whose value depends on the other parameters) such that y_I^* is decreasing in θ for $\theta < \theta_I$ and increasing in θ for $\theta > \theta_I$.*

Proof of Theorem 6. By a similar argument as in the proof of Theorem 5 we have for all $x \in (\rho,\kappa,\theta,\mu,\beta,\delta,\lambda)$: $\operatorname{sign}(\frac{dy_I^*(r_0)}{dx}) = \operatorname{sign}(G_x(y_0^*, r_0))$ where now $G(y,r)$ is given by

$$G(y,r) = 1+\left(\rho\left(\frac{1}{2}-\delta\theta\right)-1-\kappa\right)y-\rho\left(\frac{1}{2}-\delta\theta\right)y^2+\rho y(\delta\theta(1-y))^2\frac{(\lambda\mu-(1-\lambda))}{\beta(y+2(1-y)\delta)}.$$

We now solve for the sign of each of the partial derivatives of G.

$\boldsymbol{\rho}$: $G_\rho(y,r) = y(1-y)\left[\frac{1}{2}-\delta\theta\left(1-\frac{(1-y)\delta\theta(\lambda\mu-(1-\lambda))}{\beta(y+2(1-y)\delta)}\right)\right]$.

Let $s(y,r)$ denote the expression in square brackets. Then $\operatorname{sign}(G_\rho(y,r)) = \operatorname{sign}(s(y,r))$ so y_I^* is increasing in ρ if $s(y,r) > 0$ and decreasing in ρ if $s(y,r) < 0$. In Appendix D we show that both are possible and can occur when $y_I^* \in I$.

$\boldsymbol{\theta}$: $G_\theta(y,r) = \delta\rho y(1-y)\left(\frac{2\delta\theta(1-y)(\lambda\mu-(1-\lambda))}{\beta(y+2\delta(1-y))}-1\right)$.

So, $G_\theta(y,r) > 0$ if and only if

$$\theta > \frac{\beta(y + 2\delta(1-y))}{2\delta(1-y)(\lambda\mu - (1-\lambda))}.$$

Note that the RHS is always positive, so that for sufficiently small θ, y_I^* is decreasing in θ. However, it is possible that the RHS is below 1 so that for large values of θ the relationship reverses. See Appendix D for an example.
$\boldsymbol{\kappa}$: $G_\kappa(y,r) = -y < 0$.
$\boldsymbol{\mu}$: $G_\mu(y,r) = \frac{\rho y(\delta\theta(1-y))^2\lambda}{\beta(y+2(1-y)\delta)} > 0$.
$\boldsymbol{\beta}$: $G_\beta(y,r) = -\rho y(\delta\theta(1-y))^2 \frac{(\lambda\mu-(1-\lambda))}{\beta^2(y+2(1-y)\delta)} < 0$
$\boldsymbol{\delta}$: $G_\delta(y,r) = \theta\rho y(1-y)\left[\frac{2\delta\theta(1-y)(y+\delta(1-y))(\lambda\mu-(1-\lambda))}{\beta(y+2\delta(1-y))^2} - 1\right] < 0$.

For the inequality, let $f(y) = \frac{2\delta\theta(1-y)(y+\delta(1-y))(\lambda\mu-(1-\lambda))}{\beta(y+2\delta(1-y))^2} - 1$. It suffices to prove $f(y) < 0$ for all y. This follows from $f(0) = \frac{\theta(\lambda\mu-(1-\lambda))}{2\beta} - 1 < 0$ (by Assumption 2) and $f'(y) = -\frac{2\delta\theta y(\lambda\mu-(1-\lambda))}{\beta(y+2\delta(1-y))^3} < 0$.
$\boldsymbol{\lambda}$: $G_\lambda(y,r) = \frac{\rho y(\delta\theta(1-y))^2(1+\mu)}{\beta(y+2\delta(1-y))} > 0$. ■

Theorem 7. *The quasi steady state y_M^* is increasing in $\mu, \lambda, \rho,$ and δ and decreasing in κ and β. There exists $\theta_M \in (0,1]$ (whose value depends on the other parameters) such that y_M^* is decreasing in θ for $\theta < \theta_M$ and increasing in θ for $\theta > \theta_M$.*

Proof of Theorem 7. By a similar argument as in the proof of Theorem 5 we have for all $x \in (\rho,\kappa,\theta,\mu,\beta,\delta,\lambda)$: $\text{sign}(\frac{dy_M^*(r_0)}{dx}) = \text{sign}(G_x(y_0^*, r_0))$ where now $G(y,r)$ is given by

$$G(y,r) = 1 + \left(\rho\left(\frac{1}{2} - (1-\delta)\theta\right) - 1 - \kappa\right) y - \rho\left(\frac{1}{2} - (1-\delta)\theta\right) y^2 \\ + \frac{\rho y\left((1-\delta)\theta(1-y)\right)^2 \lambda\mu}{\beta(y + 2(1-y)(1-\delta))}.$$

We now solve for the sign of each of the partial derivatives of G.
$\boldsymbol{\rho}$: $G_\rho(y,r) = y\,(1-y)\left(\frac{1}{2} - \theta\,(1-\delta)\left(1 - \frac{\lambda\mu(1-y)(1-\delta)\theta}{\beta(y+2(1-y)(1-\delta))}\right)\right) > 0$.

Indeed, letting $s(y,r)$ denote the expression in square brackets, we have $\text{sign}(G_\rho(y,r)) = \text{sign}(s(y,r))$ and $s(y,r) > 0$ because $(1-\delta) < \frac{1}{2}$ so $s(y,r) = \frac{1}{2} - \theta\,(1-\delta)\,(1 - a(y,M)) > 0$.

$\boldsymbol{\theta}$ $G_\theta(y,r) = \rho y\,(1-\delta)\,(1-y)\left(\frac{2\lambda\mu(1-y)(1-\delta)\theta}{\beta(y+2(1-y)(1-\delta))} - 1\right)$.

So, $G_\theta(y,r) > 0$ if and only if

$$\theta > \frac{\beta\,(y + 2(1-y)(1-\delta))}{2\lambda\mu(1-y)(1-\delta)}.$$

Note that the RHS is always positive, so that for sufficiently small θ, y^*_M is decreasing in θ. However, it is possible that the RHS is below 1 so that for large values of θ the relationship reverses. See Appendix D for an example.

$\boldsymbol{\kappa}$: $G_\kappa(y,r) = -y < 0$.

$\boldsymbol{\mu}$: $G_\mu(y,r) = \frac{\rho y((1-\delta)\theta(1-y))^2\lambda}{\beta(y+2(1-y)(1-\delta))} > 0$.

$\boldsymbol{\beta}$: $G_\beta(y,r) = -\frac{\rho y((1-\delta)\theta(1-y))^2\lambda\mu}{\beta^2(y+2(1-y)(1-\delta))} < 0$.

$\boldsymbol{\delta}$: $G_\delta(y,r) = -\theta\rho y(1-y)\left[\frac{2(1-\delta)(1-y)\lambda\mu\theta(1-\delta(1-y))}{\beta(y+2(1-y)(1-\delta))^2} - 1\right] > 0$.

For the inequality, let $f(y) = \frac{2(1-\delta)(1-y)\lambda\mu\theta(1-\delta(1-y))}{\beta(y+2(1-y)(1-\delta))^2} - 1$. It suffices to prove $f(y) < 0$ for all y. This follows from $f(0) = \frac{\lambda\mu\theta}{2\beta} - 1 < 0$ (by Assumption 2) and $f'(y) = -\frac{2(1-\delta)\lambda\mu\theta y}{\beta(y+2(1-y)(1-\delta))^3} < 0$.

$\boldsymbol{\lambda}$: $G_\lambda(y,r) = \frac{\rho y((1-\delta)\theta(1-y))^2\mu}{\beta(y+2(1-y)(1-\delta))} > 0$.

■

To complete our characterization of comparative statics we now address how the thresholds $\hat{y}_I, \hat{y}_M$ change with the parameters.

Theorem 8. *The thresholds $\hat{y}_I$ and $\hat{y}_M$ are constant in κ and ρ and increasing in θ, μ, and β. Additionally, $\hat{y}_M$ is decreasing in δ and λ and $\hat{y}_I$ is increasing in δ and λ.*

Proof of Theorem 8.

For $X \in \{M, I\}$, let $r_0 = (\rho_0, \kappa_0, \theta_0, \mu_0, \beta_0, \delta_0, \lambda_0)$ be a vector of parameters and consider $V(y,X)$ as a function, $V^X(y,r) : \mathbb{R}^8 \to \mathbb{R}$ (for this proof we use a superscript to distinguish between the two value functions, and subscripts for partial derivatives). Recall that $\hat{y}_I$ is the unique solution $\hat{y}^X_0 \in (0,1)$ to

$$V^X(\hat{y}_0, r_0) = 0. \tag{18}$$

Additionally, recall that by Lemma 2 we have $V^X_y(y,r) > 0$ for $X = M$ and $X = I$. Since $V^X(\hat{y}^X_0, r_0) = 0$ and $V^X_y(\hat{y}^X_0, r_0) \neq 0$, by the implicit function theorem, (18) defines a function $\hat{y}^X(r) : \mathbb{R}^7 \to \mathbb{R}$ in some neighborhood of r_0 and

$$\nabla \hat{y}^X(r_0) = -\frac{1}{V_y^X(\hat{y}_0^X, r_0)} \nabla_r V^X(\hat{y}_0^X, r_0)$$

Furthermore, since $V_y^X(\hat{y}_0^X, r_0) > 0$, for all $m \in (\rho, \kappa, \theta, \mu, \beta, \delta, \lambda)$, $\text{sign}(\frac{d\hat{y}^X(r_0)}{dm}) = \text{sign}(-V_m^X(\hat{y}_0^X, r_0))$. We now use the functional forms of $V(y, M)$ and $V(y, I)$ ((8)) to solve for the sign of each of the partial derivatives of V^X. First, it is immediate that for $X = M, I$ we have $V_\kappa^X(y, r) = V_\rho^X(y, r) = 0$. The remaining partial derivatives are:[28]

$$\begin{aligned}
V_\theta^M(y,r) &= \frac{2\lambda\mu(1-y)(1-\delta)}{y+2(1-y)(1-\delta)} \left(\frac{\lambda\mu\theta(1-y)(1-\delta)}{\beta\left(y+2(1-y)(1-\delta)\right)} - 1 \right) < 0, \\
V_\theta^I(y,r) &= \frac{2(1-y)\delta(\lambda\mu-(1-\lambda))}{y+2(1-y)\delta} \left(\frac{(\lambda\mu-(1-\lambda))\theta(1-y)\delta}{\beta\left(y+2(1-y)\delta\right)} - 1 \right) < 0, \\
V_\mu^M(y,r) &= \frac{2\lambda(1-y)(1-\delta)\theta}{y+2(1-y)(1-\delta)} \left(\frac{\lambda\mu\theta(1-y)(1-\delta)}{\beta\left(y+2(1-y)(1-\delta)\right)} - 1 \right) < 0, \\
V_\mu^I(y,r) &= \frac{2\lambda(1-y)\delta\theta}{y+2(1-y)\delta} \left(\frac{(\lambda\mu-(1-\lambda))\theta(1-y)\delta}{\beta\left(y+2(1-y)\delta\right)} - 1 \right) < 0, \\
V_\beta^M(y,r) &= -\frac{1}{\beta^2} \left(\frac{\lambda\mu\theta(1-y)(1-\delta)}{y+2(1-y)(1-\delta)} \right)^2 < 0, \\
V_\beta^I(y,r) &= -\frac{1}{\beta^2} \left(\frac{(\lambda\mu-(1-\lambda))\theta(1-y)\delta}{y+2(1-y)\delta} \right)^2 < 0, \\
V_\delta^M(y,r) &= \frac{2\lambda(1-y)y}{(y+2(1-y)(1-\delta))^2} \left(1 + \mu\theta \left(1 - \frac{\lambda\mu\theta(1-y)(1-\delta)}{\beta\left(y+2(1-y)(1-\delta)\right)} \right) \right) > 0, \\
V_\delta^I(y,r) &= \frac{2(1-y)y}{(y+2\delta(1-y))^2} \left(\theta(\lambda\mu-(1-\lambda)) \left(\frac{(\lambda\mu-(1-\lambda))\theta(1-y)\delta}{\beta\left(y+2(1-y)\delta\right)} - 1 \right) - 1 \right) < 0, \\
V_\lambda^I(y,r) &= \frac{2(1-y)\delta\theta(1+\mu)}{y+2(1-y)\delta} \left(\frac{(\lambda\mu-(1-\lambda))\theta(1-y)\delta}{\beta\left(y+2(1-y)\delta\right)} - 1 \right) < 0.
\end{aligned} \tag{19}$$

Finally,

$$V_\lambda^M(y,r) = \frac{(y-2\mu(1-y)(1-\delta)\theta)}{y+2(1-y)(1-\delta)} + \frac{2\lambda}{\beta} \left(\frac{\mu(1-y)(1-\delta)\theta}{y+2(1-y)(1-\delta)} \right)^2.$$

[28] The inequalities in (19) hold because by Assumption 2 and $\lambda < 1$: $\frac{\lambda\mu\theta(1-y)(1-\delta)}{\beta(y+2(1-y)(1-\delta))} < \frac{2(1-y)(1-\delta)}{y+2(1-y)(1-\delta)} < 1$, and $\frac{(\lambda\mu-(1-\lambda))\theta(1-y)\delta}{\beta(y+2(1-y)\delta)} < \frac{2(1-y)\delta}{y+2(1-y)\delta} < 1$.

Recall that $V^M(y,r) = \frac{\lambda(y-2\mu(1-y)(1-\delta)\theta)}{y+2(1-y)(1-\delta)} + \frac{1}{\beta}\left(\frac{\lambda\mu(1-y)(1-\delta)\theta}{y+2(1-y)(1-\delta)}\right)^2$. Denote the first summand in this expression by $C(y,r)$ and the second by $D(y,r)$. Since $V^M(\hat{y}_0^M, r_0) = 0$ we have $-C(\hat{y}_0^M, r_0) = D(\hat{y}_0^M, r_0)$. Additionally, since $D(y,r) > 0$ for all y, r it must be the case that $C(\hat{y}_0^M, r_0) < 0$. Finally, note that $V_\lambda^M(\hat{y}_0^M, r_0) = \frac{1}{\lambda}C(\hat{y}_0^M, r_0) + \frac{2}{\lambda}D(\hat{y}_0^M, r_0)$ and since $\frac{2}{\lambda} > \frac{1}{\lambda} > 0$ we have $\frac{2}{\lambda}D(\hat{y}_0^M, r_0) > -\frac{1}{\lambda}C(\hat{y}_0^M, r_0)$ which implies $V_\lambda^M(\hat{y}_0^M, r_0) > 0$. ■

Appendix D: Online Appendix

Differentiation of the functions $a(y, I)$ and $a(y, M)$ shows that :

$$\begin{aligned}
\frac{\partial a(y,M)}{\partial y} &= -\frac{\lambda\mu(1-\delta)\theta}{\beta(y+2(1-y)(1-\delta))^2} < 0, \\
\frac{\partial a(y,I)}{\partial y} &= -\frac{\delta\theta(\lambda\mu-(1-\lambda))}{\beta(y+2(1-y)\delta)^2} < 0, \\
\frac{\partial a(y,M)}{\partial \theta} &= \frac{\lambda\mu(1-y)(1-\delta)}{\beta(y+2(1-y)(1-\delta))} > 0, \\
\frac{\partial a(y,I)}{\partial \theta} &= \frac{(1-y)\delta(\lambda\mu-(1-\lambda))}{\beta(y+2(1-y)\delta)} > 0, \\
\frac{\partial a(y,M)}{\partial \delta} &= -\frac{(1-y)y\lambda\mu\theta}{\beta(y+2(1-y)(1-\delta))^2} < 0, \\
\frac{\partial a(y,I)}{\partial \delta} &= \frac{(1-y)y\theta(\lambda\mu-(1-\lambda))}{\beta(y+2(1-y)\delta)^2} > 0, \\
\frac{\partial a(y,M)}{\partial \beta} &= -\frac{\lambda\mu(1-y)(1-\delta)\theta}{\beta^2(y+2(1-y)(1-\delta))} < 0, \\
\frac{\partial a(y,I)}{\partial \beta} &= -\frac{(1-y)\delta\theta(\lambda\mu-(1-\lambda))}{\beta^2(y+2(1-y)\delta)} < 0, \\
\frac{\partial a(y,M)}{\partial \lambda} &= \frac{\mu(1-y)(1-\delta)\theta}{\beta(y+2(1-y)(1-\delta))} > 0 \\
\frac{\partial a(y,I)}{\partial \lambda} &= \frac{(1-y)\delta\theta(1+\mu)}{\beta(y+2(1-y)\delta)} > 0, \\
\frac{\partial a(y,M)}{\partial \mu} &= \frac{\lambda(1-y)(1-\delta)\theta}{\beta(y+2(1-y)(1-\delta))} > 0, \\
\frac{\partial a(y,I)}{\partial \mu} &= \frac{(1-y)\delta\theta\lambda}{\beta(y+2(1-y)\delta)} > 0,
\end{aligned}$$

where we use Assumption 1, which implies $\lambda\mu - (1-\lambda) > 0$ to sign the partial derivatives of $a(y, I)$.

Below we present numerical examples for claims made in main text. All examples satisfy our standing parametric assumptions, i.e., all parameters are strictly positive, satisfy Assumptions 1 and 2, and $\theta, \lambda < 1$ and $\delta \in (\frac{1}{2}, 1)$.

D.1 Numerical Examples for Section 4.2.1

For a numerical example that the relationships between y_S^* and y_M^* and between y_I^* and y_N^* can go both ways fix $\mu = \beta = \kappa = \rho = 1$ and $\theta = \lambda = 0.75$. Calculations show that $y_M^* < y_S^*$ for $\delta \lessapprox 0.745$ and $y_M^* > y_S^*$ for $\delta \gtrapprox 0.745$. Additionally, $y_N^* < y_I^*$ for $\delta \lessapprox 0.751$ and $y_N^* > y_I^*$ for $\delta \gtrapprox 0.751$. Thus, Lemma 3 is "all we can know" regarding the ordering of the quasi steady states. Likewise, the relationship between the thresholds $\hat{y}_I, \hat{y}_M$ is also undetermined. Calculations with the same parameter values as above show that $\hat{y}_I < \hat{y}_M$ for $\delta \lessapprox 0.664$ and $\hat{y}_I > \hat{y}_M$ for $\delta \gtrapprox 0.664$.

We now show that both of the configurations that give rise to Case a. of Theorem 1 are possible. For an example where $\hat{y}_I < \hat{y}_M$ and $y_I^* < \hat{y}_I < y_N^*$, set $\rho = 20, \theta = 0.9, \kappa = 8, \mu = \beta = 1, \delta = 0.65, \lambda = 0.55$. For an example where $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$ set $\rho = 1, \theta = 0.9, \kappa = 2.4, \mu = \beta = 1, \delta = 0.9, \lambda = 0.65$. It can be verified that in both of these examples $\hat{y}_I$ is the unique stable steady state of the LDI.

D.2 Numerical Examples for Section 4.3

Non-monotonicity in θ:

We now show that each of the quasi steady states y_M^*, y_S^*, y_I^* can be first decreasing and then increasing in θ when it is a steady state for the LDI (and thus a limit point for the system). For y_S^*, set $\rho = 0.3, \kappa = 1.5, \mu = 0.6, \beta = 0.3, \delta = 0.55$, and $\lambda = 0.95$. With these parameters y_S^* is in the sharing region for all $\theta \in (0, 1)$ and is decreasing in θ for $\theta \lessapprox 0.95$ and then increasing.

For y_M^*, set $\rho = 1, \kappa = 8, \mu = 0.6, \beta = 0.3, \delta = 0.9, \lambda = 0.95$. With these parameters $\hat{y}_M < \hat{y}_I$ for all θ, so the intermediate region is M. Additionally, y_M^* is in region M for all $\theta \gtrapprox 0.15$ (otherwise, y_M^* is in region S), and y_M^* is decreasing in θ for $\theta \lessapprox 0.87$ and then increasing in θ. So y_M^* is both decreasing and increasing in θ in region M. Also, for $\theta \gtrapprox 0.16$, y_S^*, y_N^* are also in region M so y_M^* is the unique limit point of the system.

Finally, for y_I^*, set $\rho = 0.45, \kappa = 3, \mu = 0.6, \beta = 0.3, \delta = 0.53, \lambda = 0.9$. With these parameters $\hat{y}_I < \hat{y}_M$ for all θ so the intermediate region is region I. Additionally, y_I^* is in region I for all $\theta \gtrapprox 0.85$ (in region S for smaller θ), and y_I^* is decreasing in θ for $\theta \lessapprox 0.9$ and then increasing in θ. So y_I^* is both decreasing and increasing in θ in region I.

Non monotonicity in δ:

For an example that y_S^* can decrease and then increase in δ when it is a steady state for the LDI, again set $\mu = \beta = \kappa = \rho = 1$ and $\theta = \lambda = 0.75$. With these parameters, $y_S^* > \max\{\hat{y}_I, \hat{y}_M\}$ for all $\delta \in (\frac{1}{2}, 1)$ so that y_S^* is a steady state for the LDI for any value of δ. Additionally, we find that y_S^* is decreasing in δ for $\delta \lessapprox 0.727$ and increasing in δ for $\delta \gtrapprox 0.727$.

Dependence of y_I^* on ρ:

We now show that y_I^* can be either increasing or decreasing in ρ, and both cases can occur when y_I^* is a limit point. Set $\theta = 0.9, \kappa = 3, \mu = \beta = 1, \delta = 0.8, \lambda = 0.55$. With these parameters $\hat{y}_I < \hat{y}_M$ so the intermediate region is I (for any value of ρ). Starting with $\rho = 0$ we have that y_I^* is in region I and it is monotonically decreasing in ρ such that it goes into region N when $\rho \approx 24.5$. However, with the same parameter values but setting $\delta = 0.55$, the intermediate region is still I and we still have $y_I^* \in I$ for $\rho = 0$ but now it is increasing in ρ until and enters region S when $\rho \approx 161.5$. In this example, making false stories more likely to be very interesting reverses the effect of increasing reach.

D.3 Omitted phase diagrams

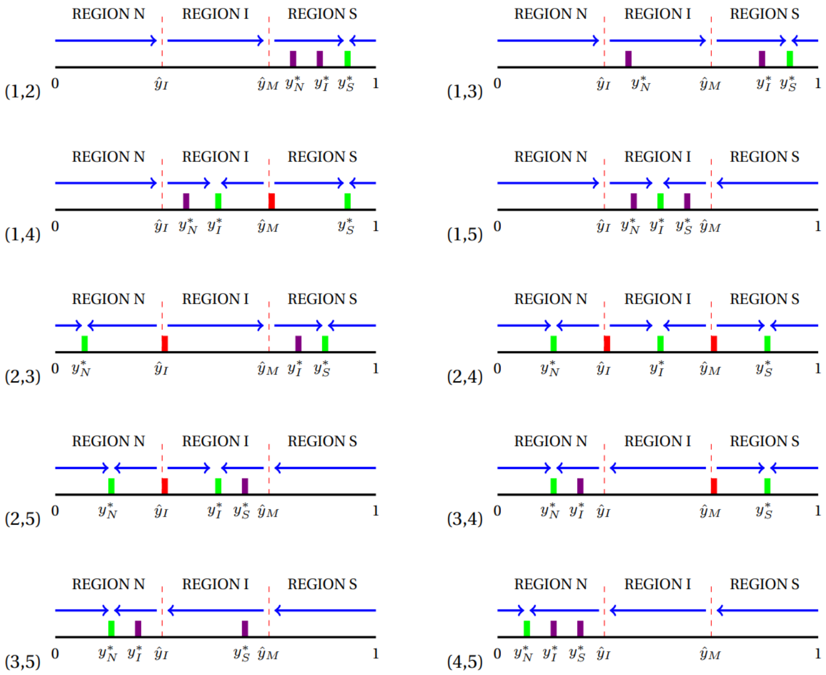
REGION N
REGION I
REGION S
(1,2) 0 $\hat{y}_I$ $\hat{y}_M$ y_N^* y_I^* y_S^* 1
(1,3) 0 $\hat{y}_I$ y_N^* $\hat{y}_M$ y_I^* y_S^* 1
(1,4) 0 $\hat{y}_I$ y_N^* y_I^* $\hat{y}_M$ y_S^* 1
(1,5) 0 $\hat{y}_I$ y_N^* y_I^* y_S^* $\hat{y}_M$ 1
(2,3) 0 y_N^* $\hat{y}_I$ $\hat{y}_M$ y_I^* y_S^* 1
(2,4) 0 y_N^* $\hat{y}_I$ y_I^* $\hat{y}_M$ y_S^* 1
(2,5) 0 y_N^* $\hat{y}_I$ y_I^* y_S^* $\hat{y}_M$ 1
(3,4) 0 y_N^* y_I^* $\hat{y}_I$ $\hat{y}_M$ y_S^* 1
(3,5) 0 y_N^* y_I^* $\hat{y}_I$ y_S^* $\hat{y}_M$ 1
(4,5) 0 y_N^* y_I^* y_S^* $\hat{y}_I$ $\hat{y}_M$ 1

Figure 3: Phase diagrams for the case $\hat{y}_I < \hat{y}_M; y_I^* > y_N^*$.

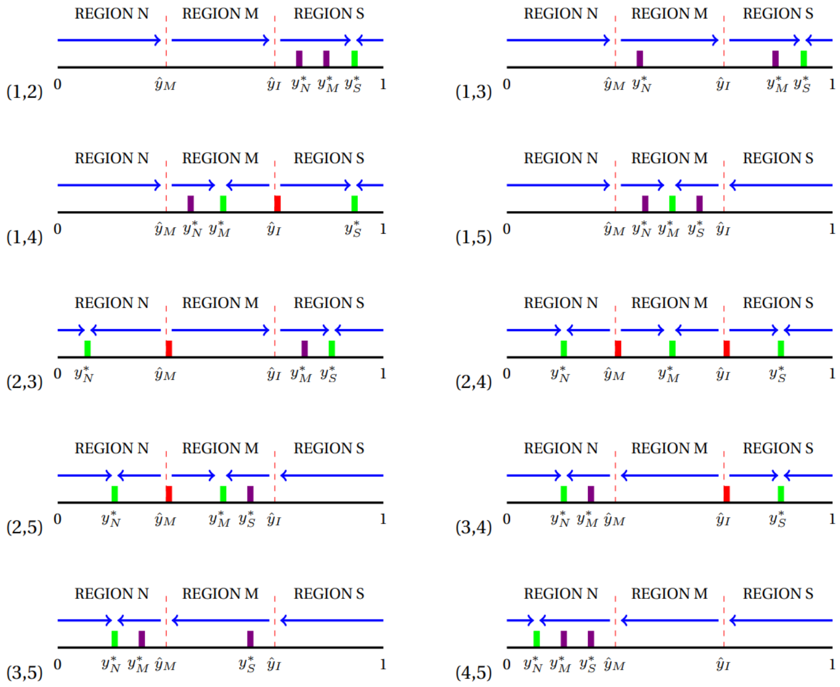
REGION N
REGION M
REGION S
(1,2) 0 $\hat{y}_M$ $\hat{y}_I$ y_N^* y_M^* y_S^* 1
(1,3) 0 $\hat{y}_M$ y_N^* $\hat{y}_I$ y_M^* y_S^* 1
(1,4) 0 $\hat{y}_M$ y_N^* y_M^* $\hat{y}_I$ y_S^* 1
(1,5) 0 $\hat{y}_M$ y_N^* y_M^* y_S^* $\hat{y}_I$ 1
(2,3) 0 y_N^* $\hat{y}_M$ $\hat{y}_I$ y_M^* y_S^* 1
(2,4) 0 y_N^* $\hat{y}_M$ y_M^* $\hat{y}_I$ y_S^* 1
(2,5) 0 y_N^* $\hat{y}_M$ y_M^* y_S^* $\hat{y}_I$ 1
(3,4) 0 y_N^* y_M^* $\hat{y}_M$ $\hat{y}_I$ y_S^* 1
(3,5) 0 y_N^* y_M^* $\hat{y}_M$ y_S^* $\hat{y}_I$ 1
(4,5) 0 y_N^* y_M^* y_S^* $\hat{y}_M$ $\hat{y}_I$ 1

Figure 4: Phase diagrams for the case $\hat{y}_I > \hat{y}_M; y_S^* > y_M^*$.

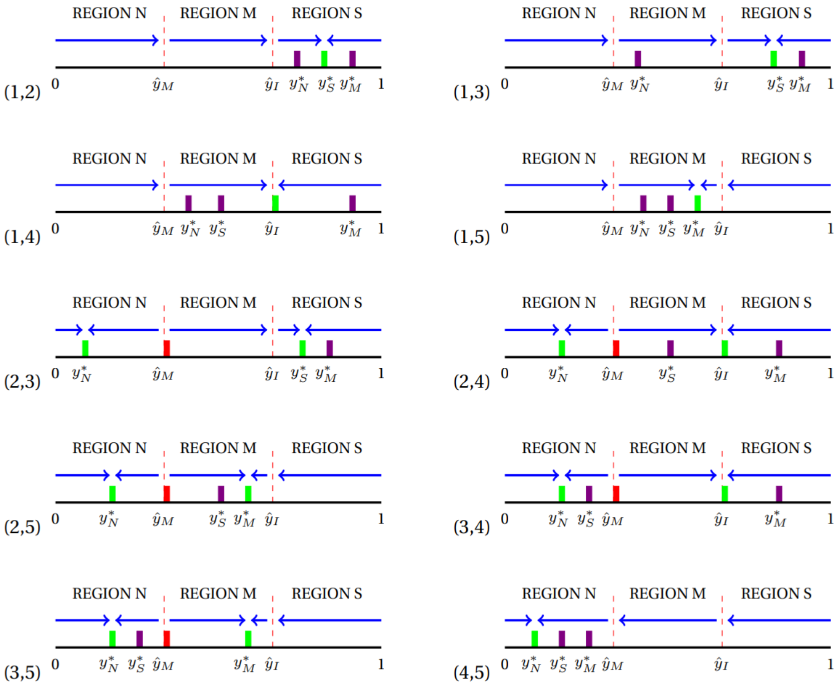

Figure 5: Phase diagrams for the case $\hat{y}_I > \hat{y}_M; y_S^* < y_M^*$.