# Endogenous Attention and the Spread of False News*

Tuval Danenberg† and Drew Fudenberg‡

August 31, 2026

**Abstract**

We study the impact of endogenous attention in a dynamic social media model. Each period, a user observes a random story and decides whether to share it. Users like sharing true and interesting stories, but identifying false stories requires costly attention. Depending on parameters, the system exhibits either a unique limit or strong path dependence. Endogenous attention responds to changes in false story credibility, so reducing credibility can boost their prevalence. Increases in the exogenous production rate of false stories can be amplified by users' sharing decisions. Increasing users' capacity to reach others amplifies both true and false stories; we identify conditions under which the net effect favors truth over falsehood.

**Keywords**: false news, endogenous attention, Polya urns, stochastic approximation, social media
**JEL codes**: D83, D91

*We thank Michel Benaim, Claire Bartolone, Yifan Dai, Krishna Dasaratha, Ben Golub, Kevin He, Navin Kartik, Rachel Kranton, Simon Loertscher, David McAdams, Reed Orchinik, David Rand, Doron Ravid, Noah Siderhurst, Philipp Strack, Alexander Wolitzky, the editor, and three referees for helpful comments and conversations. We thank NSF Grant SES-2417162 for financial support.

†Department of Economics, MIT, Cambridge, MA, 02142, tuvaldan@mit.edu

‡Department of Economics, MIT, Cambridge, MA, 02142, drew.fudenberg@gmail.com

# 1 Introduction

This paper develops a dynamic model of the spread of misinformation on social media. Vosoughi, Roy, and Aral (2018) shows that the spread of falsehoods on social media is mostly due to humans rather than bots, and Pennycook et al. (2021) attributes the sharing of false news to inattention. Motivated by these findings, our model assumes that users want to share true stories, but distinguishing true and false content requires costly attention. Users' attention depends on the prevalence and credibility of false stories: If the share of false stories in their feed is negligible, they are unwilling to spend much effort spotting them, but if that share is significant and the false stories are superficially plausible, users are willing to incur a nontrivial cost to distinguish between true and false content. In turn, attention choices affect the prevalence of false stories as more attentive users are more discerning. We study the resulting joint dynamics of users' attention and platform composition.[1]

In our model, each period a distinct user randomly draws a story from the stories on a social media platform and decides whether to share it. Users consider two factors when evaluating a story: its *veracity*, or truthfulness, and their subjective *interest* in it. Users first observe their interest in a story, and then choose their attention level and pay the corresponding cost. They then receive a binary signal of the story's veracity. One signal realization reveals that the story is false while the other may be sent for both true and false stories. False stories are characterized by a credibility measure that captures how true they appear. Signal precision increases in attention, decreases in credibility, and is supermodular in attention and credibility, so that users pay more attention when credibility is high. If the user decides to share the story, a fixed number of identical copies are added to the platform. Regardless of the sharing decision, fixed numbers of true and false stories are also exogenously added, corresponding to content creation.

We assume that users do not share boring stories, and consider two levels of interest: mildly interesting (M) and very interesting (I). While a story's veracity is fixed throughout time, each user's interest in it is drawn i.i.d (conditional on veracity). This captures the idea that different users will find different stories very interesting. We also assume that false stories are more likely to be very interesting.

[1] By "attention," we mean active scrutiny of content veracity, e.g., checking internal consistency or consulting trusted sources.

Our focus is the share of true stories in the system for each period $n \in \mathbb{N}$, which we denote by $y_n$. As a function of $y_n$, users follow one of four possible decision rules. For each interest level, the rules dictate one of two approaches: Either i) do not pay attention to stories of that level and do not share; or ii) pay attention to those stories and share them as long as they were not revealed to be false. When users do pay attention to a story, their attention level varies with $y_n$ according to a first-order condition that equates the marginal cost of attention with its marginal benefit. When $y_n$ is sufficiently high, the system is in the *sharing* region, where users pay attention to stories of both interest levels. When $y_n$ is low, the system is in the *no-sharing* region, where users do not pay attention to any story. In between, there is an intermediate region, in which the optimal decision rule is either to only pay attention to very interesting stories or only to mildly interesting ones.

By applying stochastic approximation to concatenations of generalized Polya urns, we show that $y_n$ converges almost surely and provide a complete characterization of its limit. (See the technical summary below.) 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 present in any finite platform whenever there are multiple stable steady states (excluding a special case where the platform starts in the no-sharing region and remains there).

The system converges either to a point where users strictly prefer a single decision rule or to one where they are indifferent between two rules. Comparative statics are qualitatively different in these two cases.[2] For example, in the steady states where users strictly prefer one rule, increasing the cost of attention lowers the share of true stories. However, in the steady states where users are indifferent between two rules, the limit share of true stories is increasing in the cost of attention. This occurs because the cost of attention lowers users' payoffs, so the share of true stories required for indifference increases.

False-story credibility can have a non-monotonic effect: when false stories are highly credible, users pay more attention to them. When credibility is high, user responses to an increase in credibility can outweigh the direct effect of this increase, so

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

making false stories less credible can increase their prevalence. We draw two practical implications. First, producers of false stories may choose to produce implausible stories even when credibility is free. Second, platforms that 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, because increasing the share of stories flagged as false leads users to put more trust in stories that were not flagged.

A common view is that social media platforms are hotbeds for false news because users can easily disseminate content to large audiences. However, the ability of each user to reach many others can also increase the spread of true stories. To analyze this trade-off, we define *reach* as the number of copies of a story added to the platform when a user shares it. We find that the effects of increased reach depend on users' *truth-sharing propensity* and *false-sharing propensity*, defined as the probability of sharing a true or false story, respectively, when one is drawn. We say that users are *discerning* if their truth-sharing propensity exceeds their false-sharing propensity. We show that whether users are discerning determines whether sharing increases or decreases the share of true stories. This supports the use of similar measures in Pennycook and Rand (2022) and Guriev et al. (2023) to evaluate attempts to reduce the spread of false news.

When users only share mildly interesting stories or share both mildly and very interesting stories, they are always discerning. This implies that the steady states associated with these decision rules are increasing in the reach parameter, and converge to 1 as the reach parameter increases to infinity. In contrast, when users share only very interesting stories (decision rule $I$), they may or may not be discerning, depending on the parameters. Intuitively, because users have a higher intrinsic benefit from sharing very interesting stories, they may share stories that are relatively likely to be false. If users are not discerning, then their net effect on platform composition is negative. In this case, higher reach is detrimental, degrading the quality of information on the platform. However, this negative effect of reach is self-limiting: as reach increases, the decrease in the steady-state share of true stories makes users more attentive and therefore more discerning, which partially offsets the direct negative effect. For some parameter specifications the steady state for rule $I$ converges to 0 as reach increases, while for others it converges to an interior point that equates users' truth-sharing and false-sharing propensities.

Finally, increases in the production rate of false stories can cause users to stop sharing entirely, amplifying the exogenous shock. Conversely, a significant share of false stories can persist in the limit even as their exogenous inflow vanishes.

## Technical Summary

In the Polya urn model, an urn consists of balls of various colors. In each period one ball is drawn randomly from the urn, and then returned to the urn with one additional ball of the same color. A *generalized Polya urn* (GPU) 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).[3]

If the users' decision rule were fixed instead of depending on $y_n$, our system would be a GPU. In each hypothetical system where users pick one of the four contingently-optimal decision rules and use it for all values of $y_n$, the stochastic approximation arguments in Schreiber (2001) imply that the share of true stories almost surely converges to a steady state of an associated differential equation. As we show below, each of these differential equations has a unique steady state. Hence each hypothetical system has a unique limit share, which we call a *quasi steady state*. Our system is not a GPU because its transition probabilities are discontinuous at the thresholds where the optimal decision rule changes. Therefore we study a process that follows one GPU on each part of a partition of $[0, 1]$ and switches between them at the thresholds, which we call a *piecewise generalized Polya urn* (PGPU).

To characterize our system's limit behavior, we first extend the literature on stochastic approximation of urn models to establish convergence results for two-color PGPUs that may be of independent interest. The limiting object for a PGPU is a *limit differential inclusion* (LDI) $\frac{dy}{dt} \in F(y)$, where $F$ agrees with the limit ODE of the relevant regional GPU in the interior of each part of the partition and, at a threshold, takes every value between the two regional drifts.[4] Theorem 1 shows that the share of balls of one color converges almost surely to a steady state of the LDI. The proof builds on Schreiber (2001) and Benaim, Hofbauer, and Sorin (2005) to show that the long-run behavior of a continuous-time interpolation of the share process almost surely approximates the trajectory of a solution to the LDI: over any fixed forward window, the interpolation is arbitrarily close to such a solution once the starting time

[3] Arthur and Lane (1993), Smith and Sørensen (2020), and Arieli, Babichenko, and Mueller-Frank (2024) analyze GPU models of social learning with two colors and one ball added each period.

[4] A differential inclusion is an equation of the form $\frac{dx}{dt} \in F(x)$ for a set-valued function $F$.

is large enough. Theorem 2, which follows from Pemantle (2007), complements this by showing that under a mild noise condition the share process cannot converge to a repelling steady state of the LDI.

We then apply these results to our system. Theorem 3 characterizes the stable steady states of its LDI: a quasi steady state is a stable steady state if and only if it lies in the region where its associated decision rule is uniquely optimal, and depending on the parameters there may be one additional stable steady state at the threshold where users are indifferent between sharing and not sharing very interesting stories. Theorem 4 then shows that $y_n$ converges almost surely to a stable steady state, and converges to each one with positive probability, except in the case where users never share and the system deterministically converges to the no-sharing steady state.

# 2 Related Literature

**Empirical Evidence.** Pennycook et al. (2021) argues that inattention to veracity is a key reason that users share false stories. Through a combination of survey and field experiments on Twitter (now X), it demonstrates that there is a discrepancy between users' stated preferences for accuracy and their observed sharing behavior, and that interventions designed to shift users' attention toward accuracy significantly improve the quality of the content they subsequently share.

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, and that accuracy has the most impact on sharing. The first two factors correspond to veracity and interest, with the qualification that interest in our model is subjective.

**Theory of Online Misinformation.** Much of the theoretical literature on misinformation studies sequential models in which a single story spreads through a network. In Bloch, Demange, and Kranton (2018), the network consists of biased and unbiased users, and users' evaluations of the story's veracity depend on their beliefs about the part of the network in which the story originated. In Acemoglu, Ozdaglar, and Siderius (2024), the social media platform chooses the network structure to maximize engagement, and users' sharing decisions depend on expectations about subsequent users' tastes but not past user actions. Like us, they find that flagging stories as false can backfire and increase their prevalence. These models do not feature inspection choices or endogenous attention. Papanastasiou (2020) allows costly inspection; whether users choose to inspect depends on their position on a line and their beliefs about the exogenous prevalence of false news.

Other papers have studied the spread of multiple messages about a fixed binary state. In Mostagir and Siderius (2022), each user initially gets an informative message about the state, and then repeatedly transmits their posteriors to their neighbors. Merlino, Pin, and Tabasso (2023) analyzes the mean field of an infinite-population SIS model with two messages corresponding to the two states. Users become "infected" when they encounter a message and choose how much effort to spend to verify it, a form of endogenous attention. Kranton and McAdams (2024) models the production of information, with consumers allocating attention across sources and veracity endogenously determined by producer investments.

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 only care about veracity and do not know the state of the platform. 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.

# 3 Model

We consider an infinite-horizon model of a social media platform. Each story has a fixed *veracity* $v \in \{T, F\}$, with the story being true if $v = T$ and false otherwise. When a user draws a story they observe their *interest level* $\ell$, which is either $M$ (mildly interesting) or $I$ (very interesting).[5] The probabilities of each interest level are:

$$\mathbb{P}(\ell = I | v = T) = \frac{1}{2}; \; \mathbb{P}(\ell = I | v = F) = \delta.$$

We assume that $\delta > \frac{1}{2}$, so false stories are more likely to seem very interesting, 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

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

same credibility.

The platform starts at time $n = 0$ with an initial stock of true and false stories $(T_0, F_0)$. Let $T_n$ and $F_n$ respectively denote the numbers of true and false stories on the platform at the beginning of period $n$. 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.

In each period $n \geqslant 0$, 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 observes $\ell$ and a noisy signal of $v$. The precision of this signal depends on the user's *attention* as will be explained below. The parameter $\rho$ describes the *reach* of shared stories on the platform—if the user decides to share the story, $\rho$ copies of the story are added to the platform. Regardless of the user's decision, one true story and $\kappa$ false stories are posted to the platform, corresponding to original content creation.

In summary, in each period a distinct user does the following in sequence:

1. Draws a story and observes $\ell$.

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

3. Draws a signal whose distribution depends on $a$.

4. Decides whether to share the story. If shared, $\rho$ copies of the story are added.

5. Receives payoffs.

At the end of each period, 1 new true story and $\kappa$ new false stories are added.

The cost 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 $\theta$ 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 $\theta$, users will notice the story is false only if they pay attention. 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 $\theta$, they will identify a false story as false with probability $\mathbb{P}(F'|F) = 1 - \theta + \theta a = 1 - \theta(1-a)$. If the story is true, the user receives the signal $T'$ with certainty, regardless of their attention level. Thus, signal $F'$ reveals the story is false, while after signal $T'$ the user is uncertain about the story's veracity.

Users choose their attention level after observing their interest in the story, knowing the current share of true stories $y_n$. 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. Whether or not they share, users pay the cost $\beta a^2$ of their chosen attention level. If they do not share they get no additional payoff so their total payoff is $-\beta a^2$. If they share a $(v, \ell)$-story their additional payoff is

$$u(v, \ell) = 1 - \mu \mathbb{1}(v = F) + \lambda \mathbb{1}(\ell = I),$$

where we have normalized the payoff to sharing a true, mildly interesting story to 1. The parameter $\mu$ captures the loss from sharing a false story, and the parameter $\lambda$ captures the additional gain from sharing a story that is very interesting. We assume both of these are strictly positive, so in line with the empirical evidence mentioned above, users want to share stories that are true and interesting. We also make the following parametric assumption:

**Assumption 1.** $\mu > 1 + \lambda$.

Assumption 1 implies users will not share very interesting stories they know are false, and therefore will not share any story for which they received the signal $F'$.[6] Hence, for each $\ell$, users either pay no attention and do not share, or they pay attention and share if and only if they receive the signal $T'$. In the latter case, their expected payoff to attention level $a$ is

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

Thus, if users share after $T'$ they will choose the attention level

$$a(y, \ell) := \underset{a \in [0,1]}{\operatorname{argmax}}\, U(a, y, \ell).$$

[6]This assumption is consistent with Chen, Pennycook, and Rand (2023), which finds that the content factor with the strongest positive correlation with sharing intentions is perceived accuracy.

Our second parametric assumption implies that the optimal attention levels are always strictly less than 1.

**Assumption 2.** $(\mu - 1)\theta < 2\beta$.

The unconstrained maximizer of $U(a, y, \ell)$ is highest when $y = 0$ and $\ell = M$, where it equals $(\mu - 1)\theta/(2\beta)$ (Lemma 1); Assumption 2 makes this strictly below 1 (our results continue to hold under a weak inequality). Assumption 1 ensures that the unconstrained maximizer is always nonnegative, so together these imply that optimal attention levels are given by solutions to first-order conditions.

In summary, the model parameters are $(\rho, \kappa, \theta, \mu, \beta, \delta, \lambda)$. We assume throughout that all parameters are strictly positive real numbers, satisfy Assumptions 1 and 2, and that $\theta < 1$ and $\delta \in (\frac{1}{2}, 1)$.[7]

**Modeling Assumptions.** We note three simplifying assumptions. First, users know the current share of true stories $y_n$ when choosing their attention level; this approximates a setting in which the veracity of stories shared a few periods earlier has been revealed and the mix of true and false stories changes slowly. Second, users are not forward looking: they do not internalize how their sharing today affects the platform's future composition. Users' motives to share interesting and true content can be read as a reduced form of motives that depend on subsequent users' reactions, such as sharing to influence others, or as intrinsic, e.g., due to a *self-expression* motive. Third, because interest is an i.i.d taste shock drawn each period conditional on veracity, it varies across users rather than being a fixed property of a story, so the model abstracts from characteristics that make a story interesting to many users. This modeling decision enables us to study the effects of interest and veracity while keeping the dynamics one-dimensional.

# 4 Optimal Attention and the Sharing Decision

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

[7] We treat $\kappa$ and $\rho$ as real for expositional clarity, but for their interpretation as numbers of stories and for the formal urn model they should be rational. All our results hold under this restriction. For rational $\kappa$ and $\rho$ our process is a scaled version of an integer-valued process: letting $q$ be a common denominator and rescaling the state $z_n$ by $q$ turns the per-period additions 1, $\kappa$, $\rho$ into the integers $q$, $q\kappa$, $q\rho$, which yields an integer-valued urn with the same share process $\{y_n\}$.

stories $\{y_n\}$. We first compute how user-optimal attention depends on the state.

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

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

The proof of this and all other results stated in the text are in Appendix A. By Assumptions 1 and 2, $a(y, \ell) < 1$ for all $y$ and $a(y, \ell) > 0$ if $y < 1$ (since at least 1 true and at least $\kappa$ false stories are added each period, $y_n \in (0, 1)$ for all $n \geqslant 1$). When $y = 1$ there is no need to pay attention, so $a(1, M) = a(1, I) = 0$. 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 to 0 the payoff from attention is so low that users do not share any stories and do not pay attention at all.

Our working paper Danenberg and Fudenberg (2026) shows that users pay more attention to stories of both interest levels when false stories are very credible and when the cost to sharing false stories is high. Users pay more attention to very interesting stories when false stories are more likely to be very interesting, and pay less attention to very interesting stories as the payoff to sharing them increases.

The next lemma shows that there are interior *thresholds* $\hat{y}_M, \hat{y}_I$ for each interest 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 threshold users choose the attention level given in Lemma 1 and share if and only if they received the signal $T'$.

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

Users' sharing behavior depends on the share $y_n$ of true stories as shown in Table 1. When $y_n$ is low, the expected value from sharing is negative for both interest levels so users do not share at all. When $y_n$ is high, users share both types of stories, and for intermediate values they share only one type of story. (When the state is exactly at a threshold, users are indifferent between the two associated policies; we assume that they mix between the two with some strictly interior probability.) Note that the

Table 1: **Regions and Decision Rules**

| | |
|---|---|
| $N = (0, \min\{\hat{y}_M, \hat{y}_I\})$ | Don't share any story. |
| $I = (\hat{y}_I, \hat{y}_M)$ | Share only very interesting (with signal $T'$). |
| $M = (\hat{y}_M, \hat{y}_I)$ | Share only mildly interesting (with signal $T'$). |
| $S = (\max\{\hat{y}_M, \hat{y}_I\}, 1)$ | Share both mildly and very interesting (with signal $T'$). |

system always has three regions: the extreme regions $N$ to the left and $S$ to the right, and an intermediate region that is either $I$ or $M$ depending on the ordering of $\hat{y}_I$ and $\hat{y}_M$. Numerical examples in Appendix A.1 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.

# 5 Dynamics

## 5.1 The Discrete-Time Stochastic Process

We begin the analysis of dynamics by describing how the share of true stories evolves in each region. Let $p^T_R(y), p^F_R(y)$ be the probabilities that the user shares a true or false story, respectively, when the current share of true stories is $y$ under the decision rule of region $R \in \{N, I, M, S\}$. These are given by

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

For example, $p^F_I(y) = (1-y)\delta\theta\left(1 - a(y, I)\right)$ 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 region $R \in \{N, I, M, S\}$ let $\{z_{n;R}\}$ be the Markov process that describes how the system would evolve if users followed the decision rule of region $R$ whether or not it is optimal in the current state. Let $\{y_{n;R}\}$ be the corresponding share of true stories. The laws of motion are:[8]

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

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

Each region process $\{z_{n;R}\}$ is a *generalized Polya urn* (GPU): an urn model in which a bounded number of balls is added each period, with the per-period addition probabilities depending only on the current share of color-1 balls. The actual process $\{z_n\}$ is a concatenation of these region processes: when $y_n$ lies in the interior of region $R$, the transition probabilities of $\{z_n\}$ are those of $\{z_{n;R}\}$. We call such a process a *piecewise generalized Polya urn* (PGPU). The next subsection develops the general convergence theory for PGPUs.

## 5.2 Convergence Results for PGPUs

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 $n$.

**Definition 1.** A Markov process $\{z_n\}$ as above is a (two-color) *generalized Polya urn* (GPU) if: [9]

[8] If $z_{n+1} - z_n = (a, b)$ then $y_{n+1} - y_n = \frac{y_n|z_n|+a}{|z_n|+a+b} - y_n = \frac{(1-y_n)a - y_n b}{|z_n|+a+b}$.

[9] This is a restricted version of the definition in Benaim, Schreiber, and Tarres (2004), which allows for any finite number of colors, for removal of balls, and has a weaker version of (ii).

i. Balls are never removed, and in every period at least one and at most $m$ balls are added. Formally, for all $z_n \in \mathbb{Z}_+^2$ and all $z_{n+1}$ with $\Pi(z_n, z_{n+1}) > 0$, we have $z_{n+1}^1 \geqslant z_n^1$, $z_{n+1}^2 \geqslant z_n^2$, and $1 \leqslant |z_{n+1} - z_n| \leqslant m$.

ii. For each $w \in \mathbb{Z}_+^2$ with $|w| \leqslant m$ there is a Lipschitz-continuous map $p^w : [0,1] \to [0,1]$ such that $\Pi(z, z+w) = p^w\left(\frac{z^1}{|z|}\right)$ for all nonzero $z \in \mathbb{Z}_+^2$.

Condition (ii) says that the transition probabilities depend only on the current share of color-1 balls $y = z^1/|z|$, and are Lipschitz continuous in $y$.

Previous work shows that the long-run behavior of a GPU can be analyzed through a suitably chosen deterministic *limit ODE* $\frac{dy}{dt} = g(y)$ that its share of color-1 balls tracks as the system grows. Specifically, Schreiber (2001) shows that the share process of a GPU is a *stochastic approximation* of its limit ODE: it can be written as

$$y_{n+1} - y_n = \frac{1}{|z_n|}\left(g(y_n) + \xi_{n+1} + R_n\right), \tag{6}$$

for some continuous function $g$, where $\xi_n, R_n$ are adapted to the filtration $\{\mathcal{F}_n\}$ generated by the process, $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$, and the remainder terms $R_n$ go to zero and satisfy $\sum_{n=1}^{\infty} \frac{|R_n|}{n} < \infty$ almost surely. Intuitively, as $|z_n|$ grows, the marginal impact of each period's update on the aggregate composition becomes vanishingly small: the remainder terms vanish and the bounded-increment martingale terms average out. As a result, the tail behavior of the share process is increasingly well approximated by a solution of its limit ODE, in a sense made precise below. The explicit formula for $g$ is in (12) in the appendix.

We now formalize the notion of a piecewise generalized Polya urn: a Markov process whose transition kernel, on each part of a finite partition of $[0,1]$, agrees with some GPU.

**Definition 2.** A Markov process $\{z_n\}$ with transition kernel $\Pi$ is a *piecewise generalized Polya urn* (PGPU) if there exist a finite integer $K$, GPUs $\{\{z_{n;k}\}\}_{k=1}^K$ with kernels $\Pi_k$, and an interval partition $\{I_k\}_{k=1}^K$ of $[0,1]$ such that $\Pi(z,z') = \Pi_k(z,z')$ whenever $\frac{z^1}{|z|} \in \operatorname{int}(I_k)$.[10]

The limiting object for a PGPU cannot in general be a single continuous ODE: at a boundary between two parts of the partition the limit ODEs disagree. As stan-

[10] At a boundary point $\hat{y}$ between intervals $I_k$ and $I_{k+1}$, we assume the kernel is $\alpha_{\hat{y}}\Pi_k + (1 - \alpha_{\hat{y}})\Pi_{k+1}$ for some fixed $\alpha_{\hat{y}} \in (0,1)$.

dard stochastic-approximation results for the ODE case require continuity, we instead define the limiting object to be a *limit differential inclusion* (LDI), $\frac{dy}{dt} \in F(y)$ for a set-valued function $F$.[11] This function agrees with the relevant limit ODE inside each part of the partition and, at a boundary, takes all values between the two regional drifts: writing $g_k$ for the limit ODE of GPU $k$,

$$F(y) = \{g_k(y)\} \text{ if } y \in \mathrm{int}(I_k), \qquad F(\hat{y}) = \left[\min\{g_k(\hat{y}), g_{k+1}(\hat{y})\},\ \max\{g_k(\hat{y}), g_{k+1}(\hat{y})\}\right] \tag{7}$$

if $\hat{y}$ is the boundary between adjacent parts $I_k$ and $I_{k+1}$.

The following theorem relates a PGPU to its LDI; a point $y^* \in [0,1]$ is a *steady state* for the LDI if $0 \in F(y^*)$.

**Theorem 1.** *Let $\{z_n\}$ be a PGPU, $\{y_n\}$ its share of color-1 balls, and $F$ the associated limit differential inclusion. Suppose all of the steady states of $F$ are isolated. Then $y_n$ converges almost surely to a steady state of $F$.*

The idea of the result is similar to that for the stochastic approximation of GPUs. While a PGPU is stochastic, the impact of the randomness diminishes as $|z_n|$ grows so that almost surely the tail behavior of the realized path of the PGPU is eventually arbitrarily close to a trajectory of a solution to the deterministic LDI. This implies that the system converges almost surely to a steady state of the inclusion. Still, the limit can be random, with the realized limit steady state depending on the system's initial composition and on early-period fluctuations.

The proof defines a continuous time interpolation $\mathbf{Y}(t)$ of the process $\{y_n\}$, where the time scale is normalized by the number of balls, and shows that $\mathbf{Y}$ is a *perturbed solution* to the LDI: for every fixed horizon $T$, as $t \to \infty$ its path over $[t, t+T]$ becomes arbitrarily close to a solution of the inclusion. Finally, we apply a result in BHS that says that the limit set of any bounded perturbed solution to the inclusion is internally chain transitive for the inclusion. Since the LDI defined above is a one dimensional autonomous inclusion, its chain transitive sets are steady states.

We say that $y^* \in (0,1)$ is a *stable steady state* for the LDI if it is a steady state and there exists $\epsilon > 0$ such that for all $y \in (y^* - \epsilon, y^* + \epsilon) \backslash \{y^*\}$ we have $\mathrm{sign}(x) = \mathrm{sign}(y^* - y)$ for all $x \in F(y)$. A steady state for the LDI is *globally stable*

[11] An alternative would have been to extend results in Hill, Lane, and Sudderth (1980). But it is not obvious how to generalize their arguments to processes with more than two possible increments each period, and the paper's analog of our Theorem 4 requires a continuous transition kernel.

if every solution of the LDI converges to it. A steady state is *repelling* if there exists $\epsilon > 0$ such that for all $y \in (y^* - \epsilon, y^* + \epsilon) \backslash \{y^*\}$ we have $\operatorname{sign}(x) = -\operatorname{sign}(y^* - y)$ for all $x \in F(y)$.

Theorem 1 shows that $y_n$ converges to *some* steady state. We now establish a complementary result: under a mild noise condition, $y_n$ cannot converge to a repelling steady state.

For a PGPU $\{z_n\}$, let $\mathcal{F}_n$ be the $\sigma$-algebra generated by $(z_1, \ldots, z_n)$, and let $\xi_{n+1} := (y_{n+1} - y_n - \mathbb{E}[y_{n+1} - y_n \mid z_n])|z_n|$ be the rescaled martingale increment, with $\xi_n^+ := \max\{0, \xi_n\}$ and $\xi_n^- := -\min\{0, \xi_n\}$.

**Theorem 2.** *Let $\{z_n\}$ be a PGPU with limit differential inclusion $F$, and let $\hat{y}$ be a repelling steady state of $F$. If there exist $\epsilon, r > 0$ such that $\mathbb{E}[\xi_{n+1}^+ \mid \mathcal{F}_n] \geqslant r$ whenever $y_n \in (\hat{y} - \epsilon, \hat{y} + \epsilon)$, then $\mathbb{P}(y_n \to \hat{y}) = 0$.*

Theorem 2 follows from Theorem 2.9 in Pemantle (2007).[12] Intuitively, noise jiggles $y_n$ away from the steady state, and because the steady state is repelling, the drift of the process will tend to move it further away.

## 5.3 Limit ODEs

To apply the PGPU results we identify the LDI for our system, beginning with its region-specific limit ODEs and their steady states.

**Lemma 3.** *For each $R \in \{N, I, M, S\}$, the process $\{z_{n;R}\}$ is a GPU, and $\{y_{n;R}\}$ is a stochastic approximation of the limit ODE $\frac{dy}{dt} = g_R(y)$, where*

$$g_R(y) = 1 + p_R^T(y)\rho - y\big(1 + \kappa + \rho(p_R^T(y) + p_R^F(y))\big). \tag{8}$$

For an intuition for the limit ODEs, note that in each region $R$ the expected number of incoming true stories is $1 + p_R^T(y)\rho$ and the total expected number of incoming stories is $1 + \kappa + \rho\left(p_R^T(y) + p_R^F(y)\right)$, so $g_R(y)$ equals the former minus $y$ times the latter. Thus, according to the limit ODE $\frac{dy}{dt} = g_R(y)$, the share of true stories increases 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.

[12]The result is for stochastic approximation by an ODE, but it can be applied to PGPUs as it does not require that the RHS of the ODE is continuous.

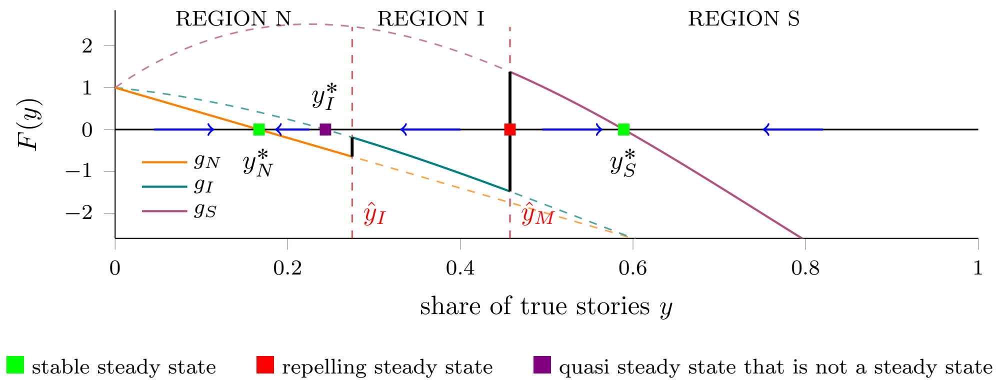


Figure 1: The LDI for $\rho = 40$, $\kappa = 5$, $\theta = 0.9$, $\mu = 2.6$, $\beta = 1$, $\delta = 0.65$, $\lambda = 0.9$. Within each region the LDI is single-valued and coincides with the regional drift $g_R$ (solid); the dashed curves continue each $g_R$ outside its region. At the thresholds the LDI is set-valued, shown by the black segments.

Our system is a PGPU whose partition has three regions: $N$ on the left, $S$ on the right, and one of $I$ or $M$ in the middle (depending on the configuration). Its limit differential inclusion is therefore

$$\frac{dy}{dt} \in F(y) = \begin{cases} \{g_R(y)\}, & y \in \operatorname{int}(R), \\ [\min\{g_R(y), g_{R'}(y)\}, \max\{g_R(y), g_{R'}(y)\}], & y \text{ threshold between } R, R'. \end{cases} \tag{9}$$

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

We call the $y_R^*$ *quasi steady states*. Figure 1 illustrates the phase diagram for the LDI. The geometry of the phase diagram is determined by the relative positions of the thresholds $\hat{y}_I, \hat{y}_M$ and the quasi steady states: the thresholds determine the system's regions, and within each region the flow is towards the corresponding quasi steady state. It is therefore important to understand the possible orderings of the four quasi steady states. Since the exogenous inflow of true and false stories is constant, any differences between the quasi steady states are due to differences in sharing behavior, so to order them we need to understand how the different decision rules lead to different mixes of true and false stories.

**Effects of Users' Sharing Decisions.** To evaluate the effects of the decision rules we introduce the following terms: For decision rule $R$ and any point $y \in (0, 1)$ we will refer to $p_R^T(y)/y$ and $p_R^F(y)/(1-y)$, respectively, as the *truth-sharing propensity* and *false-sharing propensity* for rule $R$ at $y$. These are the probabilities of sharing a true or false story, conditional on one being drawn. Let $d_R(y) := p_R^T(y)/y - p_R^F(y)/(1-y)$ denote the difference between the truth- and false-sharing propensities. We refer to $d_R(y)$ as users' *discernment* at $y$ when they follow rule $R$. We say that users are *discerning* if $d_R(y) > 0$. The following result shows that discernment is a key metric for evaluating the impact of users' sharing decisions.

**Lemma 5.** *For any two decision rules $R, W \in \{N, I, M, S\}$, $y_R^* > y_W^*$ if and only if $d_R(y_R^*) > d_W(y_R^*)$.*

In particular, this lemma shows that if one decision rule consistently implies higher discernment than another, then its corresponding quasi steady state is also higher. Comparing discernment between the different rules we arrive at the following ordering of the quasi steady states. To avoid knife-edge cases we assume that no two variables in $\{\hat{y}_I, \hat{y}_M, y_N^*, y_M^*, y_I^*, y_S^*\}$ are equal.

**Lemma 6.** $\min\{y_S^*, y_M^*\} > \max\{y_I^*, y_N^*\}$.

With decision rules $S$ and $M$ users are always discerning: With rule $M$ users share all stories that are true and mildly interesting, so the truth-sharing propensity is $1/2$, while the false-sharing propensity is below $1-\delta < 1/2$; with rule $S$ users share all true stories but only some false stories. Thus, the quasi steady states for these rules are above the quasi steady state for the no-sharing rule. Since users are discerning when sharing $M$ stories, sharing both $M$ and $I$ stories always generates a larger net increase in $y$ than sharing $I$ stories alone, so the quasi steady state for rule $S$ is above the quasi steady state for rule $I$.

Users' discernment when only sharing $I$ stories is ambiguous, which is why the relationships between $y_S^*$ and $y_M^*$ and between $y_I^*$ and $y_N^*$ cannot be signed. The ambiguity arises because with rule $I$ the truth-sharing propensity is $1/2$, and the false-sharing propensity at $y_I^*$ is $\delta\theta(1 - a(y_I^*, I))$, and either can be larger than the other.[13]

Finally, to see why $y_M^* > y_I^*$, note that under decision rule $I$, users consider sharing more false stories than under $M$ because more false stories are of type $I$.

[13]Numerical examples in Appendix A.1 show that the inequality can go both ways. We revisit this issue when discussing comparative statics of $y_I^*$ with respect to $\rho$ in Section 6.

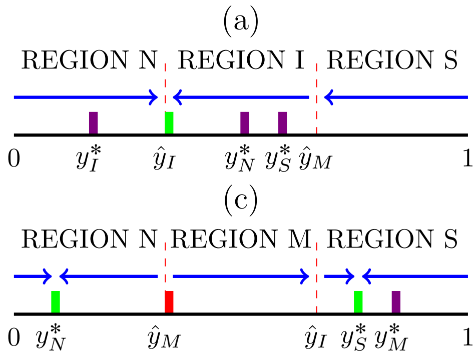


Figure 2: Examples of phase diagrams.

Additionally, they have less of an incentive to avoid sharing false stories because the payoff to sharing $I$ stories is greater. Together, these forces imply that users are more discerning with $M$ content than $I$ content.

Appendix A.1 shows that Lemma 6 is the only restriction on the ordering of the quasi steady states and thresholds. That is, if the relationship between any two of these variables is not determined by Lemma 6 then it can go both ways.[14] While the proof of Lemma 6 relies on our specific functional forms and parameter bounds, the arguments above suggest that the ordering extends to a broader class of models in which signal precision is increasing in attention and the payoff to sharing an $I$ story is greater than the payoff to sharing an $M$ story.

## 5.4 Long-Run Outcomes

We now use the properties of the limit ODEs and the theory of stochastic approximation to determine the long-run behavior of our discrete time stochastic system. Figure 2 presents four examples of phase diagrams of the LDI. As in Figure 1, stable steady states are in green, repelling steady states in red, quasi steady states that are not steady states in violet, and thresholds are marked by dashed lines.

Intuitively, one should expect that all quasi steady states that are within their regions are stable steady states for the LDI, and our next result shows that this is indeed the case. There can be anywhere from 0 to 3 such steady states, as illustrated in Figure 2. Since every limit ODE has a unique steady state, the only other candidate

[14] Danenberg and Fudenberg (2026) explains why this implies that 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 $y_I^*, y_M^*$.

steady states for the LDI are the thresholds; these are steady states where users are indifferent between two rules and randomize between them.

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 $y_I^* < \hat{y}_I < y_N^*$ (as in phase diagram (a) in Figure 2), or $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$. Appendix A.1 shows that both cases are possible, and Lemma 6 implies that flips cannot occur around $\hat{y}_M$. This implies the following characterization of the set $\mathcal{S}$ of stable steady states.

**Theorem 3** (Stable Steady States). *Either (a) $\mathcal{S} = \{y_R^* | y_R^* \in \text{region}\, R\} \cup \{\hat{y}_I\}$, or (b) $\mathcal{S} = \{y_R^* | y_R^* \in \text{region}\, R\}$. 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 $y_S^* < \hat{y}_I < y_M^*$.*

We use the term *limit points* for values to which $y_n$ converges with positive probability. Since behavior in the no-sharing region ($N$) is deterministic—exactly 1 true story and $\kappa$ 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, any stable steady state is a limit point.

**Theorem 4** (Limit Points). *$y_n$ converges almost surely to a point in $\mathcal{S}$. If $y_N^* \in N$ and $y_0 \in N$ then $y_n$ converges to $y_N^*$. Otherwise, for all $y^* \in \mathcal{S}$ there is positive probability that $y_n$ converges to $y^*$.*

Theorem 1 implies that $y_n$ almost surely converges to a steady state of the LDI, so to prove Theorem 4 it remains to show that every stable steady state has positive probability of being the limit point, and that the system almost surely does not converge to a repelling steady state. These claims are established in Lemma 10 and Lemma 11, respectively, in Appendix A. Our assumption that no two variables in $\{\hat{y}_I, \hat{y}_M, y_N^*, y_M^*, y_I^*, y_S^*\}$ are equal implies that any steady state is either stable or repelling, so these two lemmas complete the proof.[15]

To prove Lemma 10, we first show that $y_n$ has positive probability of converging to any $y_R^*$ 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 a counterfactual process that follows the

[15] If a threshold $\hat{y}$ were a steady state that is neither stable nor repelling, the flow must have the same sign on both sides of $\hat{y}$. This would mean the threshold is also a quasi steady state, which we have ruled out.

decision rule of region $R$ everywhere, because that 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) as we show, 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 the quasi steady states 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.

The proof of Lemma 11 applies Theorem 2. The required noise lower bound holds outside region $N$, where the urn updates take three distinct values with probabilities bounded away from 0 and 1. Region $N$ requires separate treatment because dynamics there are deterministic: on the event that $y_n$ enters $N$ infinitely often, the deterministic flow toward $y_N^* \neq \hat{y}$ precludes convergence to $\hat{y}$; on the complementary event, we apply Theorem 2 to an auxiliary process that agrees with $y_n$ outside $N$ and has nontrivial noise inside.

## 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 risk of sharing a false story is high), one in which the share of true stories is intermediate and users share either only mildly interesting stories or only very interesting ones, and one in which the share of true stories is high and users share both very interesting and mildly interesting stories. In other cases, the limit set consists only of the extremes, as in Figure 1, in which $y_n$ converges to either $y_N^* \approx 0.17$, where users do not share at all, or $y_S^* \approx 0.59$, where they share both types of stories. This path-dependence suggests that the long-run outcome can be influenced by shocks to the platform, such as a sudden influx of false stories. These shocks will be more likely to change limit behavior if they occur early, when the overall number of stories is small.

In this context, our model can be interpreted as tracking the entire universe of stories on a platform or channel. Alternatively, it could apply to stories on a single issue (e.g., COVID-19, a presidential election, etc.), if users take into account only the prevalence of false stories on that issue in their sharing decisions. In both cases, our results imply that the nature of early discourse on a channel or issue could have long-lasting implications. A suggestive illustration appears in Zhang et al. (2021), which considers two Reddit communities dedicated to COVID-19 that began with similar user bases and content. Following an early platform-level shock in which one was designated the official community and subjected to stricter moderation, the two rapidly diverged, with the official community attracting science enthusiasts and the other becoming a hub for conspiracy theories.

# 6 Comparative Statics

This section summarizes the comparative statics results of Appendix A.1; Danenberg and Fudenberg (2026) gives the complete set of comparative statics.[16] We begin with the effect of varying the credibility parameter $\theta$.

Table 2: **Comparative Statics for $\theta$**

There are switchpoints $\theta_M, \theta_I, \theta_S \in (0, 1]$ such that:

| | |
|---|---|
| $y^*_M$ | Decreasing for $\theta < \theta_M$ and increasing for $\theta > \theta_M$. |
| $y^*_S$ | Decreasing for $\theta < \theta_S$ and increasing for $\theta > \theta_S$. |
| $y^*_I$ | Decreasing for $\theta < \theta_I$ and increasing for $\theta > \theta_I$. |
| $\hat{y}_I$ | Increasing. |

Numerical examples in Appendix A.1 show that each of the switchpoints in Table 2 can be interior. When they are, the associated quasi steady state is decreasing in $\theta$ up to a point and then increasing.[17] Intuitively, when the credibility of false stories increases it is harder to identify false stories, but this causes users to pay more attention. This leads to two opposing forces on the limit share of true stories, and

[16] Among other results, it shows that the limit share of true stories can either increase or decrease in the probability that false stories are true ($\delta$) and the cost of attention ($\beta$).

[17] The candidate limit point $\hat{y}_I$ behaves differently: Users' payoffs from sharing are decreasing in the credibility of false stories, so $\hat{y}_I$ needs to increase to maintain indifference.

our model predicts that either one can prevail. That is, for sufficiently large values of $\theta$ the increase in attention more than compensates for the increase in credibility.

To illustrate, consider a configuration in which $y_S^*$ is the unique limit point of the system for any value of $\theta$, and the switchpoint $\theta_S$ is interior in $(0, 1)$ (as in Appendix A.1). A false news producer with costs increasing in $\theta$ and payoffs increasing in the limit share of false news would never choose $\theta > \theta_S$, even if credibility were costless. Thus, the overall prevalence of false stories may be higher if all false stories have limited credibility than if they are indistinguishable from the truth. Consequently, false news producers may choose intermediate credibility levels to benefit from the "lulling effect" associated with lower credibility.

Another interpretation of $\theta$ is that the social media platform implements a fact-checking scheme that never mislabels true stories as false, with $\theta$ the probability that a false story is *not* flagged as false. Under this interpretation, the comparative statics results imply that if flagging rates are low ($\theta$ is high), 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.

Under this interpretation, the comparative statics for the quasi steady states $\{y_S^*, y_I^*, y_M^*\}$ are closely related to what Pennycook et al. (2020) calls the "implied truth effect": the idea that, in the presence of flagging, content that is not flagged as false is considered validated. Pennycook et al. (2020) shows how this effect can arise through Bayesian updating, and demonstrates it in experiments. Instead of considering the effect of flagging on independent individual decisions, we consider its effect on limit platform composition. In our setting, the effect is mediated by attention, with users paying less attention to stories that have not been flagged. Our results show that the implied truth effect may outweigh the direct beneficial effect of flagging so that improving flagging rates may increase the prevalence of false stories on the platform. Relatedly, Acemoglu, Ozdaglar, and Siderius (2024) finds that the interaction of the implied truth effect with a platform's engagement-maximizing incentives may lead a regulator to censor less misinformation than is technologically feasible. In that model, the regulator always prefers some censorship to none. In ours, there are cases where some flagging instead of none (i.e., $\theta = 1$) increases the spread of false news.

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.

Table 3: **Comparative Statics for $\rho$**

| | |
|---|---|
| $y^*_M$ | Increasing. Converges to 1 as $\rho \to \infty$. |
| $y^*_S$ | Increasing in $\rho$. Converges to 1 as $\rho \to \infty$. |
| $y^*_I$ | Increasing if $d_I(\frac{1}{1+\kappa}) > 0$, decreasing if the inequality is reversed. Converges to either $0, 1$ or to an interior value as $\rho \to \infty$. |
| $\hat{y}_I$ | Constant. |

The reach parameter $\rho$ has no effect on the location of the threshold $\hat{y}_I$ because it is not an argument in users' payoffs. To understand the effects of $\rho$ on the quasi steady states, recall that $d_R(y)$ is the difference between users' truth- and false-sharing propensities when they follow decision rule $R$ at point $y$. Under decision rules $S$ and $M$, it is always the case that $d_R(y) > 0$: When users follow either rule, the net increase in the share of true stories is larger than when they do not share, which explains why the corresponding quasi steady states increase when the number of copies of each story shared by users increases. However, $d_I(y)$ can be either positive or negative, which is why the effect of reach on the quasi steady state for this rule is ambiguous. We explain below why it suffices to consider users' discernment with rule $I$ at the no-sharing quasi steady state $y^*_N = \frac{1}{1+\kappa}$ to sign the effect of $\rho$ on $y^*_I$.

As the reach parameter increases to infinity, the quasi steady states $y^*_S, y^*_M$ converge to 1 while $y^*_I$ can converge to 0, to 1, or to some interior value $\bar{y} \in (0, 1)$, depending on the parameters. The limit is interior if and only if there exists $\bar{y} \in (0, 1)$ at which $d_I(\bar{y}) = 0$. We find that the effect of attention on $y^*_I$ always works in the opposite direction as the effect of reach, so that the change due to $\rho$ is counterbalanced and may eventually settle down.

To illustrate the effect of $\rho$ on $y^*_I$, suppose that $d_I(y)$ is positive at $y^*_N$, and equal to zero at some interior $\bar{y} > y^*_N$. When $\rho = 0$, all quasi steady states equal $y^*_N = 1/(1 + \kappa)$. Since users are discerning with rule $I$ at $y^*_N$, a marginal increase in $\rho$ increases $y^*_I$. Since attention is decreasing in $y$, so is discernment. Thus, as $y^*_I$ increases in $\rho$, users' discernment decreases. But as long as discernment is positive, $y^*_I$ continues to increase in $\rho$. It cannot increase beyond $\bar{y}$ since at all points above $\bar{y}$ discernment is negative. Hence, $y^*_I$ monotonically converges to $\bar{y}$. This also explains why $1/(1 + \kappa)$ appears in the condition in Table 3: If $y^*_I(\rho)$ is increasing (decreasing) at $\rho = 0$, i.e., when $y_I^* = 1/(1+\kappa)$, then it is increasing (decreasing) for all $\rho$.

In contrast to the parameters discussed above, the local effects of changes in $\kappa$ are obvious: All quasi steady states are strictly decreasing in $\kappa$, converge to 0 as $\kappa \to \infty$ and both thresholds are constant in $\kappa$, since it is not an argument in users' payoffs. This implies that for sufficiently large $\kappa$, the system almost surely converges to the no sharing steady state $y_N^* = \frac{1}{1+\kappa}$. Once false stories are produced quickly enough, users stop sharing altogether and platform composition is determined by the exogenous production rates. The behavior for small $\kappa$ is more subtle, and is the content of the following result.

**Proposition 1.** *Fix all parameters other than $\kappa$ and let $\mathcal{S}_\kappa$ denote the set of stable steady states. For all sufficiently small $\kappa$:*

1. *if the intermediate region is $M$, or $\rho(\delta\theta - \frac{1}{2}) \leqslant 1$, then $\mathcal{S}_\kappa = \{y_S^*\}$;*
2. *otherwise $\mathcal{S}_\kappa$ equals $\{y_S^*\}$, $\{y_S^*, y_I^*\}$, or $\{y_S^*, \hat{y}_I\}$, and all three arise for some parameters. In the last two, a positive share of false stories survives with positive probability even as $\kappa \to 0$.*

Proposition 1 shows that increases in the exogenous production rate of false stories can be amplified by users' sharing decisions. Under the configurations of its first case, there exist $0 < \kappa_1 < \kappa_2$ such that $\mathcal{S}_\kappa = \{y_S^*\}$ for $\kappa < \kappa_1$ and $\mathcal{S}_\kappa = \{y_N^*\}$ for $\kappa > \kappa_2$, so raising the production rate from low to high shifts limit behavior from sharing stories of both interest levels to not sharing at all. The amplification then follows by combining this shift with Lemma 6: since users are discerning under rule $S$, sharing keeps the limit share of true stories above the share among exogenously produced stories, $y_S^* > y_N^* = \frac{1}{1+\kappa}$. Raising production lowers this exogenous share and at the same time drives users to stop sharing, collapsing the limit onto it, so the behavioral response compounds the direct effect.[18] This echoes the "firehose of falsehood," Paul and Matthews (2016)'s term for high-volume propaganda that floods channels with falsehood, spreading it both by crowding out competing messages and by overwhelming the attention audiences would spend scrutinizing it—two channels that arise in our model from sufficient increases in $\kappa$.

Under the first case of Proposition 1, the limit share of true stories tends to 1 as $\kappa \to 0$ and to 0 as $\kappa \to \infty$, the same limits the no-sharing benchmark $y_N^*$ delivers at

[18]Relatedly, changes in $\kappa$ lead to discontinuous jumps in the distribution of $\lim_{n\to\infty} y_n$ when a quasi steady state crosses a threshold so that it (or the threshold) is no longer a limit point.

the extremes. The second case of the proposition shows this is not the only possibility. If the intermediate region is $I$ and $\rho(\delta\theta - \frac{1}{2}) > 1$—so that reach, the credibility of false stories, and their probability of being very interesting are sufficiently large—then either $y_I^*$ or $\hat{y}_I$ may be an additional limit point for small $\kappa$. A significant share of false stories then persists in the limit even as the exogenous inflow of false stories vanishes, sustained by users' poor discernment when sharing very interesting stories.

# 7 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 typically 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.

Beyond path dependence, our comparative statics reveal an asymmetry in how endogenous attention responds to different policy levers: attention counterbalances the direct effect of reducing false-story credibility or increasing flagging rates, but can reinforce the effect of reducing false-story production. This suggests that supply-side interventions targeting producers of false news may be more effective than downstream efforts to limit the spread of false content already circulating on the platform. Our analysis also highlights that higher reach does not inherently lead to a greater spread of false news. Instead, the impact depends on how much users prioritize sharing very interesting stories and the prevalence of false stories within the system.

Our model captures many important features in a tractable framework, and departs from most of the literature by tracking the evolution of the entire platform rather than the spread of a single story. A key modeling decision is treating interest as an idiosyncratic per-period draw, which keeps the dynamics one-dimensional. It would be straightforward to analyze variations that preserve this structure. For instance, Allcott and Gentzkow (2017) shows that education, age, and total media consumption are strongly associated with discernment between true and false content. This user heterogeneity 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 by having both the user's and story's partisanship drawn every period.

Other important features of social media behavior could in principle be handled with similar techniques but a larger state space. Models where some stories are always more interesting or where users care about additional (fixed) story characteristics could be analyzed as a concatenation of urn models with more colors of balls. Extending our stochastic approximation arguments to these settings is straightforward, but analyzing the associated deterministic continuous-time dynamics is more complex as they would be described by differential inclusions in two or more dimensions. Yet other features do not fall within the urn-based formulation described here. For example, our model does not track how often each story has been shared, so it cannot capture virality of individual stories or the "illusory truth" effect of Pennycook, Cannon, and Rand (2018), where users perceive stories they have seen many times as more likely to be true.

# A Proofs

**Proof of Lemma 1.** When $v = T$, then $s = T'$ with probability 1 and $\ell = I$ with probability $\frac{1}{2}$. When $v = F$, then $\ell = I$ with probability $\delta$. 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, $\mathbb{P}_{a,y}(T', T|M) = \dfrac{y}{y + 2(1-y)(1-\delta)}$, $\mathbb{P}_{a,y}(T', F|I) = \dfrac{2(1-y)\delta\theta(1-a)}{y + 2(1-y)\delta}$, and $\mathbb{P}_{a,y}(T', F|M) = \dfrac{2(1-y)(1-\delta)\theta(1-a)}{y + 2(1-y)(1-\delta)}$. We can rewrite (2) as

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

Since $u(T, M) = 1$, and $u(F, M) = 1 - \mu$, we have

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

Similarly, $u(T,I) = 1+\lambda$ and $u(F,I) = 1+\lambda-\mu$ implies that

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

It is clear by inspection that $U(a,y,I)$ and $U(a,y,M)$ are strictly concave in $a$, and they are maximized at $a(y,I), a(y,M)$ respectively as defined in Lemma 1. Finally, $a(y,I), a(y,M) \geqslant 0$ by Assumption 1, and since each denominator is at least $2(1-y)(1-\delta)$ and $2(1-y)\delta$ respectively, both are at most $(\mu-1)\theta/(2\beta) < 1$ by Assumption 2. ∎

**Proof of Lemma 2.** Plugging in the optimal attention levels yields

$$\begin{aligned} V(y,M) &= \frac{y-2(\mu-1)(1-y)(1-\delta)\theta}{y+2(1-y)(1-\delta)} + \frac{1}{\beta}\left(\frac{(\mu-1)(1-y)(1-\delta)\theta}{y+2(1-y)(1-\delta)}\right)^2, \\ V(y,I) &= \frac{(1+\lambda)y-2(\mu-1-\lambda)(1-y)\delta\theta}{y+2(1-y)\delta} + \frac{1}{\beta}\left(\frac{(\mu-1-\lambda)(1-y)\delta\theta}{y+2(1-y)\delta}\right)^2. \end{aligned} \tag{10}$$

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 $V(y_1) = U(a(y_1),y_1) < U(a(y_1),y_2) \leqslant U(a(y_2),y_2) = V(y_2)$. These functions are increasing because

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

where the inequalities follows from Assumption 1. Both $\hat{y}_I, \hat{y}_M$ are interior, because $V(1,M) = 1 > 0, V(1,I) = 1+\lambda > 0$, and, by Assumptions 1 and 2, $V(0,M) = (\mu-1)\theta\left(\frac{(\mu-1)\theta}{4\beta}-1\right) < 0$ and $V(0,I) = (\mu-1-\lambda)\theta\left(\frac{(\mu-1-\lambda)\theta}{4\beta}-1\right) < 0$. ∎

**Proof of Theorem 1.** Let $\{z_n\}$ be a PGPU comprised of GPUs $\{\{z_{n;k}\}\}_{k=1}^K$ with associated interval partition $\{I_k\}_{k=1}^K$, and let $F$ be its limit differential inclusion as in (7). The proof proceeds in three steps. We define a continuous time interpolation of $y_n = z_n^1/|z_n|$, show that the interpolation is a bounded perturbed solution to the limit differential inclusion, and conclude by applying a result of BHS characterizing the limit sets of bounded perturbed solutions.

Define the time scale $\tau_0 = 0$, $\tau_{n+1} = \tau_n + \gamma_{n+1}$ with $\gamma_{n+1} = 1/|z_n|$. The *piecewise affine interpolation* of $\{y_n\}$ is

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

A continuous function $\mathbf{Y} : [0,\infty) \to \mathbb{R}$ is a *perturbed solution* to $F$ if it is absolutely continuous and there are a locally integrable $t \mapsto U(t)$ and a $\delta : [0,\infty) \to \mathbb{R}$ with $\lim_{t\to\infty} \delta(t) = 0$ such that $\lim_{t\to\infty} \sup_{0\leqslant h\leqslant T} |\int_t^{t+h} U(s)ds| = 0$ for all $T > 0$ and $\frac{d\mathbf{Y}(t)}{dt} - U(t) \in F^{\delta(t)}(\mathbf{Y}(t))$ for almost every $t > 0$, where $F^{\delta}(y) = \{x : |x - x'| < \delta$ for some $x' \in F(z)$ with $|z - y| < \delta\}$.[19] The following lemma plays the role for PGPUs that Theorem 2.2 of Schreiber (2001) plays for GPUs.

**Lemma 7.** *Let $\{z_n\}$ be a PGPU and $F$ its limit differential inclusion, and let $\boldsymbol{Y}$ be its piecewise affine interpolation. Then, almost surely, $\boldsymbol{Y}$ is a bounded perturbed solution to $F$.*

We now apply the following result:

**Theorem 3.6 (BHS).** *If $\mathbf{x}$ is a bounded perturbed solution to $F$, the limit set $L(\mathbf{x}) = \bigcap_{t\geqslant 0} \overline{\{\mathbf{x}(s) : s > t\}}$ is internally chain transitive for $F$.*[20]

To apply the results of BHS, we first verify that $F$ satisfies its standing assumptions: (1) $F$ has a closed graph, (2) $F(y)$ is non-empty, compact, and convex for all $y \in [0,1]$, and (3) there exists $c > 0$ such that $\sup_{x\in F(y)} |x| \leqslant c(1 + |y|)$ for all $y \in [0,1]$. Assumptions (1) and (2) are immediate from (7), and (3) follows because $g_k$ are continuous functions on compact intervals.

We now complete the proof. By Lemma 7, $\mathbf{Y}$ is almost surely a bounded perturbed solution to $F$, so Theorem 3.6 of BHS implies that $L(\mathbf{Y})$ is almost surely internally chain transitive for $F$. By the definition of the piecewise affine interpolation, $L(y_n) = L(\mathbf{Y})$, so $L(y_n)$ is almost surely internally chain transitive for $F$. Since the LDI is a one-dimensional autonomous inclusion with isolated steady states, those steady states are its internally chain transitive sets. Thus $y_n$ converges almost surely to a steady state. ∎

[19] BHS define inclusions on all of $\mathbb{R}$. Extend $F$ by $F(y) = \{g_1(0) - y\}$ for $y < 0$ and $F(y) = \{g_K(1) + 1 - y\}$ for $y > 1$; since every GPU has $g(0) \geqslant 0 \geqslant g(1)$ by (12), the extension satisfies the standing assumptions verified below and has no steady states outside $[0,1]$, so we do not distinguish between the two.

[20] BHS extend the definition of internal chain transitivity to differential inclusions.

**Proof of Lemma 7.** Define $g : [0,1] \to \mathbb{R}$ by $g(y) = g_k(y)$ for $y \in \text{int}(I_k)$ and $g(\hat{y}) = \alpha_{\hat{y}} g_k(\hat{y}) + (1-\alpha_{\hat{y}})g_{k+1}(\hat{y})$ at each boundary point $\hat{y}$ between $I_k$ and $I_{k+1}$, where $\alpha_{\hat{y}}$ is the mixing weight from Definition 2; by (7), $g(y) \in F(y)$ for all $y \in [0,1]$. Let $\xi_n$ be the rescaled martingale increments defined before Theorem 2, and let $R_n = |z_n|\mathbb{E}[y_{n+1} - y_n|z_n] - g(y_n)$. Then $\xi_n$, $R_n$ are $\mathcal{F}_n$-measurable, $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$, and

$$y_{n+1} - y_n = \frac{1}{|z_n|}(g(y_n) + \xi_{n+1} + R_n). \tag{11}$$

When $y_n \in \text{int}(I_k)$, $\{z_n\}$ follows the same law of motion as the GPU $\{z_{n;k}\}$; at a boundary point, it follows a convex combination of the two adjacent GPUs' laws of motion. In either case, Lemma 1 in Benaim, Schreiber, and Tarres (2004) yields $K_0 > 0$ such that $|R_n| \leqslant K_0/|z_n|$ and $|\xi_n| \leqslant 4m$, where $m$ is the maximal number of balls added per period.

Since $g(y_n) \in F(y_n)$, (11) gives $y_{n+1} - y_n - \gamma_{n+1}U_{n+1} \in \gamma_{n+1}F(y_n)$ with $U_{n+1} = \xi_{n+1} + R_n$ and $\gamma_{n+1} = 1/|z_n|$, and $\mathbf{Y}$ is the affine interpolation of this recursion on the time scale $\{\tau_n\}$. Because at least one and at most $m$ balls are added each period, $|z_0| + n \leqslant |z_n| \leqslant |z_0| + mn$, so $\gamma_n \to 0$ and $\sum_n \gamma_n = \infty$. Proposition 1.3 of BHS states that the affine interpolation of such a recursion is a perturbed solution to $F$ provided that (i) for all $T > 0$, $\lim_{n\to\infty} \sup\{|S_k - S_n| : k > n \text{ with } \tau_k \leqslant \tau_n + T\} = 0$, where $S_n = \sum_{i=0}^{n-1} \gamma_{i+1}U_{i+1}$, and (ii) $\sup_n |y_n| < \infty$.[21] Condition (ii) holds since $y_n \in [0,1]$. For condition (i), write

$$S_n = M_n + \sum_{i=0}^{n-1} \gamma_{i+1}R_i, \quad \text{where} \quad M_n = \sum_{i=0}^{n-1} \gamma_{i+1}\xi_{i+1}.$$

Since $\gamma_{i+1} = 1/|z_i|$ is $\mathcal{F}_i$-measurable, $\{M_n\}$ is a martingale, and it is bounded in $L^2$ because $\sum_{i=0}^{\infty} \mathbb{E}[\gamma_{i+1}^2 \xi_{i+1}^2] \leqslant 16m^2 \sum_{i=0}^{\infty}(|z_0| + i)^{-2} < \infty$. Hence $M_n$ converges almost surely, so $\sup_{k>n} |M_k - M_n| \to 0$ almost surely. For the remainder terms, $\tau_k \leqslant \tau_n + T$ gives $\sum_{i=n}^{k-1} \gamma_{i+1} = \tau_k - \tau_n \leqslant T$, and $|R_i| \leqslant K_0/(|z_0| + n)$ for all $i \geqslant n$, so $\sum_{i=n}^{k-1} \gamma_{i+1}|R_i| \leqslant K_0T/(|z_0| + n) \to 0$. Condition (i) follows, so $\mathbf{Y}$ is almost surely a perturbed solution to $F$, and it is bounded since $\mathbf{Y}(t) \in [0,1]$. ■

**Proof of Theorem 2.** The proof applies the following result.

[21] BHS take the step sizes $\gamma_n$ to be deterministic, but Proposition 1.3 is a pathwise statement, so once (i) and (ii) hold almost surely it applies along almost every sample path; here $\gamma_n \to 0$ and $\sum_n \gamma_n = \infty$ hold for every path.

**Theorem 2.9 (Pemantle (2007)).** *Suppose $\{x_n\}$ is a stochastic approximation as in (6) except that $g$ need not be continuous.*[22] *Assume that for some $p \in (0,1)$ and $\epsilon > 0$: $\operatorname{sign}(g(x)) = -\operatorname{sign}(p-x)$ for all $x \in (p-\epsilon, p+\epsilon)$. Suppose further that the martingale terms $\xi_n$ are such that $\mathbb{E}[\xi_{n+1}^+|\mathcal{F}_n]$ and $\mathbb{E}[\xi_{n+1}^-|\mathcal{F}_n]$ are bounded above and below by positive numbers when $x_n \in (p-\epsilon, p+\epsilon)$. Then $\mathbb{P}(x_n \to p) = 0$.*

Let $g$ and $R_n$ be as in the proof of Lemma 7, so that (11) holds with $|R_n| \leqslant K_0/|z_n|$ and $|\xi_n| \leqslant 4m$. This has the form of the stochastic approximation (6), except that $g$ is generally not continuous. Since $|z_n| \geqslant |z_0| + n$, we have $\sum_{n=1}^{\infty} |R_n|/n \leqslant K_0 \sum_{n=1}^{\infty} 1/(n(|z_0|+n)) < \infty$, so $\{y_n\}$ is a stochastic approximation. The bound $|\xi_n| \leqslant 4m$ gives $\mathbb{E}[\xi_{n+1}^+|\mathcal{F}_n], \mathbb{E}[\xi_{n+1}^-|\mathcal{F}_n] \leqslant 4m$.

It remains to verify the hypotheses of Pemantle's Theorem 2.9 at $p = \hat{y}$. The sign condition follows from $\hat{y}$ being repelling: shrinking $\epsilon$ if necessary, $\operatorname{sign}(g(y)) = -\operatorname{sign}(\hat{y}-y)$ for $y \in (\hat{y}-\epsilon, \hat{y}+\epsilon)\backslash\{\hat{y}\}$. This outward sign in the punctured neighborhood is sufficient for nonconvergence; any non-zero drift exactly at $\hat{y}$ simply provides an additional force pushing the process away. For the noise condition, the hypothesis gives $\mathbb{E}[\xi_{n+1}^+|\mathcal{F}_n] \geqslant r$ when $y_n \in (\hat{y}-\epsilon, \hat{y}+\epsilon)$. Since $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$, $\mathbb{E}[\xi_{n+1}^+|\mathcal{F}_n] = \mathbb{E}[\xi_{n+1}^-|\mathcal{F}_n]$, so the same lower bound holds for $\mathbb{E}[\xi_{n+1}^-|\mathcal{F}_n]$. Pemantle's Theorem 2.9 then yields $\mathbb{P}(y_n \to \hat{y}) = 0$. ∎

**Proof of Lemma 3.** We verify the conditions of Definition 1 for each region $R$. Condition (i) follows directly from (4): each increment vector adds at least one and at most $m = 1 + \kappa + \rho$ balls. For condition (ii), the law of motion (4) gives non-zero transition probabilities only for the three increment vectors $w_1 = (1+\rho, \kappa)$, $w_2 = (1, \kappa+\rho)$, and $w_3 = (1, \kappa)$, with the corresponding maps $p_R^T$, $p_R^F$, and $1 - p_R^T - p_R^F$ from (3). These maps are Lipschitz-continuous and depend only on $y = z^1/|z|$, so condition (ii) holds.

That $\{y_{n;R}\}$ is a stochastic approximation of the limit ODE $\frac{dy}{dt} = g_R(y)$ then follows from Schreiber (2001), with $g_R$ given by

$$g_R(y) = \sum_{w \in \mathbb{Z}_+^2} p_R^w(y) \left(w^1 - y|w|\right), \tag{12}$$

where $p_R^w$ denotes the Lipschitz map associated with increment $w$ (we use the repre-

[22] In Pemantle (2007) the result is stated for a process with step size $1/n$ but the proof extends to any step size that is $\Theta(1/n)$ almost surely.

sentation of the limit ODE from Benaim, Schreiber, and Tarres (2004)). Substituting $w_1, w_2, w_3$ and their associated maps and rearranging gives (8). ■

**Proof of Lemma 4.** First, note that by the definition of $g_R(y)$ in (8), for all $R \in \{N, I, M, S\}$ 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 $y_N^* = \frac{1}{1+\kappa}$. 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

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

By Assumption 1, 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 $R$. ■

**Proof of Lemma 5.** Fix two decision rules $R, W \in \{N, I, M, S\}$. By Lemma 4, $y_R^* > y_W^*$ if and only if $g_R(y_R^*) = 0 > g_W(y_R^*)$. By (8), for all $y \in [0,1]$,

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

So $g_R(y_R^*) > g_W(y_R^*)$ if and only if $(1-y_R^*)\left(p_R^T(y_R^*) - p_W^T(y_R^*)\right) > y_R^*\left(p_R^F(y_R^*) - p_W^F(y_R^*)\right)$. By Lemma 4, all quasi steady states are interior in $[0,1]$; rearranging shows $y_R^* > y_W^* \iff g_R(y_R^*) > g_W(y_R^*) \iff d_R(y_R^*) > d_W(y_R^*)$. ■

**Proof of Lemma 6.** By Lemma 4, to prove $y_R^* > y_W^*$ for $R, W \in \{N, I, M, S\}$, it suffices to prove that for all $y \in (0,1)$: $d_R(y) > d_W(y)$. Fix $y \in (0,1)$. We will show that $\min\{d_S(y), d_M(y)\} > \max\{d_I(y), d_N(y)\}$. By (3), $d_S(y) = 1 - \theta\left(1 - \delta a(y,I) - (1-\delta)a(y,M)\right)$, $d_M(y) = \frac{1}{2} - (1-\delta)\theta\left(1 - a(y,M)\right), d_I(y) = \frac{1}{2} - \delta\theta\left(1 - a(y,I)\right)$ and $d_N(y) = 0$.

So $\min\{d_S(y), d_M(y)\} > d_N(y)$ follows from the assumptions $\delta > \frac{1}{2}, \theta < 1$ and since by Lemma 1 both attention levels are nonnegative. Similarly, $d_S(y) > d_I(y)$ if and only if $\frac{1}{2} > \theta(1-\delta)(1-a(y,M))$, which always holds. Finally, to prove that $d_M(y) > d_I(y)$ it suffices to prove $(1-\delta)(1-a(y,M)) < \delta(1-a(y,I))$. Let $q(\delta) = (1-\delta)(1-a(y,M))\,; r(\delta) = \delta(1-a(y,I))$. We will prove $q(\delta) < r(\delta)$ for all $\delta \in [\frac{1}{2}, 1)$ by showing that $q(\frac{1}{2}) < r(\frac{1}{2})$ and $q(\delta)$ is decreasing in $\delta$ while $r(\delta)$ is increasing in $\delta$. First,

$$r(1/2) = \frac{1}{4}\left(2 - \frac{\theta(1-y)(\mu-1-\lambda)}{\beta}\right) > \frac{1}{4}\left(2 - \frac{\theta(1-y)(\mu-1)}{\beta}\right) = q(1/2).$$

Next,

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

Assumption 2 and $\lambda > 0$ imply that $\theta(\mu-1-\lambda) < \theta(\mu-1) < 2\beta$; simple algebra then shows that $\frac{\partial q(\delta)}{\partial \delta} < 0$ and $\frac{\partial r(\delta)}{\partial \delta} > 0$. ∎

**Proof of Theorem 3.** Any quasi steady state that lies in its associated region is a stable steady state of the LDI; the only other candidates are thresholds. A threshold $\hat{y}$ is stable iff $\text{sign}(x) = \text{sign}(\hat{y} - y)$ for all $x \in F(y)$ near $\hat{y}$.

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 of 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 A.1 we show that both are possible. We now show that flips cannot occur around $\hat{y}_M$ so $\hat{y}_M \notin \mathcal{S}$. 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 the 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 6, $y_S^* > y_I^*$. In Case 2 a flip cannot occur because by Lemma 6, $y_M^* > y_N^*$. ∎

**Proof of Theorem 4.** When $y_N^* \in N$ and $y_0 \in N$, the system follows the law of motion $z_{n+1} = z_n + (1, \kappa)$, so it never leaves the region $N$ and converges deterministically to $y_N^* = \frac{1}{1+\kappa}$. We henceforth assume that $y_N^* \notin N$ and/or $y_0 \notin N$. By Theorem 1, $y_n$ converges almost surely to a steady state of the LDI. By Lemma 10 below, when $y_N^* \notin N$ and/or $y_0 \notin N$ there is positive probability of convergence to any stable steady state, and by Lemma 11 there is zero probability of convergence to any repelling steady state. ∎

Lemma 10 and Lemma 11 below are used to prove Theorem 4, and Lemmas 8 and 9 are used to prove Lemma 10.

**Lemma 8.** *Let $\epsilon > 0$ and $y \notin N$ such that $y \in (\frac{1}{1+\kappa+\rho}, \frac{1+\rho}{1+\kappa+\rho})$. Starting from any state $z_n$ with $y_n \notin N$, the system has positive probability of arriving at some $y_m \in B_\epsilon(y)$.*

**Proof.** Since at most $1+\kappa+\rho$ stories are added each period, $|y_{n+1} - y_n| \leqslant (1+\kappa+\rho)/|z_n|$, so there exists $n_\epsilon \in \mathbb{N}$ such that $|y_{n+1} - y_n| < \epsilon$ whenever $|z_n| > n_\epsilon$. Since $|z_n| \to \infty$, we may assume $|z_n| > n_\epsilon$ (any finite sequence of consecutive true-story shares occurs with positive probability and keeps $y_n$ outside $N$ while $|z_n|$ grows: each share moves $y_n$ toward $\frac{1+\rho}{1+\kappa+\rho}$ without crossing it, and $\frac{1+\rho}{1+\kappa+\rho} > y \notin N$). Suppose that every subsequent user shares a true story if $y_n < y$ and a false story if $y_n > y$. In the first case $y_n < y < \frac{1+\rho}{1+\kappa+\rho}$, so $y_k$ increases monotonically toward $\frac{1+\rho}{1+\kappa+\rho} > y$; in the second case $y_n > y > \frac{1}{1+\kappa+\rho}$, so $y_k$ decreases monotonically toward $\frac{1}{1+\kappa+\rho} < y$. In either case, since per-period increments are below $\epsilon$, the trajectory enters $B_\epsilon(y)$ at some finite period; let $n+T$ be the first such period. Up to period $n+T$ the trajectory remains in the interval between $y_n$ and $y$, and since $y_n \notin N$, $y \notin N$, and $N$ is the leftmost region, this interval is disjoint from $N$. At any state outside $N$ where $y < 1$, there is positive probability of drawing and sharing both true and false stories (at the threshold adjacent to region $N$ this is true by the strictly interior mixing weight in Definition 2). Thus, the required $T$ consecutive shares occur with positive probability (recall that exogenous additions ensure $y_n \in (0,1)$ for $n \geqslant 1$). ∎

**Lemma 9.** *Let $\{z_n\}$ be a PGPU, let $F$ be its limit differential inclusion, and let $\psi \in (0,1)$ and $\epsilon > 0$ be such that $\operatorname{sign}(x) = \operatorname{sign}(\psi - y)$ for all $y \in B_\epsilon(\psi)\backslash\{\psi\}$ and all $x \in F(y)$. Then there exists $m_\psi \in \mathbb{N}$ such that for every $n$ and every state $z$ with $|z| > m_\psi$ and $z^1/|z| \in B_{\epsilon/2}(\psi)$,*

$$\mathbb{P}(y_m \in B_\epsilon(\psi)\ \forall m \geqslant n \mid z_n = z) \geqslant 1/2.$$

**Proof.** Since $\{z_n\}$ is a time-homogeneous Markov chain, it suffices to consider $n = 0$ and a fixed initial state $z_0 = z$ with $y_0 \in B_{\epsilon/2}(\psi)$ and $|z| > m_\psi$. Because at least one ball is added every period, $|z_k| \geqslant |z| + k$ for all $k$.

By (11), $y_{k+1} - y_k = \frac{1}{|z_k|}(g(y_k) + \xi_{k+1} + R_k)$, where, as in the proof of Lemma 7, $\mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0$, $|\xi_{k+1}| \leqslant 4m$ for $m$ the maximal number of balls added per period, $|R_k| \leqslant K_0/|z_k|$, and $g(y) \in F(y)$ for all $y$. Since at most $m$ balls are added per period, each period the share moves by at most $m/|z_k|$.

Let $M_0 = 0$ and $M_j = \sum_{k=0}^{j-1} \xi_{k+1}/|z_k|$ for $j \geqslant 1$. Then $\{M_j\}$ is a martingale with $\sup_j \mathbb{E}[M_j^2] = \sum_{k=0}^{\infty} \mathbb{E}[(\xi_{k+1}/|z_k|)^2] \leqslant 16m^2 \sum_{k=0}^{\infty}(|z| + k)^{-2} \leqslant 16m^2/(|z| - 1)$, and similarly $\sum_{k=0}^{\infty} |R_k|/|z_k| \leqslant K_0 \sum_{k=0}^{\infty}(|z| + k)^{-2} \leqslant K_0/(|z| - 1)$. Choose $m_\psi$ large enough that $|z| > m_\psi$ implies: (a) $m/|z| < \epsilon/16$; (b) $K_0/(|z| - 1) < \epsilon/16$; and (c) $(16/\epsilon)^2 \cdot 16m^2/(|z| - 1) \leqslant 1/2$. By Doob's maximal inequality and (c), the event $G = \{\sup_{j \geqslant 0} |M_j| < \epsilon/16\}$ satisfies $\mathbb{P}(G) \geqslant 1/2$.

We claim that on $G$ we have $y_k \in B_\epsilon(\psi)$ for all $k \geqslant 0$, which proves the lemma. If not, let $\tau = \min\{k : |y_k - \psi| \geqslant \epsilon\}$, and let $\sigma = \max\{k < \tau : |y_k - \psi| \leqslant \epsilon/2\}$, which is well defined because $y_0 \in B_{\epsilon/2}(\psi)$. Assume w.l.o.g. that $y_{\sigma+1} > \psi$. By the definitions of $\sigma$ and $\tau$, $\epsilon/2 < |y_k - \psi| < \epsilon$ for all $\sigma < k < \tau$, and by (a) all these $y_k$ lie on the same side of $\psi$ as $y_{\sigma+1}$, since crossing from one side of $B_{\epsilon/2}(\psi)$ to the other would require a single-period move larger than $\epsilon$. So $y_k \in (\psi + \epsilon/2, \psi + \epsilon)$, where $g(y_k) < 0$ by hypothesis, for all $\sigma < k < \tau$. Summing the increments from period $\sigma + 1$ to period $\tau$ on the event $G$,

$$y_\tau = y_{\sigma+1} + \sum_{k=\sigma+1}^{\tau-1} \frac{g(y_k)}{|z_k|} + M_\tau - M_{\sigma+1} + \sum_{k=\sigma+1}^{\tau-1} \frac{R_k}{|z_k|} < \psi + \frac{\epsilon}{2} + \frac{\epsilon}{16} + \frac{2\epsilon}{16} + \frac{\epsilon}{16} < \psi + \epsilon,$$

where the inequality uses $y_{\sigma+1} \leqslant y_\sigma + \epsilon/16 \leqslant \psi + \epsilon/2 + \epsilon/16$ by (a), that every drift term is negative, that $|M_\tau - M_{\sigma+1}| < 2\epsilon/16$ on $G$, and (b). Moreover, $y_{\tau-1} \geqslant \psi - \epsilon/2$ because $y_{\tau-1}$ either equals $y_\sigma$ or lies in $(\psi + \epsilon/2, \psi + \epsilon)$, so $y_\tau > \psi - \epsilon/2 - \epsilon/16 > \psi - \epsilon$ by (a). Thus $|y_\tau - \psi| < \epsilon$, contradicting the definition of $\tau$. ■

**Lemma 10.** *Assume that $y_N^* \notin N$ and/or $y_0 \notin N$. If $\psi$ is a stable steady state, there is positive probability that $y_n \to \psi$.*

**Proof.** Let $\psi$ be a stable steady state, and pick $\epsilon > 0$ small enough that $B_\epsilon(y_R^*) \subset R$ if $\psi = y_R^*$ for some region $R$, and $B_\epsilon(\hat{y}_I) \subset L \cup \{\hat{y}_I\} \cup Q$ if $\psi = \hat{y}_I$, where $L$ and $Q$ are the regions adjacent to $\hat{y}_I$ from the left and from the right.

*Step 1: An auxiliary process.* Define a PGPU $\{z_{n;\psi}\}$, with share of true stories $y_{n;\psi} = z^1_{n;\psi}/|z_{n;\psi}|$ and limit differential inclusion $F_\psi$ as in (7), as follows. If $\psi = y^*_R$ for a region $R$, let $\{z_{n;\psi}\}$ be the process $\{z_{n;R}\}$ defined in (4), so that $F_\psi(y) = \{g_R(y)\}$ for all $y$. If $\psi = \hat{y}_I$, let $\{z_{n;\psi}\}$ follow the law of motion of $L$ and $Q$ in those regions, while in the third region $O$ exactly one story is added each period—a false story if $O$ lies to the right of $Q$, and a true story if $O$ lies to the left of $L$—so that in $O$ the share moves deterministically toward the nearest other region.

We claim that $\text{sign}(x) = \text{sign}(\psi - y)$ for every $y \neq \psi$ and every $x \in F_\psi(y)$, and that $\psi$ is a steady state of $F_\psi$; together these imply that $\psi$ is the unique steady state of $F_\psi$. If $\psi = y^*_R$, both parts hold by Lemma 4, since global stability of the one-dimensional ODE $\frac{dy}{dt} = g_R(y)$ implies $\text{sign}(g_R(y)) = \text{sign}(y^*_R - y)$ for all $y \neq y^*_R$. If $\psi = \hat{y}_I$, then since $\hat{y}_I$ is a stable steady state of $F$ we have $y^*_Q < \hat{y}_I < y^*_L$, so by Lemma 4, $g_L > 0$ on $L$ and at $\hat{y}_I$, and $g_Q < 0$ on $Q$ and at $\hat{y}_I$; in $O$ the drift is $-y < 0$ if $O$ lies to the right of $Q$ and $1 - y > 0$ if it lies to the left of $L$; and at each threshold $F_\psi(y)$ is the interval spanned by the two adjacent drifts. The sign condition follows for every $y \neq \hat{y}_I$, and $0 \in F_\psi(\hat{y}_I)$ because $g_Q(\hat{y}_I) < 0 < g_L(\hat{y}_I)$.

By Theorem 1, $y_{n;\psi}$ converges to $\psi$ almost surely. By the sign condition, Lemma 9 yields $m_\psi \in \mathbb{N}$ such that for every $n$ and every state $z$ with $|z| > m_\psi$ and $z^1/|z| \in B_{\epsilon/2}(\psi)$, $\mathbb{P}(y_{m;\psi} \in B_\epsilon(\psi)\ \forall m \geqslant n \mid z_{n;\psi} = z) \geqslant 1/2$.

*Step 2: Positive probability of convergence after arriving near $\psi$.* By the choice of $\epsilon$ and the construction of $\{z_{n;\psi}\}$, the transition kernels of $\{z_n\}$ and $\{z_{n;\psi}\}$ coincide at every state whose share lies in $B_\epsilon(\psi)$. Hence, conditional on $z_n = z$ with $|z| > m_\psi$ and $z^1/|z| \in B_{\epsilon/2}(\psi)$, the two processes can be coupled to coincide until the share first exits $B_\epsilon(\psi)$. By Step 1, with probability at least 1/2 no exit occurs; on this event the processes coincide forever, and since $y_{m;\psi}$ converges to $\psi$ almost surely, $y_m \to \psi$.

*Step 3: Positive probability of arriving near $\psi$.* It remains to show that with positive probability the system arrives at a state $z_n$ with $|z_n| > m_\psi$ and $y_n \in B_{\epsilon/2}(\psi)$. By (8), for any region $R$,

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

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 $\rho(\kappa(1 - p^T_R(y^*_R)) + p^F_R(y^*_R)(1 + \rho)) > 0$, which always holds. Since any stable steady state is either a quasi steady state or a threshold bounded above and below by quasi steady states,

$$\frac{1}{1+\kappa+\rho} < \psi < \frac{1+\rho}{1+\kappa+\rho} \quad \forall \psi \in \mathcal{S}. \tag{13}$$

First assume $y_0 \notin N$. If $\psi \notin N$, then as in the proof of Lemma 8, the system reaches a state with $|z_n| > m_\psi$ and $y_n \notin N$ with positive probability, and the claim follows from (13) and Lemma 8, together with the fact that $|z_n|$ is nondecreasing. If $\psi \in N$ (so that $\psi = y_N^*$), an argument analogous to the proof of Lemma 8 shows that consecutive false-story shares, which have positive probability at any state outside $N$, drive the share into $N$ in finite time; from that point on the system never leaves $N$ and converges deterministically to $y_N^* = \psi$. Now assume $y_0 \in N$. Then by hypothesis $y_N^* \notin N$, so in region $N$ the share moves deterministically toward $y_N^* \notin N$ and exits $N$ in finite time, after which the previous case applies. ∎

**Lemma 11.** *The system almost surely does not converge to a repelling steady state.*

**Proof.** Since by Lemma 4 all quasi steady states are stable for their associated ODEs, the only possible repelling steady states for the LDI are the thresholds $\hat{y}_I, \hat{y}_M$. Let $\hat{y}$ be a repelling steady state. Partition into two complementary events:

**Event $A = \{y_n \in N$ infinitely often$\}$.**

If $\hat{y}$ is not adjacent to $N$, then $y_n \in N$ i.o. prevents convergence to $\hat{y}$. If $\hat{y}$ is adjacent to $N$: since $\hat{y}$ is repelling, the repelling property forces $y_N^* \in N$ and $y_N^* \neq \hat{y}$. When $y_n$ enters $N$, the dynamics are deterministic and $y_n \to y_N^*$, precluding convergence to $\hat{y}$. Thus $\mathbb{P}(A \cap \{y_n \to \hat{y}\}) = 0$.

**Event $A^C = \{y_n \in N$ finitely often$\}$.**

To show $\mathbb{P}(A^C \cap \{y_n \to \hat{y}\}) = 0$, we apply Theorem 2, which requires that $\mathbb{E}[\xi_{n+1}^+ \mid \mathcal{F}_n] \geqslant r > 0$ for all $y_n$ near $\hat{y}$. However, when $\hat{y}$ is adjacent to $N$, the noise terms $\xi_n$ vanish on the $N$ side of $\hat{y}$. We therefore apply the theorem to an auxiliary process that has nontrivial noise inside $N$. Write $A^C = \bigcup_{m \in \mathbb{N}} C_m$ where $C_m := \{y_n \notin N$ for all $n \geqslant m\}$; it suffices to show that $\mathbb{P}(C_m \cap \{y_n \to \hat{y}\}) = 0$ for every $m$.

For small $\epsilon > 0$, let $U := (\hat{y} - \epsilon, \hat{y} + \epsilon)$ intersect only the two regions adjacent to $\hat{y}$; when $\hat{y}$ is adjacent to $N$, take $\epsilon$ small enough that also $U \subset (y_N^*, 1)$, which is possible since $y_N^* \in N$ as noted above. Fix $m$ and let $\{z_n'\}_{n \geqslant m}$ be an auxiliary PGPU with $z_m' = z_m$, where $y_n'$, $\mathcal{F}_n'$, and $\xi_n'$ are defined as for $\{z_n\}$. Its transition probabilities agree with those of $\{z_n\}$ outside $N$; inside $N$, they are such that the right-hand side of its limit ODE in $N$ is negative on $N \cap U$ and $\mathbb{E}[\xi'^{+}_{n+1} \mid \mathcal{F}'_n]$ is bounded below by a positive constant whenever $y'_n \in N \cap U$, for all sufficiently large $n$ (e.g., adding $w_2 = (1, \kappa + \rho)$ or $w_3 = (1, \kappa)$ with probability $\frac{1}{2}$ each). Since the transition probabilities of the two processes coincide outside $N$, we can couple them so that $z'_n = z_n$ until the first time $n \geqslant m$ at which $y_n \in N$. On $C_m$ no such time exists, so $z'_n = z_n$ for all $n \geqslant m$, and hence $C_m \cap \{y_n \to \hat{y}\} \subseteq \{y'_n \to \hat{y}\}$. It therefore suffices to show that $\mathbb{P}(y'_n \to \hat{y}) = 0$, which we do by verifying the hypotheses of Theorem 2 for $\{z'_n\}$: (i) $\hat{y}$ is a repelling steady state of its limit differential inclusion, and (ii) the noise bound holds near $\hat{y}$.

For (i), let $F'$ be the limit differential inclusion of $\{z'_n\}$. Outside the closure of $N$, $F'$ coincides with $F$, and on $N \cap U$ it is the singleton given by the auxiliary limit ODE, which is negative by construction. Since $\hat{y}$ is repelling for $F$ and $N \cap U$, when nonempty, lies to the left of $\hat{y}$, it follows that $\operatorname{sign}(x) = -\operatorname{sign}(\hat{y} - y)$ for all $y \in U \backslash \{\hat{y}\}$ and all $x \in F'(y)$, and that $0 \in F'(\hat{y})$. Hence $\hat{y}$ is a repelling steady state of $F'$. For (ii), with $\Delta_T > \Delta_O > \Delta_F$ the three increments from (5), $\mathbb{E}[\xi'^{+}_{n+1} \mid \mathcal{F}'_n] \geqslant p^T_R(y'_n)(1 - p^T_R(y'_n))(\Delta_T - \Delta_O)|z'_n|$ when $y'_n \in U$ is in a region $R \neq N$. Since $p^T_R \in \{y'_n, y'_n/2\}$ is bounded away from 0 and 1 on $U$, and $(\Delta_T - \Delta_O)|z'_n| \geqslant (1 - y'_n)\rho/4$ for large $|z'_n|$, this is bounded below by some $r > 0$. When $y'_n = \hat{y}$, the probability of sharing a true story lies in $(0, 1)$ by Definition 2, so $\mathbb{E}[\xi'^{+}_{n+1} \mid \mathcal{F}'_n]$ is again bounded below by a positive constant. On $N \cap U$ the bound holds by construction. Theorem 2 yields $\mathbb{P}(y'_n \to \hat{y}) = 0$. ∎

## A.1 Comparative Statics

**Proposition 2.** *The quasi steady state $y^*_S$ is increasing in $\rho$. There exists $\theta_S \in (0, 1]$ (whose value depends on the other parameters) such that $y^*_S$ is decreasing in $\theta$ for $\theta < \theta_S$ and increasing in $\theta$ for $\theta > \theta_S$.*

**Proof.** 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

$$G(y^*_0, r_0) = 0. \tag{14}$$

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 (14) locally 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

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

Furthermore, since $G_y(y_0^*, r_0) < 0$, for all $x \in (\rho, \kappa, \theta, \mu, \beta, \delta, \lambda)$: $\text{sign}(\frac{dy_S^*(r_0)}{dx}) = \text{sign}(G_x(y_0^*, r_0))$. Plugging $p_S^T, p_S^F$ from (3) into (8) and rearranging yields

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

We now solve for the sign of the partial derivatives of $G$ with respect to $\rho$ and $\theta$.

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

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

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

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

$H_S$ is continuously differentiable and bounded away from zero on $(0, 1)$ so for sufficiently small $\theta$, $\theta - H_S(y_S^*(\theta)) < 0$, and thus $y_S^*$ is decreasing in $\theta$. To see that $K_S(\theta) := \theta - H_S(y_S^*(\theta))$ can change sign at most once, note that $\frac{dy_S^*(\bar{\theta})}{d\theta} = 0$ at any $\bar{\theta}$ such that $\bar{\theta} = H_S(y_S^*(\bar{\theta}))$ so $K_S'(\bar{\theta}) = 1 - H_S'(y_S^*(\bar{\theta}))\frac{dy_S^*(\bar{\theta})}{d\theta} = 1 > 0$. Let $\theta_S$ be this upcrossing if it exists, and set $\theta_S = 1$ otherwise. Then $y_S^*$ is decreasing in $\theta$ for $\theta < \theta_S$ and increasing in $\theta$ for $\theta > \theta_S$. An example in Appendix A.1 shows that the cutoff can lie below 1. ∎

**Proposition 3.** *The quasi steady state $y_I^*$ is increasing in $\rho$ if $d_I(y_I^*) > 0$ and decreas-*

*ing in $\rho$ when the sign is reversed, and both cases can arise in region $I$.*[23] *There exists $\theta_I \in (0,1]$ (whose value depends on the other parameters) such that $y_I^*$ is decreasing in $\theta$ for $\theta < \theta_I$ and increasing in $\theta$ for $\theta > \theta_I$.*

**Proof.** By a similar argument as in the proof of Proposition 2, for all $x \in (\rho, \kappa, \theta, \mu, \beta, \delta, \lambda)$: $\text{sign}(\frac{dy_I^*(r_0)}{dx}) = \text{sign}(G_x(y_0^*, r_0))$ where now $G(y, r)$ is given by

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

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

Note that the expression in square brackets is exactly $d_I(y)$, so $\text{sign}(G_\rho(y_I^*, r)) = \text{sign}(d_I(y_I^*))$. In Appendix A.1 we show that both $d_I(y_I^*) > 0$ and $d_I(y_I^*) < 0$ are possible and can occur when $y_I^* \in I$.
$\boldsymbol{\theta}$: $G_\theta(y,r) = \delta\rho y(1-y)\left(\frac{2\delta\theta(1-y)(\mu-1-\lambda)}{\beta(y+2\delta(1-y))}-1\right)$.

So, $G_\theta(y,r) > 0$ if and only if $\theta > \frac{\beta(y+2\delta(1-y))}{2\delta(1-y)(\mu-1-\lambda)}$. By a similar argument as in the proof of Proposition 2, there exists $\theta_I \in (0,1]$ such that $y_I^*$ is decreasing in $\theta$ for $\theta < \theta_I$ and increasing for $\theta > \theta_I$. An example in Appendix A.1 shows that the cutoff can lie below 1. ■

**Proposition 4.** *The quasi steady state $y_M^*$ is increasing in $\rho$. There exists $\theta_M \in (0,1]$ (whose value depends on the other parameters) such that $y_M^*$ is decreasing in $\theta$ for $\theta < \theta_M$ and increasing in $\theta$ for $\theta > \theta_M$.*

**Proof.** By a similar argument as in the proof of Proposition 2 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

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

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

[23] Part 4 of Proposition 6 shows we can replace the condition $d_I(y_I^*) > 0$ here with $d_I(\frac{1}{1+\kappa})$, which is the condition presented in the main text.

For the inequality, let $s(y,r)$ denote the expression in square brackets. Then $\text{sign}(G_\rho(y,r)) = \text{sign}(s(y,r))$ and, $s(y,r) = \frac{1}{2} - (1-\delta)\,\theta\,(1-a(y,M)) > 0$, because $1-\delta < \frac{1}{2}$.

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

So, $G_\theta(y,r) > 0$ if and only if $\theta > \frac{\beta(y+2(1-y)(1-\delta))}{2(1-\delta)(1-y)(\mu-1)}$. By a similar argument as in the proof of Proposition 2, there exists $\theta_M \in (0,1]$ such that $y_M^*$ is decreasing in $\theta$ for $\theta < \theta_M$ and increasing for $\theta > \theta_M$. An example in Appendix A.1 shows that the cutoff can lie below 1. ■

**Proposition 5.** *The threshold $\hat{y}_I$ is constant in $\rho$ and increasing in $\theta$.*

**Proof.** Let $r_0 = (\rho_0, \kappa_0, \theta_0, \mu_0, \beta_0, \delta_0, \lambda_0)$ be a vector of parameters and consider $V(y,I)$ as a function $V(y,r) : \mathbb{R}^8 \to \mathbb{R}$ (here we only consider region $I$, so we drop the region from the function's argument). Recall that $\hat{y}_I$ is the unique solution $\hat{y}_0 \in (0,1)$ to $V(\hat{y}_0, r_0) = 0$. By Lemma 2, the partial derivative of $V$ w.r.t $y$ satisfies $V_y(y,r) > 0$ for all $y \in [0,1]$. Thus, $V_y(\hat{y}_0, r_0) \neq 0$, so by the implicit function theorem the equation $V(\hat{y}_0, r_0) = 0$ defines a function $\hat{y}(r) : \mathbb{R}^7 \to \mathbb{R}$ in some neighborhood of $r_0$ and $\nabla \hat{y}(r_0) = -\frac{1}{V_y(\hat{y}_0, r_0)} \nabla_r V(\hat{y}_0, r_0)$. Furthermore, since $V_y(\hat{y}_0, r_0) > 0$, for all $m \in (\rho, \kappa, \theta, \mu, \beta, \delta, \lambda)$, $\text{sign}(\frac{d\hat{y}(r_0)}{dm}) = \text{sign}(-V_m(\hat{y}_0, r_0))$. It is immediate that $V_\rho(y,r) = 0$ since $\rho$ is not an argument in the users' value functions. Finally

$$V_\theta(y,r) = \frac{2(\mu-1-\lambda)(1-y)\delta}{y+2(1-y)\delta}\left(\frac{(\mu-1-\lambda)(1-y)\delta\theta}{\beta(y+2(1-y)\delta)} - 1\right) < 0,$$

because

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

by Assumption 2 and $\lambda > 0$. ■

**Proposition 6.** *Fix values for the other parameters and let $y_R^*(\rho)$ be the quasi steady state for region $R \in \{I, M, S\}$ as a function of $\rho$, and $y_R^\infty := \lim_{\rho\to\infty} y_R^*(\rho)$. This limit exists for all regions $R$ and:*[24]

1. $y_S^\infty = y_M^\infty = 1$.

[24] Numerical examples below show that all cases described in statements 2-4 are possible, i.e., $y_I^*(\rho)$ can converge to either $0, 1$ or to an interior point, and when the limit is interior it can be either increasing or decreasing towards its limit.

2. *If $d_I(y) > 0$ ($< 0$) for all $y \in (0,1)$, then $y_I^\infty = 1$ ($y_I^\infty = 0$).*

3. *If there exists $\bar{y} \in (0,1)$ such that $d_I(\bar{y}) = 0$, then $y_I^\infty = \bar{y}$.*

4. *If $d_I(\frac{1}{1+\kappa}) > 0$ ($< 0$), then $y_I^*(\rho)$ is strictly increasing (decreasing) toward its limit.*

**Proof.** The following observations hold for all $R \in \{I, M, S\}$: By (8),

$$y_R^*(\rho) = \frac{1 + p_R^T(y_R^*(\rho))\rho}{1 + \kappa + \rho(p_R^T(y_R^*(\rho)) + p_R^F(y_R^*(\rho)))}. \tag{15}$$

and by (3), $p_R^T(y) + p_R^F(y) > 0$ for all $y \in (0,1)$. So if the limit $y_R^\infty$ exists and is in $(0,1)$, then $y_R^\infty = \frac{p_R^T(y_R^\infty)}{p_R^T(y_R^\infty)+p_R^F(y_R^\infty)}$, which simplifies to $d_R(y_R^\infty) = 0$. Thus, if the limit defining $y_R^\infty$ exists then either $y_R^\infty \in \{0,1\}$ or $y_R^\infty \in (0,1)$ and satisfies $d_R(y_R^\infty) = 0$. Additionally, by (15), $y_R^*(0) = \frac{1}{1+\kappa}$.

**Claim 1:** By Propositions 2 and 4, both $y_S^*$ and $y_M^*$ are monotonically increasing in $\rho$, so the limit exists for both of these decision rules. As explained in the main text, for $R = M, S$ users are always discerning, i.e., $d_R(y) > 0$ for all $y \in (0,1)$. So for both of these decision rules $y_R^\infty > y_R^*(0) > 0$ and $y_R^\infty \in \{0,1\}$, which implies $y_S^\infty = y_M^\infty = 1$.

**Claim 2:** By Proposition 3, if $d_I(y) > 0$ ($d_I(y) < 0$) for all $y$ then $y_I^*(\rho)$ is monotone increasing (decreasing). Thus, in both cases the limit exists, and cannot be interior since there does not exist an interior $\bar{y}$ with $d_I(\bar{y}) = 0$. Now, if $d_I(y) > 0$ for all $y$ then $y_I^\infty > y_I^*(0) > 0$ so $y_I^\infty = 1$. If $d_I(y) < 0$ for all $y$ then $y_I^\infty < y_I^*(0) < 1$, so $y_I^\infty = 0$.

**Claims 3 and 4:** Assume that there exists $\bar{y} \in (0,1)$ such that $d_I(\bar{y}) = 0$. We will prove that $y_I^*(\rho)$ converges to $\bar{y}$, and that $y_I^*(\rho)$ is monotone increasing (decreasing) if $d_I(\frac{1}{1+\kappa}) > 0$ ($< 0$). First, consider the case where $d_I(\frac{1}{1+\kappa}) > 0$. Since attention is decreasing in $y$, $d_I(y) = \frac{1}{2} - \delta\theta\,(1 - a(y, I))$ is also decreasing in $y$. By Proposition 3, $y_I^*(\rho)$ is increasing at all $\rho$ for which $d_I(y_I^*(\rho)) > 0$, and decreasing when the inequality is reversed. Since $y_I^*(0) = \frac{1}{1+\kappa}$ the assumption $d_I(\frac{1}{1+\kappa}) > 0$ implies that $y_I^*(\rho)$ is bounded above by $\bar{y}$ and increasing for all $\rho$. Hence, the limit defining $y_I^\infty$ exists and is equal to $\bar{y}$. A symmetric argument shows that if $d_I(\frac{1}{1+\kappa}) < 0$ then $y_I^*(\rho)$ is strictly decreasing and converges to $\bar{y}$. ■

**Proof of Proposition 1.** By Lemma 6, $\min\{y_S^*, y_M^*\} > y_N^* = 1/(1+\kappa)$, so $y_S^*, y_M^*, y_N^* \to 1$ as $\kappa \to 0$; since the thresholds do not depend on $\kappa$, all three lie in region $S$ for small $\kappa$. By Theorem 3 this gives $y_S^* \in \mathcal{S}_\kappa$ and $y_M^*, y_N^* \notin \mathcal{S}_\kappa$; the same theorem implies $y_I^*$ and $\hat{y}_I$, the only remaining candidates, are never both stable, so $\mathcal{S}_\kappa$ is one of $\{y_S^*\}$, $\{y_S^*, y_I^*\}$, or $\{y_S^*, \hat{y}_I\}$.

By (8), the ODE for region $I$ can be written as $g_I(y) = (1-y)(1+\rho y d_I(y)) - \kappa y$. Since $d_I$ is decreasing in $y$, the factor $1+\rho y d_I(y)$ stays at or above 1 while $d_I \geqslant 0$ and is strictly decreasing thereafter; as it equals 1 at $y=0$ and $1-\rho(\delta\theta-\frac{1}{2})$ at $y=1$, it has a unique interior zero $\bar{y}$ iff $\rho(\delta\theta-\frac{1}{2}) > 1$. As $g_I$ is decreasing in $\kappa$, $y_I^*(\kappa)$ increases to a limit $y_I^*(0)$ as $\kappa \to 0$ with $g_I(y_I^*(0); 0) = 0$; since $g_I(\bar{y}; \kappa) = -\kappa\bar{y} < 0$ whenever $\bar{y}$ exists, $y_I^*(0) = \bar{y}$ if $\rho(\delta\theta - \frac{1}{2}) > 1$, and $y_I^*(0) = 1$ otherwise.

Suppose $\rho(\delta\theta-\frac{1}{2}) \leqslant 1$ or the intermediate region is $M$. Then $y_I^* \notin \mathcal{S}_\kappa$ for small $\kappa$: in the latter case region $I$ is empty, and in the former $y_I^*(0) = 1$, so $y_I^*(\kappa)$ lies in region $S$. Moreover $\hat{y}_I$ is not stable, because the smaller quasi steady state in the applicable condition of Theorem 3—$y_S^*$ when the intermediate region is $M$, $y_I^*$ when it is $I$—tends to $1 > \hat{y}_I$ as $\kappa \to 0$. Hence $\mathcal{S}_\kappa = \{y_S^*\}$, which proves the first claim.

Otherwise the intermediate region is $I$ and $\rho(\delta\theta - \frac{1}{2}) > 1$, so $y_I^*(0) = \bar{y} \in (0,1)$. Numerical examples in Danenberg and Fudenberg (2026) show that all three sets $\{y_S^*\}$, $\{y_S^*, y_I^*\}$, and $\{y_S^*, \hat{y}_I\}$ then arise for some parameters, when $\bar{y}$ lies in $S$, $I$, or $N$, respectively. In the last two cases the additional steady state, $y_I^*$ or $\hat{y}_I$, lies strictly below $\hat{y}_M$, which is constant in $\kappa$, so the share of false stories at it is at least $1-\hat{y}_M > 0$. Since $y_N^* = 1/(1+\kappa) \notin N$, Theorem 4 makes this steady state a limit point with positive probability for all sufficiently small $\kappa$. ∎

**Examples for Sections 4 and 5.** To show that the relationships between $y_S^*$ and $y_M^*$ and between $y_I^*$ and $y_N^*$ can go both ways fix $\beta = \kappa = \rho = 1$, $\mu = 1.75$, $\lambda = 0.25$, and $\theta = 0.75$. 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 $\delta \gtrapprox 0.751$. Likewise, the relationship between the thresholds $\hat{y}_I, \hat{y}_M$ is undetermined. Calculations with the same parameter values as above show that $\hat{y}_I < \hat{y}_M$ for $\delta \lessapprox 0.647$ and $\hat{y}_I > \hat{y}_M$ for $\delta \gtrapprox 0.647$. Finally, the relationship between the thresholds and quasi steady states is also not determined: Both $\max\{y_I^*, y_M^*, y_N^*, y_S^*\} < \min\{\hat{y}_I, \hat{y}_M\}$ and $\min\{y_I^*, y_M^*, y_N^*, y_S^*\} > \max\{\hat{y}_I, \hat{y}_M\}$ are possible (see the numerical examples for Proposition 1 in Danenberg and Fudenberg (2026)).

Both of the configurations that give rise to case (a) of Theorem 3 are possible: For an example where $\hat{y}_I < \hat{y}_M$ and $y_I^* < \hat{y}_I < y_N^*$, set $\rho = 20, \theta = 0.9, \kappa = 12, \mu = 1.55, \beta = 1, \delta = 0.65, \lambda = 0.45$. For an example where $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$, set $\rho = 1, \theta = 0.9, \kappa = 2.55, \mu = 1.65, \beta = 1, \delta = 0.8, \lambda = 0.25$. It can be verified that in both of these examples $\hat{y}_I$ is the unique stable steady state of the LDI.

**Non-monotonicity in $\theta$.** Each quasi steady state $y^*_M, y^*_S, y^*_I$ can be first decreasing and then increasing in $\theta$ when it is a steady state for the LDI (and thus a limit point for the system). For $y^*_S$, set $\rho = 0.3, \kappa = 1.5, \mu = 1.57, \beta = 0.3, \delta = 0.55, \lambda = 0.05$. With these parameters, $y^*_S$ is in region $S$ for all $\theta \in (0, 1)$ and is decreasing in $\theta$ for $\theta \lessapprox 0.95$ and then increasing. Furthermore, for all values of $\theta \in (0, 1)$, all other candidate limit points are also in region $S$, so $y^*_S$ is the unique limit point of the system for any value of $\theta$.

For $y^*_M$, set $\rho = 1, \kappa = 8, \mu = 1.57, \beta = 0.3, \delta = 0.9, \lambda = 0.05$. With these parameters, $\hat{y}_M < \hat{y}_I$ for all $\theta \in (0, 1)$, so the intermediate region is $M$. Additionally, $y^*_M$ is in region $M$ for all $\theta \gtrapprox 0.16$ (otherwise, $y^*_M$ is in region $S$), and $y^*_M$ is decreasing in $\theta$ for $\theta \lessapprox 0.87$ and then increasing in $\theta$. So $y^*_M$ is both decreasing and increasing in $\theta$ in region $M$.

Finally, for $y^*_I$, set $\rho = 0.45, \kappa = 3.34, \mu = 1.54, \beta = 0.3, \delta = 0.53, \lambda = 0.1$. With these parameters, $\hat{y}_I < \hat{y}_M$ for all $\theta \in (0, 1)$, so the intermediate region is $I$. Additionally, $y^*_I$ is in region $I$ for all $\theta \gtrapprox 0.85$, and $y^*_I$ is decreasing in $\theta$ for $\theta \lessapprox 0.88$ and then increasing in $\theta$. So $y^*_I$ is non-monotone in $\theta$ in region $I$.

**Dependence of $y^*_I$ on $\rho$.** We show that $d_I(y^*_I)$ can be either positive or negative, which (by Proposition 3) implies that $y^*_I$ can be either increasing or decreasing in $\rho$, and that both cases can occur when $y^*_I$ is a limit point. We also show that all cases described by Proposition 6 are possible. In the following examples we fix values for all parameters except $\rho$ and $\delta$ and consider comparative statics with respect to $\rho$ at four different values of $\delta$. In the first two specifications, discernment with decision rule $I$ is negative, so $y^*_I$ is decreasing in $\rho$. In the third and fourth specifications discernment is positive so $y^*_I$ increases in $\rho$.

Set $\theta = 0.9, \kappa = 3, \mu = 1.55, \beta = 1, \lambda = 0.45$, and $\delta = 0.8$. With these parameters $\hat{y}_I < \hat{y}_M$, so the intermediate region is $I$ (for any value of $\rho$). Additionally, $d_I(y) < 0$ for all $y \in (0, 1)$ so, by Proposition 6, $y^*_I(\rho) \to 0$. Starting with $\rho = 0$, we have $y^*_I$ in region $S$, and $y^*_I$ is decreasing in $\rho$ such that it enters region $I$ when $\rho \approx 15.6$, and enters region $N$ when $\rho \approx 42.2$ (so it is a limit point when $\rho$ is between those values). With the same parameter values but setting $\delta = 0.576$, the intermediate region is again $I$, but $y_I^* \in I$ for $\rho = 0$. It is still the case that $d_I(y_I^*(0)) = d_I(\frac{1}{1+\kappa}) < 0$, so $y_I^*$ is still decreasing in $\rho$, but now for $\bar{y} \approx 0.24$ we have $d_I(\bar{y}) = 0$ so $y_I^*(\rho) \to \bar{y}$ and is decreasing towards its limit. In this case, $\bar{y} \in I$ so $y_I^*$ converges to an interior point and remains in region $I$ for any value of $\rho$. Note that in this specification and the previous one $d_I(y_I^*(\rho)) < 0$ for all $\rho$. We now present two examples where this inequality is reversed. If we further decrease $\delta$ to $\delta = 0.57$, the intermediate region is again $I$, and $y_I^* \in I$ for $\rho = 0$. However, now $d_I(\frac{1}{1+\kappa}) > 0$ so $y_I^*$ is increasing in $\rho$ and enters region $S$ when $\rho \approx 142.7$. For these parameters $d_I(\tilde{y}) = 0$ for $\tilde{y} \approx 0.47$ so $y_I^*(\rho) \to \tilde{y}$ and is increasing towards its limit. Finally, setting $\delta = 0.55$, we get that $d_I(y) > 0$ for all $y \in (0, 1)$, so that $y_I^*(\rho) \to 1$.

# References


Acemoglu, Daron, Asuman Ozdaglar, and James Siderius (2024). "A model of online misinformation". *The Review of Economic Studies* 91, pp. 3117–3150.

Allcott, Hunt and Matthew Gentzkow (2017). "Social media and fake news in the 2016 election". *Journal of Economic Perspectives* 31.2, pp. 211–236.

Arieli, Itai, Yakov Babichenko, and Manuel Mueller-Frank (2024). "Sequential naive learning". *Available at SSRN 3753401.*

Arthur, W Brian and David A Lane (1993). "Information contagion". *Structural Change and Economic Dynamics* 4.1, pp. 81–104.

Benaim, Michel, Josef Hofbauer, and Sylvain Sorin (2005). "Stochastic approximations and differential inclusions". *SIAM Journal on Control and Optimization* 44.1, pp. 328–348.

Benaim, Michel, Sebastian J Schreiber, and Pierre Tarres (2004). "Generalized urn models of evolutionary processes". *Annals of Applied Probability*, pp. 1455–1478.

Bloch, Francis, Gabrielle Demange, and Rachel Kranton (2018). "Rumors and social networks". *International Economic Review* 59.2, pp. 421–448.

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

Danenberg, Tuval and Drew Fudenberg (2026). *Endogenous Attention and the Spread of False News (Extended Cut).*

Dasaratha, Krishna and Kevin He (2023). "Learning from viral content". *arXiv preprint arXiv:2210.01267.*

Guriev, Sergei, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya (2023). "Curtailing False News, Amplifying Truth".

Hill, Bruce M, David Lane, and William Sudderth (1980). "A strong law for some generalized urn processes". *The Annals of Probability*, pp. 214–226.

Kranton, Rachel and David McAdams (2024). "Social connectedness and information markets". *American Economic Journal: Microeconomics* 16.1, pp. 33–62.

Mahmoud, Hosam (2008). *Pólya urn models*. CRC Press.

Merlino, Luca P, Paolo Pin, and Nicole Tabasso (2023). "Debunking rumors in networks". *American Economic Journal: Microeconomics* 15.1, pp. 467–496.

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

Papanastasiou, Yiangos (2020). "Fake news propagation and detection: A sequential model". *Management Science* 66.5, pp. 1826–1846.

Paul, Christopher and Miriam Matthews (2016). "The Russian "firehose of falsehood" propaganda model". *Rand Corporation* 2.7, pp. 1–10.

Pemantle, Robin (2007). "A survey of random processes with reinforcement". *Probability Surveys* 4, pp. 1–79.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David 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* 66.11, pp. 4944–4957.

Pennycook, Gordon, Tyrone D Cannon, and David G Rand (2018). "Prior exposure increases perceived accuracy of fake news." *Journal of Experimental Psychology: general* 147.12, p. 1865.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand (2021). "Shifting attention to accuracy can reduce misinformation online". *Nature* 592.7855, pp. 590–595.

Pennycook, Gordon and David G Rand (2022). "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation". *Nature communications* 13.1, p. 2333.

Schreiber, Sebastian J (2001). "Urn models, replicator processes, and random genetic drift". *SIAM Journal on Applied Mathematics* 61.6, pp. 2148–2167.

Smith, Lones and Peter Norman Sørensen (2020). "Rational Social Learning with Random Sampling".

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). "The spread of true and false news online". *Science* 359.6380, pp. 1146–1151.

Zhang, Jason Shuo, Brian Keegan, Qin Lv, and Chenhao Tan (2021). "Understanding the diverging user trajectories in highly-related online communities during the COVID-19 pandemic". *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15, pp. 888–899.