Endogenous Attention and the Spread of False News*

Tuval Danenberg† and Drew Fudenberg‡

February 23, 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, and Alexander Wolitzky 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.

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 its *evocativeness*, or how interesting and stimulating it is. Users first observe the story's evocativeness level, 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 evocativeness: mildly interesting (M) and very interesting (I). While a story's veracity is fixed throughout time, evocativeness is drawn i.i.d (conditional on veracity) for each user. 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.

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 evocativeness 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 evocativeness 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.

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.[1] 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 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

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

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).[2]

If the users' decision rule were fixed instead of depending on y_n, our system would be a GPU. The stochastic approximation arguments of Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) imply that the hypothetical systems where users pick one of the four contingently-optimal decision rules and use it for all values of y_n have unique limit shares of true stories. These limits or *quasi steady states* are the unique steady states of associated differential equations. Our system is not a GPU because the optimal sharing and attention rules are not continuous, so we extend the literature on stochastic approximation of urn models to concatenations of a finite number of GPUs. This lets us relate the long-run behavior of the system to the stable steady states of the associated *limit differential inclusion* (LDI), which concatenates the differential equations associated with the GPUs.[3]

The first step in our analysis of the dynamics of the share of true stories is Theorem 1, which shows that a quasi steady state is a stable steady state for the LDI if and only if it is within the region where its associated decision rule is uniquely optimal. Depending on the parameters, there may or may not be one additional stable steady state, the *threshold* where the user is just indifferent between sharing and not sharing very interesting stories.

Next, Theorem 9 in Appendix B uses results from Benaim, Hofbauer, and Sorin (2005) (henceforth BHS) to show the system almost surely converges to a steady state of the LDI. Lemma 7 then gives a direct proof that all of the stable steady states of the limit differential inclusion have positive probability. Lemma 8 complements this by using a result of Pemantle (2007) to show the system has probability 0 of converging to an unstable steady state. Together these results imply Theorem 2, which shows that the system almost surely converges to a stable steady state of the limit differential

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

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

inclusion, and converges to each stable steady state 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. Guriev et al. (2023) structurally estimates a model of sharing decisions, and finds that users' perception of content veracity has a significant positive effect on their sharing decisions.

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. The platform contains stories with two characteristics (v, e). A story's *veracity* is $v \in \{T, F\}$, with the story being true if $v = T$ and false otherwise. A story's *evocativeness* is $e \in \{M, I\}$, with the story being mildly interesting if $e = M$ and very interesting if $e = I$. While a story's veracity is fixed (the story is either always true or always false), a story might be mildly interesting to one user and very interesting to another.[4] When a user draws a story, the probabilities of each evocativeness level are:

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

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 same credibility.

The platform begins operating 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 $z_n := (T_n, F_n)$

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

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 sees the story's evocativeness level and a noisy signal of its veracity. The precision of this signal depends on the user's *attention* as will be explained below. The parameter ρ describes the *reach* of shared stories on the platform—if the user decides to share the story, ρ copies of the story are added to the platform. Regardless of the user's decision, one true story and κ false stories are posted to the platform, corresponding to original content creation.

In summary, the sequence of events in each period is:

1. A distinct user draws a story and observes its realized evocativeness.
2. Chooses an attention level $a \in [0, 1]$.
3. Draws a signal whose distribution depends on a.
4. Decides whether to share the story. If shared, ρ copies of the story are added.
5. Receives payoffs.
6. Finally, 1 new true story and κ new false stories are exogenously added.

After drawing a story and observing its evocativeness level e, the user chooses a level of attention a, which will determine the precision of the signals they get regarding the story's veracity. The cost of attention level a is $\beta \cdot a^2$,where $\beta > 0$. The signal is $s \in \{T', F'\}$, with probabilities given by

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

The idea behind Equation 1 is that a false story of credibility θ is *clearly false* with probability $1 - \theta$, where a clearly false story is one that users will recognize as false even when they do not pay attention. With probability θ, users will notice the story is false only if they pay attention. A user's attention level a is the probability with which they pay attention. Thus, when a user's attention level is a and the credibility of false stories is θ, they will identify a false story as false with probability

$\mathbb{P}(F'|F) = 1 - \theta + \theta a = 1 - \theta(1-a)$. 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 seeing the story's evocativeness, knowing the current share of true stories y_n.[5] 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 βa^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, e)-story their additional payoff is

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

Here we have normalized the payoff to sharing a true, mildly interesting story to 1. The parameter μ captures the loss from sharing a false story, and the parameter λ 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 evocativeness level e, users either do not share at all and pay no attention (so their expected payoff is 0), or they share stories if and only if they receive the signal T'. In the latter case, their expected payoff to attention level a is

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

Thus, if users share at all they will choose the attention level

$$a(y, e) := \operatorname*{argmax}_{a \in [0,1]} U(a, y, e),$$

[5] This approximates a scenario where the veracity of stories shared a few periods back has been revealed and the mix between true and false stories is not changing too quickly.

[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.

and share only signal T' stories.

Our second parametric assumption implies that the optimal attention levels are always given by solutions to first-order conditions.

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

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

Discussion Users in our model want to share content that is both subjectively interesting and objectively true. Each of these motivations can be interpreted either as intrinsic or as a reduced form approximation of a motive that depends on subsequent users' reactions. For example, users might share subjectively interesting content simply due to a *self-expression* motive, or they might share it to influence others. Incorporating these considerations would significantly complicate our model. Instead, we focus on how these empirically grounded motives, previously derived in simpler models, interact with endogenous attention on a dynamically evolving platform.

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 stories $\{y_n\}$. To begin the analysis, we 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, and the optimal attention levels (conditional on sharing T' stories) are:*

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

The proof of this and all other results stated in the text are in Appendix A. It is straightforward to verify that $a(y, e) < 1$ for all y and $a(y, e) > 0$ if $y < 1$, and that

the system can never reach a state where $y = 1$. Intuitively, 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 evocativeness 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 evocativeness level such that if the share of true stories is below the corresponding threshold then users choose $a = 0$ and do not share the story, and if the share is above this threshold users choose the attention level given in Lemma 1 and share if and only if they received the signal T'.

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

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').

Users' sharing behavior depends on the share of true stories y_n. When y_n is below both thresholds, the expected value from sharing is negative for both evocativeness levels so users do not share at all. When y_n is above both thresholds, users share both types of stories, and otherwise they share only one type of story, as shown in Table 1. When the state is exactly at a threshold, users are indifferent between the two associated policies. Note that the 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.2 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 $p_R^T(y), p_R^F(y)$ 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

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

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

The following Markov processes describe how the system would evolve if users followed the decision rule of some region $R \in \{N, I, M, S\}$, whether or not it was optimal given the current state:

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

Appendix B.3 shows these processes are *generalized Polya urns* (GPUs), which lets us apply results from Schreiber (2001) and Benaim, Schreiber, and Tarres (2004).

The law of motion for y_n in region R is[7]

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

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

We will use stochastic approximation to approximate the behavior of the discrete stochastic system $\{y_n\}_{n\geqslant 0}$ by a continuous-time deterministic system. If our system was a single GPU, we could apply results in Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) to relate its limit behavior to that of an appropriately chosen *limit differential equation*. Instead, since our system is a concatenation of the GPUs $\{z_{n;R}\}$, we relate its limit behavior to that of a differential inclusion, an equation of the form $\frac{dx}{dt} \in F(x)$ for a set-valued function F. We construct this inclusion, which we will refer to as the *limit differential inclusion* or LDI, by pasting together the limit ODEs associated with the GPUs $\{z_{n;R}\}$. In our model these ODEs are[8]

$$\frac{dy}{dt} = 1 + p_R^T(y)\rho - y\big(1 + \kappa + \rho(p_R^T(y) + p_R^F(y))\big) := g_R(y). \tag{6}$$

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,

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

Thus, according to the limit ODE $\frac{dy}{dt} = g_R(y)$, the share of true stories increases if and only if the ratio of expected incoming true stories to total expected incoming stories is greater than the current share of true stories.

Our LDI is given by

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

where $F(y) = \{g_R(y)\}$ within each region R, and at the thresholds, F takes on all

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

values in the interval between the limit ODEs: If $\hat{y}$ is the threshold between regions R and R', then $F(\hat{y}) = [\min\{g_R(\hat{y}), g_{R'}(\hat{y})\}, \max\{g_R(\hat{y}), g_{R'}(\hat{y})\}]$.

5.2 Analysis of the Limit Continuous-Time System

We say that a point $y^* \in (0,1)$ is a *steady state* for the LDI if $0 \in F(y^*)$. We say that y^* is a *stable steady state* for the LDI if it is a steady state and there exists $\epsilon > 0$ such that for all $y \in (y^* - \epsilon, y^* + \epsilon) \backslash \{y^*\}$ we have $\text{sign}(x) = \text{sign}(y^* - y)$ for all $x \in F(y)$, and that a steady state is *globally stable* if the system converges to it with probability 1 from any initial position. We say a steady state is repelling if there exists $\epsilon > 0$ such that for all $y \in (y^* - \epsilon, y^* + \epsilon) \backslash \{y^*\}$ we have $\text{sign}(x) = -\text{sign}(y^* - y)$ for all $x \in F(y)$.

We will relate the steady states of the LDI to the behavior of the ODEs in each region. First we note that each of these ODEs has a globally stable steady state.

Lemma 3. *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*. The geometry of the phase diagram for the LDI 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 4. *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 quasi steady states are equal, and likewise rule out equality between any quasi steady state and either threshold.

Lemma 5. $\min\{y_S^*, y_M^*\} > \max\{y_I^*, y_N^*\}$.

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.[9]

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. 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.

[9]Numerical examples in Appendix A.2 show that the inequality can go both ways. We revisit this issue when discussing comparative statics of y_I^* with respect to ρ in Section 6.

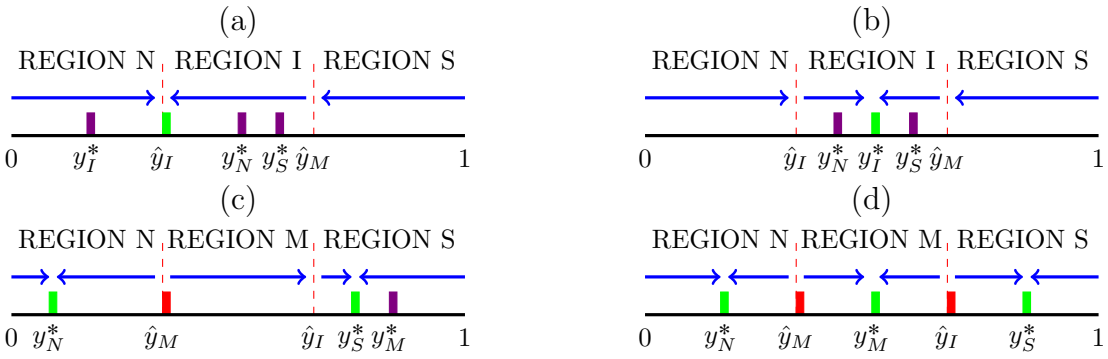

Figure 1: Examples of phase diagrams.

Appendix A.2 shows that Lemma 5 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 5 then it can go both ways.[10] The arguments for Lemma 5 do not rely on our specific choices of payoffs and signal function, so we expect that this ordering is satisfied for all specifications 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.

Figure 1 presents four examples of phase diagrams. The stable steady states of the LDI are in green, repelling steady states are in red, quasi steady states that are not steady states are in purple, 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 1. Since every limit ODE has a unique steady state, the only other candidate steady states for the LDI are the thresholds; these are steady states where users are indifferent between two rules and randomize between them.[11]

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

[10] 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^*.

[11] Here, as with mixed strategy equilibria in games, the randomizing probabilities are determined by an equilibrium condition.

diagram (a) in Figure 1), or $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$. Appendix A.2 shows that both cases are possible, and Lemma 5 implies that flips cannot occur around $\hat{y}_M$. This implies the following characterization of the set $\mathcal{S}$ of stable steady states.

Theorem 1 (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 κ false stories are added every period—if the system starts in region N and $y_N^* \in N$ then $y_n \to y_N^* = \frac{1}{1+\kappa}$ deterministically. Otherwise, any stable steady state is a limit point.

Theorem 2 (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^**.*

The proof of Theorem 2 has three parts. First, Theorem 9 in Appendix B shows that y_n almost surely converges to a steady state of the LDI. Second, Lemma 7 in Appendix A shows that every stable steady state has positive probability of being the limit point. Finally, Lemma 8 in Appendix A shows that the system almost surely does not converge to a repelling steady state. This completes the proof, because our simplifying assumption that no two variables in $\{\hat{y}_I, \hat{y}_M, y_N^*, y_I^*, y_S^*\}$ are equal means that any steady state is either stable or repelling.[12]

Detailed Proof Summary

Theorem 9 in Appendix B relates the limit behavior of concatenations of GPUs to the asymptotic behavior of the differential inclusions that concatenate the corresponding ODEs. Applied to our system, the theorem implies that the limit set of y_n, $L(y_n) = \bigcap_{m>0} \overline{\{y_n : n > m\}}$, is almost surely a steady state of the LDI.[13]

[12] If a threshold $\hat{y}$ were a steady state that is neither stable or 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.

[13] Here overline denotes the closure. The proof of Theorem 9 extends a result in Schreiber (2001) on continuous-time interpolations and perturbed solutions, and then applies a result in BHS that characterizes limits of perturbed solutions. (See appendix B for definitions of these terms.)

To prove Lemma 7, that there is positive probability of convergence to every stable steady state, 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 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.

Finally, the proof of Lemma 8, that y_n almost surely does not converge to a repelling steady state, uses Theorem 10 in Appendix B, which shows that a sufficient condition for nonconvergence to a repelling steady state is that there is a positive uniform lower bound on the noise in the stochastic process. Intuitively, noise jiggles y_n away from the steady state, and because the steady state is repelling, the drift of the process will tend to move it further away.

Discussion

Our simplified representation of platform dynamics allows for rich limit behavior. Our finding that the limit share of true stories is random, though not mathematically surprising within the context of generalized urns, has notable implications for the evolution of platform composition. It implies that starting from the same initial platform composition and parameters, the system can end up at very different limits in terms of both the share of true stories and users' limit actions. For instance, in some cases the system has positive probability of converging to any of three limits: One in which the share of true stories is low and users do not share at all (since the probability of sharing a false story is high), one in which the share of true stories is intermediate and users share only stories with one evocativeness level (very interesting/mildly interesting), and one in which the share of true stories is high and users share both very interesting and mildly interesting stories. This path-dependence suggests 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.[14] We begin with the effect of varying the credibility parameter θ.

Table 2: **Comparative Statics for θ**

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.2 show that each of the switchpoints in Table 2 can be interior. When they are, the associated quasi steady state is decreasing in θ up to a point and then increasing.[15] Intuitively, when the credibility of false stories increases it is harder to identify false stories, but this causes users to pay more

[14] 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 (δ) and the cost of attention (β).

[15] 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.

attention. This leads to two opposing forces on the limit share of true stories, and our model predicts that either one can prevail. That is, for sufficiently large values of θ 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 θ, and the switchpoint θ_S is interior in $(0, 1)$ (as in Appendix A.2). A false news producer with costs increasing in θ 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. This could provide a partial rationale for propaganda strategies such as the "firehose of falsehood," a term coined by Paul and Matthews (2016) to refer to contemporary Russian propaganda, characterized by a "shameless willingness to disseminate partial truths or outright fictions."

Another interpretation of θ is that the social media platform implements a fact-checking scheme that never mislabels true stories as false, with θ 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 (θ 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 increases the spread of false news.[16]

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 ρ**

y_M^*	Increasing. Converges to 1 as $\rho \to \infty$.
y_S^*	Increasing in ρ. 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 ρ 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 ρ 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 ρ 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 ρ is counterbalanced and may eventually settle down.

To illustrate the effect of ρ 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

[16] No flagging corresponds to $\theta = 1$, and we saw above that this can lead to a higher share of true stories than some values of $\theta < 1$.

$y_N^* = 1/(1+\kappa)$. Since users are discerning with rule I at y_N^*, a marginal increase in ρ increases y_I^*. Since attention is decreasing in y, so is discernment. Thus, as y_I^* increases in ρ, users' discernment decreases. But as long as discernment is positive, y_I^* continues to increase in ρ. 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 ρ.

In contrast to the parameters discussed above, the local effects of changes in κ are obvious: All quasi steady states are strictly decreasing in κ, and both thresholds are constant in κ, since it is not an argument in users' payoffs (see Danenberg and Fudenberg (2026)). The effect of κ on which steady states are stable is more interesting.

Theorem 3. *Fix values for the other parameters and let $\mathcal{S}_\kappa$ be the set of stable steady states as a function of κ. There exist $0 < \kappa_1 < \kappa_2 < \infty$ such that*

1. *$\mathcal{S}_\kappa$ is equal to $\{y_S^*\} \cup A$ for all $\kappa < \kappa_1$, where $A = \varnothing$ or $\{y_I^*\}$ or $\{\hat{y}_I\}$, depending on the other parameters.*

2. *$\mathcal{S}_\kappa = \{y_N^*\}$ for all $\kappa > \kappa_2$.*

As $\kappa \to \infty$, the exogenous inflow of false stories dominates any inflow due to user sharing, pushing all quasi steady states to 0. Eventually, false stories become so prevalent that users stop sharing altogether, so the unique limit point is y_N^*. For sufficiently small κ, the proof of Theorem 3 shows that when users are discerning with decision rule I, or when the intermediate region is M, then $\mathcal{S}_\kappa = \{y_S^*\}$ for all sufficiently small κ. In these cases, increasing the production rate of false stories from low to high shifts limit behavior from sharing both very interesting and mildly interesting stories to not sharing at all. Users are discerning when they share stories of both evocativeness levels, so here the exogenous decrease in the share of incoming stories that are true is amplified by user behavior.[17] However, if users are not discerning when they follow decision rule I, then either y_I^* or $\hat{y}_I$ may be an additional limit point for small κ, in which case a significant share of false stories can persist in

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

the limit even as the exogenous inflow of false stories vanishes. See Danenberg and Fudenberg (2026) for examples.

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 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 highly evocative 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. Its key simplifying feature is that it has a one-dimensional state space. We maintain this feature while considering two-dimensional story characteristics by assuming that only a story's veracity is fixed while its evocativeness is drawn every period. It would be straightforward to analyze variations that preserve this structure. For instance, Allcott and Gentzkow (2017) 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 the number of times an individual story has been shared, so it does not capture the "illusory truth" effect described in 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 $e = I$ with probability $\frac{1}{2}$. When $v = F$, then $e = I$ with probability δ. Thus,

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

Similarly, $\mathbb{P}_{a,y}(T', T|M) = \dfrac{y}{y + 2(1-y)(1-\delta)}$, $\mathbb{P}_{a,y}(T', F|I) = \dfrac{2(1-y)\delta\theta(1-a)}{y + 2(1-y)\delta}$, 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

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

The functions $U(a,y,I)$ 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, using Assumptions 1 and 2 it is straightforward to verify that $a(y,I), a(y,M) \in [0,1]$. ∎

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

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

To prove that these value functions are strictly increasing in y, it suffices to show that $U(a,y,M), U(a,y,I)$ are strictly increasing in y for all a, as then for $y_2 > y_1$ we have $V(y_1) = U(a(y_1),y_1) < U(a(y_1),y_2) \leqslant U(a(y_2),y_2) = V(y_2)$. 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 Lemma 3. First, note that by the definition of $g_R(y)$ in (6), 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

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

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 4. Fix two decision rules $R, W \in \{N, I, M, S\}$. By Lemma 3, $y_R^* > y_W^*$ if and only if $g_R(y_R^*) = 0 > g_W(y_R^*)$. By (6), 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 3, 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 5. By Lemma 3, 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 bounded above by 1. 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 $(1-\delta)\left(1-a(y,M)\right) < \delta\left(1-a(y,I)\right)$. Let $\ell(\delta) = (1-\delta)\left(1-a(y,M)\right); r(\delta) = \delta\left(1-a(y,I)\right)$. We will prove $\ell(\delta) < r(\delta)$ for all $\delta \in [\frac{1}{2}, 1)$ by showing that $\ell(\frac{1}{2}) < r(\frac{1}{2})$ and $\ell(\delta)$ is decreasing in δ while $r(\delta)$ is increasing in δ. First,

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

Next,

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

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

Proof of Theorem 1. 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 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.2 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 5, $y^*_S > y^*_I$. In Case 2 a flip cannot occur because by Lemma 5, $y^*_M > y^*_N$. ∎

Proof of Theorem 2. When $y^*_N \in N$ and $y_0 \in N$, the system follows the law of motion $z_{n+1} = z_n + \begin{pmatrix} 1 \\ \kappa \end{pmatrix}$, so it never leaves the region N and converges deterministically to $y^*_N = \frac{1}{1+\kappa}$. We henceforth assume that $y^*_N \notin N$ and/or $y_0 \notin N$. By Theorem 9 in Appendix B, the limit set of y_n is almost surely internally chain transitive for the LDI (7). Since the LDI is a one-dimensional autonomous inclusion, its internally chain transitive sets are its steady states, so y_n converges almost surely to a steady state of the LDI. By Lemma 7 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 8 there is zero probability of convergence to any repelling steady state. ∎

Lemma 7 and Lemma 8 below are used to prove Theorem 2, and Lemma 6 is used to prove Lemma 7.

Lemma 6. *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 the number of stories added each period is bounded, there exists some $n_\epsilon \in \mathbb{N}$ such that $|y_{n+1} - y_n| < \epsilon$ whenever $|z_n| > n_\epsilon$. Since $|z_n| \to \infty$ we can assume w.l.o.g. that the initial state z_n satisfies $|z_n| > n_\epsilon$. For such z_n, we consider two possible cases: $y_n < y$ and $y_n > y$.

Suppose first that $y_n < y < \frac{1+\rho}{1+\kappa+\rho}$. If the user shares a true story in period n then $1+\rho$ true and κ false stories are added to the platform, so $y_n < y_{n+1} < \frac{1+\rho}{1+\kappa+\rho}$, and if all subsequent users share true stories then $y_n \to \frac{1+\rho}{1+\kappa+\rho}$. Thus, there exists a finite $T = T(y_n) > 0$ such that if users share a true story every period for T periods then $y_{n+T} \in B_\epsilon(y)$. By a similar argument, if $y_n > y > \frac{1}{1+\kappa+\rho}$ then there is a finite $T' = T'(y_n) > 0$ such that if users share false stories for T' periods then $y_{n+T'} \in B_\epsilon(y)$. At any $y_m \notin N$ there is positive probability of drawing and sharing a true story and positive probability of drawing and sharing a false story. Also, since region N is always the leftmost region and $y \notin N$ then starting from $y_n > y$ and drawing T' false stories or starting from $y_n < y$ and drawing T true stories will not lead the system to enter region N. Thus, if $y_n < y$ ($y_n > y$) there is positive probability of sharing T (T') true (false) stories consecutively so there is positive probability of $y_m \in B_\epsilon(y)$ for some $m > n$. ∎

Lemma 7. *Assume that $y_N^* \notin N$ and/or $y_0 \notin N$. If ψ is a stable steady state, there is positive probability that $y_n \to \psi$.*

Proof. Let ψ be a stable steady state, and pick any $\epsilon > 0$.

Step 1: Defining five auxiliary processes.

The first four auxiliary processes are $\{z_{n;R}\}$ for $R \in \{N, I, M, S\}$ as defined in (4). Let $y_{n;R}$ be the share of true stories in period n for the process $\{z_{n;R}\}$. The differential inclusion associated with $\{z_{n;R}\}$ is $\frac{dy}{dt} \in \{g_R(y)\}$. By Lemma 3, this inclusion has a unique steady state y_R^*, so by Theorem 9, $y_{n;R}$ converges almost surely to y_R^*. In particular, for any $\epsilon > 0$ there exists $m_R \in \mathbb{N}$ such that starting from any y in the open ball $B_\epsilon(y_R^*)$, if the total number of stories is greater than m_R, then $y_{n;R}$ has positive probability of remaining in $B_\epsilon(y_R^*)$ forever, i.e., $\mathbb{P}\left(y_{m;R} \in B_\epsilon(y_R^*)\, \forall m > n \,|\, y_{n;R} \in B_\epsilon(y_R^*), |z_{n;R}| > m_R\right) > 0$.

The fifth auxiliary process is used to prove convergence to $\hat{y}_I$ when it is a stable steady state so we define it only for that case. Let L be the region to the left of $\hat{y}_I$ and R the region to the right of $\hat{y}_I$. Since $\hat{y}_I$ is a stable steady state, $y_R^* < \hat{y}_I < y_L^*$. Let O be the third region of the system (O is located either to the right of R or to the left of L). Define an alternative stochastic process $\{z_{n;H}\}$ with share of true stories $y_{n;H}$, where the law of motion in regions R, L is unchanged but in region O is that $y_{n;H}$ moves deterministically towards the nearest other region. (So if O is to the right of R then $y_{n;H}$ is decreasing in region O, and if O is to the left of L then it is increasing in region O). Let $\frac{dy}{dt} \in F_H(y)$ be the limit differential inclusion for this alternative process, as defined in Definition 5 in Appendix B. By construction, $\hat{y}_I$ is the unique steady state for this inclusion, so Theorem 9 implies that $y_{n;H}$ converges to $\hat{y}_I$ almost surely. In particular, there exists $m_H \in \mathbb{N}$ such that $\mathbb{P}\left(y_{m;H} \in B_\epsilon(\hat{y}_I)\, \forall m > n \,|\, y_{n;H} \in B_\epsilon(\hat{y}_I), |z_{n;H}| > m_H\right) > 0$.

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

Assume w.l.o.g. that ϵ is small enough that $B_\epsilon(y_R^*) \subset R$ if $\psi = y_R^*$ for some region R and that $B_\epsilon(\hat{y}_I) \subset [0,1] \backslash O$ if $\psi = \hat{y}_I$ (the previous step defines O as the only region not adjacent to $\hat{y}_I$). When $\psi = y_R^*$, $\mathbb{P}\big(y_m \in B_\epsilon(y_R^*) \forall m > n \,|\, y_n \in B_\epsilon(y_R^*), |z_n| > m_R\big) > 0$, since conditional on y_n remaining in $B_\epsilon(y_R^*)$ we have $y_n = y_{n;R}$. The fact that $y_n = y_{n;R}$ conditional on y_n remaining in region R implies that if the system arrives at a state z_n such that $y_n \in B_\epsilon(y_R^*)$ and $|z_n| > m_R$, then y_n converges to y_R^* with positive probability. If $\psi = \hat{y}_I$, an analogous argument (replacing $y_{n;R}$ with $y_{n;H}$), implies that if the system arrives at state z_n such that $y_n \in B_\epsilon(\hat{y}_I)$ and $|z_n| > m_H$ then y_n converges to $\hat{y}_I$ with positive probability.

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

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

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

This implies that $\frac{1}{1+\kappa+\rho} < y_R^* < \frac{1+\rho}{1+\kappa+\rho}$: the first inequality is immediate and the second is equivalent to $\rho\left(\kappa(1 - p_R^T(y_R^*)) + p_R^F(y_R^*)(1+\rho)\right) > 0$, which always holds. Since any stable steady state is either is a quasi steady state or a threshold bounded

above and below by quasi steady states, the above implies that

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

By hypothesis either $y_N^* \notin N$ or $y_0 \notin N$ (or both). First, assume $y_0 \notin N$. If $\psi \notin N$ then the claim follows immediately from (9) and Lemma 6 above, together with $|z_n| \to \infty$ surely. If $\psi \in N$ (which means $\psi = y_N^*$), then a similar argument as in the proof of Lemma 6 implies there is positive probability of arriving at some $y_m \in N$, from which point the system will converge deterministically to $\psi = y_N^*$ and in particular enter $B_\epsilon(y_N^*)$.

Now, assume $y_0 \in N$. Then, by hypothesis, $y_N^* \notin N$. Since in region N the system converges deterministically towards y_N^*, it surely arrives at $y_n \notin N$ with $|z_n| > m$ after finite time. Lemma 6 implies there is positive probability of arriving from this y_n to $B_\epsilon(\psi)$. ∎

Lemma 8. *The system almost surely does not converge to a repelling steady state.*

Proof. Since by Lemma 3 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 $\mathbb{P}(y_n \to \hat{y} \mid A) = 0$.

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

On A^C, there exists a finite stopping time τ after which $y_n \notin N$ for all $n \geqslant \tau$. To show $\mathbb{P}(y_n \to \hat{y} \mid A^C) = 0$, we apply Theorem 10, which requires: (i) $\hat{y}$ is repelling and (ii) the noise satisfies $\mathbb{E}[\xi_{n+1}^+ \mid \mathcal{F}_n] \geqslant r > 0$ for all y_n near $\hat{y}$. However, when $\hat{y}$ is adjacent to N, the martingale noise terms $\xi_n = (y_{n+1} - y_n - \mathbb{E}[y_{n+1} - y_n \mid z_n])|z_n|$ vanish on one side of $\hat{y}$. To circumvent this, we construct an auxiliary process $\{z_n'\}$ as follows: outside region N, it has the same dynamics as $\{z_n\}$; inside region N, it has negative drift with sufficient noise to satisfy Theorem 10's condition. Since τ is finite almost surely, convergence of y_n to $\hat{y}$ on A^C is equivalent to convergence of the shifted process. Hence it suffices to show $\mathbb{P}(y_n' \to \hat{y}) = 0$.

For small $\epsilon > 0$, let $U := (\hat{y} - \epsilon, \hat{y} + \epsilon)$ intersect only the two regions adjacent to $\hat{y}$. Since $\hat{y}$ is repelling, for all $y \in U \backslash \{\hat{y}\}$: $\text{sign}(x) = -\text{sign}(\hat{y} - y) \quad \text{for all } x \in F(y)$. We verify the noise condition $\mathbb{E}[\xi'^{+}_{n+1} \mid \mathcal{F}'_n] \geqslant r$ for some $r > 0$ when $y'_n \in U$. From (5), let $\Delta_T, \Delta_O, \Delta_F$ denote the three possible increments with $\Delta_T > \Delta_O > \Delta_F$. For y'_n in region $R \neq N$:

$$\mathbb{E}[\xi'^{+}_{n+1} \mid \mathcal{F}'_n] \geqslant p_R^T(y'_n)(1 - p_R^T(y'_n))(\Delta_T - \Delta_O)|z'_n|.$$

For sufficiently large $|z'_n|$:

$$(\Delta_T - \Delta_O)|z'_n| = \frac{(\kappa + |z'_n|(1 - y'_n))\rho}{(|z'_n| + 1 + \kappa)(|z'_n| + 1 + \kappa + \rho)}|z'_n| \geqslant \frac{(1 - y'_n)\rho}{4}.$$

Since $y'_n \in U \subset (0, 1)$ and $R \neq N$, we have $p_R^T(y'_n) \in \{y'_n, y'_n/2\}$ by (3), so $p_R^T(y'_n)(1 - p_R^T(y'_n))$ is bounded below by some $c_0 > 0$. Therefore for sufficiently large n:

$$\mathbb{E}[\xi'^{+}_{n+1} \mid \mathcal{F}'_n] \geqslant \frac{c_0(1 - y'_n)\rho}{4} \geqslant r > 0.$$

By Theorem 10, $\mathbb{P}(y'_n \to \hat{y}) = 0$, hence $\mathbb{P}(y_n \to \hat{y} \mid A^C) = 0$.

Since $\mathbb{P}(A) + \mathbb{P}(A^C) = 1$, we conclude $\mathbb{P}(y_n \to \hat{y}) = 0$. ∎

A.1 Comparative Statics

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

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

$$G(y_0^*, r_0) = 0. \tag{10}$$

Lemma 3 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 3 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 (10) 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 (6) 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 ρ and θ.

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

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

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

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

Note that the RHS is always positive, so that for sufficiently small θ, y_S^* is decreasing in θ. However, it is possible that the RHS is below 1 so that for large values of θ the relationship reverses. See Appendix A.2 for an example. ■

Theorem 5. *The quasi steady state y_I^* is increasing in ρ if $d_I(y_I^*) > 0$ and decreasing in ρ when the sign is reversed, and both cases can arise in region I.*[18] *There exists $\theta_I \in (0, 1]$ (whose value depends on the other parameters) such that y_I^* is decreasing in θ for $\theta < \theta_I$ and increasing in θ for $\theta > \theta_I$.*

Proof. By a similar argument as in the proof of Theorem 4, for all $x \in (\rho, \kappa, \theta, \mu, \beta, \delta, \lambda)$:

[18] Part 4 of Theorem 8 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.

$\text{sign}(\frac{dy_I^*(r_0)}{dx}) = \text{sign}(G_x(y_0^*, r_0))$ 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 ρ and θ.

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

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.2 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 $\theta > \frac{\beta(y+2\delta(1-y))}{2\delta(1-y)(\mu-1-\lambda)}$. Note that the RHS is always positive, so that for sufficiently small θ, y_I^* is decreasing in θ. However, it is possible that the RHS is below 1 so that for large values of θ the relationship reverses. See Appendix A.2 for an example. ■

Theorem 6. *The quasi steady state y_M^* is increasing in ρ. There exists $\theta_M \in (0,1]$ (whose value depends on the other parameters) such that y_M^* is decreasing in θ for $\theta < \theta_M$ and increasing in θ for $\theta > \theta_M$.*

Proof. By a similar argument as in the proof of Theorem 4 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 ρ and θ.

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

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 $\theta > \frac{\beta(y+2(1-y)(1-\delta))}{2(1-\delta)(1-y)(\mu-1)}$. Note that the RHS is always positive, so that for sufficiently small θ, y_M^* is decreasing in θ. However, it is possible

that the RHS is below 1 so that for large values of θ the relationship reverses. See Appendix A.2 for an example. ■

Theorem 7. *The threshold $\hat{y}_I$ is constant in ρ and increasing in θ.*

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 $V(y, r) : \mathbb{R}^8 \to \mathbb{R}$ (for this proof, since we only consider region I, 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 ρ is not an argument in the users' value functions. By (8), the partial derivative with respect to θ is:

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

where the inequality holds because by Assumption 2 and $\lambda > 0$,

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

■

Theorem 8. *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 ρ, and $y_R^\infty := \lim_{\rho \to \infty} y_R^*(\rho)$. This limit exists for all regions R and:*[19]

1. *$y_S^\infty = y_M^\infty = 1$.*
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 $y_I^\infty = \bar{y}$.*

[19] In Appendix A.2 we 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.

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 (6),

$$y_R^*(\rho) = \frac{1 + p_R^T(y_R^*(\rho))\rho}{1 + \kappa + \rho(p_R^T(y_R^*(\rho)) + p_R^F(y_R^*(\rho)))}. \tag{11}$$

and by (3), $p_R^T(y) + p_R^F(y) > 0$ for all $y \in (0,1)$. So if the limit y_R^∞ 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^∞ 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 (11), $y_R^*(0) = \frac{1}{1+\kappa}$.

Claim 1: By Theorems 4 and 6, both y_S^* and y_M^* are monotonically increasing in ρ, 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 Theorem 5, if $d_I(y) > 0$ $(d_I(y) < 0)$ for all y then $y_I^*(\rho)$ is monotone increasing (decreasing). Thus, in both of the cases the limit exists, and it 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 Theorem 5, $y_I^*(\rho)$ is increasing at all ρ 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 ρ. Hence, the limit defining y_I^∞ 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 Theorem 3. By Lemma 5, we have $\min\{y_S^*, y_M^*\} > y_N^* = \frac{1}{1+\kappa}$. Thus, y_S^*, y_M^*, and y_N^* all converge to 1 as κ goes to 0. Since the thresholds are constant in κ, this implies that there exists some $\kappa' > 0$ such that y_S^*, y_M^*, y_N^* are all in region S for all $\kappa < \kappa'$. Hence, $y_S^* \in \mathcal{S}_\kappa$, and $y_M^*, y_N^* \notin \mathcal{S}_\kappa$ for all $\kappa < \kappa'$. If the intermediate region is M, the above implies that $\mathcal{S}_\kappa = \{y_S^*\}$ for all $\kappa < \kappa'$ (the only other possible stable

steady state in this case is $\hat{y}_I$, which cannot be stable if y_S^* is in region S). Consider the case where the intermediate region is I. Let $y_I^*(\kappa)$ denote the quasi steady state y_I^* as a function of κ. Let $y_I^*(0) := \lim_{\kappa \to 0} y_I^*(\kappa)$. If $y_I^*(0) \in I$, then there exists $\kappa_1 > 0$ such that $\mathcal{S}_\kappa = \{y_S^*, y_I^*\}$ for all $\kappa < \kappa_1$. If $y_I^*(0) \in N$, then there exists $\kappa_1 > 0$ such that $\mathcal{S}_\kappa = \{y_S^*, \hat{y}_I\}$ for all $\kappa < \kappa_1$.Finally, if $y_I^*(0) \in S$, then there exists $\kappa_1 > 0$ such that $\mathcal{S}_\kappa = \{y_S^*\}$ for all $\kappa < \kappa_1$.[20] This completes the proof of the first claim, where we use our standing assumption that no two variables in $\{\hat{y}_I, \hat{y}_M, y_N^*, y_I^*, y_S^*\}$ are equal. To prove the second claim, note that as $\kappa \to \infty$, all quasi steady states converge to 0 (since $y_N^* = 1/(1+\kappa) \to 0$ and by (9) all others are bounded above by $\frac{1+\rho}{1+\kappa+\rho}$), while the thresholds $\hat{y}_I, \hat{y}_M$ are constant in κ. Hence there exists κ_2 large enough that all quasi steady states fall below $\min\{\hat{y}_I, \hat{y}_M\}$ and therefore lie in region N, giving $\mathcal{S}_\kappa = \{y_N^*\}$ for all $\kappa > \kappa_2$. ■

A.2 Numerical Examples

Examples for Sections 4 and 5. For an example 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 Theorem 3 in Danenberg and Fudenberg (2026)).

Both of the configurations that give rise to case (a) of Theorem 1 are possible: For an example where $\hat{y}_I < \hat{y}_M$ and $y_I^* < \hat{y}_I < y_N^*$, set $\rho = 20, \theta = 0.9, \kappa = 12, \mu = 1.55, \beta = 1, \delta = 0.65, \lambda = 0.45$. 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 θ. Each quasi steady state y_M^*, y_S^*, y_I^* can be first decreasing and then increasing in θ when it is a steady state for the LDI (and thus a limit point

[20] Numerical examples in Danenberg and Fudenberg (2026) show that all three cases are possible.

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 θ 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 θ.

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 θ for $\theta \lessapprox 0.87$ and then increasing in θ. So y_M^* is both decreasing and increasing in θ 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 θ for $\theta \lessapprox 0.88$ and then increasing in θ. So y_I^* is non-monotone in θ in region I.

Dependence of y_I^* on ρ. We show that $d_I(y_I^*)$ can be either positive or negative, which (by Theorem 5) implies that y_I^* can be either increasing or decreasing in ρ, and that both cases can occur when y_I^* is a limit point. We also show that all cases described by Theorem 8 are possible. In the following examples we fix values for all parameters except ρ and δ and consider comparative statics with respect to ρ at four different values of δ. In the first two specifications, discernment with decision rule I is negative, so y_I^* is decreasing in ρ. In the third and fourth specifications discernment is positive so y_I^* increases in ρ.

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 ρ). Additionally, $d_I(y) < 0$ for all $y \in (0,1)$ so, by Theorem 8, $y_I^*(\rho) \to 0$. Starting with $\rho = 0$, we have y_I^* in region S, and y_I^* is decreasing in ρ 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 ρ 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 ρ, 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 ρ. Note that in this specification and the previous one $d_I(y_I^*(\rho)) < 0$ for all ρ. We now present two examples where this

inequality is reversed. If we further decrease δ 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 ρ 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 $y_I^*(\rho) \to 1$.

B Urn Models

This appendix extends results from Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) about *Generalized Polya urns* (GPUs). A key feature of these urn models is that the number of balls added each period is bounded, so that as the overall number of balls grows the change in the system's composition between any two consecutive periods becomes arbitrarily small. Within each of the regions $\{N, I, M, S\}$, our system behaves like a GPU. To analyze the entire system, we define *Piecewise Generalized Polya Urns* (PGPUs), and then combine results on GPUs with results from BHS that extend the theory of stochastic approximation to cases where the continuous system is given by a solution to a differential inclusion rather than a differential equation.[21] Theorem 9 relates the limit behavior of a PGPU to the limit behavior of the associated differential inclusion; we use it in the proof of Theorem 2. Section B.3 explains why the processes $\{z_{n;R}\}$ defined in (4) are GPUs and derives the corresponding limit ODEs. Section B.4 then proves a result about repelling steady states for limit inclusions that is used in the proof of Theorem 2.

B.1 Definitions and Notation

Given a vector $w \in \mathbb{R}^2$ define $|w| = |w^1| + |w^2|$. Let $\{z_n\} = \{(z_n^1, z_n^2)\}$ be a homogeneous Markov chain with state space $\mathbb{Z}_+^2$ ($\mathbb{Z}_+$ are all the non-negative integers). Let $\Pi : \mathbb{Z}_+^2 \times \mathbb{Z}_+^2 \to [0, 1]$ denote its transition kernel, $\Pi(z, z') = \mathbb{P}(z_{n+1} = z' | z_n = z)$. We interpret the process as an urn model, with z_n^i the number of balls of color i at time step n. We now define two types of stochastic processes.

[21] We use results on stochastic approximation for differential inclusions, since existing results on stochastic approximation by differential equations do not apply to the to the discontinuous system we have here. Extending published results on stochastic approximation to discontinuous differential equations would have been more complicated, and the Hill, Lane, and Sudderth (1980) analysis of discontinuous Polya urns only covers the case where a single ball is added each period.

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

i. Balls cannot be removed and there is a maximal number of balls that can be added. Formally, for all $z_n \in \mathbb{Z}_+^2$ and all z_{n+1} such that $\Pi(z_{n+1}, z_n) > 0$: $z_{n+1}^1 \geqslant z_n^1, z_{n+1}^2 \geqslant z_n^2$ and there is a positive integer m such that $|z_{n+1} - z_n| \leqslant m$.

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

Let $y_n = \frac{z_n^1}{|z_n|}$ be the share of balls of color 1 (i.e., of true stories.)

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

$$x_{n+1} - x_n = \gamma_n \left(g(x_n) + \xi_{n+1} + R_n \right), \tag{12}$$

where $\gamma_n \in \mathcal{F}_n$ are non-negative with $\gamma_n \to 0, \sum_n \gamma_n = \infty$ almost surely, g is a Lipschitz function on $\mathbb{R}$, $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$ and the remainder terms $R_n \in \mathcal{F}_n$ go to zero and satisfy $\sum_{n=1}^{\infty} \frac{|R_n|}{n} < \infty$ almost surely.

The function g in (12) is the right hand side of the *limit ODE*, $\frac{dx}{dt} = g(x)$. Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) derive the limit ODE of a GPU and prove that with this limit ODE the sequence $\{y_n\}$ of the share of balls of color 1 is a stochastic approximation process. Since we will later consider a system that includes several GPUs we introduce the notation $\{z_{n;k}\}$ to refer to a general GPU.

Definition 3. For a GPU $\{z_{n;k}\}$ with corresponding maps p_k^w, the corresponding *limit ODE* is $\frac{dy}{dt} = g_k(y)$ where $g_k : [0,1] \to [0,1]$ is given by[22]

$$g_k(y) = \sum_{w \in \mathbb{Z}^2} p_k^w(y) \left(w^1 - y|w| \right). \tag{13}$$

B.2 Stochastic Approximation of PGPUs

This section extends the literature on GPUs to concatenations of GPUs.

Definition 4. A Markov process $\{z_n\}$ with transition kernel Π is a *piecewise generalized Polya urn* (PGPU) if there exists a finite integer K, a finite number of GPUs

[22] Definition 1.i implies that only a finite number of the summands are non-zero.

$\{\{z_{n;k}\}\}_{k=1}^K$ (each with kernel Π_k), and an interval partition $\{I_k\}_{k=1}^K$ of $[0,1]$, such that for all z', if $\frac{z^1}{|z|} \in \mathring{I}_k$ then $\Pi(z,z') = \Pi_k(z,z')$, where $\mathring{I}$ denotes the interior of I.[23]

The next definition defines the analog of a limit ODE for a PGPU.

Definition 5. For a PGPU $\{z_n\}$ the *limit differential inclusion* is $\frac{dy}{dt} \in F(y)$, where

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

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

$$\frac{dy}{dt} \in F(y) \tag{14}$$

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

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

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

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

Lemma 9. *The inclusion (14) satisfies the standing assumptions in BHS.*

Proof. Assumptions 1 and 2 follow immediately from Definition 5. Assumption 3 follows from the fact that the $g_k(y)$ are continuous functions defined over compact sets. ∎

We relate the limiting behavior of y_n to the solutions to the differential inclusion (14) using the ideas of a *perturbed solution* and a *piecewise affine interpolation.*

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

Definition 6. A continuous function $\mathbf{Y} : [0, \infty) \to \mathbb{R}$ is a *perturbed solution* to (14) (or a "perturbed solution to F") if it is absolutely continuous, and there is a locally integrable function $t \mapsto U(t)$ such that

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

Definition 7. The *piecewise affine interpolation* of y_n is

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

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

Theorem 2.2 (Schreiber (2001)). *Let $\{z_{n;k}\}$ be a GPU. Let $\mathbf{Y}^k(t)$ be the piecewise affine interpolation of $y_{n;k} = \frac{z^1_{n;k}}{|z_{n;k}|}$, and let ϕ^k be the flow of the limit ODE.*[24] *Then on the event $\{\liminf_{n\to\infty} \frac{|z_{n;k}|}{n} > 0\}$, for any $T > 0$, $\lim_{t\to\infty} \sup_{0\leqslant h\leqslant T} |\mathbf{Y}^k(t + h) - \phi^k(\mathbf{Y}^k(t), h)| = 0$.*

The next lemma extends this result from GPUs to PGPUs.

Lemma 10. *Let $\{z_n\}$ be a PGPU and (14) its limit differential inclusion, and let $\mathbf{Y}$ be its piecewise affine interpolation. Then $\mathbf{Y}$ is a bounded perturbed solution to (14).*

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

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

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

[24] The flow $\phi^k : [0, 1] \times \mathbb{R}_+ \to [0, 1]$ such that $\phi^k(x, t)$ is the time-t value of a solution to the ODE at with initial condition x. Schreiber (2001) states this theorem for piecewise constant interpolations, but it also applies to piecewise affine interpolations.

Fix $T > 0$ and $0 \leqslant h \leqslant T$. Let ϕ^k be the flow of the limit ODE $\frac{dy}{dt} = g_k(y)$. On the event "$\mathbf{Y}(s) \in I_k$ for all $s \in [t, t+h]$," we have

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

Since by Definition 4 a PGPU has a finite number of partition intervals I_k, in the interval $[t, t+h]$ the interpolation $\mathbf{Y(t)}$ transitions between intervals I_k a finite a number of times. Thus $\int_t^{t+h} U(s)ds = \sum_{j=1}^{M} \left[\mathbf{Y}^{k_j}(t_j) - \phi^{k_j}(\mathbf{Y}(t_{j-1}), h_j) \right]$, where $M > 0$ is some integer; $t = t_0 < t_1 < ... < t_M = t + h$; $h_j = t_j - t_{j-1}$, and $k_j \in 1, ..., K$ for all $1 \leqslant j \leqslant M$.[25] So from Schreiber (2001)'s Theorem 2.2, for all $T > 0$

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

■

We are now ready to state and prove Theorem 9. The proof combines the previous results with a direct application of the following theorem:

Theorem 3.6 (BHS). *If* $\mathbf{x}$ *is a bounded perturbed solution to* F*, the limit set of* $\mathbf{x}$*,* $L(\mathbf{x}) = \bigcap_{t\geqslant 0} \overline{\{\mathbf{x}(s) : s > t\}}$ *is internally chain transitive.*[26]

Theorem 9. *Let* $\{z_n\}$ *be a PGPU,* $\{y_n\}$ *the share of balls of color* 1 *and* F *the associated limit differential inclusion. Then the limit set of* $\{y_n\}$*,* $L(y_n) = \bigcap_{m>0} \overline{\{y_n : n > m\}}$*, is almost surely internally chain transitive for* F*.*

Proof. By Lemma 10, the interpolation $\mathbf{Y}$ is a perturbed solution to F. Note that it is also bounded since $\mathbf{Y}(t) \in [0, 1]$ for all $t \geqslant 0$. Thus, Theorem 3.6 in BHS implies that the limit set of $\mathbf{Y}$ is internally chain transitive for F. The asymptotic behaviors of $\mathbf{Y}(t)$ and y_n are the same by the definition of interpolation, i.e., $L(y_n) = L(\mathbf{Y})$.

■

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

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

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

This section shows that the processes $\{z_{n;R}\}$ as defined in (4) are GPUs and derive the formula for their limit ODEs.

Lemma 11. *For each $R \in \{N, I, M, S\}$, $\{z_{n;R}\}$ is a GPU with limit ODE (6).*

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

Next, (3), (4), and (13) imply that the ODE associated with $\{z_{n;R}\}$ is

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

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

B.4 Repelling Steady States

This subsection shows that if ψ is a repelling steady state for the LDI, then under a condition on the noise in the stochastic system, $\mathbb{P}(y_n \to \psi) = 0$. Consider a PGPU $\{z_n\}$, comprised of GPUs $\{z_{n;k}\}_{k=1}^K$ with associated intervals I_k, where g_k is the RHS of the limit ODE for GPU $\{z_{n;k}\}$. Let $y_{n;k} = \frac{z_{n;k}^1}{|z_{n;k}|}$. Recall that $y_n = \frac{z_n^1}{|z_n|}$ and that the LDI for this PGPU is given by (14). We now add the following assumption, which is satisfied by the PGPUs in our model:

Assumption 3. Each limit ODE $\frac{dy}{dt} = g_k(y)$ has a globally stable steady state y_k^*.

Assumption 3 implies that the only possible repelling steady states for the LDI are the thresholds between the intervals I_k. Define these these as $\hat{y}_k = \max\{I_k\}$ for

$k = 1, \ldots, K$. Finally, let $\mathcal{F}_n$ be the σ-algebra generated by $(z_1, ..., z_n)$, let $\xi_{n+1} = (y_{n+1} - y_n - \mathbb{E}[y_{n+1} - y_n | z_n])|z_n|$ and denote $\xi_n^+ = \max\{0, \xi_n\}, \xi_n^- = -\min\{0, \xi_n\}$.

Theorem 10. *Let $\hat{y}_k$ be the threshold between intervals I_k, I_{k+1} and assume that $\hat{y}_k$ is a repelling steady state for the LDI. If there exist $\epsilon, r > 0$ such that for all $n \in \mathbb{N}$: $\mathbb{E}[\xi_n^+ | \mathcal{F}_n] > r$ if $y_n \in (\hat{y}_k - \epsilon, \hat{y}_k + \epsilon)$, then $\mathbb{P}(y_n \to \hat{y}_k) = 0$.*

The proof applies the following result:

Theorem 2.9 (Pemantle (2007)). *Suppose $\{x_n\}$ is a stochastic approximation process as defined in Definition 2 except that g need not be continuous. Assume that for some $p \in (0, 1)$ and $\epsilon > 0$: $\operatorname{sign}(g(x)) = -\operatorname{sign}(p - x)$ for all $x \in (p - \epsilon, p + \epsilon)$. Suppose further that the martingale terms ξ_n in the stochastic approximation equation (12) 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 $\mathbb{P}(x_n \to p) = 0$.*

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

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

Recall $\xi_{n+1} = (y_{n+1} - y_n - \mathbb{E}[y_{n+1} - y_n | z_n])|z_n|$. Let $R_n = |z_n|\mathbb{E}[y_{n+1} - y_n | z_n] - g(y_n)$. Then ξ_n, R_n are adapted to $\mathcal{F}_n$, $\mathbb{E}[\xi_{n+1} | \mathcal{F}_n] = 0$ and

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

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

$\mathbb{E}[\xi_n|\mathcal{F}_n] = 0$, and $\mathbb{E}[\xi_n^+|\mathcal{F}_n] = \mathbb{E}[\xi_n^-|\mathcal{F}_n]$, it suffices to find a positive lower bound for $\mathbb{E}[\xi_n^+|\mathcal{F}_n]$ when $y_n \in (\hat{y} - \epsilon, \hat{y} + \epsilon)$. By assumption, $r > 0$ is such a lower bound. ■

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.