Endogenous Attention and the Spread of False News[*]

Tuval Danenberg[†] and Drew Fudenberg[‡]

First posted version: June 3, 2024
This version: August 30, 2024

Abstract

We study the impact of endogenous attention in a dynamic model of social media sharing. Each period, a user observes a random story on the platform and decides whether to share it. Users want to share stories that are true and interesting, but distinguishing true stories from false ones requires attention. Before deciding whether to share a story, users choose their level of attention based on how interesting the story is and the platform's current proportion of true stories. We characterize the long-run platform composition using stochastic approximation techniques. For some parameter specifications, the system has a unique limit. For others, the limit is random—starting from the same initial conditions, the platform may end up with different proportions of true stories and different sharing behaviors. We present comparative statics for the limit. For example, endogenous attention counterbalances shifts in the credibility of false stories but can amplify the impact of changes in their production rate.

Keywords: false news, endogenous attention, Polya urns, stochastic approximation, social media

[*] We thank Michel Benaim, Claire Bartolone, Yifan Dai, Krishna Dasaratha, Ben Golub, Kevin He, Navin Kartik, Simon Loertscher, 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 suport.

[†] 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 empirical findings, our model assumes that users want to share true stories, but distinguishing false and true content requires costly attention. Users' attention depends on the prevalence and credibility of false stories: They are not willing to spend much effort trying to spot false stories if the share of false stories in their feed is negligible, but if the share of false stories is significant and the false stories are superficially plausible, they are willing to incur a significant cost to distinguish between true and false content. In turn, users' attention choices affect the prevalence of false stories as more attentive users are better at filtering false content. Our goal is to understand the resulting joint dynamics of users' attention and platform composition.

In our model, every period, a distinct user randomly draws a story from the stories on a social media platform and decides whether or not to share it. Users consider two factors when evaluating a story: its *veracity*, or truthfulness, and its *evocativeness*, or how interesting and stimulating it is. 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. False stories are characterized by a credibility measure that captures how true they appear—when false stories are highly credible, signals about their veracity are less precise. The precision of the signal is increasing in the user's chosen attention level. We assume that the signal's precision is supermodular in credibility and attention so that users' attention is increasing in credibility. 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 exogenously added as well, which corresponds to original 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 main object of interest is the share of true stories in the system for each period $n \in \mathbb{N}$, which we denote by y_n. Users' optimal behavior depends on the value of y_n. When y_n is sufficiently high, the system is in the *sharing* region, where users share all stories for which they receive the signal suggesting the story is true. When y_n is low, the system is in the *no sharing* region, where users do not share any stories and do not pay attention. In between, there is an intermediate region, where users share either only mildly interesting stories or only very interesting stories, depending on the model parameters.

Using stochastic approximation techniques, we show that y_n converges almost surely and provide a complete characterization of its limit. (See the technical summary below for an overview of this analysis.) For some parameter values the limit is unique. For others it is random, so that starting from the same initial conditions the platform may end up with significantly different limit shares of true stories and different user behavior in the limit. This effect is most pronounced when the platform is new and the total number of stories is small, but it is still present in any finite-sized platform.

The system converges either to a point where users strictly prefer a single sharing rule or to one where they indifferent between two rules. Comparative statics are qualitatively different in these two cases.[1] For example, in the steady states where users are indifferent between two sharing rules, the limit share of true stories may be increasing in the cost of attention, because the cost of attention enters negatively into users' payoffs while the share of true stories enters positively. So when the cost of attention increases, the share of true stories required for indifference increases as well. In contrast, increasing the cost of attention lowers the share of true stories at the other limit points.

For the steady states where agents strictly prefer a single sharing rule, the share of true stories is decreasing in false story credibility for low credibility levels, but an opposite effect may arise when credibility is high. The intuition is that while false stories of high credibility are harder to identify, users also pay more attention to them. When credibility is high, user responses to an increase in credibility may more than compensate for the direct effect of this increase, thereby leading to an increase in the limit share of true stories. The comparative statics imply that producers of

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

false stories may choose low credibility levels even when credibility is free. They also imply that platforms that aim to counter the spread of false news by fact-checking false stories might be better off not fact-checking at all than fact-checking only a small share of stories, because increasing the share of stories flagged as false leads users to put more trust in stories that were not flagged.

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

Our analysis emphasizes that the effect of user sharing behavior depends on how many true stories they share as well as as how many false ones, and also on the current mix of stories on the platform. More precisely, users' sharing increases the share of false content if and only if the ratio between the probabilities of sharing a false or true story is greater than the ratio between the probabilities of drawing a false or true story. We find that this can happen if users share only very interesting stories, but not if users share all stories or only share the mildly interesting ones. Intuitively, because users have a higher intrinsic benefit from sharing very interesting stories, they may share them even if they are relatively likely to be false.

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 the ball is returned to the urn along 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).

In our model, if the users' sharing rule was fixed, instead of depending on y_n, our system would be a GPU where stories are "balls" and colors are veracity levels. Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) use stochastic approximation arguments to show that under fairly general conditions the long-run behavior of GPUs can be determined by studying the attractors of a deterministic differential equation. Their results imply that the hypothetical systems where users pick one of the four contingently-optimal sharing rules and use it for all values of y_n have unique limit shares of true stories.[2] These limits, which we call *quasi steady states*, are the unique steady states of the associated differential equations. However, because the optimal sharing and attention rules are not continuous, our system is not a GPU but a concatenation of them. For this reason, we extend the literature on the stochastic approximation of urn models to cover 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 false 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 sharing rule is 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 3 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 6 then gives a direct proof that all of the stable steady states of the limit differential inclusion have positive probability. Lemma 7 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,

[2] The contingently-optimal sharing rules are: not sharing at all, sharing only mildly interesting stories, sharing only very interesting stories, and sharing both mildly and very interesting stories. For the rules that involve sharing, users will only share if they receive a signal suggesting the story is true.

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

which shows that the system almost surely converges to a stable steady state of the limit differential inclusion, and determines which of these steady states has positive probability of being the limit as a function of the parameters and the initial state.

2 Related Literature

Empirical Evidence

In our model, inattention plays a central role in the sharing of false content. Pennycook et al. (2021) claims that inattention to veracity is one of the key mechanisms leading users to share false stories. The paper reports evidence that most people say it is important to share only accurate news, but nevertheless sometimes share false news, and finds in a combination of survey experiments and a field experiment on Twitter (now X) that shifting users' attention to accuracy increases the accuracy of the content they share. Pennycook et al. (2020b) finds similar results in the context of information about COVID-19.[4] Of course, inattention is not the sole driver of the spread of false news; the conclusion discusses how our model can be adapted to incorporate additional factors such as politically motivated reasoning and ideological alignment (e.g., Van Bavel and Pereira (2018), Allcott and Gentzkow (2017)) and digital illiteracy Guess et al. (2020).

In our model, users care about two content dimensions—veracity and evocativeness. 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 the accuracy factor has the most impact on sharing.[5] Consistent with this, we assume that users will not share stories that they know are false even if they are very interesting. Chen, Pennycook, and Rand (2023) also finds that users ratings on the evocativeness dimension are negatively correlated with stories' objective veracity. This supports our assumption that false stories are more likely to be very interesting.

[4] See Pennycook and Rand (2022) for further discussion and references on the inattention based account and the effectiveness of accuracy nudges.

[5] The evocativeness factor captures characteristics such as the extent to which content is surprising, amusing, or provokes anxiety and other negative feelings. Earlier work by Berger and Milkman (2012) also finds a positive correlation between these characteristics and sharing intentions.

Theory of Online Misinformation

Bloch, Demange, and Kranton (2018), Papanastasiou (2020), Acemoglu, Ozdaglar, and Siderius (2023), Merlino, Pin, and Tabasso (2023), and Mostagir and Siderius (2022) analyze the spread of messages about a fixed binary state across a network. In most of these papers, users only care about veracity. In Acemoglu, Ozdaglar, and Siderius (2023), users' desire to share the story depends on whether they think most subsequent users will like it, but beliefs and sharing decisions do not depend on the actions of previous users, and attention is exogenous. In Mostagir and Siderius (2022), each user initially gets an informative message about the state, and then repeatedly transmits their posteriors to their neighbors using either Bayesian updating or DeGroot learning. Merlino, Pin, and Tabasso (2023) analyzes the mean field of an infinite-population SIS model with two messages corresponding to the two states. Agents become "infected" when they encounter a message and choose how much effort to spend to verify it, so this model has a form of endogenous attention, but unlike in our model, its focus is on the proportion of users who think each state is true as opposed to the shares of true and false stories. In Kranton and McAdams (2024), one agent initially receives a story and decides whether to transmit it without inspection or inspect it and only transmit it if it is true. Agents know how often a story has been shared, and once it has been shared enough, all subsequent agents choose to share it without inspection.

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.[6] In contrast, our paper focuses on the interaction between endogenous attention and platform evolution and includes a taste for sharing more evocative stories.

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

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.[7] 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 $t = 0$ with an exogenous stock of true and false stories (T_0, F_0). In each subsequent period $n \in \mathbb{N}$, 1 true story and κ false stories are exogenously added to the platform, and T_n and F_n respectively denote the numbers of true and false stories on the platform at the beginning of period n.[8] The vector $z_n := (T_n, F_n)$ summarizes the current state of the platform; we use the notation $|z_n| := T_n + F_n$ for the total number of stories in period n, and let $y_n := \frac{T_n}{|z_n|}$ denote the share of true stories.

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

In summary, each period the current user:

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

[8] The analysis would be the same in a continuous-time model where the time the next user arrives is a random variable.

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

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

3. Draws a signal whose distribution depends on a.

4. Decides whether to share the story.

5. Receives payoffs.

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

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.[9] 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 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

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

[9] 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.

Here we have normalized the payoff to sharing a true and 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. 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 this 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

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

and share only signal T' stories. We make two parametric assumptions:

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

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

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'.[10] It remains to analyze, for each evocativeness level, when they will share stories with signal T', which we do in the beginning of the next section. Assumption 2 implies that users attention levels conditional on sharing stories with signal T' are always given by solutions to first order conditions.

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 $\delta \in (\frac{1}{2}, 1)$.

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

[10] 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.

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:*

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

The proof of this and all other results stated in the text are in Appendix A. 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 $a(1, I) = a(1, I) = 0$. As y decreases the marginal gain from paying attention increases, and since the U's are strictly concave, $da/dy < 0$. However, when y is close enough to 0 the payoff from the $a(y, e)$ is so low that users prefer not to pay attention at all. We allow users to randomize when indifferent between $a = 0$ and $a = a(y, e)$.

As shown in Online Appendix C.3, both optimal attention levels are decreasing in β and increasing in θ and μ. Attention to very interesting stories $a(y, I)$ is increasing in δ, while $a(y, M)$ is decreasing in δ, and $a(y, I)$ is decreasing in λ while $a(y, M)$ is constant in λ. That is, users pay more attention when false stories are very credible and when the cost to sharing false stories is high, and pay less attention when the share of true stories is high and when the cost of attention is high. Users pay more attention to the veracity of very interesting stories when false stories are more likely to be very interesting, and pay less attention to the veracity of very interesting stories as the payoff to sharing them increases. These observations underlie the comparative statics in Section 6.

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) $\hat{y}_M, \hat{y}_I \in (0, 1)$ s.t $V(\hat{y}_M, M) = V(\hat{y}_I, I) = 0$.*

Table 1: **Regions and Sharing 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. 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 computations show that both $\hat{y}_M < \hat{y}_I$ and $\hat{y}_M > \hat{y}_I$ are possible so the intermediate region can be either of the two.

5 Dynamics

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

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

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

The following Markov processes describe how the system would evolve if users followed the sharing rule of region $R \in \{N, I, M, S\}$ regardless of the current share of true stories:

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

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[11]

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

We will use stochastic approximation to approximate the behavior of the discrete stochastic system $\{y_n\}_{n \geqslant 0}$ by a continuous and deterministic system. If our system was a single GPU, we could apply results in Schreiber (2001) and Benaim, Schreiber, and Tarres (2004) to relate its limit behavior to that of an appropriately chosen *limit differential equation*. 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[12]

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

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

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

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

incoming stories is $1+\kappa+\rho\left(p_R^T(y)+p_R^F(y)\right)$. So,

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

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

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

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 denote the steady states of $\frac{dy}{dt}=g_R(y)$ by y_R^*, and refer to them as *quasi steady states*. We refer to $\hat{y}_I,\hat{y}_M$ as *thresholds*. The geometry of the phase diagram for the LDI is determined by the relative positions of the thresholds and quasi steady states: The thresholds determine the system's regions, and within each region the flow is towards the corresponding quasi steady state. Thus, it is important to understand the possible orderings of the four quasi steady states.

Lemma 4. $\min\{y_S^*,y_M^*\}>\max\{y_I^*,y_N^*\}$.

We summarize the arguments underlying this result here because they help explain the effect of the sharing rule on the evolution of the platform. The proof starts from the fact that since each limit ODE has a unique globally stable point, the quasi steady state y_R^* for sharing rule R is greater than the quasi steady state $y_{R'}^*$ for rule R' if and only if $g_R(y_R^*) > g_{R'}(y_R^*)$. From (6), $g_R(y) > g_{R'}(y)$ if and only if $(1-y)\left(p_R^T(y) - p_{R'}^T(y)\right) > y\left(p_R^F(y) - p_{R'}^F(y)\right)$. To understand this condition, consider a comparison between g_R^* for some region $R \in \{I, M, S\}$ and g_N^*, the right hand side of the limit ODE associated with the no-sharing rule. Here the inequality reduces to $p_R^T(y)/p_R^F(y) > y/(1-y)$. This is the case if the ratio between the probabilities of sharing a true or false story is greater than the ratio between the probabilities of drawing a true or false story. In other words, users are successfully filtering false content.

To see why $\min\{y_S^*, y_M^*\} > y_N^*$, note that $p_R^T(y)/p_R^F(y) > y/(1-y)$ is equivalent to $p_R^T(y)/y > p_R^F(y)/(1-y)$, i.e., when a true story is drawn the probability of sharing it is greater than the sharing probability for false stories. This inequality is satisfied in region S because users are sharing all true stories that they draw but only some false stories. It is also satisfied in region M because there users share 1/2 of the true stories (the stories that are both true and mildly interesting) and less than $1-\delta < 1/2$ of false stories. Thus, compared to not sharing, when users follow sharing rules M or S the net increase in the share of true stories is larger than when they do not share. And for the same reason, sharing both M and I stories generates a larger net increase in y than sharing I stories alone, which is why $y_S^* > y_I^*$.

In contrast to the conclusion for rules S and M , the effect of 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 this rule users share 1/2 of true stories and $\delta\theta(1-a(y,I))$ of false stories, and both $1/2 > \delta\theta(1-a(y,I))$ and the opposite inequality can occur (for different parameters).[13]

Finally, that $y_M^* > y_I^*$ follows from the combination of two forces. First, under sharing rule I, users consider sharing more false stories than under M because more false stories are of type I. Second, 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 successful at filtering M content than I content.[14]

Numerical calculations described in Online Appendix C.3 verify that both $y_S^* < y_M^*$

[13] We revisit this issue when discussing comparative statics of y_I^* with respect to ρ in Section 6.

[14] Note that the arguments for Lemma 4 do not rely on our specific choices of payoffs and signal function. 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.

and $y_S^* > y_M^*$ are possible and similarly that y_N^* can be either greater or less than y_I^*. Moreover, the relationship between any threshold and any quasi steady state is also undetermined, i.e., both $\max\{y_N^*, y_I^*, y_M^*, y_S^*\} < \min\{\hat{y}_I, \hat{y}_M\}$ and $\min\{y_N^*, y_I^*, y_M^*, y_S^*\} > \max\{\hat{y}_I, \hat{y}_M\}$ are possible. This means that Lemma 4 is the only restriction on the ordering of the quasi steady states and thresholds (for simplicity, we rule out the knife edge case of equality between any of these variables). Because regions M and I do not occur at the same time, for given parameters only one of y_I^* and y_M^* matters. This means there are 40 possible strict configurations for the five variables that pin down the phase diagram: the two thresholds, and the quasi steady states for the system's three regions, i.e., y_S^*, y_N^* and one of y_I^*, y_M^*.

To see why there are 40 configurations, consider the case $\hat{y}_I < \hat{y}_M$. In this case, the five variables are $\{\hat{y}_I, \hat{y}_M, y_N^*, y_I^*, y_S^*\}$. We can now count the number of orderings of these variables that satisfy our restrictions. First, we can choose the relative positions of the two thresholds, giving $\binom{5}{2} = 10$ options. Lemma 4 shows that $y_S^* > \max\{y_N^*, y_I^*\}$, and $\hat{y}_I < \hat{y}_M$ by assumption, so the only degree of freedom is the order between y_N^*, y_I^*, for a total of 20 configurations in which $\hat{y}_I < \hat{y}_M$. Similarly, there are 20 configurations with $\hat{y}_I > \hat{y}_M$.

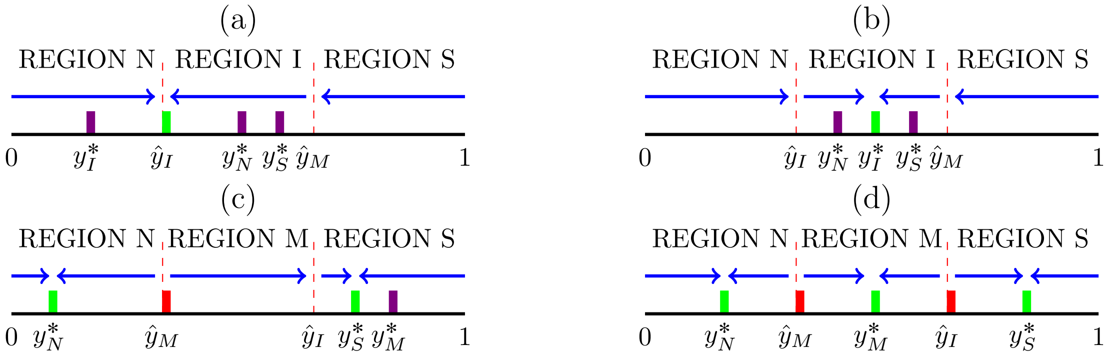

Figure 1: Examples of phase diagrams.

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. Phase diagrams for all possible configurations are presented in Figures 2, 3, 4 and 5 in Online Appendix C.4.

All quasi steady states that are within their regions are stable steady states for the LDI. As demonstrated in Figure 1, there can be anywhere from 0 to 3 such steady states; we denote this set as $\mathcal{Q} = \{y_R^* | y_R^* \in \text{region}\, R\}$. Since every limit ODE has a unique steady state, the only other candidate steady states for the LDI are the thresholds.

For a threshold $\hat{y}$ to be a stable steady state, the flow above it needs to point down and the flow below it needs to point up. This requires a "flip" of quasi steady states: Let W be the region to the left of $\hat{y}$, and Z the region to the right, a flip is $y_Z^* < \hat{y} < y_W^*$. Flips around $\hat{y}_I$ occur when $\hat{y}_I < \hat{y}_M$ and $y_I^* < \hat{y}_I < y_N^*$ (as in phase diagram (a) in Figure 1), or $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$. Online Appendix C.3 shows that both cases are possible, and Lemma 4 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) $\mathcal{S} = \mathcal{Q} \cup \{\hat{y}_I\}$, or (b) $\mathcal{S} = \mathcal{Q}$. Case (a) obtains if and only if $\hat{y}_I < \hat{y}_M$ and $y_I^* < \hat{y}_I < y_N^*$ or $\hat{y}_I > \hat{y}_M$ and $y_S^* < \hat{y}_I < y_M^*$.*

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 3 in Appendix B shows that y_n almost surely converges to a steady state of the LDI. Second, Lemma 6 in Appendix A shows that every steady state has positive probability of being the limit point. Finally, Lemma 7 in Appendix A shows that the system almost surely does not converge to a repelling 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.[15]

[15] For a threshold $\hat{y}$ to be a steady state that is neither stable or repelling, the flow must have the same sign (positive or negative) on both sides of $\hat{y}$. This is only possible at a steady state threshold in the knife-edge case where it is also a quasi steady state, which we have ruled out.

Detailed Proof Summary

Theorem 3 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.[16]

To prove Lemma 6, 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 sharing 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 7, that y_n almost surely does not converge to a repelling steady state, uses Theorem 4 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

[16] Here overline denotes the closure. The proof of Theorem 3 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.)

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 that add stories to the platform, and that such shocks will be more likely to change limit behavior if they occur early, when the overall number of stories is small.[17]

6 Comparative statics

The previous section characterized the set of limit points for every parameter specification; now we study how the the limit points change with the parameters. Since all candidate limit points are roots of continuously differentiable functions, we can apply the implicit function theorem to obtain comparative statics of these points with respect to all parameters.[18] It is straightforward to verify that $y_N^* = \frac{1}{1+\kappa}$, so this candidate limit point is decreasing in κ and constant in all other parameters. Theorems 5-8 in Online Appendix C.2 present comparative statics for each of the others. We now discuss the main takeaways from these theorems.

As one would expect, the other candidate limit points are increasing in the loss μ from sharing false stories. Additionally, any limit point that is a quasi steady state is decreasing in the exogenous inflow of false stories κ and all but y_N^* are decreasing in the cost of attention β. More surprisingly, the limit point $\hat{y}_I$ can be increasing in β or constant in κ. Recall that $\hat{y}_I$ is the point where users are exactly indifferent between sharing and not sharing very interesting stories. It is increasing in β because users' payoffs are decreasing in the cost of attention and increasing in the share of true stories. Hence, when β goes up, the share of true stories required for indifference needs to go up as well to compensate for the utility loss. $\hat{y}_I$ does not depend on

[17] The long-run outcome is not changed by these additions when there is a unique stable steady state.

[18] Each quasi steady is the root of its respective limit ODE, and the thresholds are the roots of their respective value functions.

κ, since the exogenous inflow of false stories is is not an argument in users' utility functions. However, as we show below, when κ is sufficiently large $\hat{y}_I$ will not be a limit point.

Table 2: **Comparative Statics for** κ

y_M^*, y_S^*, y_I^*	Decreasing.
$\hat{y}_I$	Constant.

Table 3: **Comparative Statics for** β

y_M^*, y_S^*, y_I^*	Decreasing.
$\hat{y}_I$	Increasing.

The quasi steady states y_N^* and y_I^* are constant in the evocativeness parameter λ; all other candidate limit points are decreasing in λ. As λ increases, users pay less attention to the veracity of very interesting stories. This leads to a decrease in the share of true stories in any limit point where users share very interesting stories, which is the case when λ is sufficiently large. Comparative statics with respect to the remaining parameters are more nuanced. We discuss each of them in turn, starting with θ, which measures the "credibility" of false stories.

The role of θ

Table 4: **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.

When θ increases it is harder to identify false stories, but users are aware of this and pay more attention (both $a(y, I)$ and $a(y, M)$ are increasing in θ). This leads to two opposing forces on the limit share of true stories, and our model predicts that either one can prevail: The candidate limit points y_S^*, y_M^* and y_I^* are decreasing in θ up to a point and then increasing in θ, so for sufficiently large values of θ the increase in attention more than compensates for the increase in credibility.[19] The candidate limit point $\hat{y}_I$ behaves differently, as it is always increasing in θ: Users' payoffs from sharing are decreasing in θ so $\hat{y}_I$ needs to increase to maintain indifference.

[19] The comparative statics in Table 4 allow for the case that a quasi steady state y_R^* is everywhere decreasing in θ (this is the case if $\theta_R = 1$). However, Online Appendix C.3 shows that all quasi steady states except y_N^* can be non-monotone in θ when they are limit points.

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 of the quasi steady states with respect to θ 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. The comparative statics for the quasi steady states $\{y_S^*, y_I^*, y_M^*\}$ are a manifestation of the "implied truth effect" empirically demonstrated in Pennycook et al. (2020a), where false content that is not flagged as false is considered validated and seen as more accurate than in the case where no content is flagged. Our results show that this effect can generate a non-monotonic relationship between flagging rates and the share of true stories.[20] Finally, the comparative statics with respect to $\hat{y}_I$ imply that the limit share of true stories may be everywhere decreasing in the flagging rate, through the constraint that users are indifferent, a mechanism distinct from the implied truth effect.

The role of δ

Table 5: **Comparative Statics for δ**

y_M^*	Increasing.
y_S^*	Decreasing for δ close to $\frac{1}{2}$, and increasing for δ close to 1.
y_I^*	Decreasing.
$\hat{y}_I$	Increasing.

Increasing δ means false stories are more likely to be very interesting, so the comparative statics for y_I^*, y_M^* are intuitive—the limit share of true stories decreases (increases) in δ when users share only very interesting (mildly interesting) stories. The quasi steady state y_S^*, where users share both types of stories, decreases in δ when δ is close to $\frac{1}{2}$, and increases in δ when δ is close to 1. Appendix C presents numerical examples where y_S^* is both decreasing and increasing in δ when it is a limit point. Intuitively, the non-monotonicity arises because when δ is close to $\frac{1}{2}$ users

[20] We find that no flagging can lead to more accurate beliefs than poor flagging. In Acemoglu, Ozdaglar, and Siderius (2023), a regulator who cares about the accuracy of users' beliefs may censor less misinformation than is technologically feasible, but will always prefer some censorship to none.

are sharing more very interesting stories than mildly interesting stories, since both types of stories are almost equally likely to be false and very interesting stories have additional value. In this case, the comparative statics with respect to δ are similar to those comparative statics for y_I^*, where users are only sharing very interesting stories. As δ moves closer to 1, the stories that users share are more likely to be mildly interesting and comparative statics with respect to δ eventually become similar to those for y_M^*. Finally, $\hat{y}_I$ is increasing in δ because for a fixed y_n, increasing δ leads to a decrease in the value from sharing very interesting stories.[21]

The role of ρ

Table 6: **Comparative Statics for ρ**

y_M^*	Increasing.
y_S^*	Increasing.
y_I^*	Increasing if $\frac{1}{2} > \delta\theta\left(1 - a(y, I)\right)$, decreasing if the inequality is reversed.
$\hat{y}_I$	Constant.

The reach parameter has no effect on the location of $\hat{y}_I$ because it is not an argument in users' payoffs. Quasi steady states are increasing in the reach parameter when users are successfully filtering false content. As mentioned in our discussion of Lemma 4 this is always the case in regions S and M but may not be the case in region I, where users are sharing $\frac{1}{2}$ of all true stories and $\delta\theta(1 - a(y, I))$ of all false stories. We find that either term can be larger when $y_I^* \in I$, so that y_I^* can be either increasing or decreasing in ρ when it is a limit point.

A common view is that social media platforms are hotbeds for false news because users can easily disseminate content to large audiences. However, an analysis of the effect of reach on the share of false stories must also consider that greater reach increases the exposure of true stories as well. Therefore, reach will only have a negative impact if users are ineffective at filtering false content. Our analysis 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.

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

The composition of $\mathcal{S}$

Making general statements about how the composition of $\mathcal{S}$ varies with parameters is challenging given the large number of possible configurations. One clear example is the effect of κ, the production rate of false stories. For sufficiently large values of κ, all quasi steady states fall in the no sharing region, and the unique limit point is y^*_N. For sufficiently small values of κ, all quasi steady states fall in the sharing region and the unique limit point is y^*_S. Thus, increasing the production rate of false stories from low to high changes limit behavior from sharing both very interesting and mildly interesting stories to not sharing at all. Since we saw above that when users are sharing stories of both evocativeness levels they are successfully filtering false content, the exogenous decrease in the share of incoming stories that are true is amplified by user behavior.[22]

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.

The limit share of true stories may be either increasing or decreasing in each of the following parameters: the cost of attention, the credibility of false stories, the probability that false stories are very interesting, and the reach of shared stories. Although endogenous attention creates a counterbalancing force to changes in the credibility/flagging of false stories, it can intensify the effect of producing more false stories. This suggests that interventions that target producers of false news might be more efficient than attempts to stop the spread of false news already on the platform.

Our model captures many important features in a tractable framework, and parts with most of the literature by tracking the evolution of the entire platform rather than

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

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.

Appendix 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 $\mathbb{P}_{a,y}(T', F|M) = \dfrac{2(1-y)(1-\delta)\theta(1-a)}{y + 2(1-y)(1-\delta)}$. By (2), the expected payoff when attention is a, evocativeness is M and the user will share the story if and only if they receive the signal T' is,

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

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

The proof of Lemma 2 is standard and relegated to the Online Appendix C.1.

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 $y_N^* = \frac{1}{1+k}$. For the other regions, it suffices to prove that $g_R'''(y) > 0$ for all $y \in [0,1]$. Indeed, for $g_R(y)$ to have more than one root in $[0,1]$ it must have a local minimum that is greater than the first root, followed by a local maximum (between the second root and $y = 1$). So, there need to be $0 < w < z < 1$ such that $g_R''(w) \geqslant 0$ while $g_R''(z) \leqslant 0$ which cannot be the case if $g_R'''(y) > 0$ for all $y \in [0,1]$. The derivatives are

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

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

Proof of Lemma 4. By (6), we have for any $R, W \in \{N, I, M, S\}$:

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

So $g_R(y) > g_W(y)$ if and only if $(1-y)\left(p_R^T(y) - p_W^T(y)\right) > y\left(p_R^F(y) - p_W^F(y)\right)$. Hence, $g_S(y) > g_I(y)$ for all $y \in (0,1)$ because, by (3),

$$\begin{aligned}
(1-y)\left(p_S^T(y) - p_I^T(y)\right) &= (1-y)\frac{y}{2},\\
y\left(p_S^F(y) - p_I^F(y)\right) &= y(1-y)\theta(1-\delta)\left(1 - a(y,M)\right),
\end{aligned}$$

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

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

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

To see that $g_M(y) > g_I(y)$ for all $y \in (0,1)$ note that $\left(p_M^T(y) - p_I^T(y)\right) = 0$ and

$y\left(p_M^F(y) - p_I^F(y)\right) = y(1-y)\theta\left((1-\delta)\left(1-a(y,M)\right) - \delta\left(1-a(y,I)\right)\right)$, so $g_M(y) > g_I(y)$ if and only if $(1-\delta)\left(1-a(y,M)\right) < \delta\left(1-a(y,I)\right)$. Fix $y \in (0,1)$ and let $\ell(\delta) = (1-\delta)\left(1-a(y,M)\right); r(\delta) = \delta\left(1-a(y,I)\right)$. 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,

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

Now,

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

Assumption 2 and $\lambda < 1$ imply that $\theta(\mu-1) < 2\beta$. Therefore, it suffices to prove $4(1-\delta)(1-y)(1-\delta(1-y)) < (y+2(1-y)(1-\delta))^2$, which simplifies to $y^2 > 0$. Hence, $\frac{\partial \ell(\delta)}{\partial \delta} < 0$. Finally, by Assumption 1,

$$\frac{\partial r(\delta)}{\partial \delta} = \frac{2\delta\theta(1-y)(\mu-1-\lambda)(\delta+y(1-\delta))}{\beta(y+2(1-y)\delta)^2} > 0,$$

which completes the proof that $\min\{y_S^*, y_M^*\} > y_I^*$.

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

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

which always holds.

Finally, $g_M(y) > g_N(y)$ if and only if

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

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

Proof of Theorem 1. That $\mathcal{Q} \subset \mathcal{S}$ follows immediately from the definitions of these sets and of F. Since each limit ODE has a unique steady state, the only other possible members of $\mathcal{S}$ are the thresholds between the regions, so $\mathcal{S} \subset \mathcal{Q} \bigcup \{\hat{y}_I, \hat{y}_M\}$. A threshold $\hat{y}$ is a stable steady state if for all $y \in (\hat{y}-\epsilon, \hat{y}+\epsilon)$ we have $\text{sign}(x) = \text{sign}(\hat{y}-y)$ for all $x \in F(y)$. This holds only if there is a "flip" of quasi steady states: Let W be the region to the left of $\hat{y}$, and Z the region to the right, a flip is: $y^*_Z < \hat{y} < y^*_W$. Flips around $\hat{y}_I$ occur if and only if one the following holds: $\hat{y}_I < \hat{y}_M$ and $y^*_I < \hat{y}_I < y^*_N$; or $\hat{y}_I > \hat{y}_M$ and $y^*_S < \hat{y}_I < y^*_M$. In Appendix C 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 4, $y^*_S > y^*_I$. In Case 2 a flip cannot occur because by Lemma 4, $y^*_M > y^*_N$. ■

Proof of Theorem 2. When $y^*_N \in N$ and $y_0 \in N$, the system follows the law of motion $z_{n+1} = z_n + \begin{pmatrix} 1 \\ k \end{pmatrix}$, so it never leaves the region N and converges deterministically to $y^*_N = \frac{1}{1+\kappa}$. We henceforth assume that $y^*_N \notin N$ and/or $y_0 \notin N$. By Theorem 3 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 simply its steady states, so y_n converges almost surely to a steady state of the LDI. By Lemma 6 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 7 there is zero probability of convergence to any repelling steady state, which completes the proof. ■

Lemma 6 and Lemma 7 below are used to prove Theorem 2, and Lemma 5 is used to prove Lemma 6.

Lemma 5. *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 6. *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 3, $y_{n;R}$ 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 3 implies that $y_{n;H}$ converges to $\hat{y}_I$ almost surely. In particular, there exists $m_H \in \mathbb{N}$ such that $\mathbb{P}\left(y_{m;H} \in B_\epsilon(\hat{y}_I)\, \forall m > n \,|\, y_{n;H} \in B_\epsilon(\hat{y}_I), |z_{n;H}| > m_H\right) > 0$.

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

Assume w.l.o.g. that ϵ is small enough that $B_\epsilon(y_R^*) \subset R$ if $\psi = y_R^*$ for some region R and that $B_\epsilon(\hat{y}_I) \subset [0,1]\backslash O$ if $\psi = \hat{y}_I$ (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

$$\frac{1}{1+\kappa+\rho} < \psi < \frac{1+\rho}{1+\kappa+\rho} \quad \forall \psi \in \mathcal{S}. \tag{8}$$

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 (8) and Lemma 5 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 5 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 5 implies there is positive probability of arriving from this y_n to $B_\epsilon(\psi)$. ■

Lemma 7. *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. Let A denote the event "$y_n \in N$ infinitely often" and let A^C denote its complement. We will prove that $\mathbb{P}(y_n \to \hat{y}) = 0$ by proving that if $\mathbb{P}(A) > 0$ then $\mathbb{P}(y_n \to \hat{y}|A) = 0$, and if $\mathbb{P}(A^C) > 0$ then $\mathbb{P}(y_n \to \hat{y}|A^C) = 0$.

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

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

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

$$\mathbb{E}[\xi_{n+1}^+|\mathcal{F}_n] \geqslant p_R^T(y_n)\left(\Delta_T - \sum_{i \in \{T,F,O\}} p_R^i(y_n)\Delta_i\right)|z_n| \geqslant p_R^T(y_n)(1-p_R^T(y_n))(\Delta_T - \Delta_O)|z_n|.$$

Now, for sufficiently large $|z_n|$,

$$(\Delta_T - \Delta_O) = \frac{(\kappa + |z_n|(1-y_n))\rho}{(|z_n|+1+\kappa)(|z_n|+1+k+\rho)} \geqslant \frac{|z_n|(1-y_n)\rho}{4|z_n|^2},$$

so $(\Delta_T - \Delta_O)|z_n| \geqslant \dfrac{(1-y_n)\rho}{4}$. Since $y_n \notin N$ from some point onward, by (3), $p_R^T(y_n) \in \{y_n, \frac{y_n}{2}\}$ for both of the adjacent regions R. Thus, for small $\epsilon > 0$, there exists $c > 0$ such that for any $y_n \in (\hat{y} - \epsilon, \hat{y} + \epsilon)$: $p_R^T(y_n)(1 - p_R^T(y_n)) \geqslant c$. So, for sufficiently large n, $\mathbb{E}[\xi_{n+1}^+ | \mathcal{F}_n] \geqslant \dfrac{c(1-y_n)\rho}{4} > 0$ for any $y_n \in (\hat{y} - \epsilon, \hat{y} + e)$. ■

Appendix 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. Theorem 3 relates the limit behavior of a PGPU to the limit behavior of the associated differential inclusion; we use it in the proof of Theorem 2. Section B.3 explains why the processes $\{z_{n;R}\}$ defined in (4) are GPUs and derives the corresponding limit ODEs. 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.

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 real number $a > 0$ such that: $\left| p^w \left(\frac{z^1}{|z|} \right) - \Pi(z, z+w) \right| \leqslant \frac{a}{|z|}$ for all nonzero $z \in \mathbb{Z}_+^2$.

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 $n \in \mathbb{N}$:

$$x_{n+1} - x_n = \gamma_n \left(g(x_n) + \xi_{n+1} + R_n \right), \tag{9}$$

where γ_n are non-negative with $\gamma_n \to 0, \sum_n \gamma_n = \infty$, 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 (9) 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[23]

$$g_k(y) = \sum_{w \in \mathbb{Z}^2} p_k^w(y) \left(w^1 - y|w| \right). \tag{10}$$

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

[23] Note that condition i. in Definition 1 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.[24]

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

$$\frac{dy}{dt} \in F(y) \tag{11}$$

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 8. *The inclusion (11) 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 (11) using the ideas of a *perturbed solution* and a *piecewise affine interpolation.*

[24] 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 (11) (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.*[25] *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 9. *Let $\{z_n\}$ be a PGPU and (11) its limit differential inclusion, and let $\mathbf{Y}$ be its piecewise affine interpolation. Then $\mathbf{Y}$ is a bounded perturbed solution to (11).*

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

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

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

[25] 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 $1 \leqslant j \leqslant M$.[26] 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 3. 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.*[27]

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

Proof. By Lemma 9, the interpolation $\mathbf{Y}$ is a perturbed solution to F. Note that it is also bounded since $\mathbf{Y}(t) \in [0, 1]$ for all $t \geqslant 0$. Thus, Theorem 3.6 in BHS implies that the limit set of $\mathbf{Y}$ is internally chain transitive for F. Note that the asymptotic behaviors of $\mathbf{Y}(t)$ and y_n are the same by the definition of interpolation, i.e., $L(y_n) = L(\mathbf{Y})$, which completes the proof.

■

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

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

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

Next, (3), (4), and (10) 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 (11). 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 4. *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 (9) 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 $g : [0, 1] \to \mathbb{R}$. By

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

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

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

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

$$y_{n+1} - y_n = \frac{1}{|z_n|} (f(y_n) + \xi_{n+1} + R_n) \tag{12}$$

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 a real number $K > 0$ such that $|R_n| \leqslant \frac{K}{|z_n|}$. Thus, $\sum_{n=1}^{\infty} \frac{|R_n|}{n} < \infty$, so $\{y_n\}$ is a stochastic approximation. By the same Lemma, $|\xi_n| \leqslant 4m$ where m is the maximal number of balls added in each period. This implies that $\mathbb{E}[\xi_n^+|\mathcal{F}_n], \mathbb{E}[\xi_n^-|\mathcal{F}_n]$ are bounded from above by $4m$. To apply Theorem 2.9, it remains to prove that $\mathbb{E}[\xi_n^+|\mathcal{F}_n], \mathbb{E}[\xi_n^-|\mathcal{F}_n]$ are bounded from below by a positive number when $y_n \in (\hat{y} - \epsilon, \hat{y} + \epsilon)$. 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 (Dec. 2023). "A model of online misinformation". *The Review of Economic Studies*, rdad111. ISSN: 0034-6527.

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

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.

Berger, Jonah and Katherine L Milkman (2012). "What makes online content viral?" *Journal of Marketing Research* 49.2, pp. 192–205.

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.

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

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar (2020). "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India". *Proceedings of the National Academy of Sciences* 117.27, pp. 15536–15545.

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.

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 (2020a). “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, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand (2020b). “Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention”. *Psychological Science* 31.7, pp. 770–780.

Pennycook, Gordon and David G Rand (2022). “Nudging social media toward accuracy”. *The Annals of the American Academy of Political and Social Science* 700.1, pp. 152–164.

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

Van Bavel, Jay J and Andrea Pereira (2018). “The partisan brain: An identity-based model of political belief”. *Trends in Cognitive Sciences* 22.3, pp. 213–224.

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