Social Attention and Information Transmission

David McAdams*

August 29, 2025

Abstract

This paper develops an equilibrium model of socially-transmitted information in which agents allocate attention between original sources and social interaction, decide which stories to retain as knowledge after observing private signals, and then socially transmit the stories they know to others. Social attention is a strategic complement: as people listen more to one another, repeated filtering makes the stories in social circulation more likely to be true. Two extensions show how this truth-promoting dynamic can be undermined. First, if some stories possess an "attractive" characteristic that induces automatic retention, then attractive stories bypass filtering and crowd out neutral stories, lowering the veracity of all stories in circulation. Second, if some agents place intrinsic value on partisan-aligned stories, they either retreat into partisan low-veracity bubbles or split their attention across partisan and non-partisan sources, depending on the strength of their partisan bias. These results explain how virality and identity-driven preferences can lead information ecosystems to become saturated with misinformation despite the filtering power of social transmission.

In the months leading up to and just after the 2020 U.S. presidential election, many Americans eagerly shared a wave of false claims about mail-in voting fraud and election-rigging (collectively, "Stop the Steal") that were being amplified by President Donald

*Fuqua School of Business and Economics Department, Duke U. Email: david.mcadams@duke.edu. Thanks Ben Golub, David Hirshleifer, Rachel Kranton, attendees 10th Annual Network Science in Economics conference.

Trump. The persistence of Stop the Steal, which more than a third of Americans continued to believe years later (Pengelly 2024), presents a puzzle. In many other settings, when individuals share information based on personal preference, social transmission tends to elevate high-quality content. For example, Trattner, Moesslang, and Elsweiler (2018) found that recipes shared more widely on the food-focused social network Allrecipes.com tended to be both tastier and healthier. Why does social transmission "work" for recipes but fail when it comes to promoting truth in politics?

This paper develops an equilibrium model in which agents choose where to seek out pieces of information called "stories," decide what stories to retain as knowledge after receiving informative private signals, and then transmit stories that they know to others during social interactions. In the plain-vanilla benchmark model, each story is simply either "true" or "false" and all agents prefer to know only true stories. Agents have an attention budget that they may allocate across original sources (like newspapers) and social interactions. By assumption, consulting an original source results in exposure to a randomly-selected "original story" while listening to someone results in exposure to a randomly-selected story from that person's store of knowledge—or nothing if that person does not currently know anything. Upon being exposed to a story, an agent then receives a private signal about whether it is true and decides whether to retain it as knowledge. Each known story is forgotten at a fixed exogenous rate.

A key finding is that agents' social-attention decisions are strategic complements: the more that agents across the population listen to each other, the more valuable that listening becomes. Intuitively, the reason is that each social-transmission event serves to *filter* the stories in social circulation, since agents will only retain (and subsequently transmit) a story if they get a sufficiently favorable private signal. If agents rarely listen to one another, most stories will be forgotten before they are ever transmitted and the transmission chains along which stories flow will be quite short. On the other hand,

if agents frequently pass along stories socially, most stories that agents encounter will have been transmitted (and filtered) many times, increasing their veracity and leading everyone ultimately to rely on social information channels.

To highlight the effects of this strategic complementarity, I assume in the analysis that no one in the population knows anything at time $t = 0$. Each agent therefore has no incentive initially to seek out social interaction[1] and gets their information exclusively from original sources. However, since agents keep only the stories that look sufficiently promising given their signals, the pool of stories they know is positively selected relative to the original-story pool and hence more likely to be true. So long as agents don't forget stories too quickly, a time will therefore come when agents find it worthwhile to start listening to others—at which point the engine of social transmission kicks into gear and social veracity begins increasing even further due to the repeated filtering of the stories in social circulation.

These benchmark findings align well with the common wisdom that truth should win out in the marketplace of ideas,[2] as repeated exposure and social filtering tend to elevate accurate stories over time. But then how should we understand the widespread diffusion and persistence of false information, such as COVID-19 vaccine misinformation during the pandemic (Pierri et al. 2023) or politically-inspired falsehoods like the Stop the Steal myth? To explore this question, I consider two extensions of the basic model.

The first extension considers the possibility that stories possess observable characteristics, beyond their truth value, that affect how people respond to them. In practice, stories are not only judged on accuracy but also on whether they are funny, emotionally resonant, or otherwise compelling. To capture this, I distinguish between "neutral"

[1]Extensions in which social interaction generates benefits *other* than being exposed to stories are discussed in the concluding remarks.

[2]Justice Oliver Wendell Holmes, dissenting in *Abrams v. United States*, 250 U.S. 616, 630 (1919): "The best test of truth is the power of the thought to get itself accepted in the competition of the market."

stories, which are filtered through agents' signals as in the benchmark model, and "attractive" stories, which agents retain automatically upon exposure. This modification highlights how emotional salience, virality, or other transmission-driving features can shape the social-informational environment.

The key finding of this *attractive-story extension* is that all stories in social circulation (even the neutral ones) are less likely to be true once attractive stories are introduced into the mix. The mechanism is straightforward. Because attractive stories are always retained, they bypass the filtering process that normally weeds out falsehoods; so, the share of true stories among those that are attractive is lower than in the benchmark. At the same time, the fact that attractive stories spread more widely "crowds out" the neutral stories in social circulation, shortening their transmission chains and weakening the filtering effect that otherwise would have sustained their higher veracity. As a result, neutral stories are in equilibrium less likely to be true than in the benchmark model, and attractive stories have the lowest veracity of all.

The second extension allows for heterogeneity across agents, recognizing that agents may differ in what sorts of information they want to know and, as a result, may choose different patterns of social interaction. Motivated by the Stop the Steal example, I divide the agent population into two types: "non-partisans" and "partisans." Non-partisans are like the agents in the benchmark model, only caring about story truth, while partisans are like the agents in the attraction-story extension, with partisan-alignment being the attractive characteristic.

Three main findings emerge in this *partisan-agent extension*. First, non-partisan agents optimally ignore the partisan population and direct all of their social attention to other non-partisans; so, their individual and collective knowledge evolves exactly as in the benchmark model. Second, as in the attractive-story extension, partisan agents' attraction to aligned content leads them to retain mostly partisan stories, which in equilibrium

are less likely to be true than those known by non-partisans. Third, whether partisans form a social-attention bubble depends on how much they value partisan-alignment for its own sake. If this preference is very strong, they focus exclusively on fellow partisans. But if it is more moderate, equilibrium requires partisans to *split their attention* between non-partisans and partisans; such equilibrium attention-mixing can leave partisans with a pool of knowledge that is markedly more accurate than if they were only able to listen to other partisans.

These findings underscore the sensitivity of social-informational environments to the transmission-driving features of the information in circulation. This point is all the more important as growing volumes of content are deliberately engineered to optimize virality, exploiting qualities that make stories more likely to spread. The rise of strong partisan and tribal identities has further expanded the menu of such "attractive" features, which information creators can leverage to capture attention and engagement. The analysis here shows that as such features become more prevalent, the equilibrium veracity of what people know is reduced—even if the stories with attractive features are ex ante just as likely to be true as others. In short, the very features that make content spread widely can simultaneously erode its reliability and weaken society's shared understanding of what is true.

Related literature. A central puzzle is why misinformation and low-quality content persist even when agents are rational and receive informative signals. Empirical work shows that false news often spreads faster and farther than true news on social media (Vosoughi, Roy, and Aral 2018; Grinberg et al. 2019), yet classical models of social learning suggest that information aggregation should drive society toward the truth. Information cascade models demonstrate how initial errors can lock in and persist (Bikhchandani, Hirshleifer, and Welch 1992), while Banerjee and Fudenberg (2004) show that

when agents only learn by word-of-mouth from a few peers, biased sampling rules can sustain inefficient outcomes. More recent network models add rumor verification and homophily, showing that truth and falsehood often coexist in steady states (Merlino, Pin, and Tabasso 2023). These approaches highlight how structural frictions in transmission and verification can sustain errors. By contrast, the present paper focuses on how the characteristics of the information itself, and the preferences of the agents who circulate it, shape long-run veracity.

A second strand of research emphasizes how selective sharing and audience misperceptions can generate persistent disagreement and polarization, even when the underlying signals are informative. Bowen, Dmitriev, and Galperti (2023) show that if some agents share selectively and others underestimate this selection, beliefs can diverge sharply, with polarization worsening as networks expand. Hsu, Ajorlou, and Jadbabaie (2021) model partisan cascades in which heterogeneous priors and persuasion motives drive the spread of low-credibility content through strategic sharing. Relatedly, Eliaz and Spiegler (2020) analyze how competing narratives can persist in equilibrium when individuals are drawn to stories that provide hopeful or identity-affirming interpretations of events. Together, these contributions highlight how the preferences and perceptions of those who transmit information, rather than the informativeness of signals alone, can shape the content that persists in circulation. My analysis builds on this insight by modeling how agents' attraction to certain types of stories alters the equilibrium filtering process of social transmission itself.

A third line of work focuses on how differences in shareability or algorithmic amplification shape long-run learning. Buechel et al. (2023) develop a model in which asymmetries in signal sharing—arising when some messages are inherently more "infectious" than others—create a substantial probability of misinformation dominating in equilibrium. Dasaratha and He (2025) show that when platform news feeds place too much

weight on viral content, rational agents can converge to misleading steady states in which most shared stories are false, even if users prefer to share truth. Complementing these theoretical contributions, empirical studies document that false or attention-grabbing content spreads more rapidly and broadly than accurate information (Vosoughi, Roy, and Aral 2018). These findings underscore the broader point that transmission-driving features—whether engineered by platforms or embedded in the content itself—can weaken the natural filtering force of social learning. This paper provides a simple microfoundation for this mechanism, showing how the rise of virally-engineered information reduces the equilibrium veracity of all the information in social circulation.

A final strand of work studies interventions designed to mitigate the spread of misinformation. Papanastasiou (2020) shows how platform fact-checking policies can alter diffusion dynamics. Experimental work by Pennycook et al. (2021) demonstrates that subtle accuracy prompts can improve the quality of what users share. These contributions suggest ways to repair information environments once distortion has taken root. My results are complementary, highlighting how interventions become necessary in the first place: when the natural filtering force of social transmission is weakened by emotionally-resonant or identity-affirming content, truth no longer prevails on its own, creating demand for external mechanisms to restore reliability. This paper's analysis also points to a novel lever for intervention: platform design or policy efforts that reduce the salience of attractive or identity-affirming features—or that strengthen cross-group interactions—can help revitalize social transmission and its ability to promote high-veracity content.

The most closely related paper is "Endogenous Attention and the Spread of False News" by Danenberg and Fudenberg (2024). Like my attractive-stories extension, their model features stories with an observable characteristic that boosts transmission, and in both settings the composition of the story pool evolves endogenously through trans-

mission. There are three main differences. First, they assume truth is negatively correlated with the transmission-driving characteristic, whereas in my model this correlation emerges endogenously in equilibrium. Second, they study a single type of agent, while I extend my analysis to a setting with heterogeneous agents. Third, "attention" captures very different things in each paper: Danenberg and Fudenberg model agents as being able to exert effort to determine whether any given story is true, while this paper models agents as deciding how much information to seek out socially. That is, their agents decide how much attention to pay to individual stories, while my agents decide how much attention to pay to social-information channels.

The rest of the paper is organized as follows. Section 1 presents the benchmark model, which has identical agents who only care about whether stories are true or false. Section 2 provides equilibrium analysis of the benchmark model. Section 3 presents the attractive-story extension. Section 4 presents the partisan-agent extension.

1 Benchmark Model

Stories. There is a unit mass of original stories, referred to as "Pool O," each of which is either true ($\omega = T$) or false ($\omega = F$).

Definition 1 (Original-story veracity)**.** *The veracity of Pool O, $p_O \in (0, 1)$, is the share of original stories that are true.*

Agents. There is a unit mass of agents. An agent who encounters a story decides whether to "retain" or "ignore" it. An agent who retains a story gets payoff $\pi_T > 0$ if the story is true or $\pi_F < 0$ if it is false; ignoring a story guarantees zero payoff.

Knowledge. Let $\mathcal{K}^i(t)$ be the stories known by agent i at time t. Let $N_m(t)$ be the mass of agents who know $m \geq 0$ stories at time t. Let $\mathcal{K}(t) = \uplus_i \mathcal{K}^i(t)$ be the multi-set union of the stories known by all agents, counting each instance of the same story as a distinct copy. This set of stories, which may then subsequently be shared socially, is referred to as "Pool S."

Definition 2 (Knowledge prevalence)**.** *Let $K(t) \equiv 1 - N_0(t)$ be the share of agents who know at least one story at time t, referred to as "knowledge prevalence."*

Let $M_T(t), M_F(t)$ be the masses of true and false stories in Pool S at time t, respectively. Define shorthand $\Sigma M(t) = M_T(t) + M_F(t)$ for the volume of agents' collective knowledge. $M(t) \equiv (M_T(t), M_F(t))$ is the *story-pool state*. $\mathbf{M} \equiv (M(t) : t \geq 0)$ is the *story-pool trajectory*.

Definition 3 (Social veracity)**.** *The veracity of Pool S, $p_S(t) \equiv \frac{M_T(t)}{\Sigma M(t)}$, is the share of stories known by agents at time t that are true. $p_S(t)$ is referred to as "social veracity."*

Story exposures and social attention. Agents are exposed to stories either through original sources or by listening to / following others. *Non-social exposures:* Each agent i encounters an original source at rate $\lambda_O^i(t)$, whereupon i is exposed to a randomly-selected story from Pool O. *Social exposures:* Each agent i encounters another randomly-selected agent i' at rate $\lambda_S^i(t)$, at which point agent i "listens" to agent i'. With probability $N_0(t)$, agent i' knows nothing and agent i is not exposed to a story; otherwise, agent i is exposed to a randomly-selected story from $\mathcal{K}^{i'}(t)$.

Upon encountering a story, agent i observes its source and receives a conditionally i.i.d. private signal $s^i \in [0, \infty)$ drawn from a distribution with continuous p.d.f. $g(s^i|\omega)$ and c.d.f. $G(s^i|\omega)$. The likelihood ratio $L(s^i) \equiv \frac{g(s^i|\omega=T)}{g(s^i|\omega=F)}$ is assumed strictly increasing in s_i, with $L(0) = 0$ and $\lim_{s^i \to \infty} L(s^i) = \infty$. Define shorthand $L^i \equiv L(s^i)$ and let $h(L^i|\omega)$

and $H(L^i|\omega)$ be the p.d.f. and c.d.f. of L^i. Since L^i is the realized likelihood ratio of the underlying private signal, $\frac{h(L^i|\omega=T)}{h(L^i|\omega=F)} = L^i$.

Attention allocation. The story-encounter rates $\lambda_O^i(t)$ and $\lambda_S^i(t)$ capture agent i's "attention" to original and social sources of information. Each agent chooses these attention levels at each point in time $t \geq 0$ subject to *attention budget* $\lambda_O^i(t) + \lambda_S^i(t) \leq \overline{\lambda}$, $\lambda_O^i(t) \geq \underline{\lambda}_O > 0$, and $\lambda_S^i(t) \geq 0$. Agents "focus on original sources" if $(\lambda_O^i(t), \lambda_S^i(t)) = (\overline{\lambda}, 0)$ for all agents i, and "focus on social sources" if $(\lambda_O^i(t), \lambda_S^i(t)) = (\underline{\lambda}_O, \overline{\lambda} - \underline{\lambda}_O)$.

Forgetting. Agents forget each story that they know independently at rate $\gamma > 0$.

Initial condition. The analysis can be easily adapted to any initial condition. For convenience, I assume agents know nothing at time $t = 0$, i.e., $M(0) = (0, 0)$.[3]

Equilibrium. An *equilibrium trajectory* is a story-pool trajectory $\mathbf{M}^*$ that arises from the initial condition, when every agent allocates attention optimally and retains stories optimally at each point in time t.

2 Equilibrium Social Transmission

This section analyzes the benchmark model, characterizing the unique equilibrium trajectory (Theorem 1). A key insight that emerge from this analysis is that there is an underlying strategic complementarity in the social transmission of information. The more that others engage in conversation and filter stories through social information channels, the more attractive these social channels become and the more that agents rely on them. For this reason, equilibrium social-transmission dynamics also tend to exhibit a *cold start*,

[3]This assumption allows me to establish clean monotonicity results about the equilibrium trajectory.

in that agents choose not to interact socially until a critical moment (denoted $\tilde{t}$ in Theorem 1) at which the value of conversing with others passes above a threshold and the engine of social transmission discontinuously kicks into gear.

2.1 Characterizing the equilibrium trajectory

Theorem 1. *(i) There is a unique equilibrium story-pool trajectory,*[4] *which converges in the long run to an equilibrium steady state. (ii) Equilibrium social veracity $p_S^*(t)$ is strictly increasing over time to $p^{*\infty} \equiv \lim_{t\to\infty} p^*(t)$. (iii) There is a critical time $\tilde{t} \in (0, \infty]$ such that agents focus on original sources up until time $\tilde{t}$ and thereafter focus on social sources.*

The rest of this section proves Theorem 1. For ease of exposition, the proof is organized into several parts.

A. Individually-optimal story retention. Suppose that agent i believes that a story has probability p_i of being true. Retaining this story yields expected knowing payoff $\pi_T p_i + \pi_F(1 - p_i)$, which is positive if and only if $p_i > \hat{p}$:

$$\hat{p} \equiv \frac{-\pi_F}{\pi_T - \pi_F} \in (0, 1). \tag{1}$$

Definition 4 (Story-retention threshold)**.** *$\hat{p}$ is the threshold belief about story truth above which agents prefer to retain stories.*

An agent who encounters a story from a source having veracity p and gets a private signal with likelihood ratio L_i will have posterior belief above $\hat{p}$ if and only if $L_i > \hat{L}(p)$,

[4] Agents' equilibrium strategies are uniquely determined except at time $\tilde{t}$, when agents are indifferent between paying attention to Pool O or Pool S.

where $\hat{L}(p)$ is implicitly defined by

$$\frac{\hat{p}}{1-\hat{p}} = \frac{p}{1-p} \times \hat{L}(p). \tag{2}$$

(Equation (2) leverages the likelihood-ratio formulation of Bayes' Rule.)

Let $\sigma_\omega(p)$ be the probability that an agent gets a favorable enough signal to optimally retain a story with truth-value $\omega \in \{T, F\}$ from a source having veracity p:

$$\sigma_\omega(p) \equiv 1 - H\left(\hat{L}(p \,\middle|\, \omega\right). \tag{3}$$

For each story pool $j \in \{O, S\}$, let $\sigma_{j\omega}(t) \equiv \sigma_\omega(p_j(t))$ be shorthand for the story-retention likelihood from Pool j at time t. Because Pool O has unchanging veracity, $\sigma_{OT}(t) = \sigma_{OT} \equiv \sigma_T(p_O)$ and $\sigma_{OF}(t) = \sigma_{OF} \equiv \sigma_F(p_O)$ for all t.

B. Monotonicity of the story-retention rate. Given attention $\lambda_O(t) = \lambda_O$ to original sources, each agent retains stories from Pool O at rate

$$\psi_O(\lambda_O) = \lambda_O \left(p_O \sigma_{OT} + (1 - p_O)\sigma_{OF}\right). \tag{4}$$

Given attention $\lambda_S(t) = \lambda_S$ to social sources, knowledge prevalence $K(t) = K$, and social veracity $p_S(t) = p_S$, agents retain stories from Pool S at rate

$$\psi_S(\lambda_S, K, p_S) = \lambda_S K \left(p_S \sigma_T(p_S) + (1 - p_S)\sigma_F(p_S)\right). \tag{5}$$

Overall, the story-retention rate at time t is $\psi(t) \equiv \psi_O(\lambda_O(t)) + \psi_S(\lambda_S(t), K(t), p_S(t))$.

C. Dynamics of individual knowledge. The number of stories known by agent i, denoted $m^i(t)$, evolves according to a Markov process in which: (i) transitions from m to

$m+1$ occur at rate $\psi(t) > 0$ for all $m \geq 0$; and (ii) transitions from m to $m-1$ occur at rate $m\gamma > 0$ for all $m \geq 1$. Recall that $N_m(t)$ is the share of the population that knows m stories at time t. Since the agent population has unit mass, $N_m(t) = \Pr(m^i(t) = m)$.

Lemma 1 provides a closed-form expression for $N_m(t)$ and a condition under which knowledge prevalence $K(t)$ must be strictly increasing.

Lemma 1. *(i)* $N_m(t) = \frac{\mu(t)^m}{m!}\exp(-\mu(t))$ *for all* $m = 0, 1, 2, ...,$ *where*

$$\mu(t) = \int_0^t \psi(s)\exp(-\gamma(t-s))\mathrm{d}s.$$

(ii) Suppose that $\psi(t)$ *is weakly increasing. Then* $K(t) = 1 - N_0(t)$ *is strictly increasing.*

D. Dynamics of the story-pool state. Recall that $M_\omega(t)$ is the mass of stories with truth-value ω in agents' collective knowledge, counting identical copies known by different agents as different stories. $M(t) = (M_T(t), M_F(t))$ evolves according to the following system of differential equations:

$$M_T'(t) = \lambda_O(t)p_O\sigma_{OT} + \lambda_S(t)K(t)p_S(t)\sigma_T(p_S(t)) - \gamma M_T(t) \tag{6}$$

$$M_F'(t) = \lambda_O(t)(1-p_O)\sigma_{OF} + \lambda_S(t)K(t)(1-p_S(t))\sigma_F(p_S(t)) - \gamma M_F(t) \tag{7}$$

where $\lambda_j(t) \equiv E_i[\lambda_j^i(t)]$ is the average attention paid to Pool j across all agents.

In equation (6), the first term is the time-t inflow of true stories retained from original sources. Across the population, there is a flow $\lambda_O(t)$ of exposures to stories from Pool O, fraction p_O of these stories are true, and fraction σ_{OT} of those are retained. Similarly, the second term is the inflow of true stories retained from Pool S. Across the population, there is a flow $\lambda_S(t)$ of social interactions, fraction $K(t)$ of these interactions are with someone who has a story to share (Pool S), fraction $p_S(t)$ of these stories are true, and

fraction $\sigma_T(p_S(t))$ of those are retained. The last term is the outflow of true stories leaving agents' collective knowledge due to forgetting. Equation (7) is similar, but for false stories.

E. Dynamics of social veracity. Social veracity $p_S(t)$ is the average veracity of all stories in agents' collective knowledge at time t. Initially at time 0, agents are exposed to stories from Pool O at rate $\lambda_O(0)$ and to stories from Pool S at rate $K(0)\lambda_S(0)$. Since all agents initially know nothing (by the assumed initial condition), $K(0) = 0$ and all initial exposures are from Pool O. Because agents only retain a story when their posterior belief exceeds the story-retention threshold $\hat{p}$, the inflow of stories that agents retain from Pool O is positively selected, with veracity strictly higher than the original-story veracity p_O.

Definition 5 (Inflow veracity)**.** *Let $p_{in}(p)$ denote the veracity of the stories that agents choose to retain from a pool having veracity p, referred to as "inflow veracity":*

$$p_{in}(p) \equiv \frac{p\sigma_T(p)}{p\sigma_T(p) + (1-p)\sigma_F(p)}. \tag{8}$$

Lemma 2 gathers together some useful facts about the inflow-veracity function $p_{in}(p)$.

Lemma 2. *(i) $p_{in}(p) > \max\{p, \hat{p}\}$ for all $p \in (0,1)$. (ii) $p_{in}(p)$ is strictly increasing and continuous in p over the range $p \in (0,1]$. (iii) $p_{in}(p)$ is strictly increasing and continuous in $\hat{p}$ for all $p \in (0,1)$.*

Returning to the question of how social veracity evolves in equilibrium, using $M_T(t) = p_S(t)\Sigma M(t)$ and $M_F(t) = (1 - p_S(t))\Sigma M(t)$, (6-7) can be re-arranged as

$$\frac{M_T'(t)}{M_T(t)} = \frac{1}{\Sigma M(t)}\left(\frac{\lambda_O(t)p_O\sigma_{OT}}{p_S(t)} + \lambda_S(t)K(t)\sigma_T(p_S(t))\right) - \gamma \tag{9}$$

$$\frac{M_F'(t)}{M_F(t)} = \frac{1}{\Sigma M(t)}\left(\frac{\lambda_O(t)(1-p_O)\sigma_{OF}}{(1-p_S(t))} + \lambda_S(t)K(t)\sigma_F(p_S(t))\right) - \gamma \tag{10}$$

The expression $\frac{M_T'(t)}{M_T(t)} - \frac{M_F'(t)}{M_F(t)}$ plays an important role in the analysis. Whenever this expression is positive, the mass of true stories in Pool S grows at a faster percentage rate than the mass of false stories, causing social veracity $p_S(t)$ to rise.

Subtracting the equations (9-10), we see $\frac{M_T'(t)}{M_T(t)} - \frac{M_F'(t)}{M_F(t)} \gtrless 0$ and hence $p'(t) \gtrless 0$ iff $IN(t) + SF(t) \gtrless 0$, where

$$IN(t) = \underbrace{\lambda_O(t) \left(\sigma_{OT} \frac{p_O}{p_S(t)} - \sigma_{OF} \frac{1 - p_O}{1 - p_S(t)} \right)}_{\text{inflow effect}} \tag{11}$$

$$SF(t) = \underbrace{\lambda_S(t) K(t) \left(\sigma_T(p_S(t)) - \sigma_F(p_S(t)) \right)}_{\text{social-filtering effect}} \tag{12}$$

Lemma 3 provides several useful facts about the inflow and social-filtering effects.

Lemma 3. *(i) The inflow effect is continuous and strictly decreasing in $p_S(t)$ and zero when $p_S(t) = p_{in}(p_O)$. (ii) The social-filtering effect is positive for all $p_S(t) \in (0, 1)$, quasiconcave in $p_S(t)$ with a maximum at $p_S(t) = \hat{p}$, and zero at both $p_S(t) = 0$ and $p_S(t) = 1$.*

An implication of Lemma 3 is that there is a unique social-veracity level, called "the attractor veracity," at which the inflow effect and social-filtering effect sum to zero.

Definition 6 (Attractor veracity)**.** *The "attractor veracity" $p^{**}(\lambda_O, \lambda_S, K)$ is the unique social-veracity level such that $IN(t) + SF(t) = 0$ when we set $\lambda_O(t) = \lambda_O$, $\lambda_S(t) = \lambda_S$, $K(t) = K$, and $p_S(t) = p^{**}(\lambda_O, \lambda_S, K)$.*

Lemma 4. *(i) $p_S'(t) \gtrless 0$ if and only if $p_S(t) \lessgtr p^{**}(t) \equiv p^{**}(\lambda_O(t), \lambda_S(t), K(t))$. (ii) $\lim_{t \to 0} p^{**}(t) = p_{in}(p_O)$.*

Lemma 4(i) says that social veracity always evolves toward the current attractor veracity. See Figure 1, which provides a snapshot view of the inflow-effect function $IN(t)$

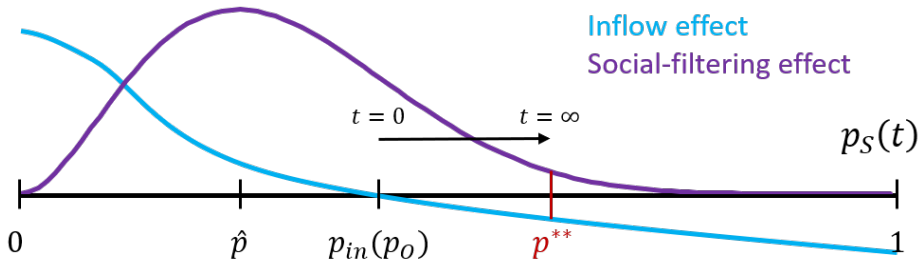

Figure 1: Illustration of the inflow and social-filtering effects (Lemma 3) and the resulting attractor veracity $p^{**}(t)$ at one point in time $t > 0$.

(in light blue), the social-filtering-effect function $SF(t)$ (in purple), and the resulting attractor veracity $p^{**}(t)$ at some point in time $t > 0$. Note that social veracity $p_S(t)$ is increasing over time so long as $p_S(t)$ remains below $p^{**}(t)$.

Lemma 4(ii) says that, at time $t = 0$, the attractor veracity starts out equal $p_{in}(p_O)$, the inflow veracity from Pool O. Indeed, as we will see later in Part G, social veracity also starts out at $p_{in}(p_O)$ and remains equal to $p_{in}(p_O)$ for an extended period of time in the unique equilibrium trajectory. Only when agents begin to pay attention to others does the attractor veracity begin rising, and pulling social veracity up along with it.

F. Individually-optimal attention. Let $E\Pi(p)$ denote the expected payoff an agent exposed to a story from a source having veracity p. Such a story is *true and retained* with probability $p\sigma_T(p)$ and *false and retained* with probability $(1-p)\sigma_F(p)$. So,

$$E\Pi(p) = \pi_T p\sigma_T(p) + \pi_F(1-p)\sigma_F(p) > 0.^5 \tag{13}$$

The marginal value of attention to original sources is $E\Pi(p_O)$. The marginal value of attention to social sources is $K(t)E\Pi(p_S(t))$, where the knowledge prevalence $K(t)$ is the

[5] $E\Pi(p) > 0$ because agents only retain stories whose conditional likelihood of being true exceeds $\hat{p}$.

likelihood that a randomly-selected agent knows at least one story.

Lemma 5 provides conditions under which the dynamics of optimal attention are especially simple.

Lemma 5. *Suppose that $p_S(t)$ is non-decreasing and $K(t)$ is strictly increasing. Then there is a critical time $\tilde{t} \in (0, \infty]$ such that agents focus on Pool O up until time $\tilde{t}$ and thereafter focus on Pool S.*

The intuition for Lemma 5 is very simple. Initially at time $t = 0$, the marginal value of social attention is zero, because other agents don't yet know anything; so, agents find it optimal initially to focus on Pool O. Over time as social veracity increases and know-nothings grow less common, the marginal value of social attention increases and agents may at some point find it optimal to switch to focus their attention socially.[6]

G. Construction of the unique equilibrium trajectory. *At time $t = 0$:* Because agents know nothing, $(N_0^*(0), N_1^*(0), N_2^*(0), ...) = (1, 0, 0, ...)$ and $(M_T^*(0), M_F^*(0)) = (0, 0)$ in any equilibrium trajectory. Moreover, by Part F, agents find it optimal to pay full attention to Pool O, i.e., $\lambda_O^*(0) = \overline{\lambda}$ and $\lambda_S^*(0) = 0$. The initial retention rate is therefore $\psi^*(0) = \psi_O\left(\overline{\lambda}\right)$. This in turn uniquely determines the initial individual-knowledge dynamics $N_m'^*(0)$ for all $m = 0, 1, 2, ...$, according to the system

$$N_0'^*(0) = -\psi^*(0) \quad \text{and} \quad N_1'^*(0) = \psi^*(0) \quad \text{and} \quad N_m'^*(0) = 0 \text{ for all } m \geq 2.$$

Similarly, the initial story-pool dynamics $M_\omega'^*(0)$ for all $\omega \in \{T, F\}$ are determined by

$$M_T'^*(0) = \overline{\lambda} p_O \sigma_{OT} \quad \text{and} \quad M_F'^*(0) = \overline{\lambda}(1 - p_O)\sigma_{OF}$$

[6] If the marginal value of social attention remains below $E\Pi(p_O)$ forever, which happens when the forgetting rate γ is sufficiently high, then $\tilde{t} = \infty$ and agents focus on Pool O forever.

Since these newly-retained stories are coming from Pool O, social veracity immediately after time $t = 0$ equals the inflow veracity from Pool O, i.e., $\lim_{t\to 0} p_S^*(0) = p_{in}(p_O)$.[7]

From time $t = 0$ until agents are indifferent between Pool O and Pool S. Let $\hat{t}$ be the first time at which $E\Pi(p_O) = (1 - N_0(\hat{t}))E\Pi(p_S(\hat{t}))$, or $\hat{t} = \infty$ if agents strictly prefer to focus on Pool O forever. Until time $\hat{t}$, agents find it optimal to focus on Pool O: $\lambda_O^*(t) = \overline{\lambda}$ and $\lambda_S^*(t) = 0$ for all $t < \hat{t}$. The story-retention rate is therefore constant at $\psi_O\left(\overline{\lambda}\right)$. This uniquely determines individual-knowledge dynamics up until time $\hat{t}$, according to the system

$$N_0'^*(t) = -\psi_O\left(\overline{\lambda}\right) N_0^*(t) + \gamma N_1^*(t) \tag{14}$$

$$N_m'^*(t) = \psi_O\left(\overline{\lambda}\right)\left(N_{m-1}^*(t) - N_m^*(t)\right) - (m+1)\gamma N_{m+1}^*(t) \text{ for all } m \geq 1 \tag{15}$$

Similarly, story-pool dynamics are uniquely determined by the system

$$M_T'^*(t) = \overline{\lambda} p_O \sigma_{OT} - \gamma M_T(t) \tag{16}$$

$$M_F'^*(t) = \overline{\lambda}(1 - p_O)\sigma_{OF} - \gamma M_F(t) \tag{17}$$

Because the story-retention rate is constant during this period, Lemma 1 implies that knowledge prevalence $K^*(t)$ is strictly increasing over time. Meanwhile, since agents are only paying attention to Pool O, social veracity remains equal to the inflow veracity from Pool O, namely, $p_S^*(t) = p_{in}(p_O)$ for all $t < \hat{t}$. So, during this period while agents focus exclusively on Pool O, the marginal value of social attention equals $K^*(t)E\Pi(p_{in}(p_O))$ and is strictly increasing over time.

[7] To verify this mathematically, note that

$$\lim_{t\to 0} p^*(t) = \lim_{t\to 0} \frac{M_T^*(t)}{M_T^*(t) + M_F^*(t)} = \lim_{t\to 0} \frac{M_T'^*(t)}{M_T'^*(t) + M_F'^*(t)} = \frac{p_O\sigma_{OT}}{p_O\sigma_{OT} + (1 - p_O)\sigma_{OF}} = p_{in}(p_O).$$

Suppose for a moment that agents were to strictly prefer Pool O forever, so that individual-knowledge dynamics remain governed by (18, 15) forever, and define $K^{*\infty} = \lim_{t\to\infty} K^*(t)$. By Lemma 1, $K^*(t) = 1 - e^{-\mu(t)}$ where $\mu(t) = \overline{\lambda}\int_0^t e^{-\gamma(t-s)}\mathrm{d}s$. Thus,

$$K^{*\infty} = 1 - e^{-\frac{\overline{\lambda}}{\gamma}}$$

If agents' forgetting rate γ is large relative to their attention budget $\overline{\lambda}$, then $K^{*\infty} \approx 0$ and social interaction remains unattractive because most other agents don't have any information to share. More precisely, agents prefer Pool O over Pool S at time t if and only if $K^*(t) > \hat{K} \equiv \frac{E\Pi(p_O)}{E\Pi(p_{in}(p_O))} \in (0,1)$. There are two basic cases

Case #1: Focus on Pool O forever. Suppose that the forgetting rate is sufficiently high and/or the attention budget is sufficiently low that $K^{*\infty} \leq \hat{K}$. In this case, agents strictly prefer to focus on Pool O forever. No one ever pays attention to others and social veracity remains equal to $p_{in}(p_O)$. Story-pool dynamics are uniquely determined by the relatively simple system (16-17).

Case #2: Eventually switch to focus on Pool S. Suppose instead that $K^{*\infty} > \hat{K}$. In this case, agents become indifferent between Pool O and Pool S at some finite time $\hat{t}$ when $K^*(t) = \hat{K}$. At this critical point in time, the inflow effect is zero (due to $p_S^*(t) = p_{in}(p_O)$ for all $t \leq \hat{t}$) and the social-filtering effect is positive. Thus, any shift of attention toward Pool S further increases the sum of these effects, causing social veracity to begin to rise while the mass of know-nothings continues to fall. The marginal value of paying attention to Pool S therefore continues to rise at time $\hat{t}$, leading agents to strictly prefer to switch focus from Pool O to Pool S immediately afterward.

As social veracity begins to rise, a virtuous cycle kicks into effect whereby (a) higher social veracity and agents' attention to this socially-shared content leads agents to retain

more stories, reducing the prevalence of know-nothing agents, which (b) boosts both the marginal value of social attention and the social-filtering effect, which in turn (c) further increases social veracity over time, making social channels even more attractive to agents.

Uniqueness of the equilibrium trajectory follows from the fact that, at each point in time during this process, individual-knowledge dynamics are governed by the system

$$N_0'^*(t) = -\psi^*(t)N_0^*(t) + \gamma N_1^*(t) \tag{18}$$

$$N_m'^*(t) = \psi^*(t)\left(N_{m-1}^*(t) - N_m^*(t)\right) - (m+1)\gamma N_{m+1}^*(t) \text{ for all } m \geq 1 \tag{19}$$

and story-pool dynamics are governed by the system

$$M_T'^*(t) = \underline{\lambda}_O p_O \sigma_{OT} + \left(\overline{\lambda} - \underline{\lambda}_O\right) p_S^*(t)\sigma_T(p_S^*(t)) - \gamma M_T(t) \tag{20}$$

$$M_F'^*(t) = \underline{\lambda}_O (1-p_O)\sigma_{OF} + \left(\overline{\lambda} - \underline{\lambda}_O\right)(1-p_S^*(t))\sigma_F(p_S^*(t)) - \gamma M_F(t) \tag{21}$$

The argument thus far provides a constructive proof that an equilibrium story-pool trajectory exists and is unique, establishing part (i) of the theorem. Moreover, I have established that equilibrium social veracity is strictly increasing over time, completing the proof of part (ii). Convergence of the equilibrium trajectory follows from the fact that the system governing story-pool dynamics is monotone and bounded; so, the equilibrium trajectory cannot cycle or diverge. Intuitively, once social veracity starts rising, the inflow and filtering effects stabilize its growth at a fixed point, ensuring the process settles into an equilibrium steady state.

Discussion: social attention as a supermodular game. To build intuition, it is helpful to consider a much-simpler "long-run attention game" in which agents care only about long-run payoffs, choose at time 0 how to divide their attention between original

and social sources, and must maintain that attention split forever. Let λ_O and λ_S denote the average attention to original and social sources, respectively.

A key feature of this game is that the long-run equilibrium social veracity, denoted $p^\infty(\lambda_O, \lambda_S)$, is strictly decreasing in λ_O (more weight on original sources amplifies the negative inflow effect) and strictly increasing in λ_S (more weight on social sources amplifies the positive social-filtering effect). This complementarity means that the value of paying attention to social sources rises as others pay more attention socially. In other words, the long-run attention game is supermodular.

2.2 Long-run equilibrium outcomes

Section 2.1 provides a constructive characterization of the unique equilibrium trajectory. This section focuses on the long-run limit, with particular attention to:

(i) long-run social veracity $p_S^{*\infty} = \lim_{t\to\infty} p_S^*(t)$;

(ii) long-run attention $(\lambda_O^{*\infty}, \lambda_S^{*\infty}) = \lim_{t\to\infty}(\lambda_O^*(t), \lambda_S^*(t))$; and

(iii) long-run knowledge prevalence $K^{*\infty} = \lim_{t\to\infty} K^*(t)$.

Long-run equilibrium attention: As explained in Part G of the proof of Theorem 1, there are two possible long-run equilibrium attention profiles. In Case #1, the marginal value of social attention never exceeds the marginal value of attention to Pool O, so agents continue to focus on Pool O forever and $(\lambda_O^{*\infty}, \lambda_S^{*\infty}) = (\overline{\lambda}, 0)$. In Case #2, the marginal value of social attention eventually overtakes that of Pool O, prompting a switch at some finite time $\tilde{t}$ to social interaction, in which case $(\lambda_O^{*\infty}, \lambda_S^{*\infty}) = (\underline{\lambda}_O, \overline{\lambda} - \underline{\lambda}_O)$.[8] Which case arises depends on how the marginal value of social attention evolves while agents all focus on Pool O.

[8]Recall that, by assumption, there is a minimal level of attention that agents must pay to original sources, corresponding to a minimal rate at which people unavoidably encounter stories non-socially.

For this purpose, let $p_S^O(t)$ and $K^O(t)$ denote the social veracity and knowledge prevalence at time t that would arise if agents were constrained to focus on Pool O forever. By the monotonicity analysis in the proof of Theorem 1, both $p_S^O(t)$ and $K^O(t)$ are strictly increasing in t. Consequently, the marginal value of social attention , $K^O(t)\, E\Pi(p_S^O(t))$, is strictly increasing in t. If

$$\lim_{t\to\infty} K^O(t)\, E\Pi(p_S^O(t)) \;\leq\; E\Pi(p_O),$$

then agents strictly prefer to remain focused on Pool O forever and $(\lambda_O^{*\infty}, \lambda_S^{*\infty}) = (\overline{\lambda}, 0)$. Otherwise, if

$$\lim_{t\to\infty} K^O(t)\, E\Pi(p_S^O(t)) \;>\; E\Pi(p_O),$$

then agents eventually switch to focus on social interaction and $(\lambda_O^{*\infty}, \lambda_S^{*\infty}) = (\underline{\lambda}_O, \overline{\lambda} - \underline{\lambda}_O)$.

This reasoning uniquely determines the long-run equilibrium attention profile. For the remainder of the analysis. I will denote this profile simply as $(\lambda_O^{*\infty}, \lambda_S^{*\infty})$ and treat it as given.

Long-run knowledge prevalence: Fix any candidate social veracity p_S and let $K^\infty(p_S)$ be the long-run knowledge prevalence that results when agents use the fixed attention profile $(\lambda_O^{*\infty}, \lambda_S^{*\infty})$ and retain stories optimally under the belief that social veracity is p_S^∞.[9] Similarly, let $\psi^\infty(p_S^\infty)$ be the long-run story-retention rate given $(p_S^\infty, \lambda_O^{*\infty}, \lambda_S^{*\infty})$.

[9]To cleanly define the function $K^\infty(p_S)$, I consider (artificial) thought experiments in which agents believe that social veracity is constant, even though it is in fact changing along the relevant dynamic path. Similarly when defining $p_S^\infty(K)$ below, I maintain the artificial assumption that each social encounter results in a story exposure with probability K, even though knowledge prevalence is in fact changing along the relevant dynamic path.

By Lemma 1,

$$K^{\infty}(p_S^{\infty}) = 1 - \exp(-\mu^{\infty}(p_S^{\infty})), \quad \text{where} \tag{22}$$
$$\mu^{\infty}(p_S^{\infty}) \equiv \lim_{t\to\infty} \int_0^t \psi^{\infty}(p_S^{\infty}) \exp\big(-\gamma(t-s)\big)\, \mathrm{d}s = \frac{\psi^{\infty}(p_S^{\infty})}{\gamma}.$$

By equations (4-5), the retention rate $\psi^{\infty}(p_S^{\infty})$ is strictly increasing in p_S^{∞} if $\lambda_S^{*\infty} > 0$ and does not depend on p_S^{∞} if $\lambda_S^{*\infty} = 0$. Thus, $K^{\infty}(p_S^{\infty})$ is non-decreasing in p_S^{∞}.

Long-run social veracity: Fix any knowledge prevalence K and let $p_S^{\infty}(K)$ be the long-run social veracity if agents maintain attention profile $(\lambda_O^{*\infty}, \lambda_S^{*\infty})$ and each social encounter generates a story exposure with probability K. By Lemma 4, the long-run social veracity p_S^{∞} must equal the long-run attractor veracity:

$$p_S^{\infty}(K) = p^{**}(\lambda_O^{*\infty}, \lambda_S^{*\infty}, K). \tag{23}$$

(The attractor veracity was originally defined and discussed in the text surrounding Figure 1.) By equations (11-12), rising knowledge prevalence does not change the inflow effect but does increase the social-filtering effect. Consequently, $p_S^{\infty}(K)$ is strictly increasing in K.

Any equilibrium steady state must correspond to a fixed point of the mapping $(p_S, K) \mapsto (p_S^{\infty}(K), K^{\infty}(p_S)$. Moreover, as established in Proposition 1, the unique equilibrium trajectory—starting from the assumed initial condition in which agents' have zero initial knowledge—converges to the minimal fixed point of this mapping.

Proposition 1. *The mapping* $(p_S, K) \mapsto (p_S^{\infty}(K), K^{\infty}(p_S)$ *has a minimal fixed point, which equals* $(p_S^{*\infty}, K^{*\infty})$.

Proposition 1, along with equations (22,23), simplify the computation of long-run

equilibrium outcomes and allow for a number of long-run equilibrium comparative statics, as discussed in the next section.

Discussion: what if agents never forget? If agents never forget, i.e., if $\gamma = 0$, then knowledge prevalence necessarily converges to 100% and this section's analysis simplifies substantially. First, because social veracity is higher than original-story veracity at each point in time, agents must eventually switch to focus on Pool S. This uniquely pins down the long-run equilibrium attention profile: $(\lambda_O^{*\infty}, \lambda_S^{*\infty}) = (\underline{\lambda}_O, \overline{\lambda} - \underline{\lambda}_O)$. With $K^{*\infty} = 1$, this in turn uniquely pins down the long-run equilibrium social veracity $p_S^{*\infty}$.

3 Extension: Attractive Stories

This section considers an extension in which stories are not only "true" ($\omega = T$) or "false" ($\omega = F$) but also may possess an observable characteristic called "attractiveness" that impacts agents' story-retention decisions. The key finding of this *attractive-story extension* is that all stories in social circulation (even the neutral ones) are less likely to be true once attractive stories are introduced into the mix (Theorem 2). The intuition for this finding is clear. Because attractive stories are always retained, they bypass the filtering process that normally weeds out falsehoods; so, the share of true stories among those that are attractive is lower than in the benchmark. At the same time, the fact that attractive stories spread more widely "crowds out" the neutral stories in social circulation, shortening their transmission chains and weakening the filtering effect that otherwise would have sustained their higher veracity. As a result, neutral stories are in equilibrium less likely to be true than in the benchmark model, and attractive stories have the lowest veracity of all.

Attractive versus neutral stories. Let q_O be the fraction of the original-story pool that is attractive ($\chi = +$) as opposed to neutral ($\chi = 0$). Let $p_{O\chi}$ be the share of original stories with attractiveness $\chi \in \{+, 0\}$ that are true.

Assumption 1. $p_{O+} = p_{O0} = p_O$ *and* $q_O = \frac{1}{2}$.

Assumption 1 says that there are no inherent differences between attractive and neutral stories: each type of story has equal inherent veracity and equal inherent prevalence. This is convenient for the presentation, as it highlights the endogenous impacts of equilibrium social transmission. But of course, there are many natural reasons why stories having some observable transmission-driving characteristic (*attractiveness*) might inherently have higher or lower veracity. For instance, marketing studies confirm that observable stylistic features-—such as emotional intensity and humor—-are strong predictors of virality, often independent of informational quality (Berger and Milkman 2012, Nelson-Field et al. 2013, Segev and Fernandes 2023). Creators can therefore potentially achieve their engagement goals without expending effort to ensure accuracy. *Being emotionally intense or funny* may therefore tend to be negatively correlated with story truth among the stories being created on sites like TikTok.[10] The point made here is that, even when there is no inherent correlation between truth and attractiveness, a negative correlation emerges endogenously due to the logic of equilibrium social transmission.

Agents' incentives to retain neutral stories are exactly as in the benchmark model: retaining a neutral story generates knowing payoff $\pi_T > 0$ if true or $\pi_F < 0$ if false. Retaining neutral stories is therefore optimal whenever a story's likelihood of being true exceeds threshold $\hat{p}$, as before. Attractive stories are different.

Assumption 2. *Agents automatically retain any attractive story that they encounter.*

[10]On the other hand, news stories published by journalists from a reputable outlet will naturally be more likely to be true, due to more intensive editing and fact-checking. *Being published by a reputable outlet* will therefore tend to be positively correlated with story truth.

Discussion: This assumption simplifies the exposition but abstracts from the mechanism underlying why agents retain attractive content. For instance, it could be that agents greatly enjoy knowing attractive content, or that such content triggers retention through psychological mechanisms such as cognitive fluency (Reber et al. 2004), memory salience (Bainbridge 2017), or emotional arousal (Revelle and Loftus 1992). In such cases, retention is not the outcome of a cost–benefit calculation but rather the automatic consequence of how people process and recall striking or emotionally-charged stimuli. This paper's findings about equilibrium social veracity apply no matter what the underlying mechanism of attraction. However, as discussed at the end of this section, welfare implications depend on whether agents' attraction is psychological ("agents are *compelled* to retain attractive stories") or economic ("agents *love* to retain attractive stories") and, in the economic case, how intensively agents benefit from the attractive characteristic.

Definition 7 (Social veracities)**.** *For each $t \geq 0$, let $p^*_{S0}(t)$ and $p^*_{S+}(t)$ denote the equilibrium social veracity at time t of neutral and attractive stories, respectively. Let $p^{*\infty}_{S0}$ and $p^{*\infty}_{S+}$ be their long-run limits.*

Definition 8 (Attractiveness prevalence)**.** *For each $t \geq 0$, let $q^*(t)$ be the equilibrium share of collectively-known stories at time t that are attractive, and let $q^{*\infty} \equiv \lim_{t\to\infty} q^*(t)$.*

Theorem 2. *So long as the forgetting rate γ is sufficiently small, $p^{*\infty}_S > p^{*\infty}_{S0} > p^{*\infty}_{S+} = p_O$ and $q^{*\infty} > \frac{1}{2}$.*

Theorem 2 has two important implications. *First*, introducing "attractive" stories that people automatically retain reduces the veracity of *all* types of stories in social circulation, even the neutral ones. Intuitively, the reason is that widespread sharing of attractive stories creates extra social-transmission traffic that crowds out the neutral

stories that are also in circulation. Neutral stories therefore tend to be less seen, and ultimately shared along shorter transmission chains, which results in less social filtering of the neutral stories that agents encounter socially.

Second, although attractiveness is inherently uncorrelated with story truth, a correlation emerges in equilibrium whereby neutral stories encountered socially are more likely to be true than attractive ones. The reason for this endogenous correlation is simple: neutral stories are more likely to be true because of the social-filtering effect that results from agents sometimes choosing *not* to share after a relatively-bad private signal.

The rest of this section provides the proof of Theorem 2.

Story retention, individual knowledge, and knowledge prevalence. Building on notation from Section 2.1, let $\psi_{j\chi}(t)$ be the rate at which each agent retains a story from Pool $j \in \{O, S\}$ having attractiveness $\chi \in \{+, 0\}$. Because attractive stories are always retained when encountered, we have

$$\psi_{O+}(t) = \frac{\lambda_O(t)}{2} \quad \text{and} \quad \psi_{S+}(t) = \lambda_S(t)K(t)q(t)$$

where $q(t)$ is the share of stories in agents' collective knowledge $\mathcal{K}(t)$ that are attractive. (Recall that, by Assumption 1, half of all stories in Pool O are attractive.) For neutral stories, the retention rates are the same as in the benchmark model, except that only half of Pool O stories and only fraction $1 - q(t)$ of the stories in Pool S are neutral:

$$\begin{aligned}
\psi_{O0}(t) &= \frac{\lambda_O(t)}{2}(p_O\sigma_{OT} + (1 - p_O)\sigma_{OF}) \\
\psi_{S0}(t) &= \lambda_S(t)K(t)q(t)\left(p_{S0}\sigma_T(p_{S0}) + (1 - p_{S0})\sigma_F(p_{S0})\right)
\end{aligned}$$

The overall story-retention rate is $\psi(t) = \psi_{O+}(t) + \psi_{O0}(t) + \psi_{S+}(t) + \psi_{S0}(t)$.

The dynamics of individual knowledge are determined exactly as in the benchmark model, with knowledge prevalence at time t depending only on the story-retention rate process up until that time. Details omitted to save space.

Story-pool dynamics and social veracities. Let $M_{T\chi}(t)$ and $M_{F\chi}(t)$ denote the masses of *true* and *false* stories in Pool S with attractiveness $\chi \in \{+, 0\}$, and define $\Sigma M_\chi(t) \equiv M_{T\chi}(t) + M_{F\chi}(t)$. So, the social veracities of attractive and neutral stories are given by

$$p_{S\chi}(t) \equiv \frac{M_{T\chi}(t)}{\Sigma M_\chi(t)} \quad \text{for} \quad \chi \in \{+, 0\} \tag{24}$$

while attractiveness prevalence is

$$q(t) \equiv \frac{\Sigma M_+(t)}{\Sigma M_+(t) + \Sigma M_0(t)} \tag{25}$$

The story-pool dynamics for attractive stories are governed by the equations:

$$M'_{T+}(t) = \underbrace{\tfrac{1}{2}\,\lambda_O(t)\,p_O}_{\text{original inflow}} + \underbrace{\lambda_S(t)K(t)q(t)\,p_{S+}(t)}_{\text{social inflow}} - \gamma\, M_{T+}(t),$$

$$M'_{F+}(t) = \underbrace{\tfrac{1}{2}\,\lambda_O(t)\,(1-p_O)}_{\text{original inflow}} + \underbrace{\lambda_S(t)K(t)q(t)\,[1-p_{S+}(t)]}_{\text{social inflow}} - \gamma\, M_{F+}(t).$$

Note that

$$\frac{M'_{T+}}{M_{T+}} - \frac{M'_{F+}}{M_{F+}} = \frac{\lambda_O/2}{\Sigma M_+}\left(\frac{p_O}{p_{S+}} - \frac{1-p_O}{1-p_{S+}}\right).$$

So, the social veracity of attractive stories is itself always attracted toward the original-story veracity p_O. Indeed, along the equilibrium trajectory (starting from an initial

condition with zero knowledge), we have

$$p^*_{S+}(t) = p_O \text{ for all } t$$

The story-pool dynamics for neutral stories are governed by the equations:

$$M'_{T0}(t) = \frac{\lambda_O(t)}{2} p_O \sigma_{OT} + \lambda_S(t) K(t)(1 - q(t)) p_{S0}(t) \sigma_T(p_{S0}(t)) - \gamma M_{T0}(t) \tag{26}$$

$$M'_{F0}(t) = \frac{\lambda_O(t)}{2}(1 - p_O)\sigma_{OF} + \lambda_S(t) K(t)(1 - q(t))(1 - p_{S0}(t)) \sigma_F(p_{S0}(t)) - \gamma M_{F0}(t) \tag{27}$$

These differential equations are very similar to (6,7) in the benchmark model, with one key difference: only half of the stories in Pool O and only fraction $1 - q(t)$ of the stories in Pool S are neutral. Consequently, only a fraction of story encounters are with neutral stories. Moreover, because $q^*(t) > \frac{1}{2}$ (shown later), neutral stories are encountered disproportionately less during social interactions than in the original-story pool.

Note that

$$\frac{M'_{T0}(t)}{M_{T0}(t)} = \frac{1}{\Sigma M_0(t)} \left(\frac{\lambda_O(t) p_O \sigma_{OT}}{p_S(t)} + \lambda_S(t) K(t)(1 - q(t)) \sigma_T(p_S(t)) \right) - \gamma \tag{28}$$

$$\frac{M'_{F0}(t)}{M_{F0}(t)} = \frac{1}{\Sigma M_0(t)} \left(\frac{\lambda_O(t)(1 - p_O)\sigma_{OF}}{(1 - p_S(t))} + \lambda_S(t) K(t)(1 - q(t)) \sigma_F(p_S(t)) \right) - \gamma \tag{29}$$

Within the large parentheses, the first expressions in each of these equations differ from those in (9,10) only in that they are multiplied by $\frac{1}{2}$, while the second expressions differ only in that they are multiplied by $(1 - q(t))$.

As in the benchmark model, we can cleanly characterize social-veracity dynamics for neutral stories in terms of an "inflow effect" and "social-filtering effect." In particular,

$p'(t) \gtrless 0$ iff $IN(t) + SF(t) \gtrless 0$, where

$$IN_0(t) = \underbrace{\lambda_O(t)\frac{1}{2}\left(\sigma_{OT}\frac{p_O}{p_{S0}(t)} - \sigma_{OF}\frac{1-p_O}{1-p_{S0}(t)}\right)}_{\text{inflow effect}} \tag{30}$$

$$SF_0(t) = \underbrace{\lambda_S(t)K(t)(1-q(t)\left(\sigma_T(p_S(t)) - \sigma_F(p_S(t))\right)}_{\text{social-filtering effect}} \tag{31}$$

Because these expressions only differ from those in the benchmark model via a multiplicative factor, all findings from Lemmas 3-4 carry over with only slight modifications. In particular, social veracity always evolves toward an "attractor veracity," $p^{**}(\lambda_0, \lambda_S, K, q)$, which now also depends on (and is decreasing in) the attractiveness prevalence q. Moreover, as in the benchmark model, the long-run equilibrium social veracity, $p_{S0}^{*\infty}$, equals the long-run attractor veracity, $p^{**}(\lambda_0^{*\infty}, \lambda_S^{*\infty}, K^{*\infty}, q^{*\infty})$.

Attractiveness prevalence. Even though attractive stories only account for half of all original stories, the inflow of retained stories from Pool O is mostly attractive. In particular, there is a constant inflow $\frac{\lambda_O}{2}$ of attractive stories and a constant inflow $\frac{\lambda_O}{2}\left[p_O\sigma_{OT} + (1-p_O)\sigma_{OF}\right]$ of neutral stories from Pool O. Consequently, equilibrium attractiveness prevalence immediately after time $t = 0$ is equals

$$q^*(0+) = \frac{1}{1 + p_O\sigma_{OT} + (1-p_O)\sigma_{OF}} > \frac{1}{2}$$

Moreover, so long as most stories in Pool S are attractive, the inflow of retained stories from Pool S also must be mostly attractive. We conclude that $q^*(t) > \frac{1}{2}$ at all times t.

The impact of rising attractiveness prevalence. The fact that agents retain attractive stories more frequently than neutral stories impacts long-run equilibrium social veracity in two basic ways. *First*, more story retention leads to higher knowledge preva-

lence. This first effect serves to *increase* long-run equilibrium social veracity, by increasing the $K^{\infty}(p_{S0}^{\infty})$ mapping corresponding to (22). However, when the forgetting rate $\gamma \approx 0$, knowledge prevalence is approximately 100% in the long run, i.e., everyone eventually knows at least one story, and this first effect is not relevant.

Second, increasing the share $q(t)$ of attractive stories has no impact on the inflow effect in equation (30) but decreases the social-filtering effect. This second effect serves to *decrease* long-run equilibrium social veracity, by decreasing the $p_{S0}^{\infty}(K^{\infty})$ mapping corresponding to (22). Moreover, this second effect does not depend on the forgetting rate γ. All together, then, the long-run equilibrium social veracity of neutral stories is lower in this extension than in the benchmark so long as γ is sufficiently small. □

Discussion: equilibrium welfare. The welfare implications of this section's analysis depend on *why* agents retain attractive content—whether retention is a deliberate (economic) choice or an automatic (psychological) reaction. Consider first a psychological model in which agents automatically retain attractive stories, as if compelled, but attractiveness *per se* has no impact on agents' knowing payoffs. In this case, long-run equilibrium welfare is unambiguously lower than in the benchmark because (i) equilibrium veracity is lower for all stories and (ii) agents are compelled to retain some stories that they would otherwise have chosen not to retain.

More interesting is the case in which agents retain false attractive stories because they enjoy knowing them. For concreteness, suppose that attractiveness raises agents' knowing payoffs to $\pi_T + b_T$ for true attractive stories and $\pi_F + b_F$ for false attractive stories, where $b_T \geq 0$ and $b_F \geq -\pi_F$. If b_T and b_F are sufficiently large (e.g., if $b_F + \pi_F > \pi_T$), then introducing attractiveness obviously raises agent welfare. However, when the payoff boost associated with attractiveness is more moderate, long-run equilibrium welfare may be lower than in a world where "attractiveness" does not exist.

Numerical example: Suppose that $\pi_T = 1$, $\pi_F = -1$, $b_T = 0$, $b_F = 1$, $p_O \approx 0$, and $\gamma = 0$. Because agents will always retain any attractive story, $p^*_{S+}(t) = p_O \approx 0$ at all times t and agents get approximately zero expected welfare from the attractive stories that they encounter. On the other hand, because agents only retain neutral stories in the rare situations when private signal is favorable enough that their posterior exceeds the threshold $\hat{p} = \frac{-\pi_F}{\pi_T - \pi_F} = \frac{1}{2}$, the neutral stories that agents know have veracity $p^*_{S0}(t) \geq p_{in}(p_O) > \frac{1}{2}$. When agents encounter a story socially, they therefore hope that it will be a neutral one, since those generated expected payoff $E\Pi(p^*_{S0}(t)) > E\Pi(\frac{1}{2}) > 0$. However, by [[cite result]], the long-run social veracity of neutral stories is strictly lower here than in the benchmark model; so, an agent here who is lucky enough to encounter a neutral story is still worse off in the long run (per story encounter) than agents in the benchmark model. Note to conference organizers: I did not have time prior to submission to program this numerical simulation. I plan to do that before the conference, which will allow me to quantify the magnitude of the lost welfare. But as you should be able to tell, the welfare loss here can be extreme—agents get basically zero welfare because neutral stories are only rarely retained while attractive ones are constantly streaming into agents' minds ... Even if agents eventually switch to focus on social interaction, that won't substantially change the problem since attractive stories will still be retained more frequently (and agents' minds are already dominated by attractive content when the social phase starts).

4 Extension: Partisan Agents

This section extends the model further to allow for multiple types of agents, only some of whom care about story attractiveness. Motivated by the *Stop the Steal* example from the introduction, I will refer to story attractiveness as "partisan alignment" and to the agents who care about that as "partisans." Unlike in Section 3, here I explicitly model

why partisan agents always retain partisan-aligned content—they enjoy knowing such stories even if false.

There are two types of agents:[11]

- "non-partisans" with knowing payoffs $\pi_T > 0$ for any true story and $\pi_F < 0$ for any false story regardless of partisanship, exactly as in the benchmark model;
- "partisans" with the same knowing payoffs as non-partisans for neutral stories but $\pi_{T+} = \pi_T + b$ for true-aligned stories and $\pi_{F+} = \pi_F + b \geq 0$ for false-aligned stories.[12]

Assumption 3. *Agents never forget stories, i.e.,* $\gamma = 0$.

This no-forgetting assumption is made for analytical convenience. As discussed at the end of Section 2.2, this assumption ensures that long-run equilibrium social veracity is uniquely determined in the benchmark model.

Suppose for a moment that agents were only able to listen to others of the same type. Non-partisans and partisans would then automatically live in distinct social-attention bubbles, each of which would evolve independently. In particular: (i) the non-partisan bubble would evolve exactly as in the benchmark model of Section 2, leading to long-run equilibrium social veracity $p_S^{*\infty}$, while (ii) the partisan bubble would evolve exactly as in the attractive-story extension, leading to long-run equilibrium social veracity $p_{S0}^{*\infty} < p_S^{*\infty}$ for neutral stories and $p_{S+}^{*\infty} = p_O < p_{S0}^{*\infty}$ for aligned stories.

Now suppose instead that agents are free to divide their social attention however they like, choosing what type(s) of agents to "follow" at any given point in time. In particular, each agent i who chooses time-t attention profile $\lambda(t) = (\lambda_O(t), \lambda_{NP}(t), \lambda_P(t))$ will encounter an original story at rate $\lambda_O(t)$, a randomly selected story from the knowledge of

[11]The mass of each type of agent does not matter to the analysis and is ignored.

[12]The fact that partisan alignment increases partisan agents' knowing payoff for all aligned stories by the same amount is not essential but simplifies the presentation.

a randomly selected non-partisan agent at rate $\lambda_{NP}(t)$, and similarly a randomly selected story from a partisan agent at rate $\lambda_P(t)$. Let $\lambda^j(t) = \left(\lambda^j_O(t), \lambda^j_j(t), \lambda^j_{j'}(t)\right)$ denote the average attention profile of all type-j agents, where $j' \neq j$ and $j, j' \in \{NP, P\}$. Let $\lambda^{j*\infty}$ be the corresponding long-run equilibrium attention profile.

Will non-partisans ever have an incentive to follow partisans? What about partisans following non-partisans? The rest of this section's analysis explores this and related questions, with Proposition 2 summarizing the main findings.

Proposition 2. *In any equilibrium of the partisan-types extension: (i) Non-partisans pay zero attention to partisans in the long run. (ii) Among non-partisans, neutral and aligned stories both have long-run social veracity $p_S^{*\infty}$, as in the benchmark model. (iii) There exists a threshold $\tilde{b} > 0$ such that partisan agents pay zero attention to non-partisans if and only if $b > \tilde{b}$.*

The rest of this section proves Proposition 2, along with some additional discussion.

Non-partisans might follow partisans during an initial phase. An advantage of paying attention to partisan agents is that they know more stories than non-partisans and hence are more likely to have a story to share. In this way, one group of people loving to know everything possible about some topic—for instance, early ham operators' enthusiastic community embrace of short-wave radio during the 1920s—can trigger others with less intense interest to start paying attention and socially filtering the information in circulation. Such transient effects can be important if they trigger broader waves of social transmission (as well as complementary investments and experimentation to generate information, not modeled in this paper) and a maturation of the social-informational ecosystem. However, because of my simplifying assumption here that there is no forgetting, this reason for following partisan agents eventually disappears.

Non-partisans do not follow partisans in the long run. In the long run, non-partisans unambiguously prefer not to pay any attention to partisans, since the stories in partisan agents' minds are (i) less likely to be true and (ii) especially likely to be false and partisan-aligned due to the endogenous negative correlation that emerges in equilibrium between truth and partisan alignment among the stories that partisan agents know (Theorem 2). Long-run equilibrium outcomes for non-partisan agents therefore arise according to the unique equilibrium trajectory in the benchmark model, derived in Section 2.1. This completes the proof of parts (i,ii).

Threshold condition for partisans not to follow non-partisans. Consider the expected marginal value to a partisan agent of each story encountered socially, depending on whether it comes from a partisan or non-partisan source. Stories from partisan sources are (in the long run) attractive with probability $q^{*\infty}$, true with probability p_O conditional on being attractive, and true with probability $p_{S0}^{*\infty}$ conditional on being neutral. Overall, the marginal value to a partisan agent from encountering a story with this distribution of types is

$$q^{*\infty}\left(b + p_O\pi_T + (1-p_O)\pi_F\right) + (1-q^{*\infty})\,E\Pi(p_{S0}^{*\infty}),$$

where $E\Pi(p)$ is the expected value to a *non-partisan* agent of encountering a story from a source having veracity p. On the other hand, stories from non-partisan sources are attractive with probability $\frac{1}{2}$ and true with probability $p_S^{*\infty}$ regardless of attractiveness. The marginal value to a partisan agent from encountering a story with this distribution of types is

$$\tfrac{1}{2}\left(b + p_S^{*\infty}\pi_T + (1-p_S^{*\infty})\pi_F\right) + \tfrac{1}{2}E\Pi(p_S^{*\infty}).$$

Note that $E\Pi(p)$ is strictly increasing in p (shown in the proof of Theorem 1), that $p_S^{*\infty} > p_{S0}^{*\infty} > p_O$, and that $E\Pi(p_O) > p_O\pi_T + (1-p_O)\pi_F$.

We conclude that there exists a unique threshold $\tilde{b} > 0$ that equates these two marginal values. If $b > \tilde{b}$, then partisan agents strictly prefer not to break out of a partisan-only bubble to follow non-partisans. In that case, the only long-run equilibrium possibility is that partisan agents also pay zero attention to non-partisans.

On the other hand, if $b < \tilde{b}$, then any partisan would strictly prefer to deviate and focus on non-partisans if all other partisans were focusing solely on partisans. So, there cannot be an equilibrium in this case in which partisan agents exist in a social-attention bubble. This completes the proof of part (iii). □

References

Bainbridge, Wilma A., "The Resiliency of Memorability: A Predictor of Memory Separate from Attention and Priming," *Royal Society Open Science*, 2017, *4* (3).

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

Berger, Jonah and Katherine L. Milkman, "What Makes Online Content Viral?," *Journal of Marketing Research*, 2012, *49* (2), 192–205.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch, "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 1992, *100* (5), 992–1026.

Bowen, T Renee, Danil Dmitriev, and Simone Galperti, "Learning from shared news: When abundant information leads to belief polarization," *The Quarterly Journal of Economics*, 2023, *138* (2), 955–1000.

Buechel, Berno, Stefan Klößner, Fanyuan Meng, and Anis Nassar, "Misinformation due to asymmetric information sharing," *Journal of Economic Dynamics and Control*, 2023, *150*, 104641.

Danenberg, Tuval and Drew Fudenberg, "Endogenous Attention and the Spread of False News," *arXiv preprint arXiv:2406.11024*, 2024.

Dasaratha, Krishna and Kevin He, "Learning from Viral Content," 2025. Working paper, August 2025.

Eliaz, Kör and Ran Spiegler, "A Model of Competing Narratives," *Theoretical Economics*, 2020, *15* (1), 1–34.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer, "Fake news on Twitter during the 2016 U.S. presidential election," *Science*, 2019, *363* (6425), 374–378.

Hsu, Chin-Chia, Amir Ajorlou, and Ali Jadbabaie, "Persuasion, News Sharing, and Cascades on Social Networks," in "Proceedings of the 60th IEEE Conference on Decision and Control (CDC)" 2021, pp. 4970–4975.

Merlino, Luca Paolo, Paolo Pin, and Nicole Tabasso, "Debunking Rumors in Networks," *American Economic Journal: Microeconomics*, 2023, *15* (1), 467–496.

Nelson-Field, Karen, Erica Riebe, and Kate Newstead, "The Emotions that Drive Viral Video," *Australasian Marketing Journal (AMJ)*, 2013, *21* (4), 205–211.

Papanastasiou, Yiangos, "Fake News Propagation and Detection: A Sequential Model," *Management Science*, 2020, *66* (5), 1826–1846.

Pengelly, Martin, "More than a third of US adults say Biden's 2020 victory was not legitimate," *The Guardian*, 2024. Accessed: 2025-01-29.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand, "Shifting attention to accuracy can reduce misinformation online," *Nature*, 2021, *592*, 590–595.

Pierri, Francesco, Matthew R. DeVerna, Kai-Cheng Yang, David Axelrod, John Bryden, and Filippo Menczer, "One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study," *Journal of Medical Internet Research*, 2023, *25*, e42227.

Reber, Rolf, Peter Wurtz, and Friedrich G. Zimmermann, "Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience?," *Personality and Social Psychology Review*, 2004, *8* (4), 364–382.

Revelle, William and Debra A. Loftus, "The implications of arousal effects for the study of affect and memory," in "Handbook of Emotion and Memory," Psychology Press, 1992, pp. 37–59.

Segev, Shira and Juliana Fernandes, "The Anatomy of Viral Advertising: A Content Analysis of Viral Advertising from the Elaboration Likelihood Model Perspective," *Journal of Promotion Management*, 2023, *29* (1), 125–154.

Shaked, Moshe and J George Shanthikumar, *Stochastic orders*, Springer, 2007.

Trattner, Christoph, Dominik Moesslang, and David Elsweiler, "On the predictability of the popularity of online recipes," *EPJ Data Science*, 2018, *7* (1), 1–39.

Vosoughi, Soroush, Deb Roy, and Sinan Aral, "The spread of true and false news online," *science*, 2018, *359* (6380), 1146–1151.

A Proofs

Proof of Lemma 1. (i) Fix an agent and time t. Retained stories arrive at each time $s < t$ at rate $\psi(s)$ and are still known at t with probability $e^{-\gamma(t-s)}$. Hence the number of stories known at t is Poisson with mean

$$\mu(t) = \int_0^t \psi(s)\, e^{-\gamma(t-s)}\, ds. \tag{32}$$

In a continuum of identical agents, the cross-sectional distribution is therefore $N_m(t) = \frac{\mu(t)^m}{m!} e^{-\mu(t)}$ for $m = 0, 1, 2, \ldots$.

(ii) $K(t) = 1 - N_0(t) = 1 - e^{-\mu(t)}$; so, it suffices to show that $\mu(t)$ is strictly increasing. Differentiating (32) and leveraging the assumptions that $\psi(s) \leq \psi(t)$ for all $s < t$ and $\psi(t) > 0$, we have

$$\mu'(t) = \psi(t) - \gamma \int_0^t \psi(s) e^{-\gamma(t-s)} ds \geq \psi(t) - \gamma \int_0^t \psi(t)\, e^{-\gamma(t-s)} ds = \psi(t) e^{-\gamma t} > 0.$$

□

Proof of Lemma 2. (i) Let $x_i \equiv \Pr(T \mid s^i, p_O)$ denote agent i's posterior belief about an original story after observing signal s^i, and let $\hat{p} \equiv \frac{\pi_F}{\pi_T + \pi_F}$ be the retention threshold. By definition,

$$p^{in}(p_O) = \mathbb{E}[\, x_i \mid x_i > \hat{p}\,]\,. \tag{33}$$

The conditioning in (33) immediately gives $p^{in}(p_O) > \hat{p}$. Moreover, because retention selects only signals with $x_i > \hat{p}$, the distribution of retained posteriors first–order stochastically dominates the unconditional posterior distribution (whose mean is p_O); so, $p^{in}(p_O) > p_O$ as well.

(ii) By assumption, the likelihood ratio $\Lambda(s) \equiv g(s \mid T)/g(s \mid F)$ is strictly increasing

in s. Let $g_{p_O}(s) = p_O g(s \mid T) + (1 - p_O) g(s \mid F)$ denote the unconditional signal density given prior p_O. The family $\{g_{p_O}(s) : 0 < p_O < 1\}$ satisfies the monotone likelihood ratio property in s as p_O increases, since

$$\frac{g_{p_2}(s)}{g_{p_1}(s)} \;=\; \frac{p_2 \Lambda(s) + (1 - p_2)}{p_1 \Lambda(s) + (1 - p_1)}$$

is strictly increasing in s whenever $p_2 > p_1$. Moreover, the mapping $s \mapsto x_i(p_O) \equiv \Pr(T \mid s, p_O)$ of signals to posterior beliefs is strictly increasing in s; so, the distribution of posteriors $x_i(p_O)$ is increasing in the likelihood–ratio order as p_O rises. It follows that the truncated distribution $x_i(p_O) \mid x_i(p_O) > \hat{p}$ is increasing in the FOSD order as p_O increases (see Shaked and Shanthikumar (2007, Thm. 1.C.5)). Therefore $p^{in}(p_O) = \mathbb{E}[\, x_i(p_O) \mid x_i(p_O) > \hat{p}\,]$ is strictly increasing in p_O. Continuity in p_O follows from continuity of the densities $g(\cdot \mid T), g(\cdot \mid F)$.

(iii) Let $\kappa \equiv \pi_F / \pi_T$, so that $\hat{p} = \kappa/(1+\kappa)$. Since the conditioning threshold $\frac{\kappa}{1+\kappa} \in (0,1)$ is strictly increasing in κ, and since $E[x_i | x_i > a]$ is strictly increasing in $a \in (0,1)$, it follows that $p_{in}(p_O)$ is strictly increasing in κ for any p_O, as desired. □

Proof of Lemma 3. (i) Fix $p_S \in (0,1)$ and treat λ_O, σ_{OT}, σ_{OF}, and p_O as parameters. The inflow effect is then

$$IN(p_S) \;=\; \lambda_O \left(\sigma_{OT} \frac{p_O}{p_S} \;-\; \sigma_{OF} \frac{1 - p_O}{1 - p_S} \right), \tag{34}$$

which is continuous on $(0,1)$. Differentiating with respect to p_S,

$$\frac{\partial}{\partial p_S} IN(p_S) \;=\; \lambda_O \left(-\sigma_{OT} \frac{p_O}{p_S^2} \;-\; \sigma_{OF} \frac{1 - p_O}{(1 - p_S)^2} \right) \;<\; 0,$$

(Strict inequality follows from $\lambda_O \geq \underline{\lambda}_O > 0$ and $\sigma_{OT}, \sigma_{OF} > 0$.) Thus, $IN(p_S)$ is strictly decreasing in p_S. Finally, by equation (8), $p^{in}(p_O) = \frac{\sigma_{OT}\, p_O}{\sigma_{OT}\, p_O + \sigma_{OF}\,(1-p_O)}$. Substituting $p_S = p^{in}(p_O)$ in (34) yields $IN(p^{in}(p_O)) = 0$, as desired. □

(ii) Write the social–filtering effect (cf. (12)) as

$$SF(p_S) = \lambda_S\, K\, \big[\sigma_T(p_S) - \sigma_F(p_S)\big],$$

Continuity of $SF(p_S)$ is immediate, and the fact that $SF(0) = SF(1)$ follows from the fact that $\sigma_T(0) = \sigma_F(0) = 0$ and $\sigma_T(1) = \sigma_F(1) = 1$.

An agent retains a story iff the posterior exceeds the threshold $\hat{p}$, which is equivalent to the likelihood–ratio condition

$$\Lambda(s) \equiv \frac{g(s \mid T)}{g(s \mid F)} \geq \tau(p_S) \quad \text{where} \quad \tau(p_S) = \frac{\hat{p}(1-p_S)}{(1-\hat{p})\, p_S}.$$

Let $\hat{s}(p_S)$ be the unique cutoff such that $\Lambda\big(\hat{s}(p_S)\big) = \tau(p_S)$. By MLRP, $\Lambda(\cdot)$ is strictly increasing, hence $\hat{s}(p_S)$ is well defined and strictly decreasing in p_S. Since $\sigma_T(p_S) = 1 - G_T\big(\hat{s}(p_S)\big)$ and $\sigma_F(p_S) = 1 - G_F\big(\hat{s}(p_S)\big)$, we have

$$\sigma_T(p_S) - \sigma_F(p_S) = G_F\big(\hat{s}(p_S)\big) - G_T\big(\hat{s}(p_S)\big).$$

Continuity of $\sigma_T(p_S) - \sigma_F(p_S)$ is immediate, and the fact that

Positivity. For each $p_S \in (0,1)$ the set $\{s \geq \hat{s}(p_S)\}$ is an increasing set, and under MLRP we have $\Pr_T\{s \geq \hat{s}\} > \Pr_F\{s \geq \hat{s}\}$, hence $\sigma_T(p_S) - \sigma_F(p_S) > 0$.

Quasiconcavity and maximizer. Let $\Delta\sigma(p_S) \equiv \sigma_T(p_S) - \sigma_F(p_S)$. By the chain rule,

$$\Delta\sigma'(p_S) = -\big[g_T\big(\hat{s}(p_S)\big) - g_F\big(\hat{s}(p_S)\big)\big]\ \hat{s}'(p_S).$$

Since $\hat{s}'(p_S) < 0$, the sign of $\Delta\sigma'(p_S)$ equals the sign of $g_T\big(\hat{s}(p_S)\big) - g_F\big(\hat{s}(p_S)\big)$. Under MLRP the densities g_T and g_F cross exactly once at the unique $s^\star$ with $\Lambda(s^\star) = 1$ (so $g_T(s^\star) = g_F(s^\star)$). Hence $\Delta\sigma'(p_S) > 0$ when $\hat{s}(p_S) > s^\star$ and $\Delta\sigma'(p_S) < 0$ when $\hat{s}(p_S) < s^\star$, implying that $\Delta\sigma$ is quasiconcave with a unique maximum attained at the p_S such that $\hat{s}(p_S) = s^\star$. Using $\Lambda\big(\hat{s}(p_S)\big) = \tau(p_S)$, the maximizer solves $\tau(p_S) = 1$, i.e.

$$\frac{\hat{p}(1-p_S)}{(1-\hat{p})\,p_S} = 1 \quad \Longleftrightarrow \quad p_S = \hat{p}.$$

Since $SF(p_S) = \lambda_S(1-N_0)\Delta\sigma(p_S)$, $SF(p_S)$ inherits the same properties: it is positive on $(0,1)$, quasiconcave, and uniquely maximized at $p_S = \hat{p}$. □

Proof of Lemma 4. (i) By construction, $p^{**}(t)$ is the unique veracity level such that $p'_S(t) = 0$. By Lemma 3, $IN(p_S) \geq 0$ for all $p_S \leq \hat{p}$ and $SF(p_S) > 0$ for all p_S. Thus, $p^{**}(t) > \hat{p}$. Moreover, $IN(p_S)$ is strictly decreasing for all p_S, while $SF(p_S)$ is strictly decreasing once $p_S > \hat{p}$ (since it is quasiconcave with a unique maximum at $\hat{p}$). Thus, $IN(p_S) + SF(p_S)$ is positive for all $p_S < p^{**}$ and negative for all $p_S > p^{**}$. We conclude as desired that $p'_S(t) > 0$ whenever $p_S(t) < p^{**}(t)$ and $p'_S(t) < 0$ whenever $p_S(t) > p^{**}(t)$.

(ii) At $t = 0$, $K(0) = 0$ since agents know nothing and hence the social-filtering effect is zero, i.e., $SF(p_S; 0) = 0$ for all p_S. Moreover, at all times $t \approx 0$, $K(t) \approx 0$ and agents prefer to focus their attention entirely on Pool O. Thus, we have $\lambda_S(t) = 0$ for all $t \approx 0$, implying that $SF(p_S; t) = 0$ for all $t \approx 0$. The attractor veracity $p^{**}(t)$ therefore must solve $IN(p_S; t) = 0$ for all $t \approx 0$. But recall that $p_S = p^{in}(p_O)$ is the unique solution to this equation for all t. Thus, $p^{**}(t) = p^{in}(p_O)$ for all $t \approx 0$, which implies the desired limiting result. □

Proof of Lemma 5. Let $V_O \equiv E\Pi(p_O)$ and $V_S(t) \equiv K(t)\,E\Pi\big(p_S(t)\big)$ denote the marginal value of attention to original and social sources, respectively. This is a standard

Bayesian cutoff problem: agents retain only when the posterior on the good state exceeds a threshold. In such problems, expected payoffs are strictly increasing in the likelihood of the good state (see, e.g., Shaked and Shanthikumar, 2007); hence $E\Pi(p)$ is strictly increasing in p. Since $p_S(t)$ is non-decreasing and $K(t)$ is strictly increasing, it follows that $V_S(t)$ is strictly increasing. Note that $V_S(0) = 0 < V_O$ because $K(0) = 0$; so, agents initially prefer to focus on original sources. As $V_S(t)$ is strictly increasing and V_O is a constant, we conclude that *either* $V_S(t) < V_O$ forever (so $\tilde{t} = \infty$) *or* there is a unique $\tilde{t} < \infty$ with $V_S(\tilde{t}) = V_O$, and $V_S(t) \gtrless V_O$ iff $t \gtrless \tilde{t}$, as desired. □

Proof of Proposition 1 Existence of a minimal fixed point, denoted $(p_S^{\min}, K^{\min})$, follows immediately from Tarski's Theorem. To prove that $(p_S^*(t), K^*(t))$ converges to $(p_S^{\min}, K^{\min})$, it suffices to show that $p_S^*(t) \leq p_S^{\min}$ and $K^*(t) \leq K^{\min}$ for all t, i.e., there is no "overshooting" on either the p_S- or K-dimension.

Part One: No overshooting of social veracity. As shown in the proof of Theorem 1, social veracity $p_S^*(t)$ is increasing over time and hence is always strictly below the attractor veracity $p^{**}\big(\lambda_O^{*\infty}, \lambda_S^{*\infty}, K^*(t)\big)$, i.e., $p_S^*(t) \leq p^{**}$ for all t. Because p^{**} is increasing in K and $p_S^\infty(K) = p^{**}(\lambda_O^{*\infty}, \lambda_S^{*\infty}, K)$, it follows that $p^{**} \leq p_S^{\min}$ so long as $K^*(t) \leq K^{\min}$. Hence $p_S^*(t) < p_S^{\min}$ whenever $K^*(t) \leq K^{\min}$.

Part Two: No overshooting of knowledge prevalence. Knowledge prevalence at time t depends only on past retention rates. By Lemma 1,

$$K^*(t) = 1 - \exp\big(-\mu^*(t)\big), \quad \text{where} \;\; \mu^*(t) = \int_0^t \psi^*(s)\, e^{-\gamma(t-s)}\, \mathrm{d}s.$$

Suppose, for contradiction, that $K^*(\tau) = K^{\min}$ for some $\tau < \infty$. Then $K^*(t) \leq K^{\min}$ for all $t \leq \tau$, implying by Part One that $p_S^*(t) \leq p_S^{\min}$ for all $t \leq \tau$. But then, since the retention rate is non-decreasing in social veracity, we have $\psi^*(t) \leq \psi^\infty(p_S^{\min})$ for all $t \leq \tau$.

This in turn implies

$$\mu^*(\tau) \;=\; \int_0^\tau \psi^*(s)\, e^{-\gamma(\tau-s)}\,\mathrm{d}s \;\le\; \int_0^\tau \psi^\infty(p_S^{\min})\, e^{-\gamma(\tau-s)}\,\mathrm{d}s \;=\; \frac{\psi^\infty(p_S^{\min})}{\gamma},$$

hence

$$K^*(\tau) \;=\; 1-\exp\big(-\mu^*(\tau)\big) \;\le\; 1-\exp\Big(-\frac{\psi^\infty(p_S^{\min})}{\gamma}\Big) \;=\; K^{\min},$$

a contradiction. Thus $K^*(t) \le K^{\min}$ for all t and, by Part One, $p_S^*(t) \le p_S^{\min}$ for all t.

□