Misinformation as Lemons: How Credibility Unravels on Social Media

Yizhou Jin
University of Toronto

November 3, 2026

Preliminary draft. Please do not cite or circulate.

yizhoujin.com

Abstract

Where in a social network does the credibility of posts hold? We model posting about a news event as a market for lemons. Producers who check facts pay more to post than producers who do not, and readers, who cannot tell whether a post is accurate even after reading it, reward each post for the average credibility of the producers they cannot tell apart. Because readers also amplify, selection works on who posts and on who amplifies. In the model, pooling can price out every producer who checks facts, and the network sets both the pool of producers readers draw on, whose composition sets the sign of selection, and the mix of readers who reward credibility. We test the model's predictions on viral posts about the 2022 mpox outbreak on Twitter and other posts by the same producers, labeled as of the day each was posted, with producers placed in network cells by whether they follow their followers back and whether their followers lead to distinct accounts. A producer's cell predicts what its posts are amplified for: among producers who post in the outbreak's first two weeks, 94% of amplification goes to accurate posts in one-way diverse cells, against 32% in mutual overlapping cells. In the one-way diverse cell, the pool of producers that posts while checking facts is costly is more accurate than the producers who post only later, and in the mutual overlapping cell it is less accurate. Mutual overlapping cells reward visibly misleading posts. Engagement is therefore not a measure of credibility, and systems that learn from engagement would inherit the equilibrium of the part of the network it comes from.

1 Introduction

On social media, the credibility of a post is often hard for readers to discern, while the producer often has superior knowledge of whether the information it contains has been properly verified. This paper studies how this information asymmetry problem affects the production and amplification of credible posts on social media, and how network structures affect equilibrium credibility.

We aim to contribute to three gaps in the literature. The first gap is in theory. Although the logic of adverse selection is well understood and often hypothesized (Akerlof 1970; Allcott and Gentzkow 2017; Kranton and McAdams 2024),[1] we lack an empirical framework that endogenizes both the supply of misinformation and how it is differentially amplified in the social network.[2] An important reason is that the social context makes homophily important to model, where echo chambers and bubbles "insulate agents from corrective signals" (Moldoveanu and Baum 2026), such as being called out (Acemoglu et al. 2024). Neither approach shows how bundling social and news preferences shapes the reward to credibility and selects who posts and who amplifies.

The second gap is in the data. Existing studies rely on proxies of whether a social media post contains misinformation. Vosoughi et al. (2018) use whether a tweet has a reply that links to a URL from one of six fact-checking organizations, but this only focuses on a small portion of "contested news" (Altay et al. 2023) as opposed to fake news.[3] More importantly, they miss misleading news that is not factually inaccurate, which can have an outsized effect on readers: unflagged vaccine-skeptical content had an estimated effect on vaccination intentions 46-fold that of flagged misinformation (Allen et al. 2024). A second approach is to directly survey readers of platforms, including in experiments (Allcott and Gentzkow 2017; Pennycook and Rand 2019; Allen et al. 2024), which captures how their beliefs are moved but does not speak to the veracity of any individual social media post. We conduct human fact-checking on a sample of posts at the time of posting, and develop a large-language-model protocol that scales fact-checking with high alignment with human judgment.

The third gap is in the empirics. The industrial organization of social media has largely focused on the production, consumption, and advertising incentives without taking a stand on accuracy (Filippas et al. 2025; Beknazar-Yuzbashev et al. 2026). On the other hand, research on social-media misinformation has largely focused only on the demand side, frequently featuring interventions that encourage readers to verify the accuracy before amplifying (Aridor et al. 2024; Pennycook et al. 2021; Guriev et al. 2026). If the logic of information asymmetry applies, there is an opportunity to bring in the selection-market toolkit (Einav et al. 2010; Einav and Finkelstein 2011) to understand the supply side and how selection and unraveling can occur in equilibrium.

We start by laying out our model in three steps. First, we introduce a simple baseline that features adverse selection: producers are heterogeneous in accuracy and incur a verification cost to generate accurate content. They are rewarded based on amplification, which allows out-of-network exposure and follower growth. Amplification is based on the consumption value of a post as well as a social component, whereas the consumption portion may or may not reward more credible posts. Under information asymmetry, amplifiers can judge the credibility of a post only by the average credibility of the market as opposed to that of each producer (Akerlof 1970). This generates rich predictions of potential unraveling and tipping of the market between credible and uncredible equilibria.

[1] Allcott and Gentzkow (2017): "Fake news arises in equilibrium because it is cheaper to provide than precise signals, because consumers cannot costlessly infer accuracy, and because consumers may enjoy partisan news."

[2] Kranton and McAdams (2024) endogenize both suppliers' story quality and consumers' sharing, as a fixed number of suppliers paid by visibility each decide whether to incur a cost to produce a high-quality story and consumers receive a private signal of unobserved quality before deciding whether to share the story with their followers. In their model, however, consumers care solely about quality without social preferences, the number of suppliers is fixed without entry, and misinformation originates from exogenous third-party agents. Acemoglu et al. (2024) model users who decide whether to share, ignore, or dislike an article and receive social feedback through reshares and call-outs that penalize sharing misinformation, but their setting has no producer and draws the veracity of each article randomly given its reliability.

[3] Researchers "often focus on URL content (i.e., 'news') that has been flagged as false or misleading by professional fact-checkers" (Allen et al. 2024), and on Facebook such flagged misinformation "received 8.7 million views, accounting for only 0.3% of the 2.7 billion vaccine-related URL views during this time period" (Allen et al. 2024).

Second, we introduce heterogeneity on the demand side across readers that amplify posts. Some are certifiers, amplifying a post only if its perceived credibility is above a threshold. Others are social readers, who give credibility little weight and amplify mainly for social value and other post attributes, such as belief alignment and humor. The presence of social readers generates a double selection margin that guarantees the presence of a low-credibility *Belief Equilibrium*, in which only low-accuracy producers post and only social readers amplify. A high-credibility *News Equilibrium*, on the other hand, may or may not exist depending on the prevalence and preference of certifiers.

Third, we propose a simple network sharing game to link realistic network features we observe in the data to these selection primitives. On the demand side, amplifier utility follows the linear-quadratic form of Ballester et al. (2006), featuring an inherent preference for credibility, a peer effect, a cost of sharing each post, and an attention constraint in the form of a quadratic cost of sharing. In this game, overlap is how far a producer's followers are followed by the same accounts. Here, equilibrium sharing preferences—both for social value and for credibility—depend on network reciprocity and overlap measures. Higher reciprocity indicates higher weight for social value, while less overlap, as in a star, increases the reach of amplification and hence new-follower growth.

We label every viral post in our corpus about the 2022 mpox outbreak on Twitter, and posts by the same producers below the viral cut, on four blocks, each judged as of the day the post was posted against a dated record of what was known and what was still open (Section 4). Truth and Appeal are vertical attributes, the quality and form of a post, and Belief and Position are horizontal attributes, the belief a post leads to and the position it takes. The labels keep three kinds of misinformation apart. On Truth, a false post states a claim that the record contradicts or that has nothing behind it, and a premature post presents an open claim as settled; on Belief, a misleading post is accurate in its words, or states no checkable fact, but leads a general reader to an untrue claim. Appeal and Position are coded apart from truth. We wrote the labeling rules human first, tuned them against a human rater's answer key, and applied them with three language models from different families, each told to use only the post and the dated record. The human rater labels a random holdout of 50 viral posts blind under the final rules, and the models' majority matches the rater on 84% of answers.

Our focus is networks of influencer-producers who posted viral tweets during the news cycle, whose mutual amplification was critical to virality. The amplification records cover every account that amplified a labeled post. We test the model's predictions on these networks (Sections 3 and 4). Two further records carry the network. The producers' follow graph places each producer whose followers include at least three accounts with collected following lists in one of four cells by the reciprocity and the overlap of her followers: one-way and diverse (G_1), one-way and overlapping (G_2), mutual and diverse (G_3), and mutual and overlapping (G_4). The retweets, quote tweets and replies on the labeled posts show who amplifies each post, and the same three models label whether a quote tweet or reply endorses the post or calls it out. We read the Breaking Window, 7 to 20 May 2022, as the regime in which checking facts is costly, a maintained assumption, and compare the producers and amplifiers active in it with those active only after it. The tests use 1,995 labeled posts by 402 producers, 240 of whom fall in a cell.

A producer's cell predicts what its posts are amplified for (Section 5.1). Among the producers who post in the Breaking Window, and over all their posts, 94% of amplification in G_1 goes to accurate posts, against 32% in G_4. The rest of G_4's amplification goes mostly to accurate but misleading posts (29%) and to false or premature posts (32%). Holding the producer's follower count fixed, the accurate share of amplification in G_4 differs from that in G_1 by -47.8 percentage points (SE 10.4). The cells are not assigned, so the difference describes how the cells' equilibria differ, not what would happen to a producer moved from one cell to another.

Mutual overlapping cells reward visibly misleading posts (Section 5.2). In G_4, the reach of a producer's accurate but misleading posts, the log of one plus their retweets and quote tweets, differs from that of her accurate posts by $+1.30$ (SE 0.63, $p = 0.044$), net of the date of posting. In the model, readers respond to what they can see of a post and not to accuracy they cannot see. A misleading post differs from an accurate one in its wording, so its reward can differ in either direction, while a false or premature post with the same public record as an accurate one earns the same expected amplification (Proposition B.7). False or premature posts in G_4 also reach more than the same producer's accurate posts, by $+0.85$ (SE 0.50, $p = 0.097$), significant at the 10 percent level. Neither estimate is distinguishable from zero once the posts' sensational and slanted

labels are held fixed. In G_4 the reward to these posts therefore cannot be separated from their sensational and slanted packaging, which readers also see.

In the one-way diverse cell, the pool of producers that posts while checking is costly is more accurate than the producers who post only after it, and in the mutual overlapping cell it is less accurate (Section 5.3). Holding the producer's follower count and profile fixed, the share of accurate posts of the producers who post in the Breaking Window differs from that of the later producers of the same cell by $+0.14$ in G_1 (SE 0.06) and by -0.25 in G_4 (SE 0.08). The size of the contrast depends on how entry is dated: the difference between the two cells is -0.40 (SE 0.10) in the baseline and -0.15 (SE 0.12), not distinguishable from zero, when entry is dated by a second collection of the producers' posts. Read through the model, with the Breaking Window as the regime in which checking is costly, the contrast is consistent with advantageous selection in the producer pool of G_1 and adverse selection in that of G_4.

Amplifiers are least discerning in G_4 (Section 5.4). Misleading posts in G_4 draw amplifiers whose amplification elsewhere is less accurate: the accurate share of what they amplify from other producers differs by -0.31 (SE 0.05) from a mean of 0.72 among the amplifiers of accurate posts. The accurate share of what amplifiers active in the Breaking Window amplify afterward differs from that of amplifiers active only after it by -0.28 in G_4 (SE 0.02) and by -0.03 in G_1 (SE 0.01). These patterns are consistent with the model's reading of the network, in which the reader mix leans most toward social readers in mutual overlapping cells. The supply and amplifier contrasts rest on the producers on whom the cells were first measured: adding the producers of the posts found in 2026 leaves the composition gap in place, but the supply contrast is then no longer distinguishable from zero, and recomputing every cell on a follow graph that adds their follows also removes the amplifier contrast (Appendix C.6).

Within one platform and one news cycle, the evidence is consistent with credibility holding in one-way diverse cells and not in mutual overlapping cells. In one-way diverse cells, amplification goes to accurate posts and the pool that posts while checking is costly is the more accurate one; in mutual overlapping cells, amplification goes mostly to misinformation and that pool is the less accurate one. The same amplification thus signals accuracy in one part of the network and misinformation in another, so engagement is not a measure of credibility. A feed ranked by engagement, or an AI system trained on what social media amplifies, would inherit the equilibrium of the cells its posts come from.

2 Model

We model posting about one news event as a market for posts with three features. Producers choose whether to post, and readers cannot recognize producer type, so every post in a cell earns a reward set by the cell's average credibility. Readers are also amplifiers: a post's reward is the amplification it receives, so selection can work on who amplifies as well as on who posts. Posts compete for one feed, so each additional post makes every post less visible.

We draw the market as Einav and Finkelstein (2011) draw insurance markets. In Figures 1 to 4, two panels share a horizontal axis: the share Q of potential producers who post, ranked by their cost of posting. The bottom panel is in units of reward, the amplification a post earns and what posting costs. The top panel is in units of credibility, the accuracy of the producer at rank Q and of everyone who posts up to Q.

2.1 Producers and the marginal cost of posting

Figure 1 lines up the potential producers of one cell by their cost of posting. The bottom panel is the marginal cost curve $\mathrm{MC}(Q)$, the cost of the Q-th producer in line. Slop producers stand at the front because they skip the verification cost, and the curve jumps by that cost at q_0, where investigative producers begin. The top panel shows what each producer adds to the pool of producers who post. The marginal accuracy $\mathrm{MA}(Q)$ is the accuracy of the Q-th producer: zero for slop and one for an investigative producer. The average accuracy $\mathrm{AA}(Q)$ is the credibility of everyone who posts up to Q, and it is what readers reward when they cannot tell producers apart (Section 2.2). Above q_0, the producer who would post next is more accurate than the pool. Selection is adverse.

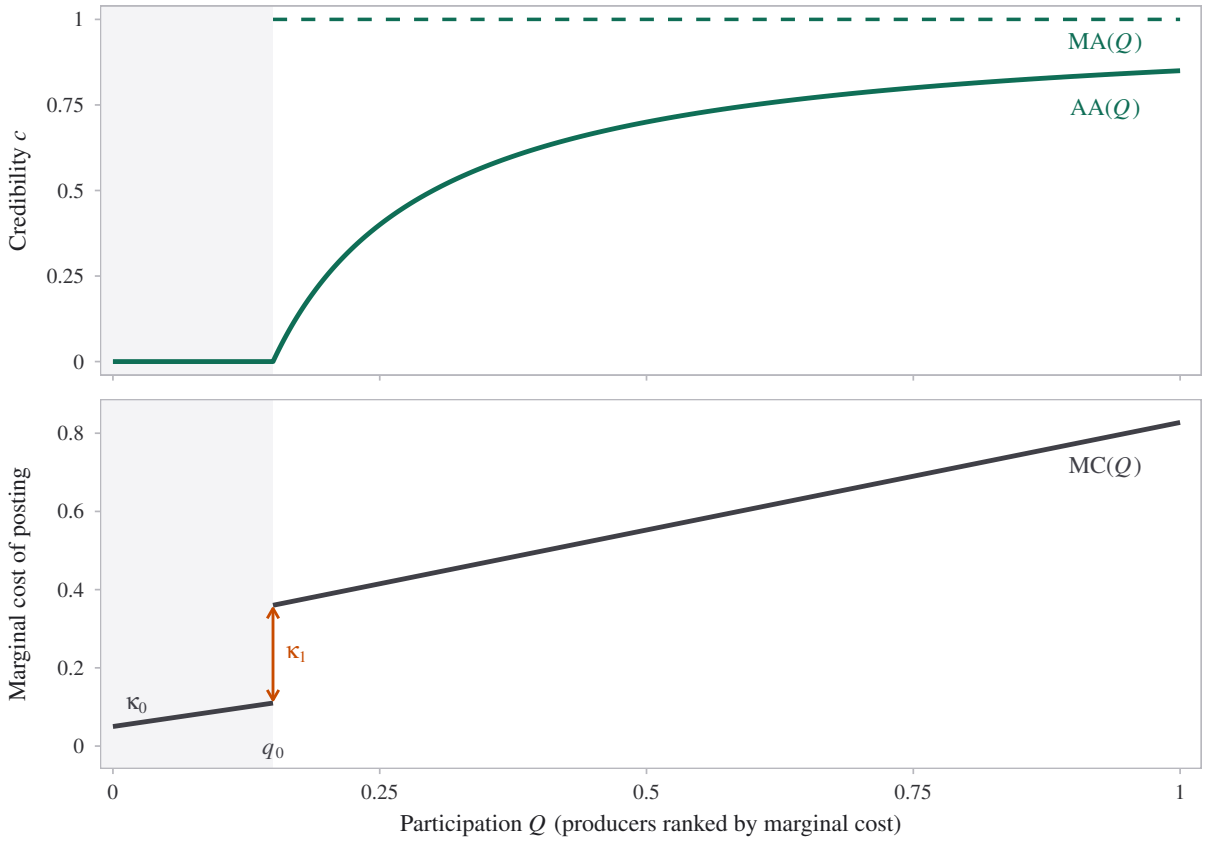


Figure 1: Producers lined up by their cost of posting. Bottom panel: the marginal cost $\mathrm{MC}(Q)$ of the producer at rank Q, which jumps by the verification cost at q_0, where investigative producers begin. Top panel: the marginal accuracy $\mathrm{MA}(Q)$ of the producer at rank Q and the average accuracy $\mathrm{AA}(Q)$ of the producers who post up to Q. The grey band marks the slop producers, $Q \leq q_0$. The curves are drawn at illustrative parameter values (Appendix A.3).

A unit mass of potential producers can post about the event, and each producer j is one of two types. An investigative producer ($\alpha_j = 1$) checks primary facts before posting, and her posts are accurate. A slop producer ($\alpha_j = 0$) posts uninformative content that is accurate only by chance, half the time. A share q_0 of potential producers are slop. A post by producer j is accurate with probability $(1+\alpha_j)/2$, so α_j indexes accuracy from uninformative to always accurate. The type is a fixed property of the producer, not an effort chosen post by post.

Posting costs producer j

$$\mathrm{mc}_j = \kappa_0 + \kappa_1 \alpha_j - \varepsilon_j, \qquad \kappa_1 > 0. \tag{1}$$

Every producer pays the base cost κ_0, investigative producers also pay the verification cost κ_1, and the cost noise ε_j is spread uniformly within each type. The cost is net of any reward the producer receives outside the amplification of her post, so a costly campaign that pays well elsewhere has a low net cost.

Assumption 1 (Separated costs) *Every investigative producer's cost exceeds every slop producer's cost. For example, with one noise range $[\underline{\varepsilon}, \bar{\varepsilon}]$ for both types, this is $\kappa_1 > \bar{\varepsilon} - \underline{\varepsilon}$.*

Lining up producers from the lowest to the highest cost along $Q \in [0,1]$ gives the marginal cost curve,

the cost of the Q-th producer in line:

$$\mathrm{MC}(Q) = \kappa_0 + \sigma Q + \kappa_1 \mathbf{1}\{Q > q_0\}. \tag{2}$$

Slop producers occupy $[0, q_0]$ and investigative producers $(q_0, 1]$. The slope σ tracks the spread of the cost noise within a type, and the curve jumps by the verification cost at q_0.[4] In this form, separated costs are the jump $\kappa_1 > 0$. The index Q ranks producers by cost; it is not time. The results below use only the following properties of MC, which equation (2) satisfies.

Assumption 2 (Marginal cost) MC *is continuous, strictly increasing and convex within each type and jumps up at* q_0.

Assumption 2 is maintained throughout this section.

Two types with separated costs are the benchmark of this section. The curves below extend to any joint law of accuracy and cost with atomless costs (Assumption 9 and Proposition B.2). Selection is adverse at every rank in an interval $(0, \bar{Q}]$ when producers' mean accuracy rises with their rank in the cost order throughout that interval, and advantageous when it falls throughout (Proposition B.3). At a single rank, selection is adverse exactly where the average accuracy of the producers up to that rank rises (Proposition B.2(iii)). Credibility is exactly zero with positive participation only if a positive mass of producers is uninformative (Proposition B.4). Without such a mass, every equilibrium has positive credibility or is silence. Without the cost jump, pooling can still price out every accurate producer when the marginal cost curve is continuous (Proposition B.5), and three equilibria, two of them stable, can coexist when accuracy is smooth and the marginal cost is linear (Proposition B.6).

2.2 Readers, the pooling cell and credibility

Readers condition on what is public about a post, not on its producer's type. They see the producer's position G in the network of who follows whom and her public signal s, such as her follower count, bio or badge. A cell (G, s) collects the posts that readers cannot tell apart on this public information.

Assumption 3 (Information asymmetry) *Within a cell, readers cannot recognize producer type: no signal of a producer's accuracy is available beyond the cell.*

Each cell is a separate market with its own producers, readers and parameters, so we fix one cell and suppress (G, s).

The order in which producers line up by cost sets the credibility of the posts in a cell. If the producers with the Q lowest costs post, the marginal accuracy $\mathrm{MA}(Q)$ is the accuracy of the Q-th producer, and the average accuracy $\mathrm{AA}(Q) = Q^{-1} \int_0^Q \mathrm{MA}(u)\, du$ is the pool's average credibility. In the baseline,

$$\mathrm{MA}(Q) = \mathbf{1}\{Q > q_0\}, \qquad \mathrm{AA}(Q) = \begin{cases} 0, & 0 < Q \le q_0, \\ 1 - q_0/Q, & q_0 < Q \le 1. \end{cases} \tag{3}$$

A post drawn from a pool with credibility c is accurate with probability $(1 + c)/2$, so $c = 0$ means no better than a coin flip. Slop dilutes the pool by q_0/Q, and average accuracy rises with participation at the rate

$$\mathrm{AA}'(Q) = \frac{\mathrm{MA}(Q) - \mathrm{AA}(Q)}{Q} = \frac{q_0}{Q^2} > 0 \qquad (Q > q_0). \tag{4}$$

A post's reward depends on how many producers post and on the credibility readers expect. When a share Q of producers post and readers expect credibility c, the reward per post is

$$R(Q, c) = \tau(Q)\big(x_0 + x_1 c\big), \qquad \tau(Q) = \tau_0 - \tau_1 Q. \tag{5}$$

[4] With ε_j uniform on $[\bar{\varepsilon}_0 - w_0, \bar{\varepsilon}_0]$ for slop producers and on $[\bar{\varepsilon}_1 - w_1, \bar{\varepsilon}_1]$ for investigative producers, MC rises at slope w_0/q_0 on $[0, q_0]$ and $w_1/(1 - q_0)$ on $(q_0, 1]$, and it jumps at q_0 by $\kappa_1 - w_0 + \bar{\varepsilon}_0 - \bar{\varepsilon}_1$. Equation (2) is the case $\bar{\varepsilon}_0 = 0$, $w_0 = \sigma q_0$, $\bar{\varepsilon}_1 = -\sigma q_0$ and $w_1 = \sigma(1 - q_0)$.

Feed visibility $\tau(Q)$ is the chance that a post lands in a reader's limited attention, and congestion lowers it at rate τ_1 as more producers post.[5] Baseline appeal x_0 is sharing for reasons other than accuracy, and accuracy scrutiny x_1 is the extra sharing of more credible posts.

Assumption 4 (Reward) *$x_0 > 0$, $x_1 > 0$ and $\tau_0 > \tau_1 \geq 0$.*

A producer values each amplification of her post at one unit, so R is both her reward and her post's expected amplification. Appendix A.1 derives equation (5) from readers who amplify when the expected payoff of sharing exceeds a cost and the state a post reports has a prior of one half, with $x_0 = \bar{a} + (u^+ + u^-)/2$ and $x_1 = (u^+ - u^-)/2$; here u^+ and u^- are a reader's payoffs from amplifying an accurate and an inaccurate post, and $\bar{a}$ is appeal unrelated to accuracy. Platform ranking enters through feed visibility, which is common to every post in the cell. Section 2.6 derives feed visibility from a game in which a post's readers share it, and gives what the network does to appeal and accuracy scrutiny.

Assumption 5 (Variety) *Each post reports a different development of the news cycle, so one accurate post does not settle another.*

Under variety, more posts lower a post's reward only by competing for readers' attention, through $\tau(Q)$.

Which credibility a post is rewarded for depends on what readers can see. When readers can recognize producer type, the Q-th producer is rewarded for her own accuracy, and the marginal reward is

$$\mathrm{MR}(Q) = \tau(Q)\left[x_0 + x_1 \mathrm{MA}(Q)\right]. \tag{6}$$

Under information asymmetry, readers cannot recognize producer type, so every post is rewarded for the pool's credibility, and the average reward is

$$\mathrm{AR}(Q) = \tau(Q)\left[x_0 + x_1 \mathrm{AA}(Q)\right]. \tag{7}$$

Because slop dilutes the pool, the average reward lies below the marginal reward of every investigative producer, by

$$\mathrm{MR}(Q) - \mathrm{AR}(Q) = \tau(Q) x_1 \frac{q_0}{Q} \qquad (Q > q_0). \tag{8}$$

This gap is a pooling tax on investigative producers that subsidizes slop. In Figure 2, it is the vertical distance between the marginal and the average reward above q_0.

Participation moves the average reward through two forces. On $Q > q_0$,

$$\mathrm{AR}'(Q) = \underbrace{\tau(Q) x_1 \frac{q_0}{Q^2}}_{\text{composition lift}} - \underbrace{\tau_1\left(x_0 + x_1 \mathrm{AA}(Q)\right)}_{\text{congestion drag}}. \tag{9}$$

The composition lift is largest just above q_0, where each investigative producer dilutes a pool of slop, and it shrinks as Q rises.

The timing is as follows. Nature draws each producer's type and cost noise, which only the producer knows. Producers decide simultaneously whether to post. Readers see the posts that reach their feed and decide whether to amplify each, given the credibility they expect of the cell's posts. Each producer receives the amplification of her post and pays her cost. Because every post in a cell earns the same reward, the producers who post are those with the lowest costs, and the share Q who post summarizes entry.

2.3 Equilibrium and two forms of unraveling

Figure 2 adds the two reward curves to the bottom panel. Under pooling, entry stops where the average reward AR meets the marginal cost MC. If readers could recognize producer type, each producer would be paid for her own accuracy, and entry would stop where the marginal reward MR meets MC, at the full-information entry benchmark Q^{FI}. The shaded band marks the investigative producers ranked between the two points.

[5] With one reading slot and one rival post present with probability Q, τ_0 is the value of reading and $\tau_1 = \tau_0^2/2$ (Lemma A.3).

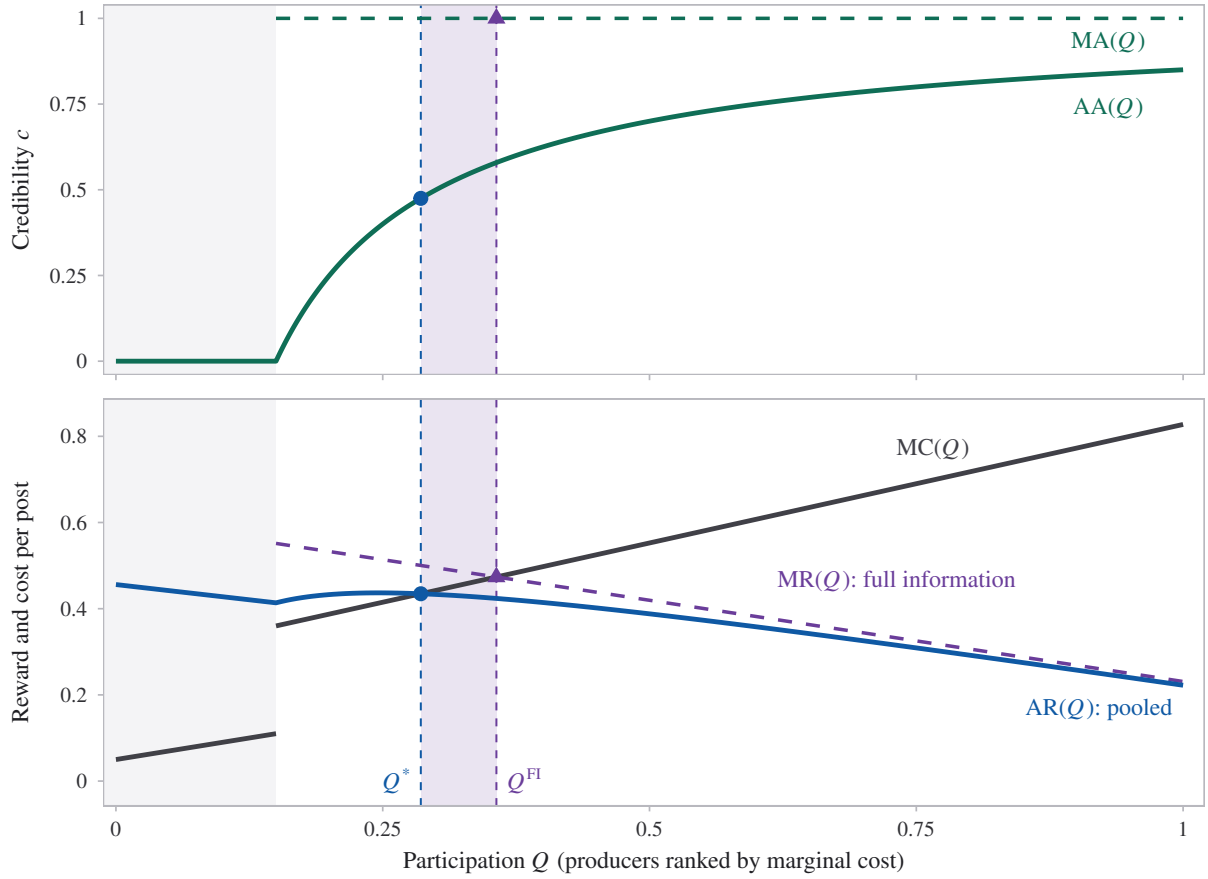


Figure 2: Entry under pooling and under full information. Bottom panel: the marginal cost MC, the average reward AR that every post earns when readers cannot recognize producer type, and the marginal reward MR that the producer at rank Q would earn if they could. Entry under pooling stops at Q^*, where AR meets MC, and the full-information entry benchmark Q^{FI} is where MR meets MC. The shaded band marks the investigative producers, ranked between Q^* and Q^{FI}, who post under full information but not under pooling. Top panel: the marginal and average accuracy of Figure 1. The grey band marks the slop producers, $Q \leq q_0$. Illustrative parameter values (Appendix A.3).

An equilibrium requires readers' credibility to be right and every producer's posting decision to be a best response.

Definition 2.1 (Equilibrium) *An equilibrium with participation $Q^* \in (0,1]$ is a pair (Q^*, c^*) such that readers' credibility equals the pool's average accuracy, $c^* = \text{AA}(Q^*)$, and, at the reward $R^* = R(Q^*, c^*)$, every producer with $\text{mc}_j < R^*$ posts and every producer with $\text{mc}_j > R^*$ does not. Silence, $Q^* = 0$, is an equilibrium if some credibility between the lowest and the highest accuracy, assigned by readers to a post that no producer is expected to make, yields a reward of at most $\text{MC}(0)$ at $Q = 0$.*

Write $D(Q) = \text{AR}(Q) - \text{MC}(Q)$ for the gap between the average reward and the marginal cost, and $\text{MC}(q_0^-)$ and $\text{MC}(q_0^+)$ for the cost of the last slop producer and of the first investigative producer. The first investigative producer's marginal reward is $\text{MR}(q_0^+) = \tau(q_0)(x_0 + x_1)$, the limit of $\text{MR}(Q)$ as Q falls to q_0.

Lemma 2.2 (Equilibrium conditions) *Let the reward $R(Q,c)$ be continuous in Q and nondecreasing in c, as in equation* (5)*, and let $\text{AR}(Q) = R(Q, \text{AA}(Q))$. Under Assumptions 1 and 2, $(Q^*, \text{AA}(Q^*))$ with*

$Q^* \in (0,1]$ is an equilibrium if and only if one of the following holds: (i) $Q^* \in (0,1)$, $Q^* \neq q_0$ and $D(Q^*) = 0$; (ii) $Q^* = q_0$ and $\mathrm{MC}(q_0^-) \leq \mathrm{AR}(q_0) \leq \mathrm{MC}(q_0^+)$; (iii) $Q^* = 1$ and $D(1) \geq 0$. Silence is an equilibrium if and only if $R(0,0) \leq \mathrm{MC}(0)$, which under equation (5) is $\tau_0 x_0 \leq \mathrm{MC}(0)$.

At the cost jump a whole range of rewards supports participation q_0, at which slop posts, no investigative producer does, and credibility is zero.

The same equilibrium can be read on the credibility axis. The credibility that the Q-th producer needs to break even is $\mathrm{SA}(Q) = [\mathrm{MC}(Q)/\tau(Q) - x_0]/x_1$, and

$$\mathrm{AR}(Q) - \mathrm{MC}(Q) = \tau(Q)\, x_1 \left[\mathrm{AA}(Q) - \mathrm{SA}(Q)\right]. \tag{10}$$

The average reward covers the marginal cost exactly where the pool has the credibility that the marginal producer needs.

Several equilibria can coexist, and we compare them by their stability under the adjustment $\dot{Q} = D(Q)$. Participation rises when the marginal producer's reward exceeds her cost and falls when it does not. An equilibrium is stable when D is positive just below it and negative just above it, so that the adjustment returns participation to it. Full entry, which participation cannot exceed, is stable when D is positive just below one. The adjustment ranks equilibria by stability and is not a model of time.

Readers who can recognize producer type would bring in more investigative producers. Under full information each post is rewarded for its producer's own accuracy, and an equilibrium is a participation Q at which every producer with $\mathrm{mc}_j < \tau(Q)(x_0 + x_1\alpha_j)$ posts and every producer with $\mathrm{mc}_j > \tau(Q)(x_0 + x_1\alpha_j)$ does not.

Proposition 2.3 (Full information: more investigative entry) *Suppose $\mathrm{MR}(q_0^+) > \mathrm{MC}(q_0^+)$ and $\mathrm{MR}(1) < \mathrm{MC}(1)$.*

(i) A unique $Q^{\mathrm{FI}} \in (q_0, 1)$ solves $\mathrm{MR}(Q^{\mathrm{FI}}) = \mathrm{MC}(Q^{\mathrm{FI}})$, and every equilibrium has $Q^* < Q^{\mathrm{FI}}$.

(ii) If readers can recognize producer type and $\mathrm{MC}(q_0^-) \leq \tau(Q^{\mathrm{FI}})\, x_0$, it is an equilibrium for the producers with the Q^{FI} lowest costs to post.

The full-information entry benchmark Q^{FI} is the participation at which the marginal producer would stop posting if she were paid for her own accuracy. When the inequality in (ii) holds, the producers ranked between Q^* and Q^{FI}, all of them investigative, post in the full-information equilibrium of (ii) but not under pooling.

Figure 3 shows the two forms of unraveling. In panel (a), the average reward lies below the marginal cost over every investigative producer, although the marginal reward lies above it. Pooling then prices out every investigative producer, and only slop posts. In panel (b), the cost jump at q_0 is an entry barrier, and the average reward meets the marginal cost three times: at a low-credibility equilibrium with slop alone, at a tipping point, and at a high-credibility equilibrium.

Proposition 2.4 (Unraveling 1: pooling prices out investigative producers) *Suppose slop producers cover their cost at zero credibility, $\mathrm{MC}(q_0^-) \leq \tau(q_0)\, x_0$, and $\mathrm{AR}(Q) < \mathrm{MC}(Q)$ for every $Q \in (q_0, 1]$. Then the unique equilibrium is $Q^* = q_0$, $c^* = 0$, with $R^* = \tau(q_0)\, x_0 > 0$. The condition on* AR *holds whenever $\mathrm{AR}(Q) < \mathrm{MC}(Q) < \mathrm{MR}(Q)$ for every $Q \in (q_0, Q^{\mathrm{FI}})$, under the hypotheses of Proposition 2.3; investigative producers then cover their cost under full information but not under pooling.*

Pooling prices out every investigative producer, while baseline appeal still covers the cost of slop. Participation unravels to slop and credibility to zero, yet slop still goes viral: $R^* = \tau(q_0)\, x_0 > 0$. Posting continues, but credibility disappears. On the credibility axis, the wedge reads $\mathrm{AA}(Q) < \mathrm{SA}(Q) < \mathrm{MA}(Q)$: the pool has less credibility than the marginal investigative producer needs, although her own accuracy would suffice.

The cost jump can also sustain a low-credibility equilibrium next to a high-credibility one.

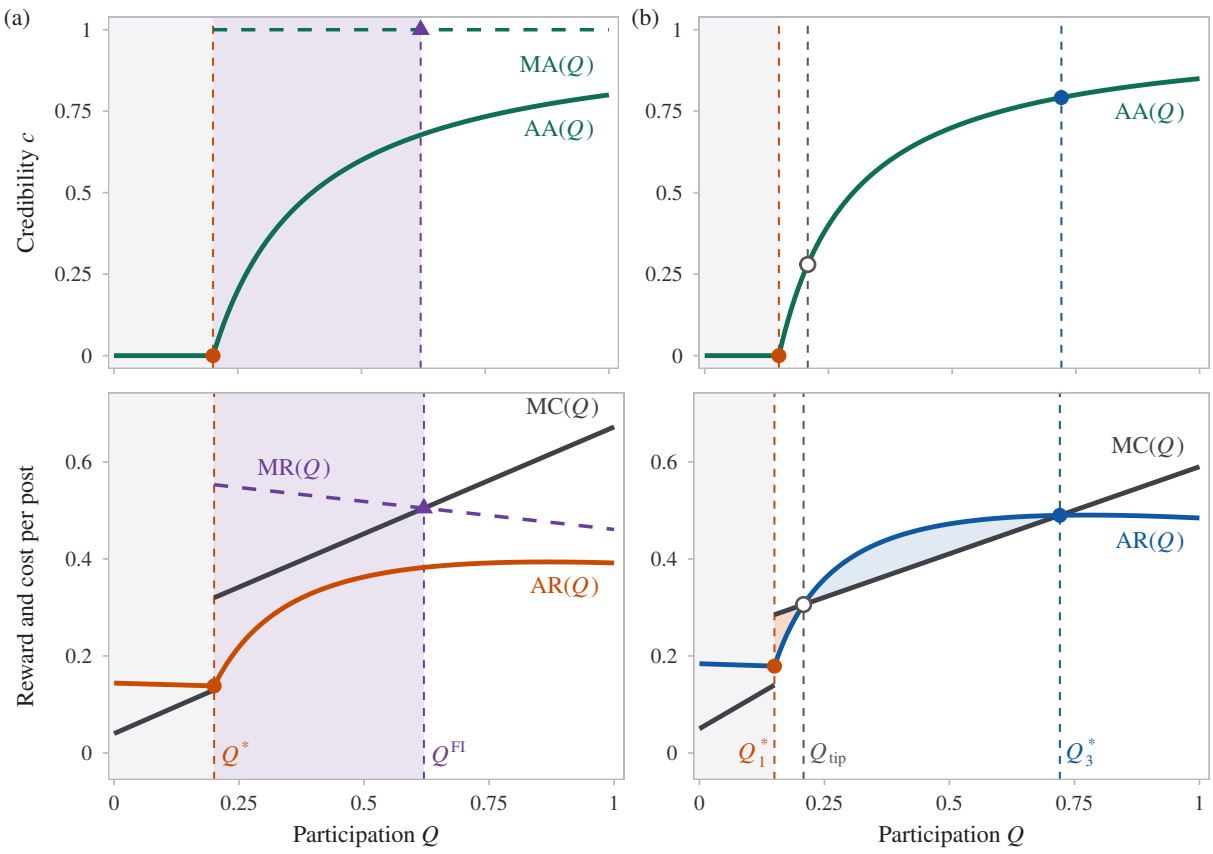


Figure 3: Two forms of unraveling. (a) Pooling prices out investigative producers: the average reward AR lies below the marginal cost MC over every investigative producer, so only slop posts and $Q^* = q_0$. The marginal reward MR lies above MC up to the full-information entry benchmark Q^{FI}; the shaded band between Q^* and Q^{FI} marks the investigative producers whom pooling prices out. (b) Multiple equilibria: the cost jump at q_0 is an entry barrier, and AR meets MC at a low-credibility equilibrium with slop alone (Q_1^*), at an unstable tipping point Q_{tip}, and at a high-credibility equilibrium (Q_3^*). The shaded areas mark where AR lies below MC, between q_0 and Q_{tip}, and where it exceeds MC, between Q_{tip} and Q_3^*, so entry beyond the tipping point carries the market to the high-credibility equilibrium. Grey bands mark the slop producers, $Q \le q_0$. Illustrative parameter values (Appendix A.3).

Proposition 2.5 (Unraveling 2: multiple equilibria) *Under Assumption 2, suppose slop producers cover their cost at zero credibility,* $\text{MC}(q_0^-) \le \tau(q_0)\,x_0$.

(i) Without an entry barrier, $\text{AR}(q_0) > \text{MC}(q_0^+)$*, the equilibrium is unique: it is interior and stable if* $D(1) < 0$*, and it is full entry otherwise.*

(ii) With an entry barrier, $\text{AR}(q_0) = \tau(q_0)\,x_0 < \text{MC}(q_0^+)$*, and* $\text{AR}(Q_a) > \text{MC}(Q_a)$ *at some* $Q_a \in (q_0, 1]$*, there are exactly three equilibria: a low-credibility equilibrium* $Q_1^* = q_0$ *with* $c_1^* = 0$*; a tipping point* $Q_{\text{tip}} \in (q_0, Q_a)$*; and a high-credibility equilibrium* $Q_3^* \ge Q_a$*, interior if* $D(1) < 0$ *and full entry otherwise, with* $c_3^* > \text{AA}(Q_{\text{tip}}) > 0$*. The low- and high-credibility equilibria are stable and the tipping point is not. Under the adjustment, D points toward the low-credibility equilibrium at every participation below* Q_{tip}*, and participation that starts above* Q_{tip} *converges to the high-credibility equilibrium.*

The entry barrier is the cost jump at q_0: the first investigative producer costs more than the reward at zero

credibility. The composition lift can still raise the average reward above the marginal cost further to the right, and the curves then cross at a tipping point and, unless entry is full, again at the high-credibility equilibrium. The low-credibility equilibrium is stable because an investigative producer who posts just above q_0 loses money, and the adjustment reaches the high-credibility equilibrium only from participation above the tipping point. Numerical result A.1 gives parameter values for each case of Propositions 2.4 and 2.5, including cases with $\tau_1 = 0$: neither form of unraveling needs congestion.

2.4 Credibility and amplification: when they co-move, when they diverge

Whether the market rewards credibility is a comparison across equilibria: whether amplification per post, R^*, rises with credibility, c^*. The answer turns on the sign of selection, so we first let selection be advantageous. Suppose the cheapest producers are specialists, who are accurate ($\alpha_j = 1$) and make up the first share q_1 of the cost order, and generalists, with accuracy $\alpha_g \in [0,1)$, follow; every specialist's cost is below every generalist's. Average accuracy then falls with participation on $Q > q_1$:

$$\mathrm{AA}(Q) = \alpha_g + (1-\alpha_g)\frac{q_1}{Q}, \qquad \mathrm{AA}'(Q) = -(1-\alpha_g)\frac{q_1}{Q^2} < 0. \tag{11}$$

The producer who would post next is less accurate than the pool. We call this the specialist variant; Lemma 2.2 holds in it with q_1 in place of q_0, and with $\tau_0(x_0 + x_1\alpha_g) \leq \mathrm{MC}(0)$ as the condition for silence.

Proposition 2.6 (Credibility and amplification) *In the baseline or the specialist variant, let (Q^*, c^*) be an interior equilibrium with $Q^* > q_0$ ($Q^* > q_1$ in the specialist variant) and $D'(Q^*) < 0$.*

(i) Along a change in costs that moves MC *at Q^* but preserves the cost order of producers, so that* AA *and* AR *stay in place,*

$$\frac{dR^*}{dc^*} = \frac{\mathrm{AR}'(Q^*)}{\mathrm{AA}'(Q^*)} = \underbrace{\tau(Q^*)x_1}_{\textit{direct trust}} - \underbrace{\tau_1\left(x_0 + x_1 c^*\right)}_{\textit{congestion}} \cdot \underbrace{\frac{1}{\mathrm{AA}'(Q^*)}}_{\textit{selection}}. \tag{12}$$

(ii) Along a change in readers' reward that moves AR *at Q^* but leaves* MC *in place,*

$$\frac{dR^*}{dc^*} = \frac{\mathrm{MC}'(Q^*)}{\mathrm{AA}'(Q^*)}. \tag{13}$$

Along a change in costs, credibility and amplification co-move under advantageous selection, because both terms of equation (12) are positive. Under adverse selection the selection term is negative, and amplification falls as credibility rises exactly when congestion drag exceeds the composition lift at Q^*, that is, when $\mathrm{AR}'(Q^*) < 0$. Congestion is what lets the two diverge: with $\tau_1 = 0$, $dR^*/dc^* = \tau_0 x_1 > 0$ along any change in costs that preserves the cost order. A change in readers' reward reverses the pattern, since $\mathrm{MC}' > 0$: credibility and amplification co-move under adverse selection and diverge under advantageous selection.

Proposition 2.6 is local, and the same comparison holds between two cost regimes, one in which checking facts is costly and one in which it is cheap. Write q for the share of the first type in the cost order, q_0 in the baseline and q_1 in the specialist variant.

Proposition 2.7 (Cheaper checking) *Consider two regimes, h and l, with the same readers and the same costs for the first type, in which the second type's marginal cost is higher in regime h: $\mathrm{MC}_h(Q) > \mathrm{MC}_l(Q)$ for every $Q \in (q,1]$. Suppose that in both regimes the marginal cost satisfies Assumption 2 with q in place of q_0, there is no entry barrier, $\mathrm{AR}(q) > \mathrm{MC}_h(q^+)$ and $\mathrm{AR}(q) > \mathrm{MC}_l(q^+)$, and the gaps $D_h = \mathrm{AR} - \mathrm{MC}_h$ and $D_l = \mathrm{AR} - \mathrm{MC}_l$ are negative at $Q = 1$. Then each regime has a unique equilibrium with positive participation, and $Q_h^* < Q_l^*$.*

(i) In the baseline, $c_h^* < c_l^*$; every producer who posts only in regime l is investigative, and so more accurate than the pool of regime h; and $R_h^* > R_l^*$ if $\mathrm{AR}' < 0$ on $[Q_h^*, Q_l^*]$.

(ii) In the specialist variant, $c_h^* > c_l^*$; every producer who posts only in regime l is a generalist, and so less accurate than the pool of regime h; and $R_h^* > R_l^*$.

(iii) If each producer keeps her rank among the producers of her type in both regimes (Assumption 6), every producer who posts in regime h also posts in regime l, and the producers who post only in regime l are those ranked in $(Q_h^*, Q_l^*]$.

Proposition 2.7 compares the pool that posts while checking is costly with the producers who post only once it is cheap. This reading needs each producer to keep her place in the cost order of her type.

Assumption 6 (Rank preservation) *Each producer has the same rank among the producers of her type in regimes h and l.*

In the baseline with separated costs, a change in the verification cost κ_1 moves every investigative producer's cost by the same amount, so it satisfies Assumption 6. Under the assumption, Proposition 2.7(iii) shows that every producer who posts while checking is costly also posts once it is cheap, and that the producers who post only once it is cheap are those ranked in $(Q_h^*, Q_l^*]$. Under adverse selection they are investigative and more accurate than the pool that posts while checking is costly, and that pool earns more amplification per post when congestion drag exceeds the composition lift throughout $[Q_h^*, Q_l^*]$. Under advantageous selection they are generalists and less accurate than that pool, which also earns more amplification per post.

Readers' payoffs move the reward through x_0 and x_1, and only some changes in them lower credibility.

Proposition 2.8 (Tolerance and rotation) *In the baseline, let (Q^*, c^*) be an interior equilibrium with $Q^* > q_0$ and $D'(Q^*) < 0$, and let readers' payoff from amplifying an inaccurate post, u^-, rise.*

(i) If u^+ and $\bar{a}$ stay fixed, participation and credibility rise: $dQ^*/du^- > 0$ and $dc^*/du^- > 0$.

(ii) If $\bar{a}$ stays fixed and u^+ falls by $\zeta \geq 0$ for each unit rise in u^-, participation and credibility fall if and only if $c^* > \tilde{c} = (1-\zeta)/(1+\zeta)$.

A higher tolerance for inaccurate posts raises x_0 by as much as it lowers x_1, so it raises the reward at every credibility below one and, at an interior equilibrium with $Q^* > q_0$ and $D'(Q^*) < 0$, draws in more producers. Among changes that raise u^-, a population of readers that rewards accuracy less therefore sustains lower credibility only through a rotation, which also lowers its payoff from accurate posts, and only where credibility already exceeds the pivot $\tilde{c}$.

2.5 Double selection: readers are also amplifiers

So far selection has worked on one margin, which producers post. Readers also amplify, and the average reward adds up their amplification decisions, so which readers choose to amplify can itself be selected. Double selection is selection on production, who posts, and on amplification, who amplifies.

Figure 4 draws two kinds of readers, defined below. Social readers share for appeal, and their reward is nearly flat in credibility. Certifiers share a post for its value to a decision, and their reward is zero until the pool's credibility clears their prior threshold, at Q_{cert}, and rises after it. Added together, the two give a pooled reward that kinks up at Q_{cert} and can meet the marginal cost three times: at a Belief Equilibrium below Q_{cert}, at a tipping point, and at a News Equilibrium above it.

A post is worth amplifying for some readers only if it can change a decision.

Lemma 2.9 (Prior threshold) *A reader who chooses an action $a \in \{0,1\}$ to match a binary state, earning one if the action matches the state and zero otherwise, with prior b that the state is one, gains from one post drawn from a pool with credibility c, whose report is correct with probability $(1+c)/2$ in either state, the decision value*

$$\frac{\left[c - \bar{c}(b)\right]_+}{2}, \qquad \bar{c}(b) = |2b-1|. \tag{14}$$

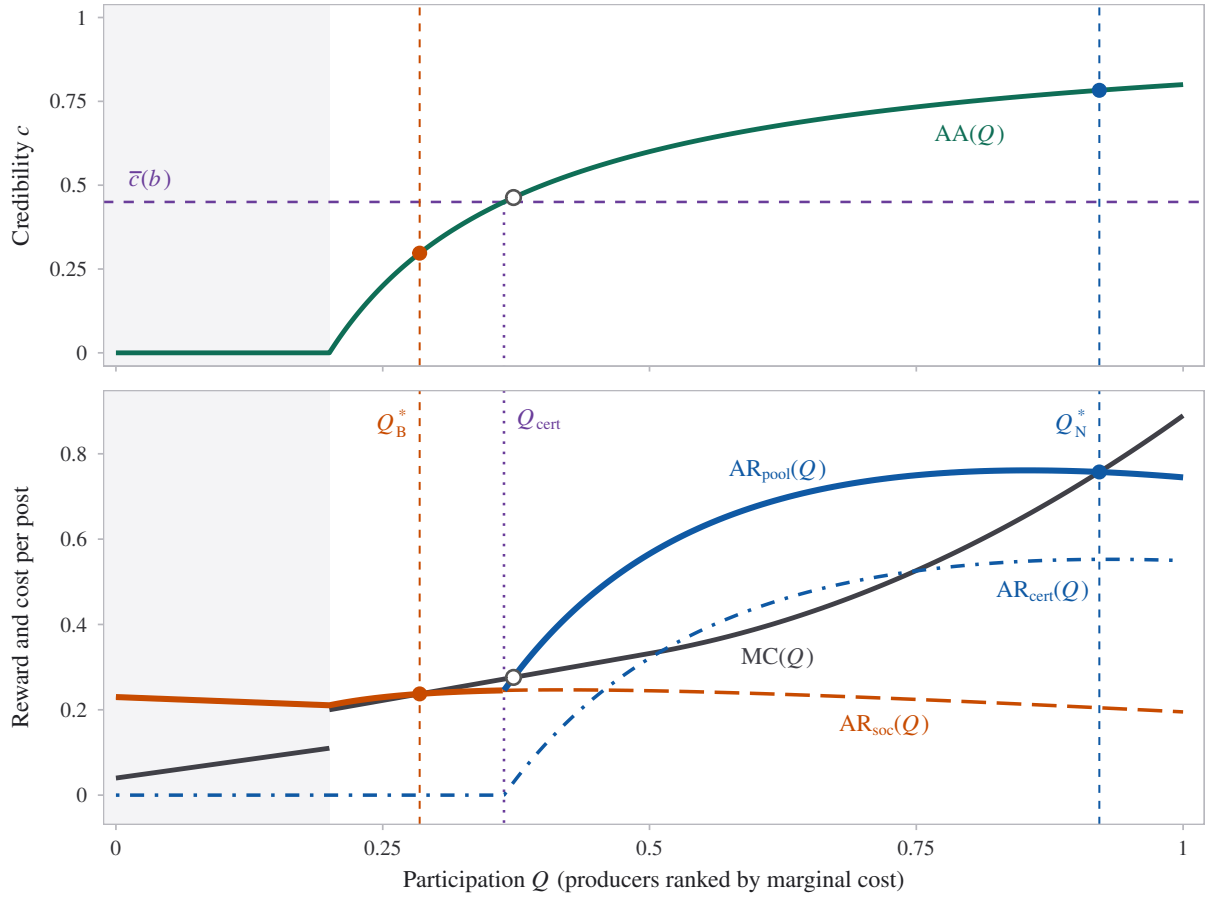


Figure 4: Two kinds of readers and two equilibria. Bottom panel: the reward from social readers, who share for appeal, the reward from certifiers, who amplify only once the pool's credibility clears their prior threshold at Q_{cert}, their sum AR_{pool}, and the marginal cost MC. The pooled reward meets MC at a Belief Equilibrium, in which only social readers amplify, at a tipping point, and at a News Equilibrium, in which certifiers amplify too. Top panel: the average accuracy AA and the prior threshold $\bar{c}(b)$. The grey band marks the slop producers, $Q \leq q_0$. Illustrative parameter values (Appendix A.3).

A post changes a reader's decision only if its credibility clears her prior threshold $\bar{c}(b)$, which is zero for a reader with no leaning and rises with the strength of her prior.

Assumption 7 (Social readers and certifiers) *A cell has two kinds of readers, and the reward is the sum of their amplification,* $\text{AR}_{\text{pool}} = \text{AR}_{\text{soc}} + \text{AR}_{\text{cert}}$. *Social readers share for appeal:* $\text{AR}_{\text{soc}}(Q) = \tau(Q)\,[x_{0,\text{soc}} + x_{1,\text{soc}}\text{AA}(Q)]$, *with* $x_{0,\text{soc}} > 0$ *and* $x_{1,\text{soc}} \geq 0$. *Certifiers share a post for its value to a decision about the development it reports, which no other post settles (Assumption 5): they hold a common prior* $b \neq 1/2$ *of their own, while the state's prior stays one half, and* $\text{AR}_{\text{cert}}(Q) = \tau(Q)\,x_{1,\text{cert}}\,[\text{AA}(Q) - \bar{c}(b)]_+$, *with* $x_{1,\text{cert}} > 0$. *Both depend on a post only through the cell's credibility.*

Social readers share mainly for appeal: $x_{1,\text{soc}}$ is small, and their reward curve is nearly flat in credibility. Certifiers' reward is proportional to the decision value of Lemma 2.9; the proportionality is an assumption. In the baseline, certifiers stay out until average accuracy reaches $\bar{c}(b)$, at $Q_{\text{cert}} = q_0/(1-\bar{c}(b))$, and the pooled reward kinks up there.

Proposition 2.10 (Double selection) *In the baseline with the readers of Assumption 7 and a marginal cost that satisfies Assumption 2, suppose* $Q_{\text{cert}} < 1$ *and*

(a) social readers alone cover the first investigative producer's cost: $\tau(q_0)x_{0,\text{soc}} > \text{MC}(q_0^+)$*;*

(b) the reward falls short of the marginal cost where certifiers would join: $\text{AR}_{\text{soc}}(Q_{\text{cert}}) < \text{MC}(Q_{\text{cert}})$*;*

(c) the pooled reward exceeds the marginal cost somewhere beyond: $\text{AR}_{\text{pool}}(Q_a) > \text{MC}(Q_a)$ *at some* $Q_a \in (Q_{\text{cert}}, 1]$.

Then there are exactly three equilibria: a Belief Equilibrium $Q_B^* \in (q_0, Q_{\text{cert}})$ *with* $0 < c_B^* < \bar{c}(b)$*, in which certifiers do not amplify; a tipping point in* (Q_{cert}, Q_a)*; and a News Equilibrium* $Q_N^* \geq Q_a$ *with* $c_N^* > \bar{c}(b)$*, in which certifiers amplify. The Belief and News Equilibria are stable and the tipping point is not. Without certifiers, the Belief Equilibrium is the unique equilibrium.*

The same producers and the same readers support two stable equilibria. In the Belief Equilibrium the feed changes no certifier's decision, certifiers sit out, and only social readers amplify. In the News Equilibrium credibility clears the prior threshold, certifiers join, and the pooled reward rises steeply with credibility. Selection on amplification thus reinforces selection on production: certifiers amplify only where enough investigative producers post, and enough investigative producers post only where certifiers amplify. Numerical result A.1 gives parameter values that satisfy conditions (a) to (c).

2.6 Networks

Every parameter above may differ across cells, because each cell is a separate market with its own producers and readers. The network plays two roles, and readers' double role as amplifiers links them. Figure 5 draws the two roles on the curves of Figures 1 and 4. In panel (a), the cell's producer pool sets the shape of the average accuracy curve: it rises with participation where slop producers have the lowest costs, as in the baseline, and falls where accurate specialists do, as in the specialist variant of Section 2.4. In panel (b), the cell's reader mix sets the shape of the reward: nearly flat in credibility where social readers dominate, and kinked up where certifiers carry weight.

First, the network defines the cell. Readers condition on the producer's position G and her signal s, so credibility is set cell by cell. Each cell has its own pool of potential producers, and the pool sets the sign of selection. Selection is adverse where slop producers have the lowest costs, as in the baseline, and advantageous where accurate specialists do, as in the specialist variant.

Second, two features of a producer's followers set the cell's mix of readers, between certifiers, who reward credibility, and social readers, who share for appeal (Figure 6). Reciprocity is the share of a producer's followers whom she follows back. Overlap is the degree to which her followers are followed by the same accounts. Low overlap marks a diverse audience: each follower who amplifies her post opens a different audience. In the first step of a post's spread, from the producer to her followers and from them to the accounts that follow them, a diverse audience is a star: each follower's reposts reach accounts of their own. An overlapping audience is a chamber: her followers are followed by the same accounts, so a second repost reaches accounts that a first has already reached. Overlap is a directed analog of one minus the reach efficiency of Borgatti (1997), which is large when a node's contacts connect to different third parties. Section 3 computes both on the following lists we collected.

The reward of equation (5) is the equilibrium of a game among the readers a post reaches in this first step (Appendix A.4). Each reader chooses how much to share each of the cell's posts. Sharing a post of credibility c is worth $x_0 + x_1 c$ to her, her expected payoff from amplifying it in Lemma A.1, and it is worth more when the accounts she follows share it too. The matrix Ω holds these gains: its entry $\omega_{kk'}$ is positive when reader k follows reader k' and zero otherwise. A reader who shares more posts in total pays more for her attention, at the rate ν. As in the games of Ballester et al. (2006), each reader's equilibrium sharing is proportional to her Bonacich centrality in Ω, and the reward is

$$R(Q,c) = \tau(Q)\left(x_0 + x_1 c\right), \qquad \tau(Q) = \mathbf{1}^\top \left[(1+\nu Q)I - \Omega\right]^{-1}\mathbf{1}. \tag{15}$$

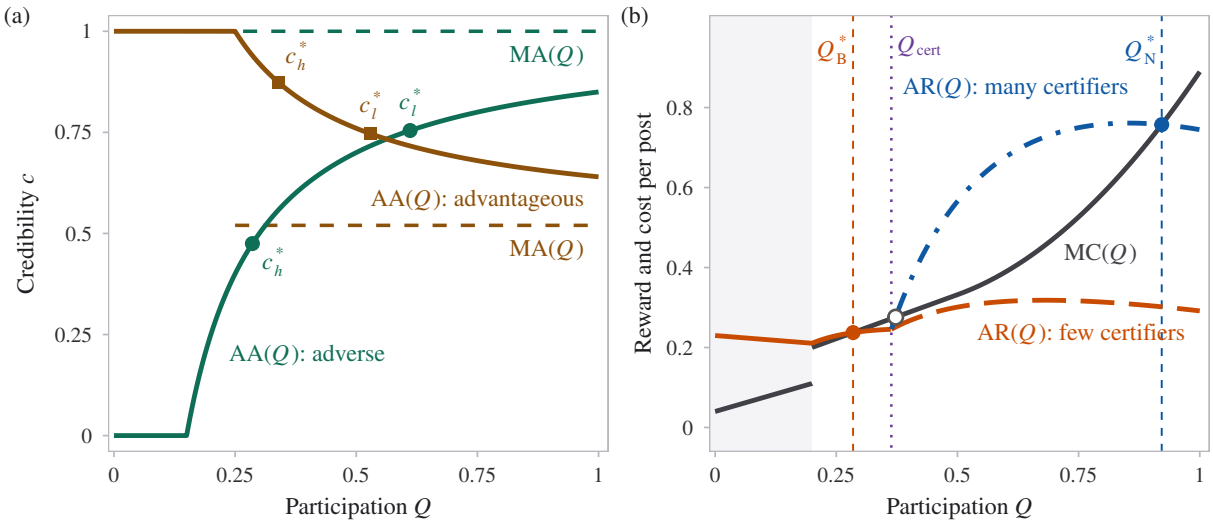


Figure 5: The network's two roles in the model. (a) The producer pool: the average accuracy AA(Q) rises with participation where slop producers have the lowest costs (adverse selection) and falls where accurate specialists do (advantageous selection). On each curve the points mark the equilibrium credibility when checking is costly, c_h^*, and when it is cheap, c_l^*: the pool that posts while checking is costly is the less accurate one under adverse selection and the more accurate one under advantageous selection (Proposition 2.7). (b) The reader mix: two cells with the same producers and the same social readers, which differ only in the weight of certifiers. Below Q_{cert} the two rewards coincide. With little certifier weight (few certifiers) the Belief Equilibrium is the only equilibrium; with more (many certifiers), a News Equilibrium coexists with it, and a tipping point (open circle) separates them. In (b), the grey band marks the slop producers, $Q \leq q_0$. Illustrative parameter values (Appendix A.3).

Feed visibility at $Q = 0$ is the sum of the readers' centralities. It falls as more producers post, because attention a reader spends on one post is not spent on another, which lowers her sharing of every post and the sharing that hers sets off among the readers who follow her. The line $\tau_0 - \tau_1 Q$ of equation (5) is this visibility to first order in Q, and the results above use the line.

Congestion bites hardest where readers follow one another densely. When the gains from the accounts each reader follows sum to the same $\omega_+ < 1$ for every reader, visibility is $N/(1-\omega_+ + \nu Q)$ for N readers, and its rate of fall, $\nu/(1-\omega_+ + \nu Q)$, rises with ω_+. We do not measure how densely a producer's readers follow one another. If chambers are also the denser audiences, which we assume rather than measure, congestion bites harder in them.

In Assumption 7, the mix is the size of the certifiers' coefficient $x_{1,\text{cert}}$ relative to the social readers' coefficients $x_{0,\text{soc}}$ and $x_{1,\text{soc}}$. Overlap lowers the certifiers' weight and leaves the social readers' appeal in place. When a producer's followers are followed by the same accounts, amplification by different followers reaches the same readers. Her post then reaches fewer distinct readers beyond her followers, and so fewer readers who can act on its credibility. The sharing game gives this effect in the overlap we measure. Call the accounts that follow her followers but not her the further readers, and let the follows among readers run only from further readers to her followers and among her followers. Each further reader then adds her own value of sharing once, and her response to the followers' sharing once for each follower she follows. For a given number of second-degree arrows into each follower, the number of distinct further readers is the number of arrows times $1 - O_j$, by equation (17). Overlap therefore scales the further readers' own values by $1 - O_j$ and leaves the followers' values, and the sharing these set off, in place (Lemma A.6). A restricted benchmark derives the size of this effect. Suppose two followers each amplify a post with probability one half, each to one further reader, and a further reader who reads the post follows the producer with probability $c/2$, the

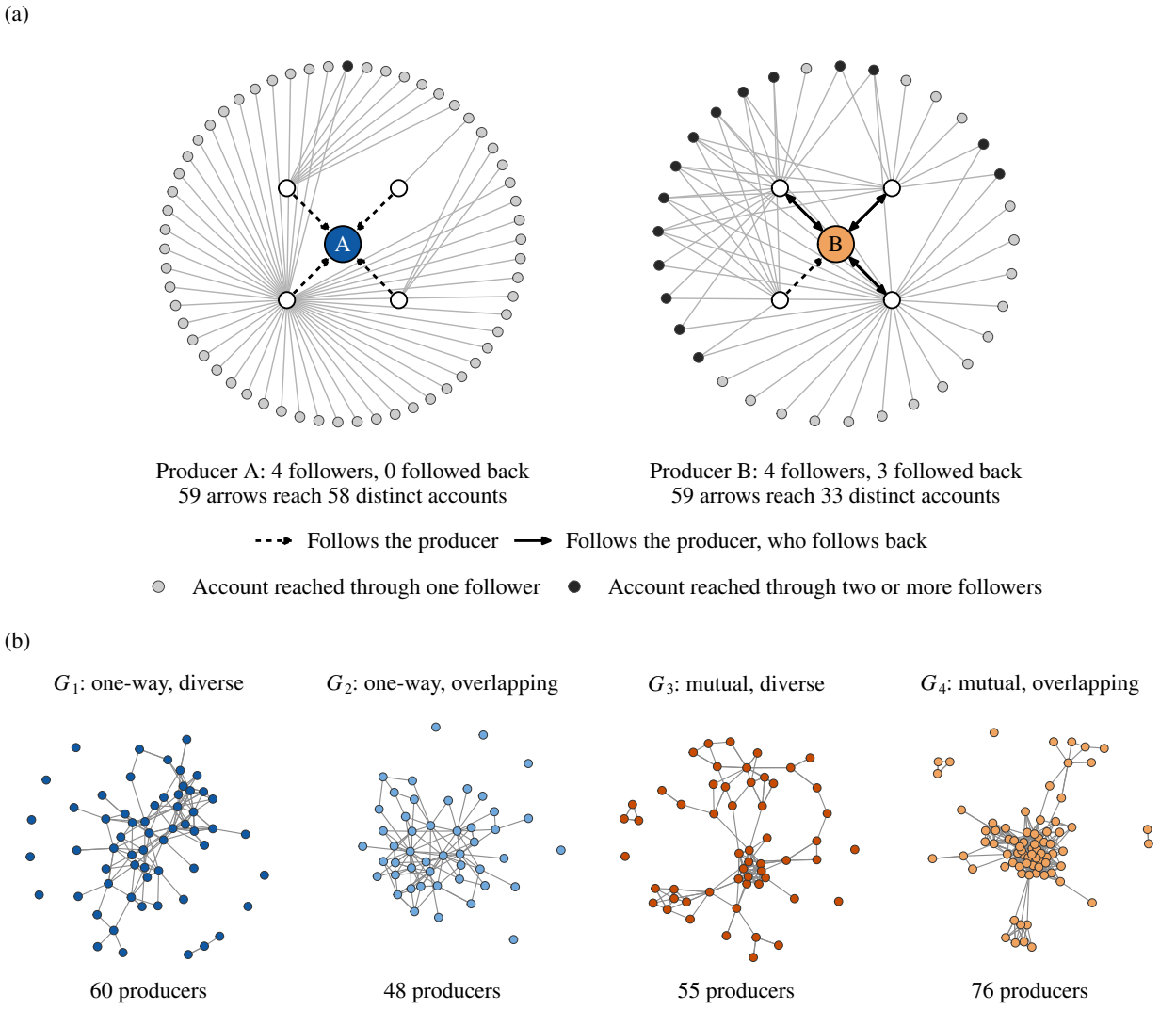


Figure 6: Reciprocity, overlap and the four network cells. (a) Two producers, each followed by four of the accounts whose following lists we collected. Producer B follows back three of her followers and producer A none, so reciprocity is high for B and low for A. The followers of producer A are followed by different accounts, and those of producer B by the same accounts, so amplification by A's followers reaches more distinct accounts; overlap is low for A and high for B. Producer A's audience is a star and producer B's a chamber. (b) The producers on the follow graph, by network cell: one-way and diverse (G_1), one-way and overlapping (G_2), mutual and diverse (G_3), and mutual and overlapping (G_4). Lines are follows among a cell's producers; each cell has its own force-directed layout, in which position shows only who is near whom. Reciprocity and overlap are defined in equations (16) and (17). Source: following lists collected in July 2026.

decision value of her next report at a prior of one half (Lemma 2.9). Further readers respond to the post only by following, so only through its credibility. In this benchmark the producer's reward is the value of what her post gains her rather than its amplification: a new follower is worth v to her, and x_0 is her value of access to her two followers' own content, less a posting cost common to every producer. Credibility's weight x_1 is then $v/2$ when the two further readers are distinct and $3v/8$ when they are the same reader, and x_0 is the same in

both.[6]

At a given number of second-degree arrows, overlap therefore sets the size of the pooling tax. The weight of credibility in the reward, τx_1, includes the further readers' own values, so it rises as overlap falls whenever further readers respond to credibility at all. So the pooling tax $\tau x_1 q_0/Q$ of equation (8) and the direct-trust term τx_1 of Proposition 2.6 are larger in a star than in a chamber with the same producers, the same followers and as many second-degree arrows into each follower. If further readers respond only to credibility, as in the benchmark, overlap lowers the weight of credibility and leaves the weight of appeal, τx_0, unchanged.

We read high reciprocity as marking a horizontal audience: a mutual follow is a relationship, and followers who follow back care about position more than news. The model states this reading as an assumption, under which reciprocity shifts weight from certifiers to social readers.

Assumption 8 (Reciprocity) *A follower whom the producer follows back is in a relationship with her rather than subscribed to her news, so the follower shares the producer's posts for appeal rather than for their value to a decision.*

Reciprocity thus sets the values her followers bring through this assumption, and overlap, in the sharing game, sets how many further readers bring values of their own.

Because readers are also amplifiers, the two roles are linked: the reader mix also selects who amplifies, and which readers amplify depends in turn on the credibility the cell's pool delivers. Without certifiers, the Belief Equilibrium is the unique equilibrium, and only social readers amplify (Proposition 2.10). When certifiers carry enough weight to satisfy condition (c) of that proposition, a News Equilibrium coexists with it, and in it certifiers amplify too. The reader mix therefore sets both how strongly the reward responds to credibility and which readers amplify.

The reader mix cannot reverse the sign of selection when misinformation risk, the mean chance that a post by a producer of a given cost rank is inaccurate, moves in one direction along the cost order, as with two types and with continuous risk types in the textbook insurance model, where willingness to pay rises with risk (Proposition B.3 and Corollary B.9.1).[7] Within a cell, readers reward every producer alike (Lemma A.2), so a change in readers moves the reward but does not reorder producers by cost. The cost order, and with it average accuracy, belongs to the cell's producer pool. Under that condition, a difference in the sign of selection across cells comes from the cells' producer pools.

2.7 Testable implications

Table 1 ties four features of the data to the model, each through a role of the network. The cells we measure are network positions, and the tests hold the producer's signal fixed through controls, such as her follower count, and through comparisons within a producer.

First, readers cannot recognize producer type, so a post's expected reward depends on its cell's credibility and not on whether the post is accurate. Within a producer and a cell, a false and an accurate post therefore earn the same expected amplification when they have the same public record and readers act only on it (Proposition B.7(i)), and whatever their public records when readers respond only to the cell's credibility (Proposition B.7(iii)). A misleading post differs from an accurate one in content readers can see, so its expected amplification can differ in either direction (Proposition B.7(ii) and (iv)). This implication rests on the first role of the network, the cell as what readers condition on.

[6] A further reader follows when this decision value exceeds a cost drawn uniformly from $[0,1]$. The expected number of distinct further readers reached is $2 \cdot \frac{1}{2} = 1$ when they are distinct and $1-(1-\frac{1}{2})^2 = \frac{3}{4}$ when they are the same reader, and each reader reached adds $vc/2$ to the producer's reward in expectation. If further readers also shared the post for appeal, overlap would scale that appeal by $3/4$ as well. Further readers who hold the certifiers' prior b follow with probability $[c-\bar{c}(b)]_+/2$, and overlap then scales $x_{1,\text{cert}}$ by the same factors, one and $3/4$.

[7] In insurance markets, a buyer's risk type affects her willingness to pay and determines her expected cost, and the marginal cost curve slopes down under adverse selection and up under advantageous selection (Einav and Finkelstein 2011), as misinformation risk falls or rises along the cost order here. When misinformation risk rises over some ranks and falls over others, the reader mix can also set the sign, because it sets where in the cost order the equilibria fall (Appendix B.5). This is the counterpart of a marginal cost curve that is not monotone, which in insurance arises once buyers' preferences differ (Einav and Finkelstein 2011).

Table 1: What the model predicts and how the data test it

Implication	In the model	Test in the data	Section
1. Readers reward what they can see of a post, not accuracy they cannot see	Proposition B.7	Reach of a producer's misleading and false or premature posts relative to her accurate posts	5.2
2. The pool that posts while checking is costly is less accurate than later producers under adverse selection and more accurate under advantageous selection	Figure 5(a); Proposition 2.7	Accuracy of the producers who post in the Breaking Window against the later producers of the same cell	5.3
3. Along a change in costs, amplification and credibility co-move under advantageous selection	Proposition 2.6	A producer's reach against her cell's accurate share across the windows of the cycle	Appendix C.2
4. Who amplifies differs between the Belief and the News Equilibrium	Figures 4 and 5(b); Proposition 2.10	Stance of quote tweets and replies; amplifiers' accuracy elsewhere; amplifiers active in the Breaking Window against later ones	5.4

Second, the composition of the pool that posts while checking is costly signs selection (Proposition 2.7). We read the Breaking Window as the regime in which checking is costly, a maintained assumption. Under Assumption 6, the comparison of the producers of a cell who post in it with those who post only after it is the comparison of Proposition 2.7: by its part (iii), every producer who posts in the Breaking Window also posts after it. When misinformation risk moves in one direction along the cost order, the reader mix cannot reverse the sign of this comparison (Section 2.6), so a difference in its sign across cells comes from the cells' producer pools.

Third, how amplification moves with credibility depends on the sign of selection (Proposition 2.6). Along a change in costs, credibility and amplification co-move under advantageous selection, and they can diverge under adverse selection only through congestion. Under the same condition on misinformation risk, the sign of selection again belongs to the cell's producer pool (Proposition B.3).

Fourth, who amplifies differs between the Belief and the News Equilibrium (Proposition 2.10). Readers who amplify posts for their value to decisions do so only where credibility clears their prior threshold. By the second role of the network, the reader mix leans most toward certifiers in one-way diverse cells (G_1) and most toward social readers in mutual overlapping cells (G_4). Because readers are also amplifiers, the mix also selects who amplifies.

Because the model can have several equilibria, a cell's outcome reflects both its primitives and the equilibrium it is in. In the model, the accurate share of a cell's amplification is $(1+c^*)/2$, so the composition of amplification by cell (Section 5.1) reflects the cell's producer pool, its reader mix and its equilibrium together; the data's accurate share is a different object, compared across cells only. Differences across cells do not by themselves show that a cell has several equilibria (Proposition B.9).

3 Data

We study one news cycle: the 2022 mpox outbreak on Twitter.[8] In May 2022 a long-known disease spread across many countries at once: by 21 May, 12 countries where it is not endemic had reported 92 confirmed and 28 suspected cases to the World Health Organization (World Health Organization 2022a). In the United States, the CDC confirmed a case on 18 May, in a Massachusetts resident who had recently traveled to

[8]On 28 November 2022 the World Health Organization recommended mpox as the new name of the disease, then called monkeypox, with both names in use for a year (World Health Organization 2022c). The keyword search that defines our data uses the 2022 names.

Table 2: Viral posts and producers by window

	W1 7–20 May	W2 21 May–3 Jun	W3 4–17 Jun	W4 18 Jun–1 Jul	All
Viral posts	555	767	250	158	1,730
with an image or video	218	347	119	92	776
text only	337	420	131	66	954
Producers with a viral post	382	517	166	115	958
first viral post in the window	382	419	83	74	958
Posts below the viral cut, same producers	100	251	232	496	1,079

Notes: Windows are 14 days long and start on 7 May 2022; W1 is the Breaking Window. A producer is the account that posted a viral post. The last column counts distinct producers, so the producer rows do not sum across windows. Posts below the viral cut are posts on the outbreak by producers of viral posts that fall short of the virality bar (Section 3).

Canada (Centers for Disease Control and Prevention 2022a). Within a few weeks of the first reported cases, "an avalanche of mostly false, misleading, irrelevant, or damaging information started to circulate on social media," and official institutions that provided accurate information "were overshadowed by rapidly spreading social media chatter" (Edinger et al. 2023). The facts changed from day to day, so a claim that could not be checked in the first days might be settled a few weeks later. The first viral post in our data is dated 7 May 2022 and the last 30 June 2022. Our population is the viral posts about the outbreak on Twitter; the tests in Section 5 also use posts by the same producers that fall below the viral cut, unweighted.

We divide the cycle into four 14-day windows, W1 to W4, starting on 7 May 2022 (Table 2). W1 runs from 7 to 20 May, W2 from 21 May to 3 June, W3 from 4 to 17 June and W4 from 18 June to 1 July. We call W1 the Breaking Window: the first two weeks of the cycle, which end as the CDC issued guidance on the outbreak to U.S. clinicians, in a media statement on 18 May and a health advisory on 20 May (Centers for Disease Control and Prevention 2022a; Centers for Disease Control and Prevention 2022b). That checking facts was costly in the Breaking Window and cheap afterward is a maintained assumption. Viral posts thin out after the first four weeks, and 84% of the producers post their first viral post in those four weeks.

The core data are the 1,730 viral posts about the outbreak. A post enters if it matches the outbreak's keywords, in English or in another language, or names a monkeypox vaccine, and if more than 1,000 distinct users retweeted it or more than 100 quoted it or replied to it. Each is an original post, not a retweet. Of these, 1,022 come from two 2022 collections, and 708 are posts those collections miss that we found in 2026. The 756 posts with an image or video come from a query of the full 2022 Twitter stream, which holds nearly every such post in the window. The 266 text-only posts come from a second 2022 collection that captured about 27% of viral text-only posts, so each carries a weight of 3.70 in totals over the corpus. No estimate applies this weight. The posts found in 2026 come from reading posts by their identifiers from a third public collection, gathered through Twitter's Academic API (Movahedi Nia et al. 2023), and from searching X: 281 are posts by our producers in that collection, 409 come from a keyword search of X for May and June 2022, and 18 come from the first sixth, in time order, of that collection's other posts in the window. Each passes the same keyword and viral rules (Appendix D.1). With them, the data hold 954 text-only viral posts, close to the roughly 985 that the second collection's share implies. Each post carries its counts of retweets, quotes and replies, and the collections record no likes. For posts with media the counts are those of the 2022 window; for text-only posts and the posts found in 2026 they are lifetime counts read in 2026. The 1,730 posts were posted by 958 accounts, which we call producers; 319 of them post only posts found in 2026.

We add posts by the same producers that fall below the viral cut, so that a producer's accuracy rests on more than its viral posts. These are 1,079 posts by 197 of the producers, posted between 7 May and 1 July 2022 and retrieved in 2026 (Appendix D.1). Of these, 660 were drawn with recorded probabilities, and 419 were added in a second labeling run whose rule for choosing posts was not recorded. With them, 2,809 posts carry labels (Section 4). The tests in Section 5 start from the 2,181 labeled posts, 1,120 of them viral, by the producers in the follow graph described below, and leave out the 186 N/A posts among them. They include the 305 posts found in 2026 by producers in the graph frame; the other 403, by 332 producers outside it, enter no network test, and Appendix C.6 reports the tests with them.

For 61 of the viral posts we also observe each engagement by a sample of users in 2022. The sample is 20% of the followers of the accounts that posted these posts, drawn by user identifier. The records give the time of every like, retweet, quote and reply by a sampled follower on these posts. They cover 59,348 sampled followers and 76,581 follower–post pairs. The 61 posts all carry an image or video and are among the most engaged in the corpus: each has more than 5,000 retweets or more than 500 quotes or replies, as do 136 of the 756 viral posts with an image or video in the 2022 collections. The rule that selected them was not recorded.

In 2026 we also collected the amplification records that remained public: the quote tweets and replies on the labeled posts, and the accounts that retweeted them. They are 40,764 quote tweets and 69,295 replies on 1,724 of the labeled posts, and lists of the accounts that retweeted 2,075 of the 2,181 posts the tests start from; the other 106 of these posts were no longer available. Section 5 uses them to measure who amplifies a post and whether a quote tweet endorses the post or calls it out.

We read the producers' profiles, with their follower counts, in 2026. By then the accounts of 58 of the 958 producers had been deleted or suspended.

The network is the follow graph among the producers and a set of other accounts. In July 2026 we collected the following lists of 575 accounts, 501 producers and 74 accounts outside the current corpus, and kept the follows from one of these accounts to another. The graph has 5,969 follows among 474 accounts, 426 of them producers, and the graph frame is the 448 producers with stored network measures. The rule that chose these 448 of the 639 producers of the 2022 collections was not recorded (Appendix D.1). Their median follower count on Twitter in 2026 was about 183,000, among the 414 producers whose count is known. The lists are current, so they carry no follow dates, and a follow may postdate the outbreak. A protected or deleted account shows no follows of its own, although the follows it receives are observed.

We place each producer in one of four network cells by two measures of its followers in the following lists we collected. Let F_j be the number of accounts whose collected following lists include producer j. Some of the accounts counted in F_j are not producers (Appendix D.6). Reciprocity is the share of those followers that j follows back:

$$r_j = \frac{\text{followers of } j \text{ that } j \text{ follows back}}{F_j}. \tag{16}$$

A producer is one-way if $r_j < 0.30$ and mutual if $r_j \geq 0.30$; the cut sits just below the median, 0.33. Overlap measures whether j's followers lead to new accounts or to the same ones. For each follower k of j, take the accounts whose collected lists include k, leaving out j and j's own followers; each such follow is a second-degree arrow of j. Overlap is one minus the ratio of the distinct accounts these arrows reach to the number of arrows:

$$O_j = 1 - \frac{\text{distinct second-degree accounts}}{\text{second-degree arrows}}. \tag{17}$$

Overlap rises with follower count, so each producer is compared with producers of a similar follower count. Within the bands F_j = 3–5, 6–10, 11–25 and 26 or more, whose median overlaps are 0.15, 0.31, 0.45 and 0.70 (taken over the 260 producers on which the cells were first computed; Appendix D.6), a producer is overlapping if O_j is at or above its band's median and diverse if below. Crossing the two splits gives four cells: G_1, one-way and diverse; G_2, one-way and overlapping; G_3, mutual and diverse; and G_4, mutual and overlapping. The cells hold the 240 producers with $F_j \geq 3$ and labeled posts that are not N/A; a producer with fewer than three such followers is in no cell. Of these producers, 60 are in G_1, 48 in G_2, 55 in G_3 and 77 in G_4.

Section 2.6 gives what the two measures read in a producer's audience, and Figure 6 draws them. The cells differ in shape as well. The median producer in G_1 is followed by 13.5 of the accounts whose following lists we collected and follows 4, while the producers of G_4 follow one another: 44% of the follows they receive come from inside G_4.

4 Labeling Misinformation

We label every post, including the text, image, or link it shows or quotes, on four blocks, each judged against what was knowable on the calendar date the post was posted. Judging on the posting date separates a post

Table 3: The four blocks of the labels

Block	Question, judged as of the posting date	Values
Vertical attributes: quality and form		
Truth	What does the post say outright? Q1: does it state a checkable fact? Q2: does the evidence that day settle it? Q3: if the claim was still open that day, does the post present it as open?	*no checkable fact* (Q1 no); *accurate* (Q2 holds, or Q3 yes); *false* (Q2 contradicted or baseless); *premature* (Q3 no)
Appeal	Why might the post spread?	*sensational*: emotion (alarm, outrage, a graphic lesion photo) or packaging (exaggeration, teasing, formatting); *humorous*: a clear sign of jest
Horizontal attributes: belief and position		
Belief	Does a general reader take away a specific untrue claim beyond the words, one that can be stated in one sentence? Asked only when Truth is accurate or no checkable fact.	*misleading*; not misleading
Position	Is the claim, stated or hinted, part of a conspiracy theory? Does the post take an explicit side?	*fringe* or *mainstream*; *slanted*

Notes: Misinformation is a post that is false or premature on Truth, or misleading on Belief. Appeal and Position are recorded for every post and do not enter the misinformation label. A post includes the text, image, or link it shows or quotes.

that got ahead of the evidence from a post that contradicted it. Labels taken from fact-checkers cannot make that separation. Researchers usually adopt the categories that fact-checkers construct (Altay et al. 2023), and a fact-check records how a claim was resolved, often after it had spread: in breaking news on Twitter, the median rumour later proven false took more than 14 hours to be resolved (Zubiaga et al. 2016).

Table 3 lists the four blocks. Truth and Appeal are vertical attributes, the quality and form of a post. Belief and Position are horizontal attributes, the belief a post leads to and the position it takes. A post is misinformation when it is *false* or *premature* on Truth, or *misleading* on Belief.

Truth records what the post says outright, in three questions asked in order. The first asks whether the post states a checkable fact; a post with none, such as an opinion, a joke or a genuinely open forecast, is coded *no checkable fact*. A claim the post asks as a question, or takes for granted, counts as stated. The second asks whether the evidence on the posting date settles the claim: a claim that holds is *accurate*, and a claim the record contradicts, or that had nothing behind it, is *false*. The third applies when something real was still open that day, such as a suspected case under test: a post that presents it as open is *accurate*, and a post that presents it as settled is *premature*. When a post makes several claims, the worst decides. A claim wrong only in degree, such as an inflated adjective on a true figure, stays accurate and is coded as exaggeration under Appeal.

Appeal records why a post may spread whatever its truth. A post is *sensational* when its emotion or its packaging goes beyond what a plain report of its facts calls for. The emotional types are alarm, outrage, and disgust or shock, which includes any close-up photo of lesions; they follow the evidence that content evoking high-arousal emotions such as anger and anxiety is more viral (Berger and Milkman 2012). The packaging types, exaggeration, teasing and dramatic formatting, follow the clickbait typology of Biyani et al. (2016). A post is *humorous* when it carries a clear sign of jest; sarcasm and ridicule are not humor. Appeal is coded for every post, whatever its Truth label: an accurate post can be sensational, and so can a false one. Truth and Appeal are therefore separate variables, and an analysis that compares appeal within a truth category does so by conditioning in the regression, not in the coding.

Belief records what a reader takes away beyond the words. A post is *misleading* when a general reader would take away a specific untrue claim that the labeler can state in one sentence and that the dated record refutes. The post need not state that claim: true facts set side by side, a leading question, a dropped condition or an old document presented as current can each plant it. Each claim is counted once, so a post already *false* or *premature* on Truth is not also coded misleading. The misleading posts are therefore the accurate but misleading posts and the misleading posts with no checkable fact. The reference reader is a general reader on the posting date, not the post's own audience. Belief therefore measures the claim a post implies, not the belief its readers came to hold.

Position records the post's horizontal alignment, in two fields. A post is *fringe* when its claim, stated or hinted, is part of a conspiracy theory: institutions lying, plotting or covering up; the outbreak planned or engineered; or the disease caused by vaccines. Every other post is *mainstream*, including harsh criticism of health agencies that alleges no plot. A post is *slanted* when it takes an explicit partisan, ideological or institutional side in its own words. Position is coded from the post alone. The producer's network is measured separately, on the follow graph (Section 3), and the identity of the producer may help a labeler understand a post but never decides its truth, slant or position.

Four posts from the corpus show the four blocks at work (Figure 7); the first two are misinformation with no false sentence. In Case 1, posted 18 May 2022, the claim is never stated. The post notes that monkeypox was identified at the Statens Serum Institut in Copenhagen in 1958, in monkeys imported from Singapore for polio vaccine research, quotes a 1972 journal abstract for the first half, sets the two details side by side and stops on an ellipsis. Both facts hold: the monkeys were incidental carriers, the virus's reservoir is African rodents, and no vaccine work created the disease. The post states none of that, yet a reader is left with vaccine research having created monkeypox, a laboratory-origin claim carried by an ellipsis (Truth: *accurate*; Appeal: plain to *sensational*, with the labelers split; Belief: *misleading*; Position: *fringe*). A coder reading sentence by sentence finds nothing to flag; Allen et al. (2024) describe the same move, in which users "present true stories out of context to support misleading narratives." In Case 2, posted 22 May 2022, every sentence is true, yet the post is misinformation. It sets Public Services and Procurement Canada's April 2022 tender for 500,000 doses of smallpox vaccine, reported in a 20 May news story, beside the WHO's report of cases in 12 member states on 21 May. Both dated facts hold, and the tender was a routine stockpile for 2023 to 2028 rather than a monkeypox purchase, but set side by side, the two dates suggest that officials knew the outbreak was coming (Truth: *accurate*; Appeal: plain; Belief: *misleading*; Position: *fringe*). Matching posts to fact-checks misses it, because no sentence is false.

Cases 1 and 2 show why our definition of misinformation follows Lazer et al. (2018), for whom misinformation is "incorrect or misleading information about the state of the world." It departs from Allcott and Gentzkow (2017), whose fake news is "intentionally and verifiably false" and leaves out "reports that are slanted or misleading but not outright false" (p. 214). The difference is large in practice. On Facebook, vaccine-skeptical content that fact-checkers had not flagged had an estimated total effect on vaccination intentions 46 times that of flagged misinformation, because it was seen far more often (Allen et al. 2024). Its authors describe this content as a gray area, misleading without being factually inaccurate (Allen et al. 2024). We do not code intent, whose ground truth is hard to obtain from content (Zhou and Zafarani 2020). The labels measure the content of a post; the producer's intent and the harm a post does are separate objects.

The last two posts in Figure 7 are not misinformation. Case 3, posted 23 June 2022, is an accurate post packaged as alarm: it frames an accurate chart of the record seven-day case average as a clear and present danger and an emergency (Truth: *accurate*; Appeal: *sensational*; Belief: not misleading; Position: *mainstream*, not slanted). Appeal moves; truth does not. Coding alarm as appeal rather than as a misleading takeaway is our rule. Case 4, posted 23 May 2022, takes a partisan side, not a falsehood: it jabs at a politician and her poll, taking an explicit partisan side while stating no checkable fact and alleging no plot (Truth: *no checkable fact*; Appeal: plain; Belief: not misleading; Position: *mainstream*, *slanted*). Slant is a side, not a falsehood: here we follow Allcott and Gentzkow (2017), who keep slanted reports that are not false out of fake news, and Gentzkow and Shapiro (2010), in which slant is a horizontal taste, a reader's distaste for slant away from the reader's own.

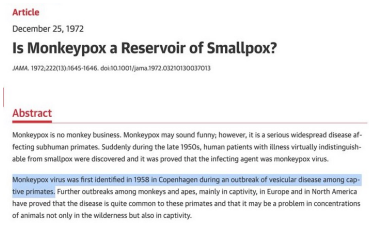
Every judgment uses only what was knowable on the posting date, and the labelers work from a dated record. For the claims that recur in the cycle, the rules carry 19 fact boxes of dated facts, among them the WHO's case counts by date, airborne spread, a laboratory origin, vaccine safety and each country's first case. A later development does not count: a case confirmed weeks after a post does not make the post's earlier claim of it accurate. A report dated a day or two after a post counts only for what the situation was on the posting date. The posting-date rule changes verdicts. In late May a post reporting a suspected case in Ghana as suspected is accurate, while a post announcing it as confirmed when the test was still pending is premature. A post of 19 May announcing a case in Chile, when officials reported no suspected case, is false. Judged after Ghana and Chile confirmed their first cases in June, both May reports of a first case would look right. For a claim outside the fact boxes, the labeler gives its own judgment as of the posting date and flags the claim for verification. At least one labeler flagged such a claim on 1,105 of the 1,591 viral posts that are not N/A.

Case 1: the claim is never stated

Posted 18 May 2022

"The 'monkeypox virus' was "first identified in 1958 in Copenhagen during an outbreak of vesicular disease among captive primates."

It's interesting to note that the monkeys in which this happened, were being used for 'polio vaccine research'. . . "


Article

December 25, 1972

Is Monkeypox a Reservoir of Smallpox?

JAMA. 1972;222(13):1645-1646. doi:10.1001/jama.1972.03210130037013

Abstract

Monkeypox is no monkey business. Monkeypox may sound funny; however, it is a serious widespread disease affecting subhuman primates. Suddenly during the late 1950s, human patients with illness virtually indistinguishable from smallpox were discovered and it was proved that the infecting agent was monkeypox virus.

Monkeypox virus was first identified in 1958 in Copenhagen during an outbreak of vesicular disease among captive primates. Further outbreaks among monkeys and apes, mainly in captivity, in Europe and in North America have proved that the disease is quite common to these primates and that it may be a problem in concentrations of animals not only in the wilderness but also in captivity.

Its image: the 1972 abstract it quotes.

1. Truth: accurate. Monkeypox was identified in Copenhagen in 1958, in monkeys imported for polio vaccine research.
2. Appeal: plain to sensational (the labelers split).
3. Belief: misleading. The reader is left with vaccine research having created monkeypox.
4. Position: fringe. A laboratory-origin story.

Misinformation carried by an ellipsis: the post never states the claim.

Case 2: every sentence true, yet misinformation

Posted 22 May 2022

"The Government of Canada tendered a contract for PHAC to acquire 500,000 doses of smallpox vaccine on April 21, 2022.

The WHO reported on May 21, 2022 that since May 13th, cases of monkeypox have been reported from 12 Member States.

Still think they are being upfront with us?"


CP24

Canada considering smallpox vaccine for monkeypox cases, says Dr. Theresa Tam

Published May 20, 2022 3:04 p.m. ET
Updated May 20, 2022 8:13 p.m. ET

Laura Osman
The Canadian Press

Share this story:

Its image: a news story from 20 May.

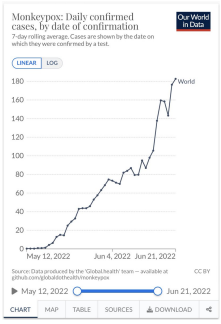
1. Truth: accurate. Canada did tender for 500,000 doses of smallpox vaccine in April 2022, and the WHO did report 12 member states on 21 May.
2. Appeal: plain.
3. Belief: misleading. Set side by side, the two dates suggest officials knew the outbreak was coming.
4. Position: fringe. Foreknowledge is a plot claim.

Misinformation, with no false sentence: matching posts to fact-checks misses it.

Case 3: an accurate post, packaged as alarm

Posted 23 June 2022

"Monkeypox – another world record of daily cases on 7-day average. It is now a clear and present danger [. . .] This is an emergency."



1. Truth: accurate. The chart is the record seven-day case average the post reports.
2. Appeal: sensational. The words around the chart turn the record into an emergency.
3. Belief: not misleading. A reader takes away nothing untrue.
4. Position: mainstream, not slanted. No plot, no partisan side.

Not misinformation. Appeal moves; truth does not.

Case 4: a partisan side, not a falsehood

Posted 23 May 2022

"Monkeypox is tailor made for Marge and other MAGA conspiracy theorists. [. . .] Speaking of crazy. . . "



1. Truth: no checkable fact. A jab at a politician and her poll.
2. Appeal: plain.
3. Belief: not misleading. No false takeaway.
4. Position: mainstream, slanted. An explicit partisan side; no plot alleged.

Not misinformation. Slant is a side, not a falsehood.

Figure 7: Four posts from the corpus labeled on the four blocks. Each panel shows a post's posting date, its text and image, its labels on Truth, Appeal, Belief and Position (Table 3), and whether it is misinformation.

The flagged claims were not checked further, so these posts keep the labelers' verdicts. Most of the labelers flagged one on 929 of these posts, 81 of them false or premature on the flagged claim. Without the false or premature posts whose claim most labelers flagged, misleading posts in the mutual, overlapping cell reach +1.16 log points more than the same producer's accurate posts (SE 0.63), against +1.30 with them (Table 9, Panel D). The 21 false or premature posts left in that cell reach +2.26 log points more (SE 0.82), against +0.85.

We built the labeling rules human first, then tuned language models to them, to address two risks of measuring with language models: that a model draws on what it learned after the date it is asked about (Ludwig et al. 2025), and that models make correlated errors when a rule is ambiguous (Dell and Rambachan 2026). First, a human rater labeled a development set of 34 posts against a dated record of what was known, and what was still open, on each day of the outbreak; the rater's labels were the answer key. Second, we drafted the written rules from the literature. Third, three language models labeled the same posts by the rules, blind to the key and to one another, under instructions to use only the post and the dated record, not their own knowledge, for every claim the record covers. Fourth, wherever a model differed from the key, or the models differed among themselves, the rater judged which reading was right and which rule had allowed the other. Fifth, we returned to the rules and sharpened that clause with invented examples: from the fourth blind run on, examples in the rules are invented, and no copy of the rules given to a model shares four consecutive words with a tested post. Sixth, we tested the final rules on a holdout of 50 viral posts labeled blind by the rater.

Three language models from different families label the posts: Claude Opus 5.5, OpenAI's gpt-6.1-sol and Google's Gemini (gemini-4-argon-high). All three label every viral post, with the same instructions. Each labeler receives the rules and a batch of posts, and nothing else: no other labeler's answers and no earlier labels. It returns every field with the claim it judged, the untrue claim when a post is misleading, a confidence score and a short rationale. Each batch's labels are stored with the model, its reasoning effort and the run date, and its prompt can be rebuilt from the code that wrote it, with two exceptions among the posts below the viral cut. The code that wrote the prompts of the second labeling run is not on file, and Gemini's labels of 147 posts in the first run carry no record of model version, settings or date (Appendix D.3). The labelers read each post's text. For a viral post they also see its images, a video or an animated image by its first frame, and the producer's handle. For the 18 viral posts that quote a post still available in 2026, they also read the quoted post's text and see its images. The posts below the viral cut are labeled from their text alone, without images, quoted text or the producer's handle. A post whose verdict depends on content the labelers were not given, such as an image or video, is N/A and receives no verdict; 139 of the 1,730 viral posts are N/A. An N/A post is left out of every estimate. Misinformation is the majority verdict of a post's labelers. On a post below the viral cut, where Claude Opus 5.5 and gpt-6.1-sol split, Gemini's label decides (Appendix D.3). Of the 1,079 posts below the cut, 513 carry only gpt-6.1-sol's label, made at low reasoning effort, and on the 13 where gpt-6.1-sol and Gemini split, which have no third label, gpt-6.1-sol's labels stand. A post's category is the one most labelers in that majority chose, ties going to Claude Opus 5.5, then gpt-6.1-sol, then Gemini, and its other labels, such as sensational, are those of the first labeler in that order who chose the category. The amplifier tests of Section 5.4 instead average each block over the post's labelers (Appendix C.5).

We test how consistently the rules are applied, on the development set and on posts the rules had not seen. Both tests used version 2.9 of the rules, which preceded the version that made the labels (Table 15). Development alignment, the share of settled answers that all three labelers matched, was 133 of 157 (85%) in the last of four blind runs on the development set. On new posts, all three agreed on 260 of 280 answers for 40 posts drawn at random (93%), and on 236 of 280 answers for 40 posts chosen to be hard (84%).[9] On the paper's own labels, made under version 3.1, all three labelers give the same answer on 10,103 of the 11,137 answers on the 1,591 viral posts that are not N/A (90.7%). They agree on the misinformation verdict for 1,422 of these posts (89.4%; Appendix D.4). The benchmark is the human rater. The rater labeled the holdout of 50 viral posts blind, under version 3.1 of the rules and with the dated record. The posts were drawn at random from the viral posts that are not N/A and that no development set contains. The labelers' majority matches the rater on 84% of answers. Agreement among the models shows that the rules are applied consistently, not that

[9]Each post has seven answers: accuracy, misleading, sensational, humorous, slanted, fringe or mainstream, and misinformation. In the development runs the labelers were Claude Opus 5.5, Claude Sonnet 5.5 and Claude Fable 5.1; in the tests on new posts they were Claude Opus 5.5, Claude Fable 5.1 and gpt-6-astra.

the labels are right: language models' errors are correlated, more so among more capable models (Dell and Rambachan 2026). Nor does agreement, among the models or with the rater, measure what readers believed. Estimates that use the labels can be corrected with a hand-labeled sample drawn with known probabilities, and the corrected estimates are consistent however biased the model labels are (Egami et al. 2023). The model labels then amplify a small validation sample rather than replace it (Ludwig et al. 2025). No estimate in Section 5 applies this correction.

Two features of the labels carry into the tests of Section 5. Appeal is measured apart from truth because in the model appeal raises sharing even at zero credibility (Section 2). Misinformation splits by what a reader needs in order to catch it. A misleading post's steering is visible in its wording, while telling a false or premature post apart requires checking outside sources. The demand tests compare the reach of each kind with that of the same producer's accurate posts.

5 Empirics

This section reports four sets of estimates, each by network cell. Section 5.1 asks what the posts of each cell are amplified for. Section 5.2 asks how a post's reach changes, within producer, when the post is misleading or false. Section 5.3 asks how accurate the producers who post while checking is costly are, as a pool, compared with the producers of the same cell who join after. Section 5.4 asks who amplifies misleading and false posts. Readers pool on the cell together with the producer's public signal, such as its follower count, profile and account age (Section 2). The tests hold that signal fixed, through controls for the producer's profile (Sections 5.1 and 5.3) or by comparing posts of the same producer (Section 5.2). Each subsection opens with the implication of the model it tests (Table 1) and reports its estimates in a figure; Appendix C.1 gives every coefficient in tables.

The producer tests use the 1,995 labeled posts, 1,037 of them viral, by 402 producers (Section 3). Of these producers, 240 have at least three followers in the following lists we collected and so fall in a network cell: 60 in G_1, 48 in G_2, 55 in G_3 and 77 in G_4. The posts of the other 162 producers enter the reward regression of Section 5.3 and the amplifiers' accuracy in Section 5.4, but no estimate for a cell. Of the labeled posts, 1,829 state a checkable fact, and the producers' accuracy below is computed on these posts. A further 186 labeled posts are N/A, because their verdict depends on content the labelers could not see; they are left out of these counts and of every estimate (Section 4). The labeled posts combine viral posts with 958 posts below the viral cut by the same producers. Of these, 577 are among the 660 posts drawn with recorded probabilities and 381 among the 419 added in a second labeling run whose selection was not recorded (Appendix D.1). No estimate reweights the posts. A post's reach is $\log(1+\mathrm{RT}+\mathrm{QT})$, where RT and QT are its retweets and quote tweets. For viral posts with an image or video the counts are those of the 2022 window; for text-only viral posts, the posts found in 2026 and for posts below the viral cut they are lifetime counts read in 2026.

5.1 What each cell amplifies

In the model, a cell's producer pool, its reader mix and the equilibrium it is in together set what its posts are amplified for (Figures 4 and 5). The first question is whether the posts that collect amplification differ across cells. The population is the producers in a cell whose first post on the outbreak falls in the Breaking Window, with all their posts from W1 to W4: 1,117 posts by 139 producers. For each cell G and category n we compute the share of the cell's amplification that goes to posts of that category:

$$s_{nG} = \frac{\sum_{i\in\mathscr{P}_G} w_i \mathbf{1}\{n_i = n\}}{\sum_{i\in\mathscr{P}_G} w_i}, \qquad w_i = \min\left(\mathrm{RT}_i + \mathrm{QT}_i,\ \bar{w}\right) + 1, \tag{18}$$

where $\mathscr{P}_G$ is the set of these posts whose producer is in cell G, and n_i is the category of post i: accurate, accurate but misleading, false or premature, or no checkable fact. The weight w_i counts the post's retweets and quote tweets, capped at $\bar{w} = 7{,}191$, the 99th percentile over the checkable labeled posts, plus one. A slanted post that is accurate counts as accurate. The accurate but misleading category also holds the 23 misleading posts with no checkable fact. Unlike the model's credibility c, under which a post is accurate with probability $(1+c)/2$, at least one half (Section 2.2), the accurate share weights posts by their amplification

and counts posts with no checkable fact and accurate but misleading posts as not accurate, so it compares cells and is not a measure of c. To attach a standard error to the gap between cells and to hold a producer's audience size fixed, we also estimate, on the same posts and weights,

$$\mathbf{1}\{n_i = \text{accurate}\} = a + \sum_{G \neq G_1} b_G D_{iG} + f \log_{10} \text{Fol}_{j(i)} + e_i, \tag{19}$$

where D_{iG} indicates that the producer $j(i)$ of post i is in cell G, Fol_j is producer j's follower count on Twitter, read in 2026, and e_i is the error. Without the follower term, b_G equals s_{nG} minus its value in G_1 for the accurate category; a is the G_1 share and f the slope on follower count. Standard errors are clustered by producer, because a producer's posts share its audience. The cells are not assigned, so b_G describes how the cells' equilibria differ, not what would happen to a producer moved from one cell to another.

The cell predicts what is amplified (Figure 8; Appendix Table 5, Panel A). In G_1, 94% of amplification goes to accurate posts; in G_4, 32% does. The rest of G_4's amplification goes to accurate but misleading posts (29%), false or premature posts (32%) and posts with no checkable fact (6%). The two middle cells lie between: 79% in G_2 and 59% in G_3. By equation (19), the accurate share of amplification in G_4 differs from that in G_1 by -61.3 percentage points (SE 8.3; Appendix Table 5, Panel C). An aggregate share can hide differences of this size: during the 2016 U.S. election, sources of fake news, classified by publisher, supplied 6.7% of the political URLs that a panel of Twitter users shared, and 0.1% of the panel accounted for 79.8% of these shares (Grinberg et al. 2019).

Robustness. Holding the producer's follower count fixed in equation (19) narrows the difference between G_4 and G_1 to -47.8 percentage points (SE 10.4; Appendix Table 5, Panel C), while larger producers' amplification is more accurate, by 12.0 percentage points per tenfold rise in followers (SE 4.2). At the median follower count of these posts, about 460,000, the predicted accurate share is 86% in G_1 and 39% in G_4 (Appendix Table 10, Panel A). News outlets are concentrated in the one-way cells, and G_1's share does not rest on them: it is 92.3% among the other producers and 94.3% among outlets. G_2's share does rest on outlets, which carry 72% of its amplification: it is 86.3% among outlets and 61.3% among the other producers (Appendix Table 10, Panel B).

5.2 Reach within producer: misleading and false posts

In the model, readers cannot see whether a post is accurate, so a producer's false and accurate posts with the same public record earn the same expected amplification, while a misleading post, whose steering readers can see, can earn more or less (Table 1, implication 1). The second question is whether, holding the producer fixed, an accurate but misleading post, or a false or premature one, reaches more or less than the producer's accurate posts. The two categories differ in what a reader can see: misleading wording is in the post itself, while telling a false or premature claim from an accurate one requires sources outside the post (Section 4). The population is the checkable labeled posts, from W1 to W4, by producers in a cell with at least two checkable posts: 1,193 posts by 140 producers. We estimate one regression per cell,

$$R_{jt} = \mu_j + \lambda_{d(t)} + \delta_G^M \text{Misl}_{jt} + \delta_G^F \text{False}_{jt} + u_{jt}, \tag{20}$$

where R_{jt} is the reach of post t by producer j, μ_j is a producer fixed effect, $\lambda_{d(t)}$ is a fixed effect for the calendar date $d(t)$ of the post, Misl_{jt} indicates an accurate but misleading post, False_{jt} indicates a false or premature post, and u_{jt} is the error. The omitted category is the producer's accurate posts. The producer effect absorbs everything fixed about the producer, including its audience and its average reach; the date effect absorbs the reach common to all posts of a date as attention rises and falls over the cycle. So δ_G^M is the reach of a producer's misleading posts relative to its accurate posts, net of the date, and δ_G^F is the same for its false or premature posts. The two effects are additive, so the comparison is within producer and within date, not within producer and date jointly. Standard errors are clustered by producer. Producers choose what to write, and misleading posts differ from accurate posts in more than their takeaway: among all labeled posts, 84% of misleading posts are slanted or fringe, against 3% of accurate posts, and 57% are sensational, against 15%. The coefficient δ_G^M therefore measures the reach of misleading posts as they were written, not of the

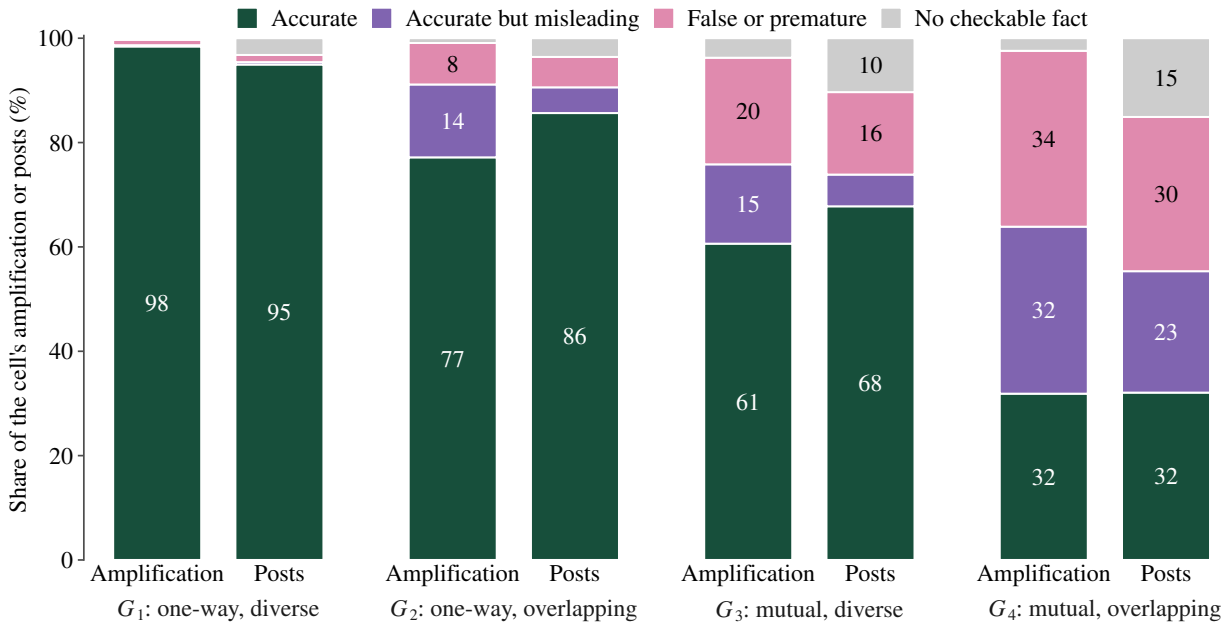


Figure 8: What each network cell posts and what it amplifies. For each cell, the share of amplification (left bar) and the share of posts (right bar) that go to accurate posts, accurate but misleading posts, false or premature posts, and posts with no checkable fact, over the W1 to W4 posts of the producers in a cell whose first post on the outbreak falls in the Breaking Window: 1,117 posts by 139 producers, with posts labeled N/A left out. Amplification is weighted as in equation (18); posts are weighted equally. Labels give the larger shares in whole percent. The shares are descriptive, so no interval is drawn; Appendix Table 5 gives the shares and the gaps between cells, with standard errors clustered by producer. Source: as in Appendix Table 5.

misleading takeaway alone.

In the mutual overlapping cell, misleading posts reach more than the same producer's accurate posts (Figure 9; Appendix Table 6). In G_4, by equation (20), the reach of a misleading post differs from that of the same producer's accurate posts by $+1.30$ (SE 0.63, $p = 0.044$), a difference of $+267\%$ in one plus its retweets and quote tweets (SE 230). In G_1 the coefficient is -0.63 (SE 0.24, $p = 0.012$) and rests on one misleading post. In G_2 and G_3 the coefficients are $+0.22$ (SE 0.61) and $+0.58$ (SE 0.48), not distinguishable from zero. In G_4 false or premature posts also reach more than the same producer's accurate posts: the coefficient is $+0.85$ (SE 0.50, $p = 0.097$), significant at the 10 percent level. In the other cells it is -0.13 in G_1 (SE 0.53), $+0.28$ in G_2 (SE 0.40) and -0.26 in G_3 (SE 0.36). Their 95 percent confidence intervals run from -1.20 to $+0.93$ in G_1, -0.54 to $+1.11$ in G_2, -1.01 to $+0.50$ in G_3 and -0.16 to $+1.87$ in G_4. Across rumor cascades on Twitter that fact-checkers had classified, Vosoughi et al. (2018) find that falsehoods were 70% more likely to be retweeted than the truth, holding fixed the original tweeter's account age, activity, followers, followees and verification. With the producer itself held fixed, false or premature posts reach more than accurate ones only in G_4, and only at the 10 percent level. The coefficients δ_G^M and δ_G^F test the first implication of the model (Section 2.7).

Robustness. Viral posts and posts below the viral cut were sampled differently, so we also add an indicator for viral posts, a control for the sampling stratum. Because the viral cut is set on a post's retweets, quotes and replies, this control conditions on reach. With it, the coefficient δ_G^M of equation (20) turns positive in G_1, $+0.70$ (SE 0.38, $p = 0.070$), while in G_4 it stays positive and significant at the 5 percent level: $+0.88$ (SE 0.39, $p = 0.028$; Appendix Table 9, Panel B).

Adding the post's sensational and slanted labels to equation (20) holds fixed cues a reader sees in the post itself (Appendix Table 9, Panel C). In G_4, 40 of the 55 false or premature posts are labeled sensational. With

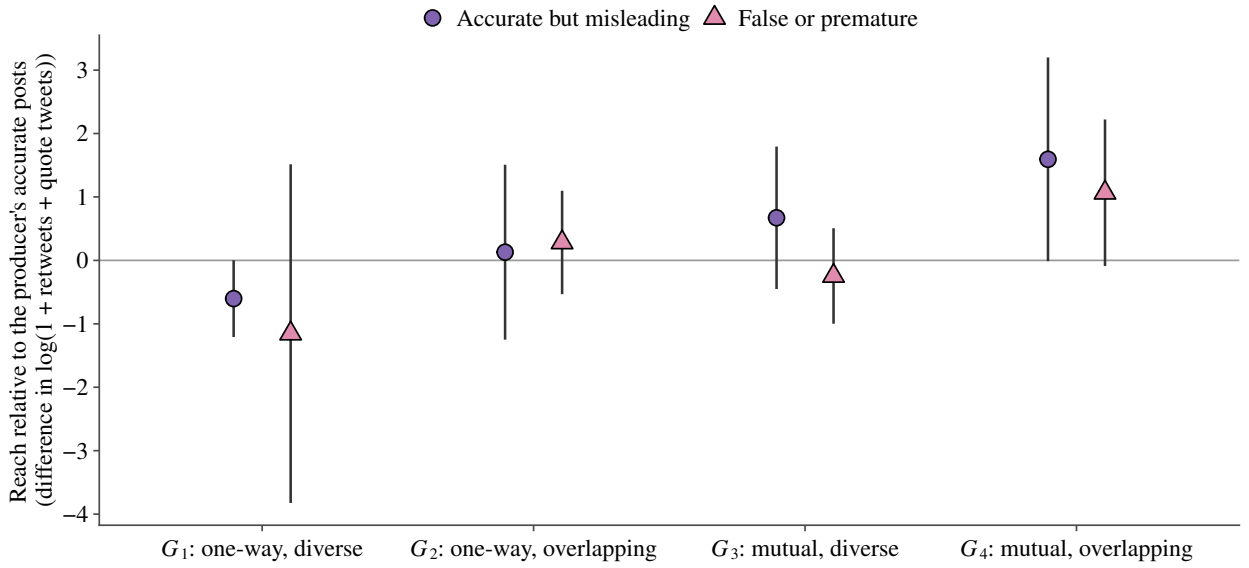


Figure 9: Reach within producer, by network cell. The coefficients δ_G^M on accurate but misleading posts and δ_G^F on false or premature posts in equation (20), relative to the same producer's accurate posts and net of the date of posting; the outcome is the log of one plus a post's retweets and quote tweets. One regression per cell, on the checkable labeled posts, W1 to W4, by producers in the cell with at least two checkable posts: 1,193 posts by 140 producers in all. The intervals are 95 percent confidence intervals from standard errors clustered by producer and the t distribution with the cell's number of producers minus one degrees of freedom. In G_1 the coefficient on misleading posts rests on one misleading post. Appendix Table 6 gives the estimates. Source: as in Appendix Table 5.

these labels added, the G_4 coefficient on false or premature posts is $+0.22$ (SE 0.68, $p = 0.745$), and the coefficient on misleading posts is $+0.75$ (SE 0.83, $p = 0.374$). Neither is distinguishable from zero once the cues are held fixed. In G_4 the reward to these posts therefore cannot be separated from their sensational and slanted packaging, which readers also see. A further measure, the co-movement of Proposition 2.6, relates a producer's reach to its cell's accurate share across the four windows (Appendix C.2); it is positive in G_1 and negative in the two mutual cells.

5.3 Who posts while checking is costly

In Figure 5(a), the pool that posts while checking is costly is less accurate than the producers who join after where selection is adverse, and more accurate where it is advantageous (Table 1, implication 2). The third question is which of the two holds in each cell. The population is the 239 producers in a cell with a checkable post, one row per producer. The indicator Breaking_j equals one if producer j's first post on the outbreak falls in W1, from 7 to 20 May 2022; 138 of the 239 producers have $\text{Breaking}_j = 1$. It is set from all of the producer's posts on the outbreak that we observe, including its posts below the viral cut and the posts found in 2026, not only from its labeled posts. We estimate one regression per cell,

$$\text{Acc}_j = a_G + \beta_G \text{Breaking}_j + \gamma_G' X_j + u_j, \tag{21}$$

where Acc_j is the share of producer j's checkable labeled posts that are accurate and not misleading, a_G is the cell's intercept, and u_j is the error. The controls X_j, with coefficients γ_G, are the $\log_{10}$ of the producer's follower count, an indicator for a missing count, an indicator for institutional accounts, and three account-age bins read from the account identifier (created before about 2011, from 2011 to 2019, and from 2020), where present in the cell. Because each cell has its own regression, β_G compares producers who post in the

Breaking Window with later entrants of the same cell and the same observable profile. Standard errors are heteroskedasticity-robust (HC1). Row (2) of Appendix Table 7 measures Acc_j on the producer's W2 to W4 posts only, so that both groups are scored on posts made after the Breaking Window. It leaves out the producers with no checkable labeled post in W2 to W4, all of them Breaking-Window producers: 37 of the 138, and 16 of the 46 in G_4. The labeled posts are only part of the producers' posts. Row (3) adds the producer's predicted reward, which comes from a regression of reach on the post's accuracy over all 1,829 checkable labeled posts:

$$R_{jt} = \nu_{W(t)} + Z_j'\pi_0 + \left(\pi_c + Z_j'\pi_1\right) A_{jt} + e_{jt}, \tag{22}$$

where $\nu_{W(t)}$ is a window effect, A_{jt} indicates that post t is accurate and not misleading, and Z_j holds the profile controls in X_j, four bio indicators (credential, media, health institution, identity), the producer's band of F_j, and three measures of its amplifiers' composition; π_0, π_c and π_1 are coefficients and e_{jt} is the error. The fit has $R^2 = 0.34$. From it we predict, for each producer, the reach of an inaccurate post in W2, $\hat{R}_j(0)$, and the gain from accuracy, $\hat{R}_j(1) - \hat{R}_j(0)$; row (3) adds both to equation (21). Row (4) instead adds the bio indicators, follower-count bands, amplifier composition, the producer's mean reach and the viral share of its posts; the last two are outcomes of its posts, so row (4) is a sensitivity check rather than a control for demand. There is no outside shift in the cost of checking: the Breaking Window also differs from later windows in the demand for news and in who is posting. That checking was costly in the Breaking Window is therefore a maintained assumption. Under it, equation (21) compares the pool that posts while checking is costly with the producers who join once it is cheaper, the comparison of Proposition 2.7. This reading also assumes that producers of a type keep their order of costs as checking becomes cheaper. The pool is less accurate under adverse selection, so that $\beta_G < 0$, and more accurate under advantageous selection, so that $\beta_G > 0$.

The sign of β_G differs between the one-way diverse and the mutual overlapping cell (Figure 10; Appendix Table 7, row 1). By equation (21), β_G is $+0.14$ in G_1 (SE 0.06), where later entrants' mean accuracy is 0.83. It is -0.25 in G_4 (SE 0.08), where later entrants' mean accuracy is 0.72. The difference between the two cells is -0.40 (SE 0.10). In G_2 and G_3 the estimates are $+0.14$ (SE 0.14, $p = 0.293$) and -0.29 (SE 0.10), and G_2's is not distinguishable from zero. The other rows keep the sign in both cells: scoring accuracy on W2 to W4 posts gives $+0.15$ in G_1 and -0.26 in G_4 (row 2), adding the predicted reward of equation (22) gives $+0.17$ and -0.21 (row 3), and the realized-reach controls give $+0.18$ and -0.23 (row 4). Across the four cells, the signs follow reciprocity rather than overlap: β_G is positive in both one-way cells and negative in both mutual cells. In the same outbreak, clusters of tweets with erroneous or irrelevant content emerged earlier and grew faster than communication by health officials, as judged by the authors' qualitative review of each cluster (Edinger et al. 2023). Split by network cell, the pool that posts while checking is costly is less accurate than the producers who join after in G_4, and more accurate in G_1.

Robustness. Two features of the entry indicator could drive the contrast: which kinds of sources post early, and how the first post is dated (Appendix Table 11). Adding the producer's role, coded from its 2026 profile, to equation (21) lowers the G_1 coefficient to $+0.08$ (SE 0.05), moves G_4 to -0.30 (SE 0.09) and G_3 to -0.17 (SE 0.12); the difference between the cells is -0.38 (SE 0.10). A second collection, a 2026 search of each producer's 2022 posts, dates the first post of 384 labeled producers; its Breaking-Window indicator agrees with ours for 303 of them. With entry dated by that collection, the difference between G_4 and G_1 is -0.15 (SE 0.12), and neither cell's coefficient is distinguishable from zero: $+0.08$ in G_1 (SE 0.09) and -0.07 in G_4 (SE 0.09). Among the producers whose two dates agree, it is -0.38 (SE 0.13). The sign of the contrast holds in every row; its size depends on how entry is dated.

5.4 Who amplifies

In the model, who amplifies differs between the Belief Equilibrium, in which only social readers amplify, and the News Equilibrium, in which certifiers amplify too (Figure 4; Table 1, implication 4). The fourth question is whether the accounts that amplify misleading and false posts differ across cells. The amplification records are the retweets and quote tweets of the labeled posts, one record per amplifier and post. Quote tweets and replies were collected in 2026; retweets come from the 2022 engagement records and from amplifiers' timelines read in 2026. Of the 53,247 records, 18,770 are retweets in the 2022 engagement records, which cover 54 posts; 34,176 are quote tweets, on 1,210 posts; and 301 are retweets read from timelines, on 228

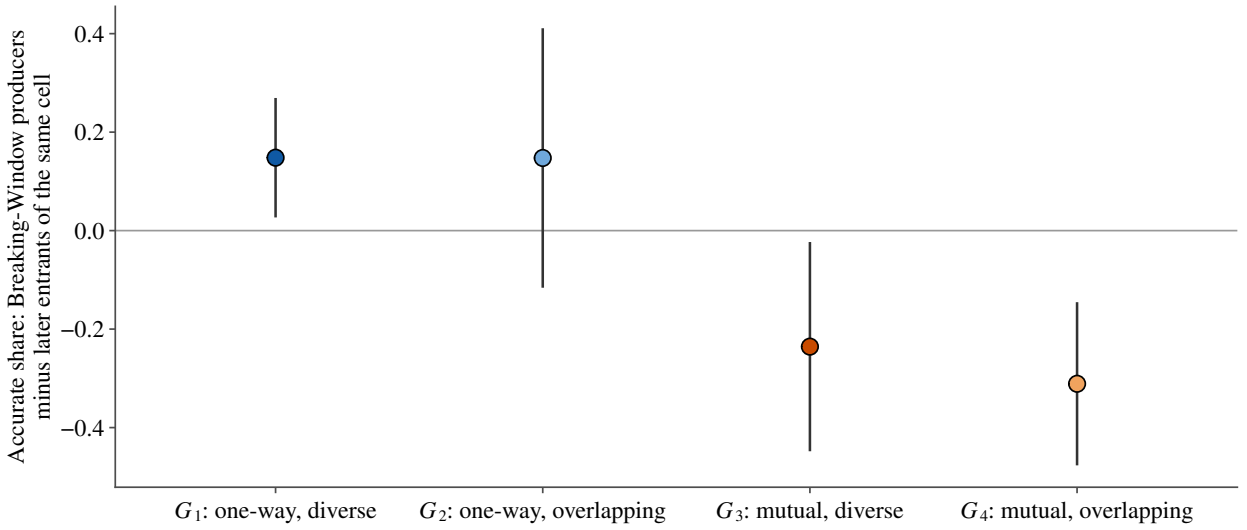


Figure 10: Who posts while checking is costly, by network cell. The coefficient β_G of equation (21): the difference in the share of checkable posts that are accurate and not misleading, on a scale from zero to one, between the producers whose first post falls in the Breaking Window and the later entrants of the same cell, holding the producer's follower count and profile fixed. One producer-level regression per cell, on the 239 producers in a cell with a checkable post, 138 of them in the Breaking Window. The intervals are 95 percent confidence intervals from heteroskedasticity-robust (HC1) standard errors and the t distribution with the cell's number of producers minus one degrees of freedom. Appendix Table 7 gives the estimates and the other rows. Source: as in Appendix Table 5.

posts. On posts by producers in a cell there are 37,618 records by 30,638 amplifiers. Each quote tweet and reply on misinformation in a cell also carries a stance label: it endorses the post, calls it out, attacks it without disputing its claim, is neutral, or is unclear. The three language models of Section 4 label each item from its text and from the text and images of the post it answers, without the post's misinformation label. The stance label is the value at least two of them choose; the 146 of the 7,822 quote tweets and replies on misinformation in the cells that have none count as unresolved. Human raters labeled 90 quote tweets and replies blind on the same codebook. The stance label matches theirs on 47, and on 26 of the 60 not used to revise the rules; of the 35 that both place as endorsing or calling out, they take opposite sides on 2. The models call items out less often than the raters do: 8 of the 90 against 13.

Most quote tweets on misinformation endorse it in every cell but G_1 (Figure 11(a); Appendix Table 8, Panel A): 85% in G_4, 79% in G_3, 73% in G_2 and 40% in G_1. G_1's lower share does not come from one post: its most-quoted post draws 240 of its 461 quote tweets, 38% of which endorse it, and on G_1's other eight posts 43% of quote tweets endorse the post. Replies follow the same order, from 59% endorsing in G_4 to 29% in G_1, and G_1's most-replied post draws 161 of G_1's 417 replies, 46% of which call it out. These shares are descriptive, and in G_1 they rest on nine posts.

To ask whether misleading and false posts draw amplifiers who are less accurate elsewhere, we take every record of an amplifier k who amplifies at least two producers and estimate, by cell,

$$\text{Acc}_{k,-j} = a_G + \phi_G^M \,\text{Misl}_{jt} + \phi_G^F \,\text{False}_{jt} + u_{kjt}, \tag{23}$$

where $\text{Acc}_{k,-j}$ is the share of accurate posts among the checkable posts that k amplifies by producers other than j, Misl_{jt} and False_{jt} describe the amplified post t by producer j as in equation (20), a_G is the mean for records on accurate posts, ϕ_G^M and ϕ_G^F are the differences from that mean for records on misleading and on false or premature posts, and u_{kjt} is the error. Leaving out producer j's posts makes the outcome a measure of the amplifier's choices elsewhere rather than of the post itself. Standard errors are clustered by producer.

Misleading posts draw amplifiers who are less accurate elsewhere in every cell (Figure 11(b); Appendix Table 8, Panel B). False or premature posts draw less accurate amplifiers only in G_4. By equation (23), the accuracy elsewhere of a misleading post's amplifiers differs by -0.30 in G_1 (SE 0.05), from a mean of 0.77 for amplifiers of accurate posts, and by -0.31 in G_4 (SE 0.05), from a mean of 0.72. In G_2 and G_3 it differs by -0.31 (SE 0.06) and -0.21 (SE 0.11). For false or premature posts the coefficient is $+0.11$ in G_1 (SE 0.05, $p = 0.027$), $+0.02$ in G_2 (SE 0.09, $p = 0.858$), -0.03 in G_3 (SE 0.06, $p = 0.614$) and -0.29 in G_4 (SE 0.08).

The last test asks the supply question of the amplifiers: are the amplifiers active while checking is costly, as a pool, more or less discerning later than the amplifiers who join after? For each amplifier k active in cell G in W2 to W4 we estimate

$$\mathrm{Acc}_k^{\mathrm{late}} = a_G + \beta_G^{\mathrm{amp}} \mathrm{Breaking}_k + \gamma_G' X_k + u_k, \tag{24}$$

where $\mathrm{Acc}_k^{\mathrm{late}}$ is the share of accurate posts among the checkable posts by producers in G that k amplifies in W2 to W4, $\mathrm{Breaking}_k$ indicates that k amplified a labeled post in W1, and u_k is the error. Measuring accuracy after the Breaking Window scores every amplifier on posts made once the facts were more settled. The controls X_k are the $\log_{10}$ of the amplifier's follower count in 2026, an indicator for a missing count, and the log of one plus its number of records in the cell in W2 to W4. Standard errors are heteroskedasticity-robust (HC1); the difference between G_4 and G_1 comes from one stacked regression with standard errors clustered by amplifier, since 126 amplifiers are active in both cells. As for producers, that checking was costly in the Breaking Window is a maintained assumption.

In the baseline, Breaking-Window amplifiers are less accurate later in G_1, G_3 and G_4, by the most in G_4, and more accurate in G_2 (Appendix Table 8, Panel C). By equation (24), β_G^{amp} is -0.03 in G_1 (SE 0.01), $+0.06$ in G_2 (SE 0.01), -0.08 in G_3 (SE 0.01) and -0.28 in G_4 (SE 0.02), where later amplifiers' mean accuracy is 0.71. The difference between G_4 and G_1 is -0.25 (SE 0.03). With the amplifier controls the G_4 coefficient is -0.29 (SE 0.02). Among the retweeters listed for each post in 2026, a smaller share carry a certifier bio in G_4 (6.0%) than in G_1 (10.4%; Appendix C.5).

Robustness. On quote tweets alone (Appendix Table 12, Panel A), the misleading coefficient of equation (23) is -0.33 in G_1 (SE 0.11) and -0.18 in G_4 (SE 0.04), and the Breaking-Window coefficient of equation (24) is -0.02 in G_1 (SE 0.01) and -0.10 in G_4 (SE 0.03).

The producers found in 2026. The tests above include the 305 posts found in 2026 by producers already in the graph frame, and leave out the other 403 and their 332 producers. Adding those producers, with every other producer kept in its cell and each new one placed by the same rule, leaves the composition gap and the amplifier test in place, while the supply estimate and the coefficient on misleading posts in G_1 are no longer distinguishable from zero (Appendix C.6). Recomputing every cell on a follow graph that adds the new producers' follows changes the cell, or the absence of one, of 101 of the 442 producers the tests start from, and also removes the amplifier result. The new producers are measured on less: 80% of those in a cell with a checkable post have a single one, against 41% of the others, whose accuracy also rests on their posts below the viral cut. Under the data appendix's other cell rules, the supply estimate and the amplifier test hold, and the negative coefficient on misleading posts in G_1 holds only with the reciprocity cut at 0.25 or at the median.

What the estimates do not identify. The cells are not assigned, so each estimate describes a cell's equilibrium, and a difference between cells can come from the producers, the readers or both. The Breaking-Window comparisons of producers and amplifiers have no outside shift in the cost of checking. The follow graph, follower counts, profiles, quote tweets, replies and retweeter lists were collected in 2026, and only what remained public then is observed. The two roles of the network in the model give two readings of a difference between cells, and each test bears on one of them. A difference in producer pools shows in the accuracy of the pool that posts while checking is costly (Section 5.3). A difference in reader mix shows in how reach responds, within a producer, to what readers can see in a post, and in who amplifies (Sections 5.2 and 5.4). A cell could also stand in for the producer's public signal, which the controls for the producer's profile and the comparisons within a producer hold fixed. The composition of amplification reflects the producer pool, the reader mix and the equilibrium together (Section 2.7), and no test above divides it among them.

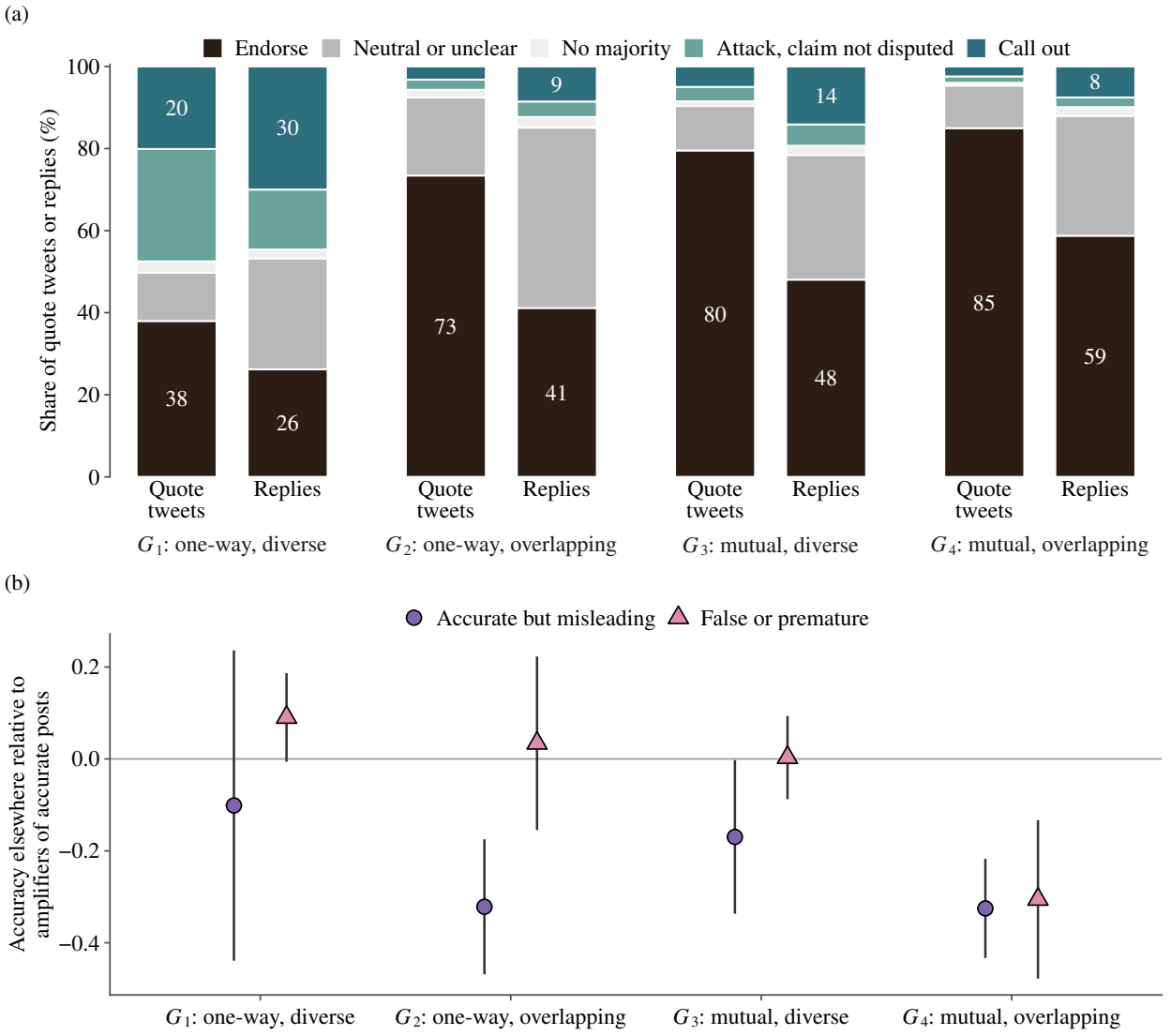


Figure 11: Who amplifies, by network cell. (a) The stance of the quote tweets (left bar) and replies (right bar) on misinformation posts by producers in each cell, 7,822 items in all: the shares that endorse the post, call it out, attack it without disputing its claim, are neutral or unclear, or have no majority label. An item's stance is the label that at least two of the three language models choose. Labels give the larger endorse and call-out shares in whole percent. (b) The coefficients ϕ_G^M and ϕ_G^F of equation (23): the difference between the accuracy, elsewhere, of the amplifiers of misleading and of false or premature posts and that of the amplifiers of accurate posts. Accuracy elsewhere is the share, from zero to one, of accurate posts among the checkable posts an amplifier amplifies by other producers; one record per retweet or quote tweet by an amplifier who amplifies at least two producers. The intervals are 95 percent confidence intervals from standard errors clustered by producer and the t distribution with the cell's number of producers minus one degrees of freedom. Appendix Table 8 gives the shares and estimates. Source: as in Appendix Table 8.

6 Related literature

Allcott and Gentzkow (2017) explain fake news partly by information asymmetry: it arises in equilibrium because it is cheaper to provide than precise signals and because consumers cannot costlessly infer accuracy.

This is the logic of the lemons market, in which "dishonest dealings tend to drive honest dealings out of the market" (Akerlof 1970). Einav and Finkelstein (2011) draw the same force with demand and cost curves: a market unravels completely when average cost lies above demand at every quantity. We write the market for posts in this form. Investigative producers pay a cost to check facts before posting that slop producers avoid, readers who cannot tell the two apart reward the pooled credibility of the producers who post, and pooling can price out the investigative producers (Section 2).

The formal models closest to ours take the set of producers as given. In Kranton and McAdams (2024), each of a fixed number of suppliers chooses whether to pay an investigation cost for a high-quality story and is paid by the story's visibility, and misinformation comes from exogenous third-party agents. In Andreottola and De Moragas (2025), outlets holding an unverified rumor publish it at once or fact-check it first, competition reduces fact-checking through the risk of being preempted (p. 1247), and fabrication is set aside (p. 1246, fn. 11). Papanastasiou (2020) assumes that fake news is generated exogenously. Entry into news is cheap, however: barriers to entry in the media industry have dropped (Allcott and Gentzkow 2017). In our model, investigative and slop producers choose whether to post, and a crowded feed lowers the visibility of each post as more producers post. When producers' costs change, the crowded feed is what lets the credibility of posts and their amplification move in opposite directions. Readers also choose whether to amplify, so selection works on production, who posts, and on amplification, who amplifies; this is double selection.

Media economics explains how readers' beliefs shape what outlets report. In Gentzkow and Shapiro (2006), a consumer judges a source to be of higher quality when its reports conform to her prior, so bias arises from firms' desire to build a reputation for accuracy, and it falls as ex post feedback about the truth becomes more likely. Mullainathan and Shleifer (2005) show that newspapers slant toward their readers' beliefs (p. 1031) by selectively omitting facts rather than by fabricating them (Section V.A and fn. 5). Across U.S. newspapers, consumer ideology explains 22.3% of the within-state variation in slant and owner ideology 3.8% (Gentzkow and Shapiro 2010). In these papers the outlets are given and choose how to slant what they report; in ours, readers reward credibility pooled across producers, so their demand also selects who posts. Our labels keep the two forms of distortion apart: a post whose words are accurate but from which a reader takes away an untrue claim is misleading, and a post that takes an explicit side is slanted (Section 4).

The structure of an audience shapes what readers see and, in theory, how they treat unreliable content. In Acemoglu et al. (2024), users who share mostly with like-minded users are less likely to be called out for spreading misinformation, so they share low-reliability content more, and an engagement-maximizing platform builds filter bubbles for such content (p. 3117). On Facebook, users' own choices and the composition of their friend networks limit exposure to cross-cutting news more than the ranking algorithm does (Bakshy et al. 2015). On Twitter, 22.1% of connected pairs follow each other (Kwak et al. 2010), and Kwak et al. (2010) read following as subscribing to a stream of tweets. Filippas et al. (2025) argue that users also follow others in order to be followed back and receive their attention, an exchange they call attention bartering. We place each producer in one of four cells by two measures of its followers in the following lists we collected, reciprocity and overlap (Section 3), and the cells are the units of our tests. In our model, reciprocity and overlap set a cell's mix of certifiers and social readers (Section 2.6).

Evidence on misinformation comes mostly from fact-checked claims and from belief surveys. Vosoughi et al. (2018) study rumor cascades on Twitter whose veracity six fact-checking organizations had classified, and find that "falsehood diffused farther and faster than the truth" (p. 1149). Altay et al. (2023) point out that this sample is contested news and leaves aside uncontested news that is extremely viral. Allcott and Gentzkow (2017) build their database of fake news from the lists of Snopes, PolitiFact and BuzzFeed, and Pennycook and Rand (2019) take the false headlines of their survey from Snopes. Fact-checks also depend on counter-evidence that may not exist: misinformation surges when demand for information outstrips the supply of credible answers, and it often builds on the absence of credible counter-evidence (Glockner et al. 2022). Content can also mislead without being false. A moderator can robustly change what receivers do only by removing information from the sender's signal (Kominers and Shapiro 2024), so moderating a misleading post built from accurate facts removes accurate information. Edinger et al. (2023) study the 2022 mpox outbreak on Twitter and find that clusters of tweets with erroneous or irrelevant content emerged earlier and grew faster than communication by health officials; the accuracy of each cluster was judged by the authors' qualitative review (p. 8). We label every viral post in our corpus, accurate or not, against what was knowable

on the day it was posted, and we keep false, premature and misleading posts apart (Section 4).

Language models make it feasible to label a large body of text, and their labels need validation. Zero-shot ChatGPT exceeded the accuracy of crowd workers by about 25 percentage points on average across four datasets (Gilardi et al. 2023), and on classification tasks language models fall short of the best fine-tuned models but reach fair agreement with human coders (Ziems et al. 2024). A measure built from text needs validation against the quantity of interest before it enters a downstream analysis (Gentzkow et al. 2019), and without a validation sample the size and pattern of a language model's errors cannot be assessed (Ludwig et al. 2025). Human judges of veracity also disagree: three professional fact-checkers all agreed on 42% of the articles they rated (Allen et al. 2021). Our protocol states explicit rules for each block, revises them against a human answer key that judges each post as of its posting date, and instructs the language models to use only the post and the record available on that date (Section 4). It codes the claim a reader takes away separately from the truth of what a post states, which is how it captures posts that are accurate but misleading.

Research on the economics of social media has mostly studied the sharing of misinformation rather than its production (Aridor et al. 2024). Experiments on misinformation intervene on readers: prompting Twitter users to think about accuracy raised the quality of the news they shared (Pennycook et al. 2021). In a survey experiment that replicated sharing on X, offering fact-checks reduced the sharing of true posts as well as false ones (Guriev et al. 2026). The empirics of selection markets supply a template for testing selection. Einav et al. (2010) trace demand and average cost with variation in prices and read adverse selection from a cost curve that rises with price. Our tests use two regimes of the cost of checking facts in place of variation in prices. In each cell, we compare the accuracy of the pool of producers that posts while checking facts is costly with that of the producers who post only once it is cheap, and we run the same comparison on the readers who amplify posts (Section 5). The two comparisons ask whether selection works on production and on amplification. That checking was costly in the first two weeks of the cycle and cheap afterward is a maintained assumption (Section 3).

7 Conclusion

This paper asked where in a social network the credibility of posts holds. In the model, readers who cannot see a post's accuracy reward it for the credibility of the producers they cannot tell apart, so credibility is set cell by cell. A cell's producer pool sets the sign of selection, and its reader mix sets how strongly the reward responds to credibility and which readers amplify. In the 2022 mpox outbreak on Twitter, cells defined by reciprocity and overlap differed on the three margins the model ties to the network: what is amplified, who posts while checking is costly, and who amplifies. These differences arose within one platform, one news event and the same weeks, so a description of credibility platform by platform misses them.

Engagement is therefore not a measure of credibility. Lazer et al. (2018) propose that source quality could be incorporated into the algorithmic rankings of content, and our evidence suggests that where a source sits in the network carries information about the accuracy of what it posts. An AI system trained on social media faces the same problem: what it learns from amplification depends on the cells its posts come from. The sign of selection also matters for policies aimed at readers. In the model, a change in readers that raises the reward to posting draws in more producers, which raises credibility where selection is adverse and lowers it where selection is advantageous (Proposition 2.6). Tools that make posts cheap to write without checking them, such as generative AI, can also add to the share of producers who post slop, and in the model a larger slop share dilutes the pool's credibility at every level of participation (equation (3)).

The evidence describes equilibria, not the effect of moving a producer from one cell to another. The cells are not assigned, so a difference between cells can come from their producers, their readers or the equilibrium each cell is in. Differences across cells alone also do not show that a cell has several equilibria (Proposition B.9). The comparisons of the Breaking Window rest on the maintained assumption that checking facts was costly in it, since no outside shift in that cost is observed, and the follow graph was collected in 2026. Three extensions would sharpen these answers: an outside shift in the cost of checking facts, a cell observed moving from one equilibrium to another over time, and a model in which the platform chooses its ranking, which here is common to every post in a cell.

References

Acemoglu, Daron et al. (2024). "A Model of Online Misinformation". In: *Review of Economic Studies* 91.6, pp. 3117–3150.

Akerlof, George A. (1970). "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism". In: *Quarterly Journal of Economics* 84.3, pp. 488–500.

Allcott, Hunt and Matthew Gentzkow (2017). "Social Media and Fake News in the 2016 Election". In: *Journal of Economic Perspectives* 31.2, pp. 211–236.

Allen, Jennifer et al. (2021). "Scaling Up Fact-Checking Using the Wisdom of Crowds". In: *Science Advances* 7.36, eabf4393.

Allen, Jennifer et al. (2024). "Quantifying the Impact of Misinformation and Vaccine-Skeptical Content on Facebook". In: *Science* 384.6699, eadk3451. Page numbers refer to the working paper of 20 May 2024.

Altay, Sacha et al. (2023). "Misinformation on Misinformation: Conceptual and Methodological Challenges". In: *Social Media + Society* 9.1, p. 20563051221150412.

Andreottola, Giovanni and Antoni-Italo De Moragas (2025). "Breaking News: A Model of Media Competition, Fact-Checking, and Political Accountability". In: *The Journal of Politics* 87.4, pp. 1243–1254.

Aridor, Guy et al. (2024). "The Economics of Social Media". In: *Journal of Economic Literature* 62.4, pp. 1422–1474.

Bakshy, Eytan et al. (2015). "Exposure to Ideologically Diverse News and Opinion on Facebook". In: *Science* 348.6239, pp. 1130–1132.

Ballester, Coralio et al. (2006). "Who's Who in Networks. Wanted: The Key Player". In: *Econometrica* 74.5, pp. 1403–1417.

Beknazar-Yuzbashev, George et al. (Feb. 2026). *Social Media Advertising Loads as Prices*. Warwick Economics Research Paper 1602. University of Warwick.

Berger, Jonah and Katherine L. Milkman (2012). "What Makes Online Content Viral?" In: *Journal of Marketing Research* 49.2, pp. 192–205. Page numbers refer to the ahead-of-print version, pp. 1–14.

Biyani, Prakhar et al. (2016). ""8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pp. 94–100.

Borgatti, Stephen P. (1997). "Structural Holes: Unpacking Burt's Redundancy Measures". In: *Connections* 20.1, pp. 35–38.

Brady, William J. et al. (2017). "Emotion Shapes the Diffusion of Moralized Content in Social Networks". In: *Proceedings of the National Academy of Sciences* 114.28, pp. 7313–7318.

Centers for Disease Control and Prevention (May 18, 2022a). *CDC and Health Partners Responding to Monkeypox Case in the U.S.* Centers for Disease Control and Prevention. URL: `https://www.cdc.gov/media/releases/2022/s0518-monkeypox-case.html` (visited on 11/01/2026).

— (May 20, 2022b). *Monkeypox Virus Infection in the United States and Other Non-Endemic Countries—2022*. Centers for Disease Control and Prevention. URL: `https://emergency.cdc.gov/han/2022/han00466.asp` (visited on 11/01/2026).

Dell, Melissa and Ashesh Rambachan (Aug. 2026). *The Measurement Revolution? Credible Measurement and Inference in the Age of AI*. Preprint 2608.23524v1. arXiv.

Edinger, Andy et al. (2023). "Misinformation and Public Health Messaging in the Early Stages of the Mpox Outbreak: Mapping the Twitter Narrative With Deep Learning". In: *Journal of Medical Internet Research* 25, e43841.

Egami, Naoki et al. (2023). "Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., pp. 68589–68601. Page numbers refer to arXiv:2306.04746v3.

Einav, Liran and Amy Finkelstein (2011). "Selection in Insurance Markets: Theory and Empirics in Pictures". In: *Journal of Economic Perspectives* 25.1, pp. 115–138.

Einav, Liran et al. (2010). "Estimating Welfare in Insurance Markets Using Variation in Prices". In: *Quarterly Journal of Economics* 125.3, pp. 877–921.

Filippas, Apostolos et al. (2025). "The Production and Consumption of Social Media". In: *Management Science*.
Gentzkow, Matthew and Jesse M. Shapiro (2006). "Media Bias and Reputation". In: *Journal of Political Economy* 114.2, pp. 280–316.
— (2010). "What Drives Media Slant? Evidence from U.S. Daily Newspapers". In: *Econometrica* 78.1, pp. 35–71.
Gentzkow, Matthew et al. (2019). "Text as Data". In: *Journal of Economic Literature* 57.3, pp. 535–574. Page numbers refer to NBER Working Paper 23276 (2017).
Gilardi, Fabrizio et al. (2023). "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks". In: *Proceedings of the National Academy of Sciences* 120.30, e2305016120. Page numbers refer to arXiv:2303.15056v2.
Glockner, Max et al. (2022). "Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5916–5936.
Grinberg, Nir et al. (2019). "Fake News on Twitter during the 2016 U.S. Presidential Election". In: *Science* 363.6425, pp. 374–378.
Guriev, Sergei et al. (2026). "Curtailing False News, Amplifying Truth". In: *Econometrica*.
Kominers, Scott Duke and Jesse M. Shapiro (2024). *Robust Content Moderation: Theory and Applications*. Working Paper 32156. National Bureau of Economic Research.
Kranton, Rachel and David McAdams (2024). "Social Connectedness and Information Markets". In: *American Economic Journal: Microeconomics* 16.1, pp. 33–62.
Kwak, Haewoon et al. (2010). "What is Twitter, a Social Network or a News Media?" In: *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. Raleigh, NC, USA: ACM, pp. 591–600.
Lazer, David M. J. et al. (2018). "The Science of Fake News". In: *Science* 359.6380, pp. 1094–1096. Page numbers refer to the accepted manuscript.
Ludwig, Jens et al. (Dec. 2025). *Large Language Models: An Applied Econometric Framework*. NBER Working Paper 33344. National Bureau of Economic Research.
Moldoveanu, Mihnea C. and Joel A. C. Baum (2026). *Adversarial Social Epistemology for Assemblies of Humans and Large Language Models*.
Movahedi Nia, Zahra et al. (2023). "Mpox Panic, Infodemic, and Stigmatization of the 2SLGBTQIAP+ Community: Geospatial Analysis, Topic Modeling, and Sentiment Analysis of a Large, Multilingual Social Media Database". In: *Journal of Medical Internet Research* 25, e45108.
Mullainathan, Sendhil and Andrei Shleifer (2005). "The Market for News". In: *American Economic Review* 95.4, pp. 1031–1053.
Papanastasiou, Yiangos (2020). "Fake News Propagation and Detection: A Sequential Model". In: *Management Science* 66.5, pp. 1826–1846.
Pennycook, Gordon and David G. Rand (2019). "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning". In: *Cognition* 188, pp. 39–50.
Pennycook, Gordon et al. (2021). "Shifting Attention to Accuracy Can Reduce Misinformation Online". In: *Nature* 592, pp. 590–595.
Vosoughi, Soroush et al. (2018). "The Spread of True and False News Online". In: *Science* 359.6380, pp. 1146–1151.
World Health Organization (May 21, 2022a). *Multi-Country Monkeypox Outbreak in Non-Endemic Countries*. World Health Organization. URL: https://www.who.int/emergencies/disease-outbreak-news/item/2022-DON385 (visited on 11/01/2026).
— (July 23, 2022b). *Second Meeting of the International Health Regulations (2005) (IHR) Emergency Committee Regarding the Multi-Country Outbreak of Monkeypox*. World Health Organization. URL: https://www.who.int/news/item/23-07-2022-second-meeting-of-the-international-health-regulations-%282005%29-%28ihr%29-emergency-committee-regarding-the-multi-country-outbreak-of-monkeypox (visited on 10/23/2026).
— (Nov. 28, 2022c). *WHO Recommends New Name for Monkeypox Disease*. World Health Organization. URL: https://www.who.int/news/item/28-11-2022-who-recommends-new-name-for-monkeypox-disease (visited on 11/01/2026).

Zhou, Xinyi and Reza Zafarani (2020). “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities”. In: *ACM Computing Surveys* 53.5, 109:1–109:40. Page numbers refer to arXiv:1812.00315v2.

Ziems, Caleb et al. (2024). “Can Large Language Models Transform Computational Social Science?” In: *Computational Linguistics* 50.1, pp. 237–291. Page numbers refer to arXiv:2305.03514v3.

Zubiaga, Arkaitz et al. (2016). “Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads”. In: *PLOS ONE* 11.3, e0150989.

A Model appendix

A.1 Microfoundations

The reward in equation (5) follows from readers who decide whether to amplify after reading.

Lemma A.1 (Affine reward) *Let a cell's readers have unit mass, and let each read a given post with probability $\tau(Q)$. After reading, a reader draws an amplification cost z, uniform on $[0,1]$ and independent across posts and of producer types, and amplifies if $z \le B^A(c)$, where*

$$B^A(c) = \bar{a} + \frac{1+c}{2}\,u^+ + \frac{1-c}{2}\,u^-$$

is her expected payoff from amplifying a post from a pool with credibility c, which by Assumption 5 does not depend on the other posts she reads. Let the state have a balanced prior, so that a post's report is correct with probability $(1+c)/2$ whichever direction it takes. If $0 \le B^A(c) \le 1$, a post's expected amplification is $\tau(Q)\,(x_0 + x_1 c)$, with $x_0 = \bar{a} + (u^+ + u^-)/2$ and $x_1 = (u^+ - u^-)/2$. With one unit of payoff per amplification, this is the reward $R(Q,c)$.

Proof. A post drawn from the pool comes from a producer of accuracy α with the pool's distribution, and it is correct with probability $(1+\alpha)/2$, so it is correct with probability $(1+c)/2$. With a balanced prior, a report of either direction has probability one half for every α, so the report's direction carries no information about the producer's type, and the correctness probability given the direction is also $(1+c)/2$. The reader's expected payoff from amplifying is therefore $B^A(c)$, and she amplifies with probability $\Pr(z \le B^A(c)) = B^A(c)$ on $[0,1]$. A post's expected amplification is $\tau(Q)\,B^A(c)$. Collecting terms, $B^A(c) = \bar{a} + (u^+ + u^-)/2 + c\,(u^+ - u^-)/2$. □

Every producer in a cell faces the same reward when amplification does not depend on the direction of a post's report or the prior is balanced.

Lemma A.2 (Common reward) *Suppose producer j decides to post before her report is drawn, the state is one with probability b, and her report is correct with probability $(1+\alpha_j)/2$ in either state. If a post that reports state one earns R^+ and a post that reports state zero earns R^-, producer j's expected reward is*

$$\frac{R^+ + R^-}{2} + \left(b - \frac{1}{2}\right)\left(R^+ - R^-\right)\alpha_j.$$

It is the same for every type if and only if $b = 1/2$ or $R^+ = R^-$.

Proof. Producer j reports state one with probability $b\,(1+\alpha_j)/2 + (1-b)(1-\alpha_j)/2 = 1/2 + (b-1/2)\,\alpha_j$. Her expected reward is this probability times R^+ plus its complement times R^-, which rearranges to the display. The coefficient on α_j is zero if and only if $b = 1/2$ or $R^+ = R^-$. □

The state's prior is one half in the baseline and with the readers of Assumption 7, whose certifiers hold a prior b of their own. By Lemma A.2, every producer in a cell therefore faces the same reward, and producers enter in the order of their costs.

Feed visibility that falls linearly in participation arises from one reading slot.

Lemma A.3 (Single-slot visibility) *Suppose a reader has one reading slot and values reading any post at $J \in [0,1]$. Reading costs are independent and uniform on $[0,1]$, and the reader reads the post with the lowest cost among those whose cost is below J. If one rival post is present with probability Q, the focal post is read with probability $\tau(Q) = J - QJ^2/2$, so $\tau_0 = J$ and $\tau_1 = J^2/2$.*

Proof. Without the rival, the focal post is read if its cost is below J, with probability J. With the rival, the focal post is read if its cost k is below J and the rival's cost is above k, with probability $\int_0^J (1-k)\,dk = J - J^2/2$. Averaging with weights $1-Q$ and Q gives $J - QJ^2/2$. □

A.2 Proofs for Section 2

Equations (4), (8), (9), (10) and (11) follow by direct calculation. Differentiating $Q\,\mathrm{AA}(Q) = \int_0^Q \mathrm{MA}(u)\,du$ gives $\mathrm{AA} + Q\,\mathrm{AA}' = \mathrm{MA}$, which is the first equality in equation (4). In the baseline, $\mathrm{MA} - \mathrm{AA} = q_0/Q$ on $Q > q_0$, which gives the second equality and equation (8). The product rule gives $\mathrm{AR}' = \tau x_1\,\mathrm{AA}' + \tau'(x_0 + x_1\mathrm{AA})$ with $\tau' = -\tau_1$, which is equation (9). Substituting SA into $\tau x_1(\mathrm{AA} - \mathrm{SA})$ gives $\tau(x_0 + x_1\mathrm{AA}) - \mathrm{MC}$, which is equation (10). In the specialist variant, $Q\,\mathrm{AA}(Q) = q_1 + \alpha_g(Q - q_1)$ on $Q > q_1$, which gives equation (11).

Most proofs rest on one property of the gap between reward and marginal cost on one type's segment of the cost order. The reward on a segment $[\underline{Q}, 1]$, $0 < \underline{Q} < 1$, has the form

$$A(Q) = \tau(Q)\,\frac{p_1 Q + p_0}{Q}, \qquad p_1 \geq 0, \tag{25}$$

with τ as in equation (5). In the baseline on $[q_0, 1]$, $Q\,\mathrm{AA}(Q) = Q - q_0$, so AR has this form with $p_1 = x_0 + x_1$ and $p_0 = -x_1 q_0$. In the specialist variant on $[q_1, 1]$, $p_1 = x_0 + x_1\alpha_g$ and $p_0 = x_1(1 - \alpha_g)\,q_1$. On a segment the marginal cost is continuous, nondecreasing and convex, extended to the segment's left end by its right limit, so that on $[q_0, 1]$ in the baseline $\mathrm{MC}(q_0)$ means $\mathrm{MC}(q_0^+)$.

Lemma A.4 (The gap on one segment) *Let A be as in equation* (25)*, let* MC *be continuous, nondecreasing and convex on $[\underline{Q}, 1]$, and let $\Delta(Q) = A(Q) - \mathrm{MC}(Q)$ and $\widetilde{D}(Q) = Q\Delta(Q)$. Then $\widetilde{D}$ is concave on $[\underline{Q}, 1]$, and Δ has the sign of $\widetilde{D}$.*

(i) If $\Delta(\underline{Q}) > 0$, then either $\Delta > 0$ on $[\underline{Q}, 1]$, or Δ has exactly one zero $r \in (\underline{Q}, 1]$, with $\Delta > 0$ on $[\underline{Q}, r)$ and $\Delta < 0$ on $(r, 1]$.

(ii) If $\Delta(\underline{Q}) < 0 < \Delta(Q_a)$ for some $Q_a \in (\underline{Q}, 1]$, then Δ has exactly one zero $r_1 \in (\underline{Q}, Q_a)$, and $\Delta < 0$ on $[\underline{Q}, r_1)$. Either $\Delta > 0$ on $(r_1, 1)$ and $\Delta(1) \geq 0$, or Δ has exactly one more zero $r_2 \in (Q_a, 1)$, with $\Delta > 0$ on (r_1, r_2) and $\Delta < 0$ on $(r_2, 1]$.

(iii) Where MC *is differentiable, $\Delta' < 0$ at the zero $r < 1$ of (i) and at the zero r_2 of (ii), and $\Delta' > 0$ at the zero r_1 of (ii).*

Proof. The function $QA(Q) = (\tau_0 - \tau_1 Q)(p_1 Q + p_0)$ has second derivative $-2\tau_1 p_1 \leq 0$, so it is concave. The function $Q\,\mathrm{MC}(Q)$ has right derivative $\mathrm{MC}(Q) + Q\,\mathrm{MC}'_+(Q)$. Both terms are nondecreasing, the second because Q is positive and increasing and the right derivative MC'_+ of a convex nondecreasing function is nonnegative and nondecreasing. So $Q\,\mathrm{MC}(Q)$ is convex, and $\widetilde{D} = QA - Q\,\mathrm{MC}$ is concave. Since $Q > 0$, Δ and $\widetilde{D}$ have the same sign.

For (i), the set $S = \{Q \in [\underline{Q}, 1] : \widetilde{D}(Q) \geq 0\}$ is convex because $\widetilde{D}$ is concave, closed because $\widetilde{D}$ is continuous, and contains $\underline{Q}$, so $S = [\underline{Q}, \bar{r}]$ for some $\bar{r} \in [\underline{Q}, 1]$, and $\bar{r} > \underline{Q}$ by continuity. At any $Q \in (\underline{Q}, \bar{r})$, concavity gives $\widetilde{D}(Q) \geq [(\bar{r} - Q)\widetilde{D}(\underline{Q}) + (Q - \underline{Q})\widetilde{D}(\bar{r})]/(\bar{r} - \underline{Q}) > 0$. If $\bar{r} = 1$ and $\widetilde{D}(1) > 0$, then $\Delta > 0$ on $[\underline{Q}, 1]$. Otherwise $\widetilde{D}(\bar{r}) = 0$, either because $\bar{r} = 1$ and $\widetilde{D}(1) = 0$ or, if $\bar{r} < 1$, by continuity, and $\widetilde{D} < 0$ on $(\bar{r}, 1]$; set $r = \bar{r}$.

For (ii), S is again a closed interval, now containing Q_a and not $\underline{Q}$, so $S = [r_1, r_2]$ with $\underline{Q} < r_1 < Q_a \leq r_2 \leq 1$ and $\widetilde{D}(r_1) = 0$ by continuity. For $Q \in (r_1, Q_a)$, concavity on $[r_1, Q_a]$ gives $\widetilde{D}(Q) \geq [(Q_a - Q)\widetilde{D}(r_1) + (Q - r_1)\widetilde{D}(Q_a)]/(Q_a - r_1) > 0$, and for $Q \in (Q_a, r_2)$ concavity on $[Q_a, r_2]$ gives $\widetilde{D}(Q) > 0$ in the same way. If $r_2 = 1$, then $\Delta > 0$ on $(r_1, 1)$ and $\Delta(1) \geq 0$. If $r_2 < 1$, then $\widetilde{D}(r_2) = 0$, $r_2 > Q_a$, and $\widetilde{D} < 0$ on $(r_2, 1]$.

For (iii), at a zero r, $\widetilde{D}'(r) = \Delta(r) + r\Delta'(r) = r\Delta'(r)$. A concave function's derivative at r is at most the slope of any chord from a point to the left of r and at least the slope of any chord to a point to the right. At the zero r of (i), the chord from $\underline{Q}$ gives $\widetilde{D}'(r) \leq -\widetilde{D}(\underline{Q})/(r - \underline{Q}) < 0$, and at r_2 the chord from Q_a gives $\widetilde{D}'(r_2) \leq -\widetilde{D}(Q_a)/(r_2 - Q_a) < 0$. At r_1, the chord to Q_a gives $\widetilde{D}'(r_1) \geq \widetilde{D}(Q_a)/(Q_a - r_1) > 0$. □

In part (ii), when $\widetilde{D}(1) = 0$ the reward exactly covers the cost of the last producer, and full entry is an equilibrium; the statements below count it as the high-credibility equilibrium.

Proof of Lemma 2.2. Let F be the distribution of costs. It has no atoms, so producers with cost exactly R^* have mass zero, and the producers with $\text{mc}_j < R^*$ have mass $F(R^*)$. Since $c^* = \text{AA}(Q^*)$ gives $R^* = \text{AR}(Q^*)$, a pair $(Q^*, \text{AA}(Q^*))$ with $Q^* > 0$ is an equilibrium if and only if $Q^* = F(\text{AR}(Q^*))$. By Assumptions 1 and 2, F is continuous and strictly increasing on the cost ranges of the two types and flat on the gap $[\text{MC}(q_0^-), \text{MC}(q_0^+)]$ between them. For $Q \in (0,1)$ with $Q \neq q_0$, $F(R) = Q$ if and only if $R = \text{MC}(Q)$, which is (i). $F(R) = q_0$ if and only if $\text{MC}(q_0^-) \leq R \leq \text{MC}(q_0^+)$, which is (ii). $F(R) = 1$ if and only if $R \geq \text{MC}(1)$, which is (iii). Silence requires some $c_\varnothing \in [0,1]$ with $R(0, c_\varnothing) \leq \text{MC}(0)$, since at such a reward no producer has a cost below it, while a reward above $\text{MC}(0)$ would draw in a positive mass. Because R is nondecreasing in c, such a $c_\varnothing$ exists if and only if it exists at $c_\varnothing = 0$, that is, if and only if $R(0,0) \leq \text{MC}(0)$, which under equation (5) is $\tau_0 x_0 \leq \text{MC}(0)$. The same argument applies in the specialist variant with q_1 in place of q_0; there the lowest accuracy is α_g, so silence requires $\tau_0(x_0 + x_1\alpha_g) \leq \text{MC}(0)$. □

Two facts about the slop segment recur. On $(0, q_0)$, $D(Q) = \tau(Q)x_0 - \text{MC}(Q)$ is strictly decreasing, because τ is nonincreasing and MC is strictly increasing. So if $\text{MC}(q_0^-) \leq \tau(q_0)x_0$, then $D > 0$ on $[0, q_0)$: there is no equilibrium with $Q^* \in (0, q_0)$, and $\tau_0 x_0 > \text{MC}(0)$ rules out silence.

Proof of Proposition 2.3. On $(q_0, 1]$, $\text{MR}(Q) = \tau(Q)(x_0 + x_1)$ is nonincreasing and MC is strictly increasing, so $\text{MR} - \text{MC}$ is continuous and strictly decreasing. It is positive just above q_0 and negative at one, so it has a unique zero $Q^{\text{FI}} \in (q_0, 1)$, positive below and negative above. For (i), an equilibrium with $Q^* \leq q_0$ has $Q^* < Q^{\text{FI}}$. An interior equilibrium with $Q^* > q_0$ has $\text{MC}(Q^*) = \text{AR}(Q^*) < \text{MR}(Q^*)$, because $\text{AA}(Q^*) < 1$, so $Q^* < Q^{\text{FI}}$. Full entry would need $\text{AR}(1) \geq \text{MC}(1)$, but $\text{AR}(1) < \text{MR}(1) < \text{MC}(1)$. For (ii), when readers can recognize producer type, producer j's reward at participation Q is $\tau(Q)(x_0 + x_1\alpha_j)$. If the producers with the Q^{FI} lowest costs post, every slop producer earns $\tau(Q^{\text{FI}})x_0 \geq \text{MC}(q_0^-)$ and covers her cost. Every investigative producer with rank in $(q_0, Q^{\text{FI}}]$ earns $\text{MR}(Q^{\text{FI}}) = \text{MC}(Q^{\text{FI}})$, at least her cost. Every producer with rank above Q^{FI} is investigative, and her cost exceeds $\text{MC}(Q^{\text{FI}})$, the reward she would earn. No producer gains by changing her decision. □

Proof of Proposition 2.4. By the slop-segment facts there is no equilibrium with $Q^* \in (0, q_0)$ and no silence. Since $D < 0$ on $(q_0, 1]$, Lemma 2.2 rules out an equilibrium with $Q^* \in (q_0, 1]$. At q_0, $\text{AR}(q_0) = \tau(q_0)x_0 \geq \text{MC}(q_0^-)$ by assumption. As Q falls to q_0, $\text{AR}(Q) \to \tau(q_0)x_0$ and $\text{MC}(Q) \to \text{MC}(q_0^+)$, so $\text{AR}(Q) < \text{MC}(Q)$ on $(q_0, 1]$ gives $\tau(q_0)x_0 \leq \text{MC}(q_0^+)$. Lemma 2.2(ii) then makes $Q^* = q_0$, $c^* = \text{AA}(q_0) = 0$ the unique equilibrium, with $R^* = \tau(q_0)x_0 > 0$. For the last claim, the wedge gives $\text{AR} < \text{MC}$ on (q_0, Q^{FI}). On $[Q^{\text{FI}}, 1]$, Proposition 2.3 gives $\text{MR} \leq \text{MC}$, and $\text{AR} < \text{MR}$ because $\text{AA} < 1$, so $\text{AR} < \text{MC}$ there as well. Under full information, the mass of producers who cover their cost at participation Q, those with $\text{mc}_j < \tau(Q)(x_0 + x_1\alpha_j)$, is continuous and nonincreasing in Q, because τ is continuous and nonincreasing and costs have no atoms by Assumption 2. So Q minus this mass is continuous and strictly increasing, nonpositive at $Q = 0$ and nonnegative at $Q = 1$, and the full-information equilibrium exists and is unique. Let $\widehat{Q}$ be its participation and $\widehat{I}$ the mass of investigative producers who post in it. Every investigative producer earns $\tau(\widehat{Q})(x_0 + x_1)$, so they post in the order of their costs, and $\widehat{Q} \leq q_0 + \widehat{I}$. If $q_0 + \widehat{I} < Q^{\text{FI}}$, the investigative producers with ranks $Q \in (q_0 + \widehat{I}, Q^{\text{FI}})$ do not post, yet by the proof of Proposition 2.3 each has cost $\text{MC}(Q) < \text{MR}(Q) \leq \tau(\widehat{Q})(x_0 + x_1)$, the last because $Q > \widehat{Q}$, a contradiction. So every investigative producer with rank in (q_0, Q^{FI}) posts and covers her cost under full information, while under pooling none posts. □

Proof of Proposition 2.5. Apply Lemma A.4 to AR on $[q_0, 1]$, where $\Delta(q_0) = \tau(q_0)x_0 - \text{MC}(q_0^+)$. The slop-segment facts rule out equilibria with $Q^* \in (0, q_0)$ and silence in both parts.

(i) Without an entry barrier, $\Delta(q_0) > 0$. Lemma 2.2(ii) fails at q_0, because $\text{AR}(q_0) > \text{MC}(q_0^+)$. By Lemma A.4(i), if $D(1) < 0$ then D has exactly one zero in $(q_0, 1)$, positive below it and negative above it, and it is the unique equilibrium and stable. If $D(1) \geq 0$, D has no zero in $(q_0, 1)$, and full entry is the unique equilibrium.

(ii) With an entry barrier, $\Delta(q_0) < 0 < \Delta(Q_a)$, and Lemma 2.2(ii) holds at q_0, so $Q_1^* = q_0$ with $c_1^* = 0$ is an equilibrium. By Lemma A.4(ii), D has exactly one zero $Q_{\text{tip}} \in (q_0, Q_a)$, and either a second zero $Q_3^* \in (Q_a, 1)$ with $D(1) < 0$, or $D > 0$ on $(Q_{\text{tip}}, 1)$ with $D(1) \geq 0$, in which case full entry is an equilibrium and we set $Q_3^* = 1$. There are no other zeros in $(q_0, 1)$, so these are all the equilibria. Since AA is strictly increasing on $(q_0, 1]$, $0 < \text{AA}(Q_{\text{tip}}) < c_3^*$. The signs of D are positive on $(0, q_0)$, negative on (q_0, Q_{tip}), positive on

(Q_{tip}, Q_3^*) and negative on $(Q_3^*, 1]$ when $Q_3^* < 1$. Hence D points toward q_0 at every participation below Q_{tip}. Participation that starts in $(Q_{\text{tip}}, 1]$ moves monotonically to Q_3^*, because D is continuous there. So q_0 and Q_3^* are stable, Q_{tip} is not, and the adjustment behaves as stated. □

Propositions 2.6 and 2.8 assume $D'(Q^*) < 0$. By Lemma A.4(iii), this holds wherever MC is differentiable at the interior equilibria of Propositions 2.5(i) and 2.7, and at an interior high-credibility equilibrium of Proposition 2.5(ii).

Proof of Proposition 2.6. (i) Let χ index the change in costs, so that $D(Q;\chi) = \text{AR}(Q) - \text{MC}(Q;\chi)$ with AR and AA independent of χ. Since $D_Q(Q^*) < 0$, the implicit function theorem gives a differentiable equilibrium path with $dQ^*/d\chi = \text{MC}_\chi(Q^*)/D_Q(Q^*)$, which is nonzero because MC moves at Q^*. Along the path $R^* = \text{AR}(Q^*)$ and $c^* = \text{AA}(Q^*)$, so $dR^*/d\chi = \text{AR}'(Q^*)\,dQ^*/d\chi$ and $dc^*/d\chi = \text{AA}'(Q^*)\,dQ^*/d\chi$. Their ratio is $\text{AR}'(Q^*)/\text{AA}'(Q^*)$, which is finite because $\text{AA}'(Q^*) \neq 0$ at $Q^* > q_0$ in the baseline and at $Q^* > q_1$ in the specialist variant, by equations (4) and (11). Substituting $\text{AR}' = \tau x_1 \text{AA}' - \tau_1(x_0 + x_1 \text{AA})$ gives equation (12). (ii) Let χ index the change in readers' reward, so that $D(Q;\chi) = \text{AR}(Q;\chi) - \text{MC}(Q)$. Readers' reward is common to all producers in the cell, so the order of entry and AA do not change. The implicit function theorem gives $dQ^*/d\chi = -\text{AR}_\chi(Q^*)/D_Q(Q^*) \neq 0$. Along the path $R^* = \text{MC}(Q^*)$ and $c^* = \text{AA}(Q^*)$, so the ratio of their derivatives is $\text{MC}'(Q^*)/\text{AA}'(Q^*)$, finite as in (i). □

Proof of Proposition 2.7. Fix one regime, h or l, and drop its subscript from D and Q^*. On $(0, q)$, AA equals the first type's accuracy, so AR is nonincreasing there and D is strictly decreasing. Its limit at q is $\text{AR}(q) - \text{MC}(q^-)$, which exceeds $\text{AR}(q) - \text{MC}(q^+) > 0$ by separated costs and no entry barrier, so $D > 0$ on $(0, q)$, and Lemma 2.2(ii) fails at q. By Lemma A.4(i) and $D(1) < 0$, D has exactly one zero $Q^* \in (q, 1)$, positive below it and negative above it, so Q^* is the unique equilibrium with positive participation. At Q_h^*, $D_l(Q_h^*) = \text{AR}(Q_h^*) - \text{MC}_l(Q_h^*) > \text{AR}(Q_h^*) - \text{MC}_h(Q_h^*) = 0$, so $Q_h^* < Q_l^*$. Every producer of the first type posts in both regimes, because her cost is the same in both and lies below every cost of the second type, and $Q_h^* > q$. So every producer who posts only in regime l is of the second type. (i) In the baseline they are investigative, AA is strictly increasing on $(q_0, 1]$, and $c_h^* < c_l^* < 1$. Both regimes share AR, so $R_h^* = \text{AR}(Q_h^*)$ and $R_l^* = \text{AR}(Q_l^*)$, and $R_h^* > R_l^*$ if and only if $\text{AR}(Q_h^*) > \text{AR}(Q_l^*)$, which holds when $\text{AR}' < 0$ on $[Q_h^*, Q_l^*]$. (ii) In the specialist variant they are generalists, with accuracy $\alpha_g < \text{AA}(Q_h^*)$ by equation (11). AA is strictly decreasing on $(q_1, 1]$, so $c_h^* > c_l^*$. On $(q_1, 1]$, $\text{AR} = \tau(x_0 + x_1 \text{AA})$ is the product of a positive nonincreasing function and a positive strictly decreasing one, so it is strictly decreasing and $R_h^* > R_l^*$. (iii) Under Assumption 6, and because the first type fills $[0, q]$ of the cost order in both regimes, every producer has the same rank in both. Then every producer who posts in regime h also posts in regime l, and those who post only in regime l are the producers with ranks in $(Q_h^*, Q_l^*]$. □

Proof of Proposition 2.8. Consider a change along which u^- rises, u^+ falls by ζ per unit of u^-, and $\bar{a}$ stays fixed. By Lemma A.1, $dx_0/du^- = (1-\zeta)/2$ and $dx_1/du^- = -(1+\zeta)/2$, with $\zeta = 0$ in part (i). Holding Q fixed, $D_{u^-}(Q^*) = \tau(Q^*)\,[(1-\zeta) - (1+\zeta)\,c^*]/2$. Since $D_Q(Q^*) < 0$, the implicit function theorem gives $dQ^*/du^- = -D_{u^-}/D_Q$, which has the sign of D_{u^-}, and $dc^*/du^- = \text{AA}'(Q^*)\,dQ^*/du^-$ has the same sign because $\text{AA}'(Q^*) = q_0/(Q^*)^2 > 0$ at $Q^* > q_0$, by equation (4). (i) With $\zeta = 0$, $D_{u^-} = \tau(Q^*)(1-c^*)/2 > 0$, because $c^* = 1 - q_0/Q^* < 1$. (ii) $D_{u^-} < 0$ if and only if $c^* > (1-\zeta)/(1+\zeta)$. □

Proof of Lemma 2.9. A post from a pool with credibility c reports the state correctly with probability $(1+c)/2 \geq 1/2$ in either state. A reader can ignore the post and choose the action her prior favors, correct with probability $\max\{b, 1-b\}$. She can follow the post, correct with probability $(1+c)/2$, or act against it, correct with probability $1 - (1+c)/2 \leq \max\{b, 1-b\}$. Her best rule is therefore correct with probability $\max\{b, 1-b, (1+c)/2\}$, and her gain from the post is $\max\{b, 1-b, (1+c)/2\} - \max\{b, 1-b\}$. Since $\max\{b, 1-b\} = (1 + |2b-1|)/2$, the gain is $\max\{0, (c - |2b-1|)/2\}$. □

Proof of Proposition 2.10. Write $D_{\text{soc}} = \text{AR}_{\text{soc}} - \text{MC}$ and $D = \text{AR}_{\text{pool}} - \text{MC}$. On $(0, q_0]$, $\text{AA} = 0 < \bar{c}(b)$, so $\text{AR}_{\text{cert}} = 0$ and $D(Q) = \tau(Q)\,x_{0,\text{soc}} - \text{MC}(Q)$, which is strictly decreasing. By (a), its limit at q_0 exceeds $\tau(q_0)\,x_{0,\text{soc}} - \text{MC}(q_0^+) > 0$, so there is no equilibrium with $Q^* \in (0, q_0]$ and no silence, and Lemma 2.2(ii) fails at q_0.

On $[q_0, Q_{\text{cert}}]$, $\text{AA} \leq \bar{c}(b)$, so $D = D_{\text{soc}}$. AR_{soc} has the form of equation (25) with $p_1 = x_{0,\text{soc}} + x_{1,\text{soc}}$ and $p_0 = -x_{1,\text{soc}}\,q_0$. By (a), $D_{\text{soc}}(q_0^+) > 0$, and by (b), $D_{\text{soc}}(Q_{\text{cert}}) < 0$, so Lemma A.4(i) on $[q_0, 1]$ gives exactly

one zero $Q_B^* \in (q_0, Q_{\text{cert}})$, with $D_{\text{soc}} > 0$ below it and $D_{\text{soc}} < 0$ on $(Q_B^*, 1]$. So Q_B^* is the only equilibrium in $(q_0, Q_{\text{cert}}]$, it is stable, and $0 < c_B^* < \bar{c}(b)$, where $\text{AR}_{\text{cert}}(Q_B^*) = 0$. The same zero is the unique equilibrium without certifiers, since then $D_{\text{soc}} < 0$ on $(Q_B^*, 1]$, including at $Q = 1$.

On $[Q_{\text{cert}}, 1]$, $\text{AA} \geq \bar{c}(b)$ and $\text{AR}_{\text{pool}} = \tau\,[(x_{0,\text{soc}} - x_{1,\text{cert}}\bar{c}(b)) + (x_{1,\text{soc}} + x_{1,\text{cert}})\text{AA}]$. This has the form of equation (25) with $p_1 = x_{0,\text{soc}} + x_{1,\text{soc}} + x_{1,\text{cert}}(1 - \bar{c}(b)) > 0$ and $p_0 = -(x_{1,\text{soc}} + x_{1,\text{cert}})\,q_0$. By (b) and (c), $D(Q_{\text{cert}}) < 0 < D(Q_a)$, so Lemma A.4(ii) gives exactly one zero in (Q_{cert}, Q_a), the tipping point, with $D < 0$ just below it and $D > 0$ just above it. It also gives either a second zero $Q_N^* \in (Q_a, 1)$, with $D < 0$ beyond it, or $D > 0$ on the rest of the segment and $D(1) \geq 0$, in which case full entry is an equilibrium and $Q_N^* = 1$. Since $Q_N^* > Q_{\text{cert}}$ and AA is strictly increasing, $c_N^* > \bar{c}(b)$, and certifiers amplify. These three are all the equilibria, and the sign pattern of D makes the Belief and News Equilibria stable and the tipping point unstable. □

A.3 Illustrative parameter values

Numerical result A.1 (Equilibria at illustrative parameter values) *At the parameter values in Table 4, the equilibria are those listed in the table. Each parameterization satisfies the hypotheses of the proposition named in its row; the row that names Section 2.6 gives the reader mix with few certifiers in Figure 5(b); this mix satisfies Assumption 7 and the hypotheses of Proposition 2.10 except condition (c).*

The equilibria were computed as follows. For each parameterization with a linear marginal cost on each type's segment, $QD(Q)$ on the second segment is a quadratic (Lemma A.4), whose roots in $(q, 1)$ were computed in closed form. For the double-selection parameters, whose marginal cost has a convex term, the zeros located by Lemma A.4 were found by bisection. The boundary equilibrium at q was checked against Lemma 2.2(ii), full entry against $D(1)$, and Q^{FI} as the zero of $\text{MR} - \text{MC}$. A brute-force search for sign changes of D on a grid of 200,000 points per segment returned the same equilibria. Values are rounded to three decimals.

A.4 A sharing game among a post's readers

Producers in a cell share a network position, and we take each post of the cell to reach the same N readers, indexed by k. Index the cell's posts by their producers, $j \in [0, Q]$ in the cost order. Reader k chooses how much to share each post, $m_{kj} \geq 0$, and her payoff is

$$U_k = \int_0^Q m_{kj}\Big[B_k^A(c) + \sum_{k'} \omega_{kk'}\, m_{k'j}\Big]\,dj \;-\; \frac{1}{2}\int_0^Q m_{kj}^2\,dj \;-\; \frac{\nu}{2}\Big(\int_0^Q m_{kj}\,dj\Big)^2.$$

Here $B_k^A(c) = x_{0k} + x_{1k}c$ is her expected payoff from amplifying a post of credibility c, as in Lemma A.1; it is the same for every post of the cell, because readers cannot tell the cell's posts apart. The gain $\omega_{kk'} \geq 0$ is positive only when reader k follows reader k', with $\omega_{kk} = 0$, and Ω is the matrix of these gains. The attention cost $\nu \geq 0$ makes the cell's posts compete for each reader. The per-post cost has coefficient one by the choice of units for sharing. Let $\mathbf{B}(c)$ be the vector of the readers' values $B_k^A(c)$, $\mathbf{1}$ the vector of ones, and $H(Q) = [(1 + \nu Q)I - \Omega]^{-1}$.

Lemma A.5 (Sharing game) *Let the largest eigenvalue of Ω be below one and $B_k^A(c) > 0$ for every reader. The game has a unique equilibrium in which every reader shares every post. In it every post is shared alike, reader by reader, with $(m_{1j}, \ldots, m_{Nj})^\top = H(Q)\,\mathbf{B}(c)$, and a post's total sharing is $R(Q,c) = \mathbf{1}^\top H(Q)\,\mathbf{B}(c)$. When every reader has the values x_0 and x_1, $R(Q,c) = \tau(Q)\,(x_0 + x_1 c)$ with $\tau(Q) = \mathbf{1}^\top H(Q)\,\mathbf{1}$, and*

$$\tau'(Q) = -\nu\,\mathbf{1}^\top H(Q)^2\,\mathbf{1} \leq 0, \qquad \tau''(Q) = 2\nu^2\,\mathbf{1}^\top H(Q)^3\,\mathbf{1} \geq 0.$$

Proof. Because Ω is nonnegative and its largest eigenvalue is below one, the series $H(Q) = \sum_{n\geq 0} \Omega^n/(1 + \nu Q)^{n+1}$ converges for every $Q \geq 0$, so $H(Q)$ exists, is nonnegative, and has a diagonal of at least $1/(1 + \nu Q)$; the same holds for $(I - \Omega)^{-1}$. Each reader's payoff is strictly concave in her own sharing, so a profile in

Table 4: Equilibria at illustrative parameter values

Case	Parameters	Equilibria (Q^*, c^*, R^*)	Q^{FI}
Multiple equilibria (Prop. 2.5(ii))	$q_0 = 0.20$; MC $= 0.04 + 0.45Q$, then $0.29 + 0.55(Q - 0.20)$; $x_0 = 0.18$, $x_1 = 0.62$; $\tau = 0.90 - 0.162Q$	low $(0.200, 0, 0.156)$; tipping $(0.337, 0.407, 0.366)$; high $(0.487, 0.589, 0.448)$	0.795
Unraveling 1 (Prop. 2.4)	$q_0 = 0.20$; MC $= 0.04 + 0.45Q$, then $0.32 + 0.44(Q - 0.20)$; $x_0 = 0.16$, $x_1 = 0.48$; $\tau = 0.90 - 0.18Q$	$(0.200, 0, 0.138)$	0.620
Unraveling 2 (Prop. 2.5(ii))	$q_0 = 0.15$; MC $= 0.05 + 0.60Q$, then $0.285 + 0.305(Q - 0.15)/0.85$; $x_0 = 0.20$, $x_1 = 0.52$; $\tau = 0.92 - 0.1656Q$	low $(0.150, 0, 0.179)$; tipping $(0.208, 0.280, 0.306)$; high $(0.721, 0.792, 0.490)$	0.902
Unraveling 1, no congestion (Prop. 2.4)	$q_0 = 0.20$; MC $= 0.10 + 0.25Q$, then $0.45 + 0.50(Q - 0.20)$; $x_0 = 0.20$, $x_1 = 0.30$; $\tau = 1$	$(0.200, 0, 0.200)$	0.300
Multiple equilibria, no congestion (Prop. 2.5(ii))	$q_0 = 0.20$; MC $= 0.10 + 0.20Q$, then $0.24 + 0.20(Q - 0.20)$; $x_0 = 0.20$, $x_1 = 0.80$; $\tau = 1$	low $(0.200, 0, 0.200)$; tipping $(0.211, 0.053, 0.242)$; high, full entry $(1, 0.800, 0.840)$	n.a.
Costly checking, baseline (Props. 2.5(i), 2.7(i))	$q_0 = 0.15$; MC $= 0.05 + 0.40Q$, then $0.36 + 0.55(Q - 0.15)$; $x_0 = 0.48$, $x_1 = 0.16$; $\tau = 0.95 - 0.589Q$	$(0.286, 0.475, 0.435)$	0.357
Cheap checking, baseline (Prop. 2.5(i); Prop. 2.7(i) with the row above)	as above, with MC $= 0.17 + 0.40(Q - 0.15)$ on $Q > 0.15$	$(0.611, 0.755, 0.354)$	0.641
Costly checking, specialist variant (Prop. 2.7(ii))	$q_1 = 0.25$, $\alpha_g = 0.52$; MC $= 0.12 + 0.55Q$, then $0.44 + 0.58(Q - 0.25)$; $x_0 = 0.08$, $x_1 = 0.58$; $\tau = 0.95 - 0.3325Q$	$(0.339, 0.874, 0.492)$	n.a.
Cheap checking, specialist variant (Prop. 2.7(ii) with the row above)	as above, with MC $= 0.28 + 0.42(Q - 0.25)$ on $Q > 0.25$	$(0.529, 0.747, 0.397)$	n.a.
Slop share 0.15 (Prop. 2.5(ii))	$q_0 = 0.15$; MC $= 0.04 + 0.40Q$, then $0.25 + 0.45(Q - 0.15)$; $x_0 = 0.16$, $x_1 = 0.58$; $\tau = 0.90 - 0.162Q$	low $(0.150, 0, 0.140)$; tipping $(0.206, 0.272, 0.275)$; high $(0.667, 0.775, 0.483)$	0.848
Slop share 0.45 (Prop. 2.4)	$q_0 = 0.45$; MC $= 0.04 + 0.20Q$, then $0.28 + 0.45(Q - 0.45)$; other parameters as above	$(0.450, 0, 0.132)$	n.a.
Double selection (Prop. 2.10)	$q_0 = 0.20$; MC $= 0.04 + 0.35Q$, then $0.20 + 0.44(Q - 0.20) + 1.35 \max\{0, Q - 0.5\}^2$; $\tau = 0.95 - 0.399Q$; $x_{0,\mathrm{soc}} = 0.242$, $x_{1,\mathrm{soc}} = 0.14$; $x_{1,\mathrm{cert}} = 2.85$, $\bar{c}(b) = 0.45$, so $Q_{\mathrm{cert}} = 0.364$	Belief $(0.285, 0.297, 0.237)$; tipping $(0.373, 0.463, 0.276)$; News $(0.922, 0.783, 0.757)$; social readers alone: $(0.285, 0.297, 0.237)$	n.a.
Few certifiers (Section 2.6)	as in the double-selection row, with $x_{1,\mathrm{cert}} = 0.50$	Belief $(0.285, 0.297, 0.237)$	n.a.

Notes: MC is given on the first type's segment and then on the second type's. "Low", "tipping" and "high" are the low-credibility equilibrium, the tipping point and the high-credibility equilibrium of Proposition 2.5(ii). Q^{FI} is reported where Proposition 2.3 applies; "n.a." marks rows where MR $\geq$ MC at $Q = 1$ or, in the specialist and double-selection rows, where the proposition is not stated. Q^{FI} is the root of MR $=$ MC; it is the participation of the full-information equilibrium only when the inequality in Proposition 2.3(ii) holds, which fails in the Unraveling 1 row. The slope τ_1 is a free parameter in every row, not $\tau_0^2/2$ as in Lemma A.3.

which every reader shares every post is an equilibrium if and only if it satisfies every reader's first-order condition. Raising m_{kj} on a set of posts of small mass around j gives the condition

$$m_{kj} = B_k^A(c) + \sum_{k'} \omega_{kk'} m_{k'j} - \nu \int_0^Q m_{kj'}\, dj',$$

whose last term, the reader's price of attention, is the same for every post she faces. Let $\mathbf{m}_j$ be the vector of sharing of post j and $\mathbf{m}^Q = \int_0^Q \mathbf{m}_j\, dj$ the vector of readers' totals. Stacking over readers, $\mathbf{m}_j = \mathbf{B}(c) + \Omega \mathbf{m}_j - \nu \mathbf{m}^Q$. Integrating over the posts gives $\mathbf{m}^Q = Q\mathbf{B}(c) + \Omega \mathbf{m}^Q - \nu Q \mathbf{m}^Q$, so $\mathbf{m}^Q = QH(Q)\mathbf{B}(c)$. Given the totals, the stacked condition has the single solution $\mathbf{m}_j = (I-\Omega)^{-1}[\mathbf{B}(c) - \nu Q H(Q)\mathbf{B}(c)]$. Since $(I-\Omega)H(Q) = I - \nu Q H(Q)$, this is $\mathbf{m}_j = H(Q)\mathbf{B}(c)$ for every post. It is positive, because $H(Q)$ is nonnegative with a positive diagonal and $\mathbf{B}(c)$ is positive, so it is an equilibrium, and the first-order conditions have no other solution. A post's total sharing is $\mathbf{1}^\top \mathbf{m}_j$. With $\mathbf{B}(c) = (x_0 + x_1 c)\mathbf{1}$ it is $\tau(Q)(x_0 + x_1 c)$. The derivatives use $dH/dQ = -H\,(dH^{-1}/dQ)\,H = -\nu H^2$. □

With equal values, a reader's sharing is $(x_0 + x_1 c)$ times her entry of $H(Q)\mathbf{1} = (1+\nu Q)^{-1}[I - \Omega/(1+\nu Q)]^{-1}\mathbf{1}$, her Bonacich centrality in Ω with decay $1/(1+\nu Q)$, as in Ballester et al. (2006). At $Q = 0$, $\tau_0 = \mathbf{1}^\top (I-\Omega)^{-1}\mathbf{1}$ is the sum of the readers' centralities, and the line $\tau_0 - \tau_1 Q$ of equation (5) is τ to first order, with $\tau_1 = \nu\, \mathbf{1}^\top (I-\Omega)^{-2}\mathbf{1}$. Because τ is convex, it lies above the line, and $\tau_0 > \tau_1$ in Assumption 4 holds when the attention cost is small enough. Under the derived visibility itself, the congestion term of Proposition 2.6 is $-\tau'(Q^*)\,(x_0 + x_1 c^*)$ in place of $\tau_1(x_0 + x_1 c^*)$. If every reader's gains sum to the same $\omega_+ < 1$, so that $\Omega\mathbf{1} = \omega_+ \mathbf{1}$, then $H(Q)\mathbf{1} = \mathbf{1}/(1 - \omega_+ + \nu Q)$ and $\tau(Q) = N/(1 - \omega_+ + \nu Q)$; its rate of fall, $-\tau'/\tau = \nu/(1 - \omega_+ + \nu Q)$, rises with ω_+. When values differ across readers, $R(Q,c) = \sum_k w_k(Q) B_k^A(c)$ with $w(Q) = H(Q)^\top \mathbf{1}$: reader k's value counts once for her own sharing and again, discounted, for the sharing it sets off among the readers who follow her, directly or through others.

The overlap of equation (17) enters the reward through the readers beyond the producer's followers. Let producer j's readers be her F_j followers and L_j further readers, the distinct accounts that follow at least one of her followers and are neither j nor her followers. By equation (17), L_j is $1 - O_j$ times the number of her second-degree arrows. Order the followers first, let Ω_{11} hold the gains among the followers and Ω_{21} the further readers' gains from the followers they follow, and let $\mathbf{B}_1(c)$ be the followers' values and $H_1(Q) = [(1+\nu Q)I - \Omega_{11}]^{-1}$.

Lemma A.6 (Overlap) *Suppose the follows among the readers run only among the followers and from further readers to followers, that every further reader has the same value $B_2^A(c)$, that every value is positive, and that the largest eigenvalue of Ω_{11} is below one. Then*

$$R(Q,c) = \mathbf{1}^\top H_1(Q)\,\mathbf{B}_1(c) + \frac{1}{1+\nu Q}\Big[L_j B_2^A(c) + \mathbf{1}^\top \Omega_{21} H_1(Q)\,\mathbf{B}_1(c)\Big].$$

If every follow of a follower by a further reader carries the same gain, then given the gains among the followers and the number of second-degree arrows into each follower, overlap enters the reward only through L_j: a rise in O_j lowers the further readers' term $L_j B_2^A(c)/(1+\nu Q)$ and leaves the other terms in place.

Proof. With the followers first, $(1+\nu Q)I - \Omega$ is block lower triangular, with diagonal blocks $(1+\nu Q)I - \Omega_{11}$ and $(1+\nu Q)I$ and lower block $-\Omega_{21}$. So is Ω, with diagonal blocks Ω_{11} and zero, so its largest eigenvalue is that of Ω_{11} and Lemma A.5 applies. The inverse has diagonal blocks $H_1(Q)$ and $(1+\nu Q)^{-1}I$ and lower block $(1+\nu Q)^{-1}\Omega_{21} H_1(Q)$, which gives the display after multiplying by the values and summing. The entry of $\mathbf{1}^\top \Omega_{21}$ for follower k is the sum of the further readers' gains from k, which is the common gain times the number of second-degree arrows into k. □

In a star the further readers are many and each follows few of the followers; in a chamber they are few and each follows many. At a given number of arrows into each follower, the followers' values and the sharing they set off among the further readers are the same in both, and only the further readers' own values, counted once per distinct further reader, differ.

B General accuracy types

This appendix replaces the two types of Section 2 with any joint distribution of accuracy and cost. Sections B.1 to B.3 show which results of Section 2 hold without two types or a cost jump, and Sections B.4 and B.5 turn to comparisons within a producer, hidden effort and comparisons across cells.

B.1 Setup

Producers differ only in accuracy and cost; readers and the reward are those of Section 2.

Assumption 9 (General types) *Producer j has accuracy $\alpha_j \in [0,1]$ and cost mc_j, and the joint distribution of $(\alpha_j, \mathrm{mc}_j)$ is otherwise unrestricted. The distribution function F of mc_j is continuous and strictly increasing on a bounded interval, which has probability one.*

The cost of equation (1) is a special case, since $\mathrm{mc}_j = \kappa_0 + \kappa_1 \alpha_j - \varepsilon_j$ restricts nothing when the cost noise ε_j may depend on α_j. The second part of the assumption rules out the baseline's cost jump and any mass of producers at one cost, so the costs of accurate and inaccurate producers may overlap.

Producer j's rank in the cost order is $U_j = F(\mathrm{mc}_j)$. The marginal cost $\mathrm{MC}(Q) = F^{-1}(Q)$ is continuous and strictly increasing on $[0,1]$. The marginal accuracy $\mathrm{MA}(Q) = E[\alpha_j \mid U_j = Q]$ is the mean accuracy of the producers of rank Q. It is defined up to a set of ranks of measure zero, and a statement that MA is monotone means that some version of it is. The average accuracy $\mathrm{AA}(Q) = E[\alpha_j \mid U_j \leq Q]$ is the mean accuracy of the producers with the Q lowest costs. As in Section 2.2, selection is adverse at Q if $\mathrm{MA}(Q) > \mathrm{AA}(Q)$, so that the producer who would post next is more accurate than the pool, and advantageous if $\mathrm{MA}(Q) < \mathrm{AA}(Q)$.

Lemmas A.1 and A.2 do not use two types. With a balanced prior, every producer in a cell therefore faces the reward $R(Q,c)$ of equation (5), and the producers who post are those with the lowest costs. The average reward $\mathrm{AR}(Q) = \tau(Q)[x_0 + x_1 \mathrm{AA}(Q)]$, the gap $D(Q) = \mathrm{AR}(Q) - \mathrm{MC}(Q)$, Definition 2.1 and the stability of Section 2.3 carry over. In Definition 2.1, the lowest accuracy is $\underline{\alpha}$, the largest α with $\Pr(\alpha_j \geq \alpha) = 1$, and the highest accuracy $\bar{\alpha}$ is the smallest α with $\Pr(\alpha_j \leq \alpha) = 1$.

Lemma B.1 (Equilibrium conditions with general types) *Under Assumptions 9 and 4, $(Q^*, \mathrm{AA}(Q^*))$ with $Q^* \in (0,1]$ is an equilibrium if and only if either $Q^* < 1$ and $D(Q^*) = 0$, or $Q^* = 1$ and $D(1) \geq 0$. Silence is an equilibrium if and only if $\tau_0(x_0 + x_1 \underline{\alpha}) \leq \mathrm{MC}(0)$.*

Proof. Because F is continuous, producers with cost exactly equal to a reward R have mass zero, and at reward R a mass $F(R)$ of producers posts. At credibility $c^* = \mathrm{AA}(Q^*)$ the reward is $\mathrm{AR}(Q^*)$, so $(Q^*, \mathrm{AA}(Q^*))$ is an equilibrium if and only if $Q^* = F(\mathrm{AR}(Q^*))$. The function F is zero up to $\mathrm{MC}(0)$, maps $[\mathrm{MC}(0), \mathrm{MC}(1)]$ continuously and one to one onto $[0,1]$, and is one from $\mathrm{MC}(1)$ on. So for $Q^* \in (0,1)$, $F(R) = Q^*$ if and only if $R = \mathrm{MC}(Q^*)$, and $F(R) = 1$ if and only if $R \geq \mathrm{MC}(1)$. With $R = \mathrm{AR}(Q^*)$, these are $D(Q^*) = 0$ and $D(1) \geq 0$. At silence, a post that no producer is expected to make earns $\tau_0(x_0 + x_1 c_\varnothing)$ for the credibility $c_\varnothing \in [\underline{\alpha}, \bar{\alpha}]$ that readers assign to it, and no producer posts if and only if this reward is at most $\mathrm{MC}(0)$. Because $x_1 > 0$, some $c_\varnothing$ in the interval qualifies if and only if $c_\varnothing = \underline{\alpha}$ does. □

B.2 Accuracy curves and the sign of selection

The two accuracy curves come from one joint distribution, so the marginal accuracy fixes the average accuracy.

Proposition B.2 (Accuracy curves) *Under Assumption 9:*

(i) *U_j is uniform on $[0,1]$, and the producers with the Q lowest costs, $\mathrm{mc}_j \leq \mathrm{MC}(Q)$, are those with $U_j \leq Q$.*

(ii) *AA is the running mean of MA: $\mathrm{AA}(Q) = Q^{-1} \int_0^Q \mathrm{MA}(u)\, du$ for $Q \in (0,1]$, and $\mathrm{AA}(1) = E[\alpha_j]$.*

(iii) AA *is continuous on* $(0,1]$*, and*

$$\mathrm{AA}'(Q) = \frac{\mathrm{MA}(Q) - \mathrm{AA}(Q)}{Q}$$

for almost every $Q \in (0,1)$ *and at every* $Q \in (0,1)$ *at which* MA *is continuous.*

Proof. (i) With probability one, mc_j lies in the interval on which F is strictly increasing, so $F(\mathrm{mc}_j) \le Q$ if and only if $\mathrm{mc}_j \le F^{-1}(Q)$. Hence $\Pr(U_j \le Q) = F(F^{-1}(Q)) = Q$, the last step because F is continuous. (ii) Since U_j is uniform, the definition of the conditional mean MA gives $E[\alpha_j \mathbf{1}\{U_j \le Q\}] = \int_0^Q \mathrm{MA}(u)\,du$. Dividing by $\Pr(U_j \le Q) = Q$ gives the running mean, and $Q = 1$ gives $E[\alpha_j]$. (iii) Since $0 \le \mathrm{MA} \le 1$ almost everywhere, $Q \mapsto \int_0^Q \mathrm{MA}(u)\,du$ is Lipschitz, so AA is continuous on $(0,1]$. By the Lebesgue differentiation theorem, the derivative of $\int_0^Q \mathrm{MA}(u)\,du$ is $\mathrm{MA}(Q)$ at almost every Q and at every point of continuity of MA. At such Q, differentiating $Q\,\mathrm{AA}(Q) = \int_0^Q \mathrm{MA}(u)\,du$ gives $\mathrm{AA}(Q) + Q\,\mathrm{AA}'(Q) = \mathrm{MA}(Q)$. □

Parts (ii) and (iii) use only that U_j is uniform, which needs only that the cost distribution has no atoms. They therefore also hold in the baseline of Section 2, whose marginal cost jumps at q_0. By part (iii), selection is adverse at Q exactly when average accuracy rises there, wherever the identity holds.

The slope of mean accuracy in the cost rank signs selection; a correlation does not.

Proposition B.3 (The sign of selection) *Under Assumption 9, let* $0 < \bar{Q} \le 1$.

(i) If MA *is nondecreasing on* $(0,\bar{Q}]$*, then* $\mathrm{MA}(Q) \ge \mathrm{AA}(Q)$ *for every* $Q \in (0,\bar{Q}]$ *and* AA *is nondecreasing there: selection is weakly adverse. If* MA *is strictly increasing on* $(0,\bar{Q}]$*, then* MA > AA *and* AA *is strictly increasing there: selection is adverse.*

(ii) If MA *is nonincreasing on* $(0,\bar{Q}]$*, the inequalities of (i) reverse, and selection is weakly advantageous there; if* MA *is strictly decreasing, selection is advantageous.*

(iii) The sign of the covariance of accuracy and cost does not determine the sign of selection. Let mc_j *be uniform on* $[0,1]$*, with* $\alpha_j = 1$ *when* $\mathrm{mc}_j \le 1/10$ *or* $\mathrm{mc}_j > 1/2$ *and* $\alpha_j = 0$ *otherwise. Then* $\mathrm{cov}(\alpha_j, \mathrm{mc}_j) = 2/25 > 0$*, but selection is advantageous at every* $Q \in (1/10, 1/2)$*. With* $1 - \alpha_j$ *in place of* α_j*, the covariance is negative and selection is adverse at every* $Q \in (1/10, 1/2)$.

(iv) Suppose the mean of accuracy given cost is linear, $E[\alpha_j \mid \mathrm{mc}_j] = m_0 + m_1 \mathrm{mc}_j$*. Then selection is adverse at every* $Q \in (0,1]$ *if and only if* $\mathrm{cov}(\alpha_j, \mathrm{mc}_j) > 0$*, and advantageous at every* $Q \in (0,1]$ *if and only if* $\mathrm{cov}(\alpha_j, \mathrm{mc}_j) < 0$*. Suppose moreover that* $\mathrm{mc}_j = \kappa_0 + \kappa_1 \alpha_j - \varepsilon_j$*, where* α_j *and* ε_j *have positive standard deviations* $\mathrm{sd}(\alpha_j)$ *and* $\mathrm{sd}(\varepsilon_j)$ *and correlation* ρ*. Then selection is adverse at every* $Q \in (0,1]$ *if and only if* $\rho < \kappa_1 \mathrm{sd}(\alpha_j)/\mathrm{sd}(\varepsilon_j)$.

Proof. (i) Fix $Q \in (0,\bar{Q}]$. Since $\mathrm{MA}(u) \le \mathrm{MA}(Q)$ for $u \le Q$, the running mean in Proposition B.2(ii) gives $\mathrm{AA}(Q) \le \mathrm{MA}(Q)$. For $Q_1 < Q_2 \le \bar{Q}$, $\mathrm{MA}(u) \ge \mathrm{MA}(Q_1) \ge \mathrm{AA}(Q_1)$ for $u \ge Q_1$, so

$$Q_2\,\mathrm{AA}(Q_2) = Q_1\,\mathrm{AA}(Q_1) + \int_{Q_1}^{Q_2} \mathrm{MA}(u)\,du \ge Q_1\,\mathrm{AA}(Q_1) + (Q_2 - Q_1)\,\mathrm{AA}(Q_1) = Q_2\,\mathrm{AA}(Q_1).$$

If MA is strictly increasing, $\mathrm{MA}(u) < \mathrm{MA}(Q)$ for every $u < Q$ and $\mathrm{MA}(u) > \mathrm{MA}(Q_1)$ for every $u > Q_1$, and both inequalities become strict, because the sets of ranks involved have positive measure. (ii) The marginal and average accuracy of $1 - \alpha_j$ are $1 - \mathrm{MA}$ and $1 - \mathrm{AA}$; apply (i) to them. (iii) Here $U_j = \mathrm{mc}_j$, and $E[\alpha_j] = 1/10 + 1/2 = 3/5$, $E[\mathrm{mc}_j] = 1/2$ and $E[\alpha_j \mathrm{mc}_j] = \int_0^{1/10} u\,du + \int_{1/2}^1 u\,du = 1/200 + 3/8 = 19/50$. So $\mathrm{cov}(\alpha_j, \mathrm{mc}_j) = 19/50 - 3/10 = 2/25$. For $Q \in (1/10, 1/2)$, $\mathrm{MA}(Q) = 0$ and $\mathrm{AA}(Q) = (1/10)/Q > 0$, so $\mathrm{MA}(Q) < \mathrm{AA}(Q)$. For $1 - \alpha_j$ the covariance is $-2/25$, and $1 - \mathrm{MA}(Q) = 1 > 1 - \mathrm{AA}(Q)$ on the same

ranks. (iv) Since F is continuous and strictly increasing on the support of mc_j, conditioning on $U_j = Q$ is conditioning on $\text{mc}_j = \text{MC}(Q)$, so $\text{MA}(Q) = m_0 + m_1 \text{MC}(Q)$ and, by Proposition B.2(ii),

$$\text{MA}(Q) - \text{AA}(Q) = m_1 \left[\text{MC}(Q) - \frac{1}{Q} \int_0^Q \text{MC}(u)\, du \right].$$

The bracket is positive for every $Q \in (0,1]$ because MC is strictly increasing, so MA $-$ AA has the sign of m_1 at every Q. By iterated expectations, $\text{cov}(\alpha_j, \text{mc}_j) = \text{cov}(m_0 + m_1 \text{mc}_j, \text{mc}_j) = m_1 \text{var}(\text{mc}_j)$, and $\text{var}(\text{mc}_j) > 0$ because F is continuous. So m_1 has the sign of the covariance. All moments are finite because α_j and mc_j, and hence ε_j, are bounded. With $\text{mc}_j = \kappa_0 + \kappa_1 \alpha_j - \varepsilon_j$, $\text{cov}(\alpha_j, \text{mc}_j) = \kappa_1 \text{var}(\alpha_j) - \rho\, \text{sd}(\alpha_j)\, \text{sd}(\varepsilon_j)$, which is positive if and only if $\rho < \kappa_1 \text{sd}(\alpha_j)/\text{sd}(\varepsilon_j)$. □

Under any joint distribution, the threshold on ρ in part (iv) is the same as a positive covariance of accuracy and cost. Part (iv) is a case in which this threshold signs selection at every rank, part (iii) shows that outside such a case it can give the wrong sign, and parts (i) and (ii) give a condition that does sign selection.

B.3 Unraveling and multiple equilibria without two types

Credibility of exactly zero needs a mass of uninformative producers, with $\alpha_j = 0$, at the bottom of the cost order.

Proposition B.4 (Zero credibility needs an atom) *Under Assumption 9:*

(i) For $Q \in (0,1]$, $\text{AA}(Q) = 0$ if and only if $\alpha_j = 0$ for almost every producer with $U_j \le Q$.

(ii) An equilibrium with $Q^* > 0$ and $c^* = 0$ therefore requires $\Pr(\alpha_j = 0) \ge Q^*$: a positive mass of uninformative producers that includes almost every producer with rank at most Q^*.

(iii) If $\Pr(\alpha_j = 0) = 0$, as when accuracy has a density, then under Assumption 4 every equilibrium either has $Q^* > 0$ and $c^* > 0$ or is silence, and silence requires $\tau_0(x_0 + x_1 \underline{\alpha}) \le \text{MC}(0)$.

Proof. (i) Since $\alpha_j \ge 0$ and $\Pr(U_j \le Q) = Q > 0$, $E[\alpha_j \mid U_j \le Q] = 0$ if and only if $\Pr(\alpha_j > 0, U_j \le Q) = 0$. (ii) If $c^* = \text{AA}(Q^*) = 0$, part (i) gives $\Pr(\alpha_j = 0, U_j \le Q^*) = \Pr(U_j \le Q^*) = Q^*$, so $\Pr(\alpha_j = 0) \ge Q^*$. (iii) By part (ii), no equilibrium has $Q^* > 0$ and $c^* = 0$, and Lemma B.1 gives the condition for silence. □

Credibility can approach zero as participation shrinks without reaching it: with $\alpha_j = U_j$, for example, $\text{AA}(Q) = Q/2 > 0$ for every $Q > 0$.

Unraveling to zero credibility does not need the cost jump. In the next law accuracy has the two types of the baseline, but the costs of the two types adjoin, so the marginal cost is continuous.

Proposition B.5 (Unraveling with a continuous marginal cost) *Let U_j be uniform on $[0,1]$, $\alpha_j = \mathbf{1}\{U_j > 3/10\}$ and $\text{mc}_j = 1/20 + U_j/5$. A share $q_0 = 3/10$ of producers is slop, and $\text{MC}(Q) = 1/20 + Q/5$. Let there be no congestion, $\tau(Q) = 1$, and let $x_0 = x_1 = 1/10$, so that $R(Q,c) = (1+c)/10$.*

(i) The unique equilibrium is $Q^* = 1/4$, $c^* = 0$, with $R^* = 1/10$. Silence is not an equilibrium for any credibility in $[0,1]$ that readers assign to an unexpected post.

(ii) If readers could recognize producer type, so that producer j earns $(1+\alpha_j)/10$, the investigative producers with ranks in $(3/10, 3/4]$, a mass of $9/20$, would post, and so would the slop producers with ranks at most $1/4$.

Proof. The cost distribution is uniform on $[1/20, 1/4]$, so Assumption 9 holds, and $\text{MA}(Q) = \mathbf{1}\{Q > 3/10\}$. By Proposition B.2(ii), $\text{AA}(Q) = 0$ on $(0, 3/10]$ and $\text{AA}(Q) = 1 - 3/(10Q)$ on $(3/10, 1]$. (i) On $(0, 3/10]$, $D(Q) = 1/10 - 1/20 - Q/5 = 1/20 - Q/5$, which is zero only at $Q = 1/4$. On $(3/10, 1]$,

$$D(Q) = \frac{1}{10}\left(2 - \frac{3}{10Q}\right) - \frac{1}{20} - \frac{Q}{5} = -\frac{20Q^2 - 15Q + 3}{100\,Q} < 0,$$

because the quadratic $20Q^2 - 15Q + 3$ has discriminant $225 - 240 < 0$ and so is positive. By Lemma B.1, $Q^* = 1/4$ is the only equilibrium with $Q^* \in (0,1)$, and $D(1) < 0$ rules out full entry. At $Q^* = 1/4$, $c^* = \mathrm{AA}(1/4) = 0$ and $R^* = R(1/4, 0) = 1/10$. At $Q = 0$ a post earns at least $1/10$ for any credibility in $[0,1]$, which exceeds $\mathrm{MC}(0) = 1/20$, so silence is not an equilibrium. (ii) With $\tau = 1$, producer j's reward under recognition is $(1+\alpha_j)/10$ whatever the participation. An investigative producer of rank $u > 3/10$ covers her cost if and only if $1/20 + u/5 \le 1/5$, that is $u \le 3/4$, and these ranks have mass $3/4 - 3/10 = 9/20$. A slop producer of rank u covers her cost if and only if $1/20 + u/5 \le 1/10$, that is $u \le 1/4$. □

Readers value accuracy here, $x_1 > 0$, and under recognition $9/20$ of all producers would post accurate content, yet under pooling none does. Not every slop producer covers her cost at zero credibility, so participation stops at $1/4$, below q_0. The slop producers with ranks in $(1/4, 3/10]$ do not post in either case, so the producers who post under recognition are not the $7/10$ with the lowest costs.

Multiple equilibria do not need two types or a cost jump either. With a continuous marginal cost, every equilibrium with participation in $(0,1)$ is a zero of D (Lemma B.1), so two stable equilibria need a third zero between them, at which D turns from negative to positive. In the next law accuracy is smooth in the cost rank, selection is adverse at every rank, the marginal cost is linear, and D has three zeros.

Proposition B.6 (Three equilibria with smooth accuracy) *Let U_j be uniform on $[0,1]$, $\mathrm{mc}_j = 3/10 + U_j/5$ and $\alpha_j = \mathrm{MA}(U_j)$, where*

$$\mathrm{MA}(Q) = \frac{26}{125} + \frac{167}{250}\,Q + \frac{9}{20}\,Q^2 - \frac{2}{5}\,Q^3.$$

Let $x_0 = 1/5$, $x_1 = 1/2$ and $\tau_0 = 1$. Then $\mathrm{MC}(Q) = 3/10 + Q/5$; MA is strictly increasing, so selection is adverse at every rank; accuracy has no atom and takes values in $[26/125, 463/500]$; and

$$\mathrm{AA}(Q) = \frac{1}{5} + \frac{2}{5}\,Q - \frac{1}{10}\,Y(Q), \qquad Y(Q) = \Big(Q - \frac{1}{5}\Big)\Big(Q - \frac{1}{2}\Big)\Big(Q - \frac{4}{5}\Big).$$

(i) Without congestion, $\tau_1 = 0$, the equilibria are exactly $Q^* = 1/5$, $1/2$ and $4/5$, with credibility $7/25$, $2/5$ and $13/25$. The first and the third are stable and the second is not. Under the adjustment $\dot{Q} = D(Q)$, participation that starts in $(0, 1/2)$ converges to $1/5$, and participation that starts in $(1/2, 1]$ converges to $4/5$.

(ii) With congestion $\tau_1 = 1/2000$, there are exactly three equilibria, one in each of $[0.15, 0.25]$, $[0.45, 0.55]$ and $[0.75, 0.85]$. The first and the third are stable and the second is not.

In both cases neither full entry nor silence is an equilibrium.

Proof. The cost distribution is uniform on $[3/10, 1/2]$, so Assumption 9 holds, U_j is the rank, and $\mathrm{MC}(Q) = 3/10 + Q/5$. Because α_j is a function of U_j, its mean given $U_j = Q$ is $\mathrm{MA}(Q)$. The derivative $\mathrm{MA}'(Q) = 167/250 + 9Q/10 - 6Q^2/5$ is concave, so its minimum on $[0,1]$ is at an endpoint: $\mathrm{MA}'(0) = 167/250$ and $\mathrm{MA}'(1) = 46/125$. So MA is strictly increasing, accuracy is continuously distributed on $[\mathrm{MA}(0), \mathrm{MA}(1)] = [26/125, 463/500] \subset [0,1]$, and by Proposition B.3(i) selection is adverse at every rank. By Proposition B.2(ii), $\mathrm{AA}(Q) = 26/125 + 167Q/500 + 3Q^2/20 - Q^3/10$, which is the display because $Y(Q) = Q^3 - 3Q^2/2 + 33Q/50 - 2/25$. So AA is strictly increasing from $\mathrm{AA}(0^+) = 26/125$ to $\mathrm{AA}(1) = 74/125$, and $\mathrm{AA}'(Q) = 167/500 + 3Q/10 - 3Q^2/10 \le 409/1000$ on $[0,1]$, with the maximum at $Q = 1/2$. Silence fails in both parts: readers assign an unexpected post credibility at least $\underline{\alpha} = 26/125$, so it earns at least $\tau_0(x_0 + x_1\underline{\alpha}) = 38/125 > 3/10 = \mathrm{MC}(0)$, and Lemma B.1 applies.

(i) With $\tau = 1$, $\mathrm{AR}(Q) = 1/5 + \mathrm{AA}(Q)/2 = 3/10 + Q/5 - Y(Q)/20$, so

$$D(Q) = -\frac{Y(Q)}{20}.$$

Its zeros in $(0,1)$ are $1/5$, $1/2$ and $4/5$, at which $\mathrm{AA} = 1/5 + 2Q/5$ equals $7/25$, $2/5$ and $13/25$. D is positive on $(0, 1/5)$, negative on $(1/5, 1/2)$, positive on $(1/2, 4/5)$ and negative on $(4/5, 1]$, so $D(1) < 0$ rules out full entry, $1/5$ and $4/5$ are stable, and $1/2$ is not. Because D is a polynomial, solutions of $\dot{Q} = D(Q)$

are unique and cannot cross a zero of D. A solution that starts in $(0,1/5)$ therefore rises and stays below $1/5$, and it converges to a zero of D, which can only be $1/5$. The same argument gives convergence to $1/5$ from $(1/5,1/2)$, and to $4/5$ from $(1/2,4/5)$ and from $(4/5,1]$, where $D(1)<0$ keeps participation below one.

(ii) Write $V(Q)=1/5+\mathrm{AA}(Q)/2$, so that with $\tau(Q)=1-Q/2000$ the gap is $D_{\mathrm{cong}}(Q)=D(Q)-QV(Q)/2000$, with D the gap of part (i). Since $\mathrm{AA}\in[26/125,74/125]$, $V\in[38/125,62/125]$, and $0\le D-D_{\mathrm{cong}}\le Q/4000$. Also $D'_{\mathrm{cong}}-D'=-[V+Q\,\mathrm{AA}'/2]/2000$, so $|D'_{\mathrm{cong}}-D'|\le(62/125+409/2000)/2000=1401/4\,000\,000$. Since $D'=-Y'/20$ with $Y'(Q)=3Q^2-3Q+33/50$, which falls on $[0,1/2]$ and is symmetric about $1/2$, $D'\le-39/8000$ on $[0.15,0.25]$ and on $[0.75,0.85]$, and $D'\ge33/8000$ on $[0.45,0.55]$. These bounds exceed $1401/4\,000\,000$, so D_{cong} is strictly decreasing on the first and third intervals and strictly increasing on the second, and has at most one zero in each. At the endpoints, $D(0.15)=91/160\,000$, $D(0.55)=7/32\,000$ and $D(0.75)=11/32\,000$ exceed the loss bound $Q/4000$ there, so $D_{\mathrm{cong}}(0.15)$, $D_{\mathrm{cong}}(0.55)$ and $D_{\mathrm{cong}}(0.75)$ are positive, while $D_{\mathrm{cong}}(0.25)\le D(0.25)<0$, $D_{\mathrm{cong}}(0.45)\le D(0.45)<0$ and $D_{\mathrm{cong}}(0.85)\le D(0.85)<0$. So D_{cong} has exactly one zero in each interval, with $D'_{\mathrm{cong}}<0$, >0 and <0 there: the first and third zeros are stable and the second is not. There are no other zeros in $(0,1]$. On $(0,0.15]$, Y is increasing and $D\ge D(0.15)=91/160\,000>0.15/4000$, so $D_{\mathrm{cong}}>0$. On $(0.25,0.45)$ and $(0.85,1]$, $D_{\mathrm{cong}}\le D<0$. On $[0.55,0.75]$, Y is convex, so $-Y$ attains its minimum at an endpoint, $D\ge7/32\,000>0.75/4000$, and $D_{\mathrm{cong}}>0$. Full entry fails because $D_{\mathrm{cong}}(1)\le D(1)<0$. □

In both parts the low equilibrium has positive credibility, as Proposition B.4 requires when there is no atom of uninformative producers. The primitives do not say which stable equilibrium a cell is in; Proposition B.9 returns to this.

B.4 Comparisons within a producer and hidden effort

Readers who act only on what they see cannot reward a post for being accurate.

Proposition B.7 (Public-record null) *Fix producer j and her cell. Let $T\in\{0,1\}$ indicate that her post is accurate, and let M be the amplification it receives. The public record $\mathscr{I}$ is what readers see of the post when they act, such as the cell, the direction of the report and its wording. Suppose $\mathscr{I}$ takes countably many values and $M=\psi(\mathscr{I},\xi)$ is integrable, where readers' private draws ξ, such as their reading and amplification costs, are independent of T given $\mathscr{I}$. Let $\bar{M}(i)=E[\psi(i,\xi)\mid\mathscr{I}=i]$. When $\Pr(T=t)>0$, let $P_t(i)=\Pr(\mathscr{I}=i\mid T=t)$.*

(i) $E[M\mid T=t,\mathscr{I}=i]=\bar{M}(i)$ whenever $\Pr(T=t,\mathscr{I}=i)>0$. In particular, a false and an accurate post with the same public record earn the same expected amplification.

(ii) If $\alpha_j<1$, so that both accurate and false posts occur,

$$E[M\mid T=0]-E[M\mid T=1]=\sum_i\bar{M}(i)\left[P_0(i)-P_1(i)\right].$$

(iii) With the readers of Lemma A.1, the false and the accurate posts of a producer with $\alpha_j<1$ earn the same expected amplification.

(iv) Suppose instead that the state is one with probability $b\in(0,1)$, that producer j, with $\alpha_j<1$, posts before the state and her report are drawn, and that her report is correct with probability $(1+\alpha_j)/2$ in either state. If the record is the report's direction, so that, as in Lemma A.2, a post that reports state one earns $\bar{M}=R^+$ and a post that reports state zero earns $\bar{M}=R^-$, then

$$E[M\mid T=0]-E[M\mid T=1]=(1-2b)\,(R^+-R^-).$$

Proof. (i) Given $\mathscr{I}=i$, ξ is independent of T, so $E[\psi(i,\xi)\mid T=t,\mathscr{I}=i]=E[\psi(i,\xi)\mid\mathscr{I}=i]=\bar{M}(i)$. (ii) By part (i), $E[M\mid T=t]=\sum_i P_t(i)\,E[M\mid T=t,\mathscr{I}=i]=\sum_i P_t(i)\,\bar{M}(i)$, where the sums converge absolutely because M is integrable. Subtracting the sum for $t=1$ from the sum for $t=0$ gives the display. (iii) In Lemma A.1, a reader reads with probability $\tau(Q)$ and amplifies when an independent cost is below $B^A(c)$,

whatever the direction or wording of the post. So $\bar{M}(i) = \tau(Q)\,B^A(c)$ for every record i in the cell, and the sum in part (ii) is $\tau(Q)\,B^A(c)\sum_i[P_0(i) - P_1(i)] = 0$. (iv) Because j posts before the draws and her report is correct with the same probability in either state, T is independent of the state. An accurate report matches the state, so it reports state one with probability b; a false report reports state one with probability $1-b$. Part (ii) then gives $R^+[(1-b)-b] + R^-[b-(1-b)] = (1-2b)(R^+ - R^-)$. □

Part (i) is the model's prediction for a producer's false and accurate posts in one cell: holding fixed what readers see, a false post earns the same expected amplification as an accurate one. A difference between them arises only through what readers see, as in part (ii), and its sign depends on the primitives: in part (iv) it has the sign of $(1-2b)(R^+ - R^-)$. With $b \neq 1/2$ and $R^+ \neq R^-$, Lemma A.2 shows that the reward differs across types, so part (iv) describes the posts of one producer and does not claim that the equilibrium of Section 2 survives.

Pooling also removes the private return to hidden effort.

Proposition B.8 (Hidden effort) *Fix a cell with participation Q and credibility c. Producer j can raise the accuracy of her post from α_j to $\alpha_j(e) \in [0,1]$ by effort $e \in [0,\bar{e}]$, at cost $K(e)$, with $\alpha_j(0) = \alpha_j$, $\alpha_j(e) \geq \alpha_j$ and $K(e) \geq K(0) = 0$. Readers observe neither her effort nor her accuracy, and producer j has mass zero, so her effort changes neither Q nor c. The state is one with probability b. Producer j decides whether to post and chooses her effort before the state and her report are drawn, and her report is correct with probability $(1+\alpha_j(e))/2$ in either state. A post that reports state one earns R^+, and a post that reports state zero earns R^-.*

(i) If $b = 1/2$ or $R^+ = R^-$, her payoff from posting with effort e is $(R^+ + R^-)/2 - \mathrm{mc}_j - K(e)$, which does not depend on her accuracy. Effort with $K(e) > 0$ strictly lowers her payoff, so she chooses an effort with $K(e) = 0$, and she posts if and only if she would without effort.

(ii) Otherwise, effort e changes her expected reward by $(b-1/2)(R^+ - R^-)\,[\alpha_j(e) - \alpha_j]$, which, when effort raises her accuracy, has the sign of $(b-1/2)(R^+ - R^-)$.

Proof. By Lemma A.2, applied to accuracy $\alpha_j(e)$, her expected reward with effort e is $(R^+ + R^-)/2 + (b - 1/2)(R^+ - R^-)\,\alpha_j(e)$, and R^+ and R^- do not depend on e because Q and c do not. (i) The second term is zero, so her payoff is $(R^+ + R^-)/2 - \mathrm{mc}_j - K(e)$, which is largest exactly at efforts with $K(e) = 0$, where it equals her payoff without effort. (ii) Subtract the expected reward at $e = 0$ from the expected reward at e. □

With a balanced prior, as in Section 2, a producer who could check facts at a cost has no private reason to do so under pooling, so hidden effort adds to the pooling problem rather than undoing it. Effort would then pay only through a reward that depends on something that reveals her accuracy, and within a cell readers see nothing that does (Assumption 3).

B.5 What comparisons across and within cells identify

Cells with the same primitives can be in different equilibria, so the distribution of credibility across cells mixes primitives with the equilibrium each cell is in.

Proposition B.9 (What comparisons identify) *Let each cell of a population satisfy Assumption 9, with the reward of equation* (5) *and its own primitives.*

(i) Let every cell have the primitives of Proposition B.6(i), and let participation in each cell follow the adjustment $\dot{Q} = D(Q)$ from an initial participation other than $1/2$. Credibility converges to $7/25$ in the cells that start below $1/2$ and to $13/25$ in the others. If all cells start on the same side of $1/2$, every cell ends with the same credibility, although every cell has two stable equilibria.

(ii) Give a cell the accuracy and readers of Proposition B.6(i) and a linear marginal cost of slope $1/4$. For every $c \in (26/125, 74/125)$ there is an intercept of the marginal cost at which the cell has a unique equilibrium, which is stable and has credibility c. So any distribution of credibility across cells on $(26/125, 74/125)$, including two atoms with nothing between them, arises from cells that each have a unique equilibrium.

(iii) Let a cell's costs change by a continuous strictly increasing transformation common to all producers. Let $0 < Q_h < Q_l \leq 1$ be the participation in an equilibrium of the cell under each of the two cost regimes, whichever equilibria the cell is in. The producers who post at Q_l but not at Q_h have mean accuracy above $\mathrm{AA}(Q_h)$ if and only if $\mathrm{AA}(Q_l) > \mathrm{AA}(Q_h)$. If MA *is nondecreasing on $(0, Q_l]$, they are at least as accurate as the pool at Q_h, and if* MA *is nonincreasing there, at most as accurate.*

(iv) Within one producer and cell, under the hypotheses of Proposition B.7, posts whose public records are i and i' differ in expected amplification by $\bar{M}(i) - \bar{M}(i')$, whatever their accuracy.

Proof. (i) This is the convergence statement of Proposition B.6(i), at whose stable equilibria credibility is $7/25$ and $13/25$. (ii) Keep $\alpha_j = \mathrm{MA}(U_j)$, $\tau = 1$, $x_0 = 1/5$ and $x_1 = 1/2$ from Proposition B.6, so that $\mathrm{AR}(Q) = 1/5 + \mathrm{AA}(Q)/2$. Since AA is continuous and strictly increasing from $26/125$ to $74/125$, there is a unique $Q_c \in (0,1)$ with $\mathrm{AA}(Q_c) = c$. Let $\mathrm{mc}_j = \mathrm{AR}(Q_c) + (U_j - Q_c)/4$, so that $\mathrm{MC}(Q) = \mathrm{AR}(Q_c) + (Q - Q_c)/4$ and Assumption 9 holds. Then $D(Q) = \mathrm{AR}(Q) - \mathrm{MC}(Q)$ has $D(Q_c) = 0$ and $D'(Q) = \mathrm{AA}'(Q)/2 - 1/4 \leq 409/2000 - 1/4 < 0$, by the bound on AA' in the proof of Proposition B.6. So Q_c is the only zero of D, it is stable, and $D(1) < 0$ rules out full entry. For silence, the function $Q \mapsto \mathrm{AR}(Q) - Q/4$ has the same negative derivative and tends to $1/5 + 13/125 = 38/125$ as $Q \to 0$, so $\mathrm{MC}(0) = \mathrm{AR}(Q_c) - Q_c/4 < 38/125 = \tau_0(x_0 + x_1\underline{\alpha})$, and Lemma B.1 rules out silence. A population of such cells, with intercepts chosen cell by cell, has any given distribution of credibility on $(26/125, 74/125)$. (iii) A common continuous strictly increasing transformation of costs keeps Assumption 9 and every producer's rank U_j, so MA and AA are the same in both regimes, and the producers who post at Q_h are among those who post at Q_l. The producers with ranks in $(Q_h, Q_l]$ have mass $Q_l - Q_h$ and mean accuracy $E[\alpha_j \mid Q_h < U_j \leq Q_l] = [Q_l \mathrm{AA}(Q_l) - Q_h \mathrm{AA}(Q_h)]/(Q_l - Q_h)$, so

$$E[\alpha_j \mid Q_h < U_j \leq Q_l] - \mathrm{AA}(Q_h) = \frac{Q_l \left[\mathrm{AA}(Q_l) - \mathrm{AA}(Q_h)\right]}{Q_l - Q_h},$$

which has the sign of $\mathrm{AA}(Q_l) - \mathrm{AA}(Q_h)$. By Proposition B.3, AA is nondecreasing on $(0, Q_l]$ if MA is, and nonincreasing if MA is. (iv) For a record i with $\Pr(\mathscr{I} = i) > 0$, $E[M \mid \mathscr{I} = i]$ averages $E[M \mid T = t, \mathscr{I} = i]$ over t, and each of these is $\bar{M}(i)$ by Proposition B.7(i). □

Parts (i) and (ii) show that the distribution of credibility across cells does not reveal whether cells have multiple equilibria. Multiple equilibria need not produce two modes, and two modes with a missing middle need not come from multiple equilibria; telling them apart requires restrictions on how primitives vary across cells and on which equilibrium each cell is in. Parts (iii) and (iv) are comparisons within one cell and hold whichever equilibrium the cell is in. Applied cell by cell, part (iii) signs selection between the two participation levels, and part (iv) measures readers' response to what they see of a post. A shift of one cell from one stable equilibrium to the other also requires a rule for which equilibrium the cell is in after its primitives change, and the static model does not supply that rule.

When marginal accuracy is monotone, the reader mix cannot change the sign of the comparison in part (iii), although it sets where the two equilibria fall.

Corollary B.9.1 (Monotone accuracy signs the comparison) *Let a cell have two cost regimes, h and l, in each of which the producers' costs have no atoms, as under Assumption 9 and with the two types of Section 2, and let every producer have the same rank U_j in both regimes. In each regime, let readers reward every post of the cell alike, with any reward $R(Q, c)$, such as the pooled reward of Assumption 7 for any mix of social readers and certifiers. Let $0 < Q_h < Q_l \leq \bar{Q} \leq 1$, where Q_h and Q_l are the participation in an equilibrium of regime h and in an equilibrium of regime l, whichever equilibria the cell is in.*

(i) If MA *is nondecreasing on $(0, \bar{Q}]$, the producers who post at Q_l but not at Q_h have mean accuracy at least $\mathrm{AA}(Q_h)$, and above $\mathrm{AA}(Q_h)$ unless* MA *is almost everywhere constant on $(0, Q_l]$.*

(ii) If MA *is nonincreasing on $(0, \bar{Q}]$, their mean accuracy is at most $\mathrm{AA}(Q_h)$, and below $\mathrm{AA}(Q_h)$ unless* MA *is almost everywhere constant on $(0, Q_l]$.*

Proof. Fix one regime and an equilibrium of it with participation Q^* and reward R^*, and let F be the distribution function of costs in that regime, so that $U_j = F(\mathrm{mc}_j)$. By Definition 2.1, the producers with cost

below R^* post and those with cost above it do not. Because F is continuous, the producers with cost exactly R^* have mass zero, and so do those with cost above R^* and $F(\mathrm{mc}_j) = F(R^*)$, whose costs lie in an interval on which F is flat. Up to a set of measure zero, the producers who post are therefore those with $U_j \le F(R^*)$, and $F(R^*) = Q^*$ is the mass that posts. Applied to both regimes, and since every producer has the same rank in both, this shows that the producers who post at Q_h are, up to a set of measure zero, among those who post at Q_l, and that those who post at Q_l but not at Q_h are those with ranks in $(Q_h, Q_l]$. The joint distribution of α_j and U_j, and with it MA and AA, is the same in both regimes. Costs without atoms make U_j uniform, so Proposition B.2(ii) holds, and the identity in the proof of Proposition B.9(iii), which uses only that part, shows that the mean accuracy of these producers exceeds $\mathrm{AA}(Q_h)$ by $Q_l\,[\mathrm{AA}(Q_l) - \mathrm{AA}(Q_h)]/(Q_l - Q_h)$. Readers enter only through Q_h and Q_l, so it remains to sign $\mathrm{AA}(Q_l) - \mathrm{AA}(Q_h)$.

(i) Take a nondecreasing version of MA on $(0, \bar{Q}]$. The proof of Proposition B.3(i), which also uses only Proposition B.2(ii), gives $\mathrm{MA}(Q_h) \ge \mathrm{AA}(Q_h)$ and $\mathrm{AA}(Q_l) \ge \mathrm{AA}(Q_h)$. Suppose $\mathrm{AA}(Q_l) = \mathrm{AA}(Q_h)$. Then $\int_{Q_h}^{Q_l} \mathrm{MA}(u)\,du = Q_l\mathrm{AA}(Q_l) - Q_h\mathrm{AA}(Q_h) = (Q_l - Q_h)\,\mathrm{AA}(Q_h)$, while $\mathrm{MA}(u) \ge \mathrm{MA}(Q_h) \ge \mathrm{AA}(Q_h)$ for $u > Q_h$, so $\mathrm{MA} = \mathrm{AA}(Q_h)$ almost everywhere on $(Q_h, Q_l]$. Hence $\mathrm{MA}(Q_h) \le \mathrm{AA}(Q_h)$, so $\mathrm{MA}(Q_h) = \mathrm{AA}(Q_h)$. Then $\int_0^{Q_h} \mathrm{MA}(u)\,du = Q_h\,\mathrm{MA}(Q_h)$, while $\mathrm{MA}(u) \le \mathrm{MA}(Q_h)$ for $u \le Q_h$, so $\mathrm{MA} = \mathrm{MA}(Q_h)$ almost everywhere on $(0, Q_h]$. So if MA is not almost everywhere constant on $(0, Q_l]$, $\mathrm{AA}(Q_l) > \mathrm{AA}(Q_h)$. (ii) The marginal and average accuracy of $1 - \alpha_j$ are $1 - \mathrm{MA}$ and $1 - \mathrm{AA}$; apply (i) to them. □

Without monotonicity, the reader mix can set the sign. When MA falls and then rises, as in Proposition B.3(iii), two cells with the same producers and the same change in costs, whose readers differ only in baseline appeal x_0, can have both equilibria where AA falls in one cell and both where it rises in the other, so that the producers who post only in regime l are less accurate than the pool at Q_h in the first cell and more accurate in the second.

C Empirical appendix

C.1 Main estimates by network cell

Tables 5 to 8 give every estimate that Section 5 reports, in the order of its subsections.

Table 5: What is amplified, by network cell: posts by producers who entered in the Breaking Window

	G_1 One-way, diverse	G_2 One-way, overlapping	G_3 Mutual, diverse	G_4 Mutual, overlapping
Panel A: Share of amplification (%)				
Accurate	93.7	79.3	59.5	32.5
Accurate but misleading	0.1	12.7	15.8	29.3
False or premature	5.9	7.1	20.8	32.0
No checkable fact	0.2	0.8	3.9	6.2
Panel B: Share of posts (%)				
Accurate	95.3	88.0	67.0	34.6
Accurate but misleading	0.3	4.3	6.5	21.4
False or premature	2.0	4.7	15.9	27.5
No checkable fact	2.4	3.0	10.6	16.5
Panel C: Accurate share of amplification relative to G_1 (percentage points)				
Unadjusted	–	-14.4^{**}	-34.3^{***}	-61.3^{***}
		(6.0)	(8.2)	(8.3)
Holding log followers fixed	–	-11.0	-24.1^{***}	-47.8^{***}
		(6.8)	(7.1)	(10.4)
Posts	296	300	339	182
Producers	34	28	30	47

Notes: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Panel C: standard errors clustered by producer in parentheses; stars from the normal approximation. Posts in W1–W4 by producers whose first post on the outbreak falls in the Breaking Window (W1), by the network cell of the producer. Panel A weights each post by its retweets plus quote tweets, capped at the 99th percentile (7,191), plus one (equation (18)); Panel B weights posts equally. Categories partition the posts: accurate (and not misleading); accurate but misleading, which also holds the misleading posts with no checkable fact; false or premature; and no checkable fact. Panel C reports b_G from the weighted regression in equation (19), without and with $\log_{10}$ of the producer's follower count. Posts labeled N/A, whose verdict depends on an image or video the labelers did not see, are left out here and in every table below. Source: viral mpox posts and posts below the viral cut by the same producers, 7 May to 1 July 2022, labeled as of the posting date; follow graph among producers.

Table 6: Reach of misleading and false posts relative to the same producer's accurate posts

	G_1 One-way, diverse	G_2 One-way, overlapping	G_3 Mutual, diverse	G_4 Mutual, overlapping
Accurate but misleading, δ_G^M	-0.63^{**}	0.22	0.58	1.30^{**}
	(0.24)	(0.61)	(0.48)	(0.63)
False or premature, δ_G^F	-0.13	0.28	-0.26	0.85^{*}
	(0.53)	(0.40)	(0.36)	(0.50)
Accurate but misleading, percent difference	-47	24	79	267
	(13)	(76)	(87)	(230)
Producer fixed effects	Yes	Yes	Yes	Yes
Calendar-date fixed effects	Yes	Yes	Yes	Yes
Posts	322	364	311	196
Producers	40	30	25	45

Notes: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Standard errors clustered by producer in parentheses; stars from the t distribution with the number of producers minus one degrees of freedom. One regression per cell, equation (20): the outcome is $\log(1 + \mathrm{RT} + \mathrm{QT})$ of a post; the omitted category is the same producer's accurate posts. Sample: checkable posts (viral and below the viral cut) by producers in the cell with at least two checkable posts. The percent difference is $100(e^{\hat{\delta}_G^M} - 1)$ in $1 + \mathrm{RT} + \mathrm{QT}$, with a delta-method standard error. Source: as in Table 5.

Table 7: Accuracy of producers who entered in the Breaking Window, by network cell

	G_1 One-way, diverse	G_2 One-way, overlapping	G_3 Mutual, diverse	G_4 Mutual, overlapping	$G_4 - G_1$
(1) Accuracy on all checkable posts	0.144^{**}	0.144	-0.291^{***}	-0.252^{***}	-0.396^{***}
	(0.061)	(0.135)	(0.101)	(0.084)	(0.104)
(2) Accuracy on W2–W4 posts only	0.154^{**}	0.201	-0.287^{**}	-0.258^{***}	-0.413^{***}
	(0.065)	(0.146)	(0.112)	(0.091)	(0.112)
(3) Row (1) + predicted reward	0.167^{**}	0.090	-0.286^{***}	-0.211^{**}	-0.379^{***}
	(0.066)	(0.140)	(0.090)	(0.083)	(0.106)
(4) Row (1) + profile, network and realized reach	0.177^{***}	0.284^{*}	-0.292^{**}	-0.227^{**}	-0.404^{***}
	(0.062)	(0.145)	(0.112)	(0.096)	(0.114)
Producers, rows (1), (3), (4)	60	48	55	76	136
of which Breaking-Window entrants	34	28	30	46	
Producers, row (2)	54	41	47	60	114
Mean accuracy, later entrants	0.833	0.604	0.784	0.717	

Notes: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Heteroskedasticity-robust (HC1) standard errors in parentheses. Each entry is $\hat{\beta}_G$ from one producer-level regression per cell, equation (21). The outcome is the producer's share of checkable posts that are accurate and not misleading, on all W1–W4 posts (rows 1, 3, 4) or on W2–W4 posts (row 2). Every row controls for $\log_{10}$ follower count, a missing-count indicator, an institutional-account indicator and three account-age bins, where present in the cell. Row (3) adds the producer's predicted reach at accuracy 0 and 1 (equation (22)). Row (4) adds bio indicators, follower-count bands, the amplifier mix, mean reach and the viral share of the producer's posts. $G_4 - G_1$: difference of the two cell estimates; its standard error treats the two cells' disjoint sets of producers as independent. Later entrants: producers whose first post falls after W1. Source: as in Table 5.

Table 8: Amplifiers by network cell: stance, selection and Breaking-Window entry

	G_1 One-way, diverse	G_2 One-way, overlapping	G_3 Mutual, diverse	G_4 Mutual, overlapping	$G_4 - G_1$
Panel A: Stance of quote tweets and replies on misinformation (%)					
Quote tweets: endorse	39.9	72.6	78.7	84.7	
Quote tweets: call out	8.9	4.5	6.3	2.6	
Quote tweets / posts quoted or replied to	461 / 9	828 / 25	633 / 41	1,099 / 44	
Replies: endorse	29.0	39.2	49.3	58.6	
Replies: call out	27.3	11.1	13.7	8.0	
Replies / posts quoted or replied to	417 / 9	740 / 22	1,522 / 38	2,122 / 44	
Panel B: Amplifier accuracy on other producers, ϕ_G					
Accurate but misleading, ϕ_G^M	-0.296^{***}	-0.314^{***}	-0.211^{**}	-0.305^{***}	
	(0.050)	(0.064)	(0.107)	(0.050)	
False or premature, ϕ_G^F	0.108^{**}	0.016	-0.029	-0.285^{***}	
	(0.049)	(0.092)	(0.058)	(0.080)	
Mean on accurate posts, a_G	0.775	0.755	0.845	0.721	
Amplifications / producers	2,305 / 56	3,444 / 46	3,049 / 51	2,106 / 69	
Panel C: Accuracy of amplifiers who entered in the Breaking Window, β_G^{amp}					
Baseline	-0.035^{**}	0.060^{***}	-0.084^{***}	-0.283^{***}	-0.249^{***}
	(0.014)	(0.013)	(0.011)	(0.022)	(0.026)
+ amplifier follower count and volume	-0.047^{***}	0.075^{***}	-0.068^{***}	-0.287^{***}	-0.239^{***}
	(0.014)	(0.014)	(0.011)	(0.022)	(0.027)
Mean accuracy, later amplifiers	0.916	0.825	0.916	0.712	
Amplifiers in W2–W4 / of which in W1	4,240 / 597	4,091 / 719	6,529 / 1,134	2,841 / 624	

Notes: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Panel A: share of the quote tweets (replies) on accurate but misleading, false or premature posts whose stance endorses or calls out the post; the remainder is neutral, an attack that leaves the post's claim standing, unclear, or without a majority of the three labelers. Panel B: equation (23), one row per amplification (a retweet or quote of a post by an amplifier who amplifies at least two producers); standard errors clustered by producer. Panel C: equation (24), one row per amplifier active in W2–W4, outcome the share of accurate posts among the checkable posts the amplifier amplifies in W2–W4; heteroskedasticity-robust (HC1) standard errors in the cell columns; $G_4 - G_1$ from one stacked regression with standard errors clustered by amplifier. Later amplifiers: amplifiers with no amplification in W1. Source: retweets in the 2022 engagement records and 2026 timelines; quote tweets and replies collected in 2026.

C.2 Co-movement of reach with the cell's credibility

This test asks whether a producer's posts reach further in windows when its cell posts more accurately. The population is the checkable labeled posts by producers in a cell with at least two checkable posts, the sample of equation (20). For each cell G and window W, $c_{G,W}$, the cell's accurate share, is the average over the cell's producers of their share of accurate checkable posts in that window. We estimate one regression per cell,

$$R_{jt} = \mu_j + \theta_G c_{G,W(t)} + u_{jt}, \tag{26}$$

where R_{jt} is the reach of post t by producer j, μ_j is a producer fixed effect, $W(t)$ is the window of the post, θ_G is the slope, and u_{jt} is the error, with standard errors clustered by producer. Within a cell, $c_{G,W}$ takes four values, one per window, so any change in reach common to all producers of the cell over the cycle loads onto θ_G.

By equation (26), θ_G is $+3.73$ in G_1 (SE 1.36), $+0.43$ in G_2 (SE 2.15), -2.62 in G_3 (SE 0.69) and -7.20 in G_4 (SE 1.83) (Table 9, Panel A).

Panel B of the same table reports equation (20) with an indicator for viral posts, the control for the sampling stratum, which Section 5.2 discusses. Panel C adds to equation (20) an indicator that the post is labeled sensational and an indicator that it is labeled slanted or fringe, on the same posts.

C.3 What is amplified at a common follower count, and by producer type

Table 10, Panel A, predicts each cell's shares of amplification from equation (19), estimated for each category's indicator in turn, at the median follower count of the posts. Panel B computes the accurate share of equation (18) separately for news outlets and for the other producers of each cell; a producer is a news outlet when two classifier passes over its 2026 profile both code it so. Section 5.1 reports what the two panels show.

C.4 Breaking-Window producers: role and the dating of entry

Table 11 re-estimates equation (21) in two ways. Rows (2) and (3) add six producer-role groups coded from the producer's 2026 profile: news outlet, journalist, health or science, government, political, and other. Rows (4) to (7) use a second collection, a 2026 search of each producer's 2022 posts on the outbreak with complete coverage up to its first post. It dates the first post of 384 labeled producers. For 42 of them our indicator places the first post in the Breaking Window and the second collection does not; for 39 the reverse holds. Row (4) is the baseline on these producers, row (5) replaces Breaking_j with the second collection's indicator, row (6) adds the role groups to row (5), and row (7) keeps the 303 producers whose two indicators agree. Section 5.3 reports the estimates.

C.5 Amplifiers: quote tweets alone, and who retweets

The amplifier tests classify a post by averaging over its labelers. A post is checkable when fewer than half of its labelers find no checkable fact. A checkable post counts as accurate when at least half find it accurate and fewer than half find it misleading, as misleading when at least half find it misleading, and as false or premature when fewer than half find it accurate. This gives 2,124 checkable posts. The producer tests instead take the majority verdict of the post's labelers on misinformation, and the category most of the labelers in that majority chose, ties going to Claude Opus 5.5, then gpt-6.1-sol, then Gemini (Appendix D.3), which gives the 1,829 checkable posts of Section 5.

Panel A of Table 12 re-estimates equations (23) and (24) on quote tweets alone. Panel B describes who retweets. For each labeled post we observe the list of accounts that retweeted it, as it stood in 2026, and each listed account's 2026 profile. The lists omit retweets undone and accounts closed since 2022. A keyword list on the profile codes an account as partisan or identity; otherwise as a certifier, for credentials, health and science terms, or journalism and news organizations; otherwise as neither. For the 1,029 posts by 229 producers in a cell with at least ten listed retweeters, we estimate

$$y_i = a + \sum_{G \neq G_1} b_G D_{iG} + e_i, \tag{27}$$

Table 9: Demand within producer: co-movement with the cell's accurate share, controls for the viral stratum and for visible cues, and claims outside the fact boxes

	G_1 One-way, diverse	G_2 One-way, overlapping	G_3 Mutual, diverse	G_4 Mutual, overlapping
Panel A: Co-movement of reach with the cell's accurate share, θ_G				
θ_G	3.73^{***}	0.43	-2.62^{***}	-7.20^{***}
	(1.36)	(2.15)	(0.69)	(1.83)
Posts / producers	322 / 40	364 / 30	311 / 25	196 / 45
Panel B: Equation (20) *with an indicator for the viral stratum*				
Accurate but misleading, δ_G^M	0.70^{*}	0.31	0.05	0.88^{**}
	(0.38)	(0.41)	(0.30)	(0.39)
False or premature, δ_G^F	0.09	0.53^{*}	-0.37	0.30
	(0.65)	(0.29)	(0.27)	(0.35)
Posts / producers	322 / 40	364 / 30	311 / 25	196 / 45
Panel C: Equation (20) *with the post's sensational and slanted labels*				
Accurate but misleading, δ_G^M	-1.22	0.11	0.48	0.75
	(1.28)	(0.62)	(0.40)	(0.83)
False or premature, δ_G^F	-0.13	0.11	-0.39	0.22
	(0.52)	(0.33)	(0.38)	(0.68)
Sensational	1.03^{***}	1.00^{***}	0.56^{*}	0.67
	(0.31)	(0.26)	(0.31)	(0.50)
Slanted or fringe	0.35	-0.24	-0.31	0.78
	(1.30)	(0.32)	(0.20)	(0.60)
Sensational / slanted or fringe posts	21 / 6	60 / 30	109 / 42	82 / 77
Posts / producers	322 / 40	364 / 30	311 / 25	196 / 45
Panel D: Equation (20) *without the false or premature posts whose claim lies outside the fact boxes*				
Accurate but misleading, δ_G^M	-0.64^{***}	0.44	0.85^{*}	1.16^{*}
	(0.24)	(0.58)	(0.42)	(0.63)
False or premature, δ_G^F	-0.28	1.07^{*}	0.74^{*}	2.26^{***}
	(1.86)	(0.56)	(0.39)	(0.82)
False or premature posts left out / kept	4 / 6	11 / 11	27 / 23	34 / 21
Posts / producers	318 / 40	353 / 30	284 / 25	162 / 45

Notes: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Standard errors clustered by producer in parentheses; stars in Panel A use the normal approximation, in Panels B to D the t distribution as in Table 6. Panel A: equation (26), one regression per cell, producers with at least two checkable posts; $c_{G,W}$ takes four values per cell. Panel B: equation (20) with an indicator for posts above the viral cut, the stratum by which posts were sampled, added; producer and calendar-date fixed effects. The cut is set on a post's retweets, quotes and replies, so the indicator conditions on reach. Panel C: equation (20) with indicators that the post is labeled sensational and that it is labeled slanted or fringe, added; producer and calendar-date fixed effects; same posts as Table 6. Panel D: equation (20) on the posts of Table 6 less the false or premature posts on which most of the post's labelers flag the claim they judged as outside the fact boxes, so that the verdict rests on the labeler's own judgment of the claim; producer and calendar-date fixed effects.

Table 10: What is amplified, by network cell, holding the producer's follower count fixed

	G_1 One-way, diverse	G_2 One-way, overlapping	G_3 Mutual, diverse	G_4 Mutual, overlapping
Panel A: Share of amplification at the median follower count (%)				
Accurate	86.3	75.3	62.2	38.5
Accurate but misleading	3.4	14.5	14.6	26.7
False or premature	10.2	9.4	19.3	28.5
No checkable fact	0.0	0.8	3.9	6.3
Panel B: Accurate share of amplification by producer type (%)				
News outlets	94.3	86.3	70.7	74.7
posts / producers	209 / 18	249 / 16	14 / 5	8 / 7
Other producers	92.3	61.3	59.2	30.3
posts / producers	87 / 16	51 / 12	325 / 25	174 / 40
News outlets' share of the cell's amplification	71.7	72.1	2.8	4.9

Notes: Predicted shares from equation (19), estimated separately for each category's indicator with $\log_{10}$ follower count, evaluated at the median follower count of these posts (about 460,000). Same posts and weights as Table 5, Panel A; the four predicted shares sum to one in each cell. Follower counts are read in 2026. Panel B: equation (18) for the accurate category, separately for producers whose 2026 profile two classifier passes both code as a news outlet and for all other producers.

Table 11: Breaking-Window producers: producer role and the dating of entry

	G_1 One-way, diverse	G_2 One-way, overlapping	G_3 Mutual, diverse	G_4 Mutual, overlapping	$G_4 - G_1$	Producers
(1) Baseline (Table 7, row 1)	0.144^{**}	0.144	-0.291^{***}	-0.252^{***}	-0.396^{***}	239
	(0.061)	(0.135)	(0.101)	(0.084)	(0.104)	
(2) Row (1) + producer role	0.079	0.075	-0.168	-0.297^{***}	-0.376^{***}	239
	(0.049)	(0.146)	(0.118)	(0.085)	(0.098)	
(3) W2–W4 accuracy + producer role	0.069	0.121	-0.175	-0.257^{**}	-0.326^{***}	202
	(0.050)	(0.165)	(0.134)	(0.102)	(0.114)	
(4) Row (1), producers dated twice	0.150^{**}	0.096	-0.288^{***}	-0.280^{***}	-0.430^{***}	229
	(0.063)	(0.138)	(0.105)	(0.081)	(0.103)	
(5) Row (4), second dating	0.081	0.090	-0.306^{***}	-0.070	-0.151	229
	(0.090)	(0.134)	(0.095)	(0.086)	(0.124)	
(6) Row (5) + producer role	0.011	-0.028	-0.174	-0.103	-0.114	229
	(0.086)	(0.159)	(0.111)	(0.084)	(0.120)	
(7) Row (1), the two dates agree	0.170^{*}	0.151	-0.403^{***}	-0.215^{**}	-0.385^{***}	178
	(0.089)	(0.171)	(0.093)	(0.093)	(0.129)	

Notes: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Heteroskedasticity-robust (HC1) standard errors in parentheses; $G_4 - G_1$ as in Table 7, with stars from the normal approximation. Each row is equation (21) with the controls of Table 7, row (1). Producer role: six groups coded from the producer's 2026 profile (news outlet, journalist, health or science, government, political, other). Rows (4) to (7) use the producers whose first post on the outbreak is observed in a second collection, a 2026 search of each producer's 2022 posts; row (5) sets the Breaking-Window indicator from that first post.

where y_i is the share of post i's listed retweeters of a given type, D_{iG} indicates that the post's producer is in cell G, a is the G_1 mean, b_G is cell G's difference from it, and e_i is the error, with standard errors clustered by producer.

By equation (27), 10.4% of a G_1 post's listed retweeters carry a certifier bio, against 6.0% in G_4, a difference of -4.4 percentage points (SE 2.5). Among the first 20 retweeters the difference is -4.1 points (SE 2.8). The share with a partisan or identity bio is 3.2 points higher in G_4 (SE 1.9).

Table 12: Amplifiers: quote tweets only, and who retweets

	G_1 One-way, diverse	G_2 One-way, overlapping	G_3 Mutual, diverse	G_4 Mutual, overlapping	$G_4 - G_1$
Panel A: Quote tweets only					
Accurate but misleading, ϕ_G^M, equation (23)	-0.325^{***}	-0.145^{*}	-0.197^{**}	-0.179^{***}	
	(0.109)	(0.080)	(0.081)	(0.040)	
Breaking-Window amplifiers, β_G^{amp}, equation (24)	-0.022	0.137^{***}	-0.124^{***}	-0.099^{***}	
	(0.014)	(0.018)	(0.023)	(0.031)	
Panel B: Composition of a post's retweeters (%), equation (27)					
Certifier bio, all listed retweeters	10.4	8.4	10.1	6.0	-4.4^{*}
					(2.5)
Certifier bio, first 20 retweeters	10.9	7.8	10.5	6.8	-4.1
					(2.8)
Partisan or identity bio, all listed retweeters	8.1	11.1	9.3	11.3	3.2^{*}
					(1.9)
Posts / producers					1,029 / 229

Notes: $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$. Panel A: equations (23) and (24) estimated on quote tweets alone; standard errors as in Table 8. Panel B: mean share of a post's listed retweeters whose 2026 profile matches a keyword list, over posts in the four cells with at least ten listed retweeters; $G_4 - G_1$ is the G_4 coefficient of equation (27), with standard errors clustered by producer and stars from the normal approximation.

C.6 The posts found in 2026, and other cell rules

Table 13 reruns the network tests with the posts found in 2026 and their producers, and under the cell rules of Appendix D.6. Row (1) is the paper's: the July cells, with the 305 posts found in 2026 by producers already in the graph frame. Row (2) adds the other 403 posts found in 2026 and their 332 producers. Every other producer keeps its cell, and each new producer is placed by the same rule, from its follows to and from the accounts of the July follow file and the July band medians, so that no producer's cell depends on which other new producers were found. Row (3) instead recomputes every cell on the follow graph that adds the new producers' follows, which changes the cell, or the absence of one, of 101 of the 442 producers that row (1) starts from. Rows (4) to (8) keep row (1)'s producers and graph and place them in cells by the rules of Table 21. The composition gap is negative in every row, and the amplifier difference in every row but (3). The supply difference and the negative coefficient on misleading posts in G_1 hold in row (1) and with the reciprocity cut at 0.25 or at the median, and are not distinguishable from zero in rows (2) and (3); under the other three rules the supply difference holds and the coefficient on misleading posts in G_1 turns positive. Of the 165 new producers in a cell in row (2) with a checkable post that is not N/A, 80% have a single one, against 41% of row (1)'s producers in a cell with such a post, and their posts are accurate 55% of the time on average, against 69%.

Table 13: The network tests with the producers found in 2026 and under other cell rules

	Composition, gap (pp)	Supply, $G_4 - G_1$	Misleading, G_1	Misleading, G_4	Amplifiers, $G_4 - G_1$	Producers in a cell
(1) Paper	−47.8 (10.4)	−0.40 (0.10)	−0.63 [0.012]	+1.30 [0.044]	−0.25 (0.03)	240
(2) All found posts, July cells	−43.2 (8.6)	−0.08 (0.11)	−0.32 [0.475]	+1.22 [0.044]	−0.20 (0.02)	464
(3) All posts, combined graph	−46.7 (10.1)	−0.05 (0.10)	−0.29 [0.437]	+1.49 [0.056]	+0.03 (0.02)	535
(4) r_j cut at 0.25	−46.6 (10.3)	−0.32 (0.10)	−0.58 [0.029]	+1.34 [0.024]	−0.15 (0.02)	240
(5) r_j cut at median	−47.8 (10.4)	−0.39 (0.10)	−0.63 [0.012]	+1.30 [0.044]	−0.25 (0.03)	240
(6) r_j cut at 0.50	−40.3 (10.9)	−0.24 (0.12)	+0.97 [0.164]	+2.44 [0.008]	−0.26 (0.03)	240
(7) Overlap vs trend	−39.6 (11.4)	−0.41 (0.10)	+1.27 [< 0.001]	+0.88 [0.050]	−0.25 (0.02)	240
(8) Burt's size	−44.6 (10.6)	−0.41 (0.11)	+0.21 [0.613]	+0.59 [0.113]	−0.17 (0.02)	240

Notes: Standard errors in parentheses, p-values in brackets, as in Tables 5 to 8. Composition gap: the difference between G_4 and G_1 in the accurate share of amplification, at a common follower count. Supply: the Breaking-Window coefficient of equation (21). Misleading: the reach of accurate but misleading posts relative to the same producer's accurate posts, equation (20). Breaking-Window amplifiers: the coefficient β_G^{amp} of equation (24). Row (1) is the paper's: the July cells, and the posts found in 2026 by producers already in the follow-graph frame. Rows (2) and (3) add the other posts found in 2026 and their producers; rows (4) to (8) keep row (1)'s producers and graph and change the rule that places them in cells (Appendix D.6).

D Data appendix

D.1 From collected posts to analysis samples

Every sample in the paper is a subset of the 2,809 labeled posts, and Table 14 gives the count at each step. The labeled posts are the 1,730 viral posts, 708 of them found in 2026, and 1,079 posts by the same producers that fall below the viral cut. A post whose verdict depends on content the labelers were not given is N/A and receives no verdict: 139 viral posts and 105 posts below the cut are N/A. An N/A post is left out of every estimate.

The 708 viral posts found in 2026 come from three reads of X. A third public collection of posts on the outbreak, gathered through Twitter's Academic API on the disease's names in English, Spanish, French and Portuguese (Movahedi Nia et al. 2023), lists 2,400,202 post identifiers from 1 May to 25 December 2022, and 790,631 of them fall in our window and are not in the second 2022 collection. Reading by identifier the 10,806 of these posted by our producers gives 281 new viral posts. Reading the other 779,825 in time order gives 18 from the first sixth, which runs to 20 May 2022 and had been read when these posts were chosen on 2 November 2026. A keyword search of X for May and June 2022 gives the other 409. A post enters if it is still readable in 2026, is an original post in the window, matches the outbreak's keywords, passes the viral cut on its 2026 counts and is not already labeled. The three model families labeled these posts as of their posting date, with the same rules, inputs and images as the other viral posts.

The posts below the viral cut come from a pool of 1,380 posts by 226 producers, posted between 7 May and 1 July 2022 and retrieved in 2026. Of the 1,079 labeled, 660 were drawn by a recorded design from the 957 posts in the pool whose text names the disease or one of its vaccines. The first 380 raise the number of labeled posts of producers with few viral posts, at most six posts per producer, and the other 280 are drawn at random within five strata of producers, three defined by their position in the follow graph and two by follower count. Each drawn post carries its probability of selection. The other 419 posts come from a second labeling run, whose rule for choosing posts was not recorded. They hold 296 of the 297 posts in the pool that name the disease or a vaccine and that the design did not draw, and 123 of the 423 posts in the pool that name neither. No estimate weights the posts by these probabilities.

The network tests start from 2,181 labeled posts, 1,120 of them viral, by 442 producers. Of these posts, 186 are N/A and are left out of the tests. Of the 402 producers left, the 240 producers with at least three followers in the following lists we collected fall in a network cell, and the other 162 fall in none. The 442 producers belong to the graph frame, the 448 producers with stored network measures; the rule that chose the

448 from the 639 producers of the 2022 collections was not recorded, and all of them are on the roster of 506 producers described in Appendix D.6.

Table 14: From collected posts to analysis samples

	Posts	Viral posts	Producers
Panel A: Labeled posts			
Viral posts	1,730	1,730	958
From the 2022 collections	1,022	1,022	639
Found in 2026	708	708	431
N/A	139	139	110
With a label	1,591	1,591	867
Posts below the viral cut, same producers	1,079	–	197
Drawn by the recorded design	660	–	161
From the second labeling run	419	–	105
N/A	105	–	66
With a label	974	–	190
All labeled posts	2,809	1,730	958
Panel B: Network tests			
Labeled posts of producers in the graph frame	2,187	1,126	448
Posts the network tests start from	2,181	1,120	442
of which N/A, left out	186	83	116
Producers in a network cell ($F_j \geq 3$), N/A left out	1,389	703	240
Producers in no cell ($F_j < 3$), N/A left out	606	334	162

Notes: Counts of posts and of distinct producers. A post is N/A when at least one labeler marks that its verdict depends on content it was not given. The graph frame is the 448 producers with stored network measures; 6 viral posts by 6 producers in the frame, each with no other labeled post, are not in the network tests. F_j is the number of accounts whose collected following lists include producer j. Posts below the viral cut drawn by the recorded design carry their sampling probabilities. Source: post records of 2022 and 2026; labels of 27 October to 2 November 2026.

D.2 Labeling rules

The labels come from version 3.1 of the rules, the last of four versions written between 18 and 27 October 2026 (Table 15). Each version was revised where a labeler differed from the settled answer, or labelers differed among themselves: we decided which reading was right and which rule had allowed the other, and restated that rule in general terms with an invented example. The last revision rests on 17 cases chosen by the kind of disagreement they showed, each sent to the human rater with the post, the dated record and every labeler's answer and reason. For each case she said which rule decided it, and where the wording left it open.

Five general rules apply before any block. A post is judged together with everything it quotes, shows or links. Content the post passes on with no sign of disagreement is the post's own claim, and naming its source is not a hedge. Content quoted to refute, correct or mock it makes the refutation the post's claim, while content quoted to question it, or set beside other evidence, is judged for what the post says and what it leads readers to believe. A fact shown in an image, such as a headline, a chart or a screenshot, counts as said by the post, and a video is judged by its first frame and its words. When the verdict depends on content the labeler was not given, the post is marked as having lost content and receives no verdict.

Truth turns on what the post states outright. No checkable fact covers opinions, policy preferences, vague accusations with nothing checkable, anecdotes nobody can verify, questions that take nothing checkable for granted, and personal jokes that riff on a headline without a new claim. A subjective wrapper still states its content, so "I think X" is judged on X, and calling the outbreak itself a hoax is judged against the case counts. Forecasts follow three rules. Predicting as future an outcome that had already happened or been ruled out by the posting date is false. Claiming that an authority is about to announce something is a claim about a present plan, false when the dated record shows nothing behind it. A genuinely open forecast has no checkable fact, and any untrue premise it implies is judged under Belief. Premature requires something real still pending on

Table 15: Versions of the labeling rules

Version	Written	Used for	Main changes from the previous version
2.8	18 Oct 2026	Earlier labels of the full corpus, not used in this paper	–
2.9	20–22 Oct	Four blind runs on the development set; two tests on new posts	Premature defined as something real still pending presented as settled; a claim with nothing behind it is false; rules from the rulings on each run's splits
3.0	22–24 Oct	The rater's check of a relabel; a trial of the dated record on 40 further posts	A post judged with everything it quotes, shows or links; one rule for fringe; posts whose content is lost set aside; labelers use only the dated record
3.1	26–27 Oct	The labels in this paper, made 27 October to 2 November 2026	Rules for forecasts, idioms beside the full facts, news shorthand, wrong in kind or only in degree, personal jokes and old clippings; each claim counted once; close-up lesion photos sensational; fringe requires alleging a plot; who posted informs context only

Notes: The development set is the 34 posts the human rater labeled first; Table 19 reports the tests. Source: dated records of the rules and of each ruling.

the posting date, such as a suspected case under test, an inquiry or a preprint. A hedge counts only when it covers the core claim, and it never rescues a claim that the record contradicts or that had nothing behind it. An unaltered historical clipping is accurate on what it says. Panel A of Table 16 separates a claim that is wrong in kind, which is false or premature, from one that is wrong only in degree, which stays accurate and counts as exaggeration under Appeal.

Table 16: Coding rules for Truth and Appeal

Panel A: Wrong in kind (false or premature) or only in degree (accurate, exaggeration)

Dimension	Wrong in kind	Wrong only in degree
Events, rules and status	Zero against one or more cases; advice against a mandate; an open question presented as settled; a real figure attached to the wrong vaccine or disease	Rounding within the same order of magnitude; a caption that slightly misreads its own chart
Past and present	One corrected past mistake presented as an ongoing practice	Inflated wording for a real change whose two sides the post states
Direction and size	Wrong direction or order of magnitude	An inflated adjective on a true figure
Scope	A literal "never", "only" or "everyone" with a counterexample on the posting date	Routine news shorthand, judged against the series the post refers to

Panel B: Types of sensational appeal

Type	Definition	Basis
Alarm	Fearmongering or worst-case framing	High-arousal emotions such as anxiety make content more viral (Berger and Milkman 2012)
Outrage	Moral anger at a person, group or institution, including anger at being deceived	Each moral-emotional word raises diffusion by about 20% (Brady et al. 2017)
Disgust or shock	Graphic images or descriptions, including any close-up photo of lesions	Graphic: salacious or disturbing subject matter (Biyani et al. 2016)
Exaggeration	Wrong only in degree, inflated for effect	Biyani et al. (2016)
Teasing	Details held back to build suspense	Biyani et al. (2016)
Formatting	Dramatizing capitals, repeated punctuation, emoji floods or urgent calls to share	Biyani et al. (2016)

Notes: A post is sensational if any type applies, and the labeler records every type that does. Wire tags such as "BREAKING" are ignored, and a frightening fact stated plainly is not sensational. Source: the rules, version 3.1.

Appeal records why a post may spread, in two fields that can take any combination. A post is humorous when it carries a clear signal of jest, such as a punchline, an absurd premise or a reaction meme; sarcasm,

ridicule, bitter irony and a jeering clown emoji are not humor.

Belief asks what a general reader on the posting date takes away beyond the words. The labeler writes that claim in one sentence and codes the post misleading only if the dated record refutes it. The general reader is neither the post's own audience nor the labeler, and the producer's intent plays no part. Five forms of steering recur in the cycle: a dropped condition or a cherry-picked fact; false context, such as an old document presented as explaining the current outbreak; true items stacked side by side under a "Coincidence?"; a leading question or an insult with a false premise; and a routine event recast as alarming. Two forms are not misleading. A hyperbolic idiom beside both sides of the actual change, such as "flipped a 180" with both statements quoted, is exaggeration under Appeal. An opinion, insult or mockery with no checkable false premise plants no claim.

Position records fringe or mainstream, and slant. Fringe covers three kinds of claim, stated or hinted: institutions lying, staging the outbreak or covering it up; the outbreak planned or engineered, including a laboratory origin; and the disease caused by vaccines or relabeled to hide vaccine harm. Scientific debate, partisan blame, mockery of political correctness and harsh criticism of agencies for incompetence or overreaction are mainstream. Slant requires an explicit marker of a partisan, ideological or institutional side, such as us-against-them vocabulary, and the known views of the producer never count.

Beyond the four blocks, each labeler records the claim it judged, the fact box it used, the untrue claim when the post is misleading, every sensational type, whether content was lost, whether the media changed its answer, the post's role (satire, personal, opinion or news) and topic, a confidence score, and a reason of at most 30 words.

The labelers judge every claim against a dated record of 19 fact boxes, each a short list of dated facts on a claim that recurs in the cycle (Table 17). They use only the post and these boxes, and a claim the boxes do not settle is flagged for verification. At least one labeler flagged such a claim on 1,105 of the 1,591 viral posts that are not N/A. The flagged claims were not checked further, and these posts keep the labelers' verdicts.

D.3 Labelers and the combined label

Three language models from different families labeled the posts (Table 18). Claude Opus 5.5, gpt-6.1-sol and Gemini each labeled all 1,730 viral posts on 1 to 2 November 2026, in batches of 20, with the same instructions. Each post came with its text, its posting date, the producer's handle and its images, a video or an animated image by its first frame. The images of 114 viral posts with an image or video could no longer be retrieved, and these posts are N/A. To find the viral posts that quote another post, we read each of the 1,022 viral posts of the 2022 collections by its identifier in 2026: 20 quote another post, and 146 could not be read, because the post was deleted or its account was closed, suspended or protected. For the 18 quote posts whose quoted post is still available, all three labelers labeled the post again with the quoted post's text and images, and these labels replace the first. The other 2 quote a post that is no longer available and keep their first labels. Of the posts found in 2026, 16 quote a post that could still be read, and were labeled with it in the same way. The posts below the viral cut keep labels made from their text alone in the earlier runs, in which each labeler received the rules and a batch of 12 or 20 posts, with each post's text, its posting date, and whether it carries media and of what kind. No labeler of these runs saw an image, a video, a quoted post or the producer's handle. In the first run gpt-6.1-sol labeled the posts drawn by the recorded design, and Gemini labeled 147 of them as well. In the second run gpt-6.1-sol and Claude Opus 5.5 labeled the other 419 posts below the cut, and Gemini later added a third label to the 36 of these on which the two split on misinformation.

The labels combine as follows. A post is N/A when any labeler marks that its verdict depends on content it was not given, or when it is a viral post whose image or video could not be retrieved. Among the viral posts, Claude Opus 5.5 marked 50, gpt-6.1-sol 44 and Gemini 35, and 129 of the 139 N/A viral posts carry an image or video. Misinformation is the majority verdict on a post with three labels. On a post with two labels it is their common verdict. Every viral post has three labels. On 13 posts below the cut gpt-6.1-sol and Gemini split, and with no third label these take gpt-6.1-sol's labels. A post's category is the one most of the labelers in the majority chose: among false, premature and misleading for a misinformation post, and among accurate and no checkable fact for any other post. A tie goes to Claude Opus 5.5, then gpt-6.1-sol, then Gemini, and the post's other labels, such as sensational and slanted, are those of the first labeler in that order who chose the category. The amplifier tests instead count a post by the share of its labelers who coded

Table 17: The dated record given to the labelers

Fact box	What it records
WHO case counts	Confirmed and suspected cases outside endemic countries by date, from the first WHO count on 21 May 2022
Airborne spread	Spread mainly by close contact; long-range airborne spread not shown; short-range aerosols still under study in late May and early June
Laboratory origin	The 2022 genome a natural West African clade, with no evidence of engineering or release
2021 tabletop exercise	A real preparedness simulation of March 2021, with no evidence that it planned or caused the outbreak
Mutations	A study of 24 June 2022 found faster evolution than expected, not engineering
Other diseases	Monkeypox an orthopoxvirus, unrelated to HIV, shingles, herpes or chickenpox
Who gets it	Spread mostly within sexual networks of gay and bisexual men but not only there; no evidence it was designed to target a group
Children	No evidence it was designed to target children, or targeted them
WHO declaration	The decision on a public health emergency open until 23 July 2022
Pandemic accord	Negotiations at an early stage in May and June 2022
Vaccines	Partial protection from smallpox vaccination; the authorized vaccine; no evidence a vaccine causes monkeypox; WHO advice against mass vaccination of 14 June 2022
A new virus	First identified in 1958 in research monkeys in Denmark; endemic in parts of Africa; earlier spread in the UK in 2018 and 2021
COVID vaccines	No evidence that COVID vaccines or "shedding" cause monkeypox
Renaming	WHO's plan of 14 June 2022 to rename the virus; the name mpox from 28 November 2022
First cases in Chile and Ghana	The dates of official statements and confirmations in each country
UK health service page	The revision of its guidance on 20–22 May 2022
CDC traveler notice	The mask advice added around 6 June 2022 and removed on 7 June
Canada travel advice	Advice, not mandatory quarantine, in June 2022
CDC safer-sex advice	Harm-reduction advice of 15 June 2022

Notes: One row per fact box, in the order of the rules; each box lists dated facts, summarized here. Facts dated after a post count only for what the situation was on its posting date. Source: the rules, version 3.1, which date each fact but carry no links to the documents behind it. The WHO's case count of 21 May 2022, its emergency declaration of 23 July 2022 and the name mpox from 28 November 2022 are in World Health Organization (2022a), World Health Organization (2022b) and World Health Organization (2022c).

Table 18: Labelers of the posts

Provider	Model	Reasoning effort	Viral posts	Posts below the cut
OpenAI	gpt-6.1-sol	low (first run), high (second run)	1,730	1,079
Anthropic	Claude Opus 5.5	high	1,730	419
Google	gemini-4-argon-high[a]	high[a]	1,730	183

Notes: Posts labeled by each model, out of 1,730 viral posts and 1,079 posts below the viral cut. The viral posts were labeled with their images, at high reasoning effort for every model; the first and second runs labeled the posts below the cut from their text alone. [a]On the viral posts and for the third labels below the cut; the model version and reasoning effort of Gemini's first-run labels of 147 posts below the cut were not recorded. Source: labels of 27 October to 2 November 2026.

it so (Appendix C.5). The 513 posts below the cut that only gpt-6.1-sol labeled carry its label.

D.4 Validation and agreement

Table 19 reports every test of the labels that the records hold, each named for what it compares. Panels A and B test version 2.9 of the rules, with labelers other than those that made the paper's labels. On the development set the benchmark is the settled answer: our ruling, an answer left standing on a case we reviewed, or an answer on which the human rater and the earlier labels agreed. Panel A reports development alignment, the share of settled answers that all three labelers matched. These posts shaped the rules, so agreement on them overstates agreement on new posts. On new posts the benchmark is agreement among the labelers, and our rulings on the answers they split on. Agreement among the three labelers fell from 260 of 280 answers on the random draw to 236 of 280 on the draw chosen to be hard.

Panel C reports agreement among the labelers on the paper's own labels, under version 3.1. On the 1,591 viral posts that are not N/A, all three labelers give the same answer on 10,103 of 11,137 answers (90.7%) and the same misinformation verdict on 1,422 posts (89.4%). Their disagreements concentrate on Truth, where all three agree on 1,347 of the 1,591 posts. Panel D gives the same labeler the same inputs twice: Gemini changed its misinformation verdict on 3 of 20 posts. Panel E compares the labelers' majority with the human rater on a holdout of 50 viral posts. The posts are drawn at random from the viral posts that are not N/A and that no development set contains. The rater labels them blind, under version 3.1 of the rules and with the dated record, and the labelers' majority matches her on 84% of answers.

D.5 Worked cases

Table 20 applies the rules to eight posts from the corpus, beyond the four cases of Section 4. Panel A shows the posting-date rule: four reports of the same kind receive three different verdicts according to what the record held on the day. Panel B shows the blocks moving one at a time: in each row one block moves while Truth stays put.

D.6 Network cells: construction and stability

In July 2026 we collected the following lists of 575 accounts. The 575 are every producer of the 2022 collections with a viral post that carries an image or video, 500 producers in all, and 75 other accounts; one of the 75 posted a post found in 2026, so 501 of the 575 are producers of the current corpus. Keeping only follows from one of these accounts to another gives 5,969 follows among 474 accounts. The 139 producers of the 2022 collections whose viral posts are all text-only, and the producers of posts found in 2026 who are not among the 575, have no observed follows in either direction, so they fall in no cell.

The cell measures of Section 3 are computed on this graph. F_j counts the accounts in the graph that follow producer j, and the second-degree arrows run through the same accounts. Reciprocity is computed only for a producer whose own list was returned, and all 240 producers in a cell have one. One of them returned an empty list; such a producer's reciprocity is zero, so it is one-way. The band medians of overlap are taken over the 260 producers with $F_j \geq 3$ on the roster on which the cells were first computed. That roster holds the 506 producers with at least one viral post that labels under version 2.8 of the rules coded as accurate and not misinformation, or as false, premature or misinformation.

Table 21 changes one rule at a time and counts the producers who stay in their cell. Moving the reciprocity cut to the median, 0.33, keeps 239 of the 240 producers in their cell, and a cut at 0.25 keeps 223. A cut at 0.50 keeps 192, because it moves producers with reciprocity between 0.30 and 0.50 from the mutual cells to the one-way cells. Measuring overlap against a log-linear trend in F_j rather than within bands keeps 201. Burt's effective size, divided by the number of followers and computed in the binary form of Borgatti (1997), which asks whether a producer's followers follow one another rather than whether they share followers, keeps 165. Appendix C.6 reports the tests of Section 5 under each of these rules (Table 13, rows (4) to (8)).

Table 19: Tests of the labels

	Posts	Agreeing answers	Answers compared
Panel A: Development alignment, rules version 2.9; settled answers matched by all three labelers			
Run 1	34	109	157
Run 2	34	123	157
Run 3	34	130	157
Run 4	34	133	157
Panel B: New posts, rules version 2.9			
Random draw: all three labelers agree	40	260	280
Split answers we ruled on, matched by:			
Claude Opus 5.5		10	19
Claude Fable 5.1		9	19
gpt-6-astra		13	19
Draw chosen to be hard: all three labelers agree	40	236	280
Split answers we ruled on, matched by:			
Claude Opus 5.5		25	31
Claude Fable 5.1		18	31
gpt-6-astra		9	31
Panel C: The labels used in this paper, rules version 3.1; viral posts that are not N/A, all three labelers agree			
Claude Opus 5.5, gpt-6.1-sol and Gemini: all answers	1,591	10,103	11,137
Truth	1,591	1,347	1,591
Misinformation	1,591	1,422	1,591
All seven answers on the post	1,591	1,023	1,591
Panel D: The same labeler twice on the same inputs, rules version 3.1; Gemini			
All answers unchanged	20	131	140
Misinformation unchanged	20	17	20
Panel E: The human rater on posts the rules have not seen			
Labelers' majority matches the rater: all answers	50	84%	

Notes: Each post has 7 answers: accuracy, misleading, sensational, humorous, slanted, fringe or mainstream, and misinformation. Panel A: the 34 posts the human rater labeled first; labelers Claude Opus 5.5, Claude Sonnet 5.5 and Claude Fable 5.1. In runs 2 and 3 the rules carried 3 examples worded from tested posts; without those answers the counts are 120 and 127 of 154. Panel B: labelers Claude Opus 5.5, Claude Fable 5.1 and gpt-6-astra; the second draw takes the most-retweeted posts that earlier labels called misinformation, fringe or sensational, the most-retweeted posts they called none of these, and text-only posts at random. Panel C: labelers as in Table 18. Source: label and ruling records, 20 October to 2 November 2026.

Table 20: The rules applied to eight posts

Posted	What the post does	Labels	Misinformation
Panel A: Judged on what was known on the posting date			
24 May 2022	Reports a suspected first case in Ahanta West, Ghana, as suspected	Truth accurate: a report the next day shows a suspected case was under investigation, and the post presents it as open	no
25 May 2022	Announces Ghana's first case in Ahanta West as confirmed	Truth premature: the case was still under test	yes
19 May 2022	Announces a detected case in Chile	Truth false: officials reported no suspected case	yes
25 June 2022	Says the WHO is about to announce a worldwide vaccination program	Truth false: nothing in the record stood behind the plan	yes
Panel B: Appeal, Belief and Position move apart from Truth			
3 June 2022	Says the WHO "flipped a 180", quoting both its May and its June statement	Truth accurate; Appeal sensational (exaggeration); Belief not misleading, since both sides of the change are stated	no
23 June 2022	Jokes that a disease spread mainly through sex makes the poster safe	No checkable fact; Appeal humorous	no
19 May 2022	Shares a real 1965 study of smallpox spreading up to nine miles through the air	Truth accurate; Belief misleading: a reader takes away that monkeypox spreads nine miles through the air	yes
21 May 2022	Presents a real 2021 tabletop exercise with a monkeypox scenario starting in May 2022 as proof the outbreak was planned	Truth accurate; Appeal sensational; Belief misleading: the outbreak was staged; Position fringe	yes

Notes: Posts paraphrased. Labels are those the rules give; blocks not listed take their default value (not sensational, not humorous, not misleading, mainstream, not slanted). Ghana confirmed its first cases on 8 June 2022 and Chile on 17 June, so with hindsight both May reports of a first case would look right. Source: posts from the corpus; the rules, version 3.1.

Table 21: Producers who stay in their network cell under other rules

Rule changed	Same cell	G_1	G_2	G_3	G_4
None (one-way if $r_j < 0.30$; overlap against band medians)	240	60	48	55	77
One-way if $r_j < 0.25$	223	54	37	61	88
One-way if r_j is below its median, 0.33	239	61	48	54	77
One-way if $r_j < 0.50$	192	84	72	31	53
Overlap against a log-linear trend in F_j	201	58	50	52	80
Burt's effective size in place of overlap	165	80	28	42	90

Notes: The 240 producers with $F_j \geq 3$ and labeled posts that are not N/A. Same cell: producers in the same cell as under the rules of Section 3. Columns G_1 to G_4: producers in each cell under the changed rule. The log-linear trend and the effective size are fitted over the 260 producers on which the band medians are taken; under the effective size a producer is overlapping when its followers are more connected than the trend predicts. Source: following lists collected in July 2026.

D.7 Producer post panel

To date each producer's first post on the outbreak apart from the labeled posts, we searched in October 2026 for the posts on the outbreak of 647 accounts: the 639 producers of the 2022 collections and 8 others, the accounts whose followers were sampled for the 2022 engagement records (Section 3) and that posted none of the viral posts. The search covered five windows from 1 May to 1 July 2022, with the outbreak's keywords in English, Spanish, Dutch and Italian and the vaccine names. A window counts as covered when the search ran to its end, and a window with no post counts as covered only when a separate check found the account active and searchable. Table 22 gives the coverage by window, and all five windows are covered for 517 of the 647 accounts. The search finds 695 of the 1,022 viral posts of the 2022 collections, and 184 of the 301 of these posted in the Breaking Window. In November 2026 the same search covered the 318 producers of the posts found in 2026 who are not among the 647, for the checks of Appendix C.6. The panel dates entry only in the robustness checks of Section 5.

Table 22: Coverage of the producer post panel

	1–15 May	15–29 May	29 May–12 Jun	12–26 Jun	26 Jun–1 Jul
Covered, with a post on the outbreak	36	447	292	262	162
Covered, no post	501	98	251	280	370
Not covered	51	43	45	46	56
Account no longer available	59	59	59	59	59
Accounts	647	647	647	647	647

Notes: Accounts by search window, 2022. Covered: the search ran to its end; a window with no post is covered only when the account was found active and searchable. Not covered: the search could not reach the account's posts, or ended early. Search windows differ from the windows W1 to W4 of the analysis. Source: searches of October 2026.

D.8 Limitations of the records

The follow graph is read from following lists of July 2026. A follow may postdate the outbreak, and a follow that ended before 2026 is missing. Of the 575 lists, 84 returned no account at all, as a protected, deleted or suspended account does; such a producer's followers among the producers are still observed, but its own follows are not. By 2026 the accounts of 58 of the 958 producers had been deleted or suspended.

The text of the viral posts of the 2022 collections was collected in 2022, so the corpus keeps posts deleted since then. Every record collected in 2026 misses them: the posts below the viral cut, the producer post panel, and the lists of accounts that retweeted each post. A post found in 2026 enters only if it could still be read in 2026. Of the 2,181 posts the network tests start from, 2,075 returned a list of retweeting accounts in October or November 2026 and 106 were unavailable.

Every share in the paper refers to posts about the outbreak on Twitter in 2022, viral or below the viral cut. The records describe what producers posted and how accounts engaged with these posts on that platform, and nothing about other platforms or about readers who left no trace.