Misinformation as Lemons: How Credibility Unravels on Social Media

Yizhou Jin*
University of Toronto

November 4, 2026

Preliminary draft. Please do not cite or circulate.

Click here for the newest version

Abstract

This paper studies how information asymmetry between producers and readers drives selection in the production and amplification of social-media posts, and how network structure shapes that selection. We first develop an empirical framework with selection on two margins: producers with heterogeneous accuracy and amplifiers with different demand for credibility. We then micro-found this demand in a network sharing game, linking it to two network features: reciprocity and overlap. Empirically, we collect viral posts about the 2022 mpox outbreak on Twitter and develop a large-language-model protocol that scales fact-checking of each post while aligning closely with human judgment. Our empirical tests demonstrate adverse selection on both production and amplification margins within reciprocal or high-overlap networks, and advantageous selection in one-way and diverse (low-overlap) networks.

*Rotman School of Management, University of Toronto (yizhoujin.com). I thank participants at the University of Toronto Summer IO Seminar and the Rotman Strategy Brown Bag for helpful comments and suggestions. I am grateful to Nimrat Grewal, Ishaan Gupta, Athena He, and Jingwen (Steven) Shi for excellent research assistance, and especially for their dedicated work on tweet labeling and verification. Financial support from the University of Toronto is gratefully acknowledged.

1 Introduction

On social media, the credibility of a post is often hard for readers to discern, while the producer has superior knowledge of whether the information it contains has been properly verified. This paper studies how this information asymmetry drives selection in the production and amplification of posts, and how network structure shapes that selection.

We aim to contribute to three gaps in the literature. The first gap is theoretical. 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 on the social network.[2] An important reason is that the social context requires modeling homophily, as 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 is in measurement. 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.[4] 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 is empirical. The industrial organization literature on social media has focused on the production, consumption, and advertising incentives without taking a stand on accuracy (Filippas et al. 2025; Beknazar-Yuzbashev et al. 2026). Research on social-media misinformation, by contrast, studies the demand side, often through 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

[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).

[4] Unflagged vaccine-skeptical content had an estimated effect on vaccination intentions 46-fold that of flagged misinformation (Allen et al. 2024).

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.

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. This heterogeneity generates a double selection margin on both production and amplification. Below the certifier threshold, only social readers amplify, which removes certifiers' contribution to the credibility premium and sustains a low-credibility *Belief Equilibrium*; above the threshold, certifiers amplify as well, which can sustain a high-credibility *News Equilibrium* when certifiers are sufficiently prevalent and responsive.

Third, we propose a simple network sharing game to link network features we observe in the data to amplification demand. 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. In a benchmark with equal influence across readers, reciprocal followers give more weight to social value, while one-way followers and further readers give more weight to accuracy. Holding follower and second-degree arrow counts fixed, higher reciprocity raises the social share of demand, and greater overlap can raise it by reducing the audience of further readers. The equilibrium reward therefore links both network measures to how strongly amplification responds to credibility. Producer pools and their accuracy profiles along the cost order may also differ across these network cells.

Our empirical setting is the 2022 mpox outbreak on Twitter: in May 2022 a long-known disease spread across many countries at once, and by 21 May, 12 countries where it is not endemic had reported cases to the WHO. Across the news cycle, from the first viral post in our data on 7 May 2022 to the last on 30 June 2022, 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. We label every viral post in our corpus, 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. 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 and refined them in a development cycle,

adding a human rater's rulings on disagreements until all three language models matched the rater on 85% of the settled answers. We apply the final rules with three language models from different families, each told to use only the post and the dated record; on a random holdout of 50 viral posts labeled blind by the rater, the models' majority matches the rater on 84% of answers. The remaining split posts feature genuine disagreements even among human raters due to differences in priors and unclear ground-truth timeline. We rule based on the majority. 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.

Our empirical focus is networks of influencer-producers who posted viral tweets during the news cycle, whose mutual amplification was critical to virality. We have 2.00 million follow-graph edges to 1.46 million distinct accounts. The 1,995 (1,037 viral) labeled posts drew 1.49 million retweets, quote tweets and replies. The producers' follow graph places each producer whose followers include at least three accounts in one of four networks by the reciprocity—how much producers follow their followers—and overlap—how much her followers share the same (2-hop) followers: one-way and diverse (G_1), one-way and overlapping (G_2), mutual and diverse (G_3), and mutual and overlapping (G_4). Within a network of connected producer-influencers, we assume that they cannot readily differentiate the content veracity between their peers. We read the Breaking Window, 7 to 20 May 2022, which ends as the CDC issued a health advisory on the outbreak to U.S. clinicians (Centers for Disease Control and Prevention 2022b), 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 network.

In the one-way diverse network, producers who post in the Breaking Window, while checking facts is costly, are more accurate than producers who first post after it; in the mutual overlapping network, they are less accurate. 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 network by $+0.14$ in G_1 (SE 0.06) and by -0.25 in G_4 (SE 0.08). Read through the model, the contrast is consistent with advantageous selection in the producer pool of G_1 and adverse selection in that of G_4. On the amplification margin, 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). Misleading posts in G_4 also 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. 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 networks.

The equilibrium relationship between reach and credibility also reverses across networks. Among producers who post in the Breaking Window, 94% of amplification goes to accurate posts in G_1, against 32% in G_4. Across the four windows, a producer's reach, the log of one plus retweets and quote tweets, moves with its network's accurate share by $+3.73$ in G_1 (SE 1.36) and by -7.20 in G_4 (SE 1.83), positive in one-way diverse networks and negative in mutual overlapping networks. Vosoughi et al. (2018) find that falsehood diffused "farther, faster, deeper, and more broadly" than the truth; in our data, that pattern holds only in some parts of the network. Within a producer, net of the date of posting, false or premature posts reach more than the same producer's accurate posts by $+0.85$ in G_4 (SE 0.50, $p = 0.097$), significant at the

10 percent level, and by -0.13 in G_1 (SE 0.53). In G_4, the reach of a producer's accurate but misleading posts differs from that of her accurate posts by $+1.30$ (SE 0.63). Neither G_4 estimate is distinguishable from zero once the posts' sensational and slanted labels are held fixed. This is consistent with the model, in which readers respond to what they can see of a post and not to accuracy they cannot see (Proposition B.6).

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 network (or pooling 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. Section 2.1 sets up a baseline with adverse selection and shows when equilibrium amplification and credibility co-move or diverge; Section 2.2 introduces heterogeneous readers and double selection; and Section 2.3 embeds the market in a network sharing game.

We draw the market as Einav and Finkelstein (2011) draw insurance markets. In Figures 1, 2 and 3, 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 A baseline with adverse selection

2.1.1 Producers and the marginal cost of posting

Figure 1(a) 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. Low-accuracy (*slop*) producers stand at the front because they skip the verification cost, and the curve jumps by that cost at q_0, where high-accuracy (*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 a low-accuracy (slop) producer and one for a high-accuracy (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 2.1.2). Above q_0, the producer who would post next is more accurate than the pool. Selection is adverse.

A unit mass of potential producers can post about the event, and each producer j is one of two types. A high-accuracy (*investigative*) producer ($\alpha_j = 1$) checks primary facts before posting, and her posts are accurate. A low-accuracy (*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 a base cost κ_0, an investigative verification cost $\kappa_1 > 0$ when $\alpha_j = 1$, and an idiosyncratic cost draw ε_j:

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

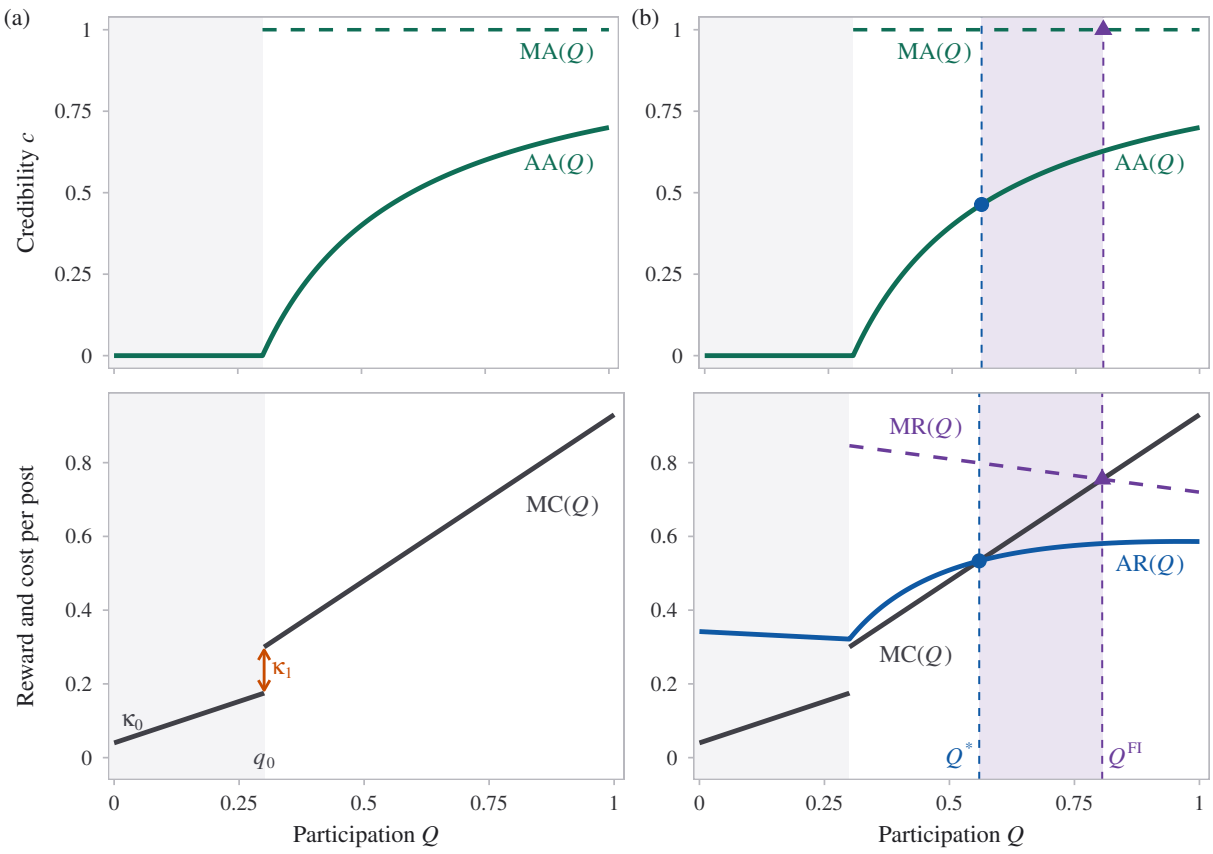


Figure 1: Supply and equilibrium. (a) Producers lined up by their cost of posting. Bottom panel: the marginal cost MC(Q) of the producer at rank Q, which jumps by the verification cost at q_0, where investigative producers begin. Top panel: the marginal accuracy MA(Q) of the producer at rank Q and the average accuracy AA(Q) of the producers who post up to Q. (b) 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 curves of panel (a), with the equilibrium credibility at Q^*. The grey bands mark the slop producers, $Q \leq q_0$. The curves are drawn at illustrative parameter values (Appendix A.3).

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. Lining up producers from the lowest to the highest cost along $Q \in [0,1]$ gives the marginal cost curve MC(Q), the cost of the Q-th producer in line; the index Q ranks producers by cost and is not time.

Assumption 1 (Marginal cost and separated types) *(i) In the baseline, every investigative producer's cost exceeds every slop producer's cost, so slop producers occupy* $[0, q_0]$ *and investigative producers* $(q_0, 1]$. *(ii) The marginal cost curve* MC(Q) *is continuous, strictly increasing, and convex within each type and jumps up at the boundary between types.*

The cost-shape restrictions in Assumption 1(ii) are maintained throughout this section; the specialist variant below changes which accuracy type comes first. In our tractable linear benchmark, where ε_j is spread uniformly within each type and κ_1 exceeds the noise range, the marginal cost curve takes the piecewise-linear form[5]

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

where σ tracks the spread of the cost noise within a type and the jump $\kappa_1 > 0$ at q_0 reflects separated verification costs.

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 B.1 and Proposition B.1). 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.2). At a single rank, selection is adverse exactly where the average accuracy of the producers up to that rank rises (Proposition B.1(iii)). Credibility is exactly zero with positive participation only if a positive mass of producers is uninformative (Proposition B.3). 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.4), and three equilibria, two of them stable, can coexist when accuracy is smooth and the marginal cost is linear (Proposition B.5).

2.1.2 Readers, the network 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 network cell (G, s)—which we also call a *network* when holding the public signal s fixed—collects the posts that readers cannot tell apart on this public information.

Assumption 2 (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,

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

[5] 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 $w_1 = \sigma(1 - q_0)$.

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

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

The visibility multiplier $\tau(Q)$ captures the attention a post receives, and congestion lowers it at rate τ_1 as more producers post.[6] Baseline appeal x_0 captures social value and sharing for reasons other than accuracy (such as belief alignment or humor), and accuracy scrutiny x_1 captures the consumption or decision value that drives extra sharing of more credible posts.

Assumption 3 (Reward and variety) *Each post reports a different development of the news cycle. A reader's standalone payoff from amplifying it depends only on its pool credibility c; other posts provide no informational substitution and affect its reward only by competing for attention through $\tau(Q)$. The reward parameters satisfy $x_0 > 0$, $x_1 > 0$, and $\tau_0 > \tau_1 \geq 0$.*

A producer values each amplification of her post at one unit (for example, from out-of-network exposure and follower growth), 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 posts report a binary state symmetrically, with prior 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 2.3 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.

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

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

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

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

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

[6] 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).

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 1(b), it is the vertical distance between the marginal and the average reward above q_0.

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

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

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.

2.1.3 Equilibrium and two forms of unraveling

Panel (b) of Figure 1 adds the two reward curves to the bottom panel. Under pooling, entry stops where the average reward AR meets the marginal cost MC. The intersection of the marginal reward MR and marginal cost MC defines the full-information entry benchmark Q^{FI}. It equals total participation under full information when all slop producers still cover their cost there, as in Proposition 1(ii). The shaded band marks the investigative producers ranked between the two points.

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

Definition 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, $c^* = \mathrm{AA}(Q^*)$, and, at the reward $R^* = R(Q^*, c^*)$, every producer with $\mathrm{mc}_j < R^*$ posts and every producer with $\mathrm{mc}_j > R^*$ 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 $\mathrm{MC}(0)$ at $Q = 0$.*

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

Lemma 1 (Equilibrium conditions) *Let the reward $R(Q,c)$ be continuous in Q and nondecreasing in c, as in equation (5), and let $\mathrm{AR}(Q) = R(Q, \mathrm{AA}(Q))$. Under Assumption 1, $(Q^*, \mathrm{AA}(Q^*))$ 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 $\mathrm{MC}(q_0^-) \le \mathrm{AR}(q_0) \le \mathrm{MC}(q_0^+)$*; (iii)* $Q^* = 1$ *and* $D(1) \ge 0$*. Silence is an equilibrium if and only if* $R(0,0) \le \mathrm{MC}(0)$*, which under equation* (5) *is* $\tau_0 x_0 \le \mathrm{MC}(0)$.

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 their local stability using entry incentives on either side. Equilibria, including those at the cost jump, are rest points of the adjustment. Away from equilibria, the cost jump and full entry, participation rises when $D(Q) > 0$ and falls when $D(Q) < 0$. An equilibrium $Q^* \in (0,1)$ is stable when $D(Q) > 0$ just below Q^* and $D(Q) < 0$ just above Q^*, so that local adjustment returns participation to Q^*. Full entry ($Q^* = 1$), which participation cannot exceed, is stable when $D(Q) > 0$ 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 1 (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 $\mathrm{MC}(q_0^-) \le \tau(Q^{\mathrm{FI}})\, x_0$*, it is an equilibrium for the producers with the* Q^{FI} *lowest costs to post.*

The full-information entry benchmark Q^{FI} is the cost rank at which the marginal reward equals the marginal cost. 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 2 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 (Unraveling and multiple equilibria) *Under Assumption 1, suppose slop producers cover their cost at zero credibility,* $\mathrm{MC}(q_0^-) \le \tau(q_0)\, x_0$.

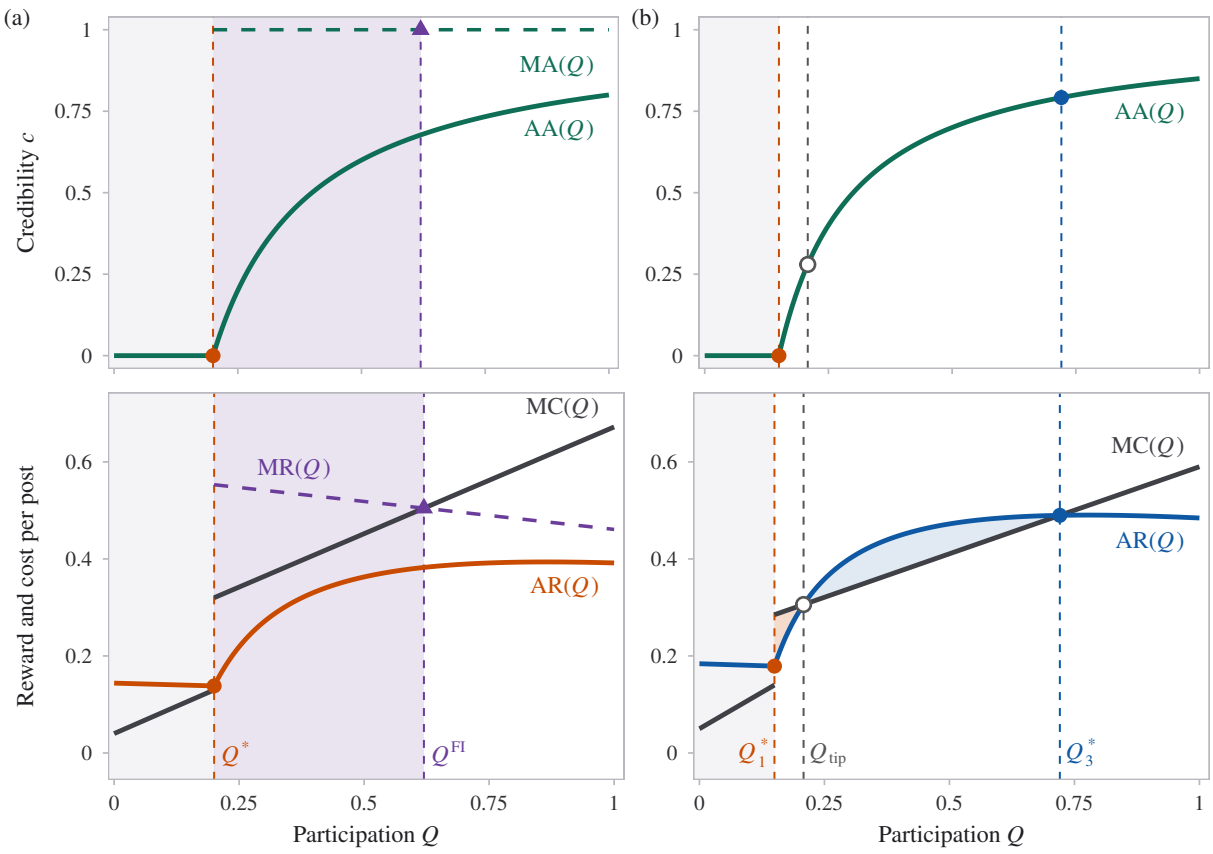


Figure 2: Two forms of unraveling. (a) Pooling prices out investigative producers: the average reward AR lies below the marginal cost MC over every investigative producer, so only slop posts and $Q^* = q_0$. The marginal reward MR lies above MC up to the full-information entry benchmark Q^{FI}; the shaded 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).

(i) Pooling prices out investigative producers. *If* $\mathrm{AR}(Q) < \mathrm{MC}(Q)$ *for every* $Q \in (q_0, 1]$, *the unique equilibrium is* $Q^* = q_0$, $c^* = 0$, *with* $R^* = \tau(q_0)\,x_0 > 0$. *This holds whenever* $\mathrm{AR}(Q) < \mathrm{MC}(Q) < \mathrm{MR}(Q)$ *for every* $Q \in (q_0, Q^{FI})$ *under the hypotheses of Proposition 1; investigative producers then cover their cost under full information but not under pooling.*

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

(iii) Entry barrier and multiple equilibria. *If* $\mathrm{AR}(q_0) = \tau(q_0)\,x_0 < \mathrm{MC}(q_0^+)$ *and* $\mathrm{AR}(Q_a) > \mathrm{MC}(Q_a)$ *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* $Q_{tip} \in (q_0, Q_a)$; *and a high-credibility equilibrium* $Q_3^* \ge Q_a$, *interior if* $D(1) < 0$ *and*

full entry otherwise, with $c_3^* > \mathrm{AA}(Q_{\mathrm{tip}}) > 0$. The low- and high-credibility equilibria are stable and the tipping point is not: D is positive on $(0, q_0)$ and negative on (q_0, Q_{tip}), so adjustment approaches q_0 from either side, while participation starting above Q_{tip} moves toward Q_3^*.

In part (i), pooling prices out every investigative producer while baseline appeal still covers the cost of slop: participation unravels to q_0 and credibility to zero, yet slop still goes viral ($R^* = \tau(q_0)x_0 > 0$). 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. In part (iii), the cost jump at q_0 acts as an entry barrier because the first investigative producer costs more than the reward at zero credibility, while the composition lift raises AR above MC further to the right; the low-credibility equilibrium is locally stable because an investigative producer who posts just above q_0 loses money, and the market reaches Q_3^* only from participation above Q_{tip}. Numerical result A.1 gives parameter values for each case of Proposition 2, including cases with $\tau_1 = 0$: neither form of unraveling needs congestion.

2.1.4 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 high-accuracy *specialists*, who are accurate ($\alpha_j = 1$) and make up the first share q_1 of the cost order, and lower-accuracy *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; under the cost-shape restrictions of Assumption 1(ii), Lemma 1 holds with q_1 in place of q_0 and $\tau_0(x_0 + x_1\alpha_g) \leq \mathrm{MC}(0)$ as the condition for silence.

Proposition 3 (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.[7]

Proposition 3 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 4 (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 the cost-shape restrictions of Assumption 1(ii), with q the first type's share, 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, 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 4 compares the pool that posts while checking is costly with the producers who post only once it is cheap. 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 each producer keeps her rank within her type as required by Proposition 4(iii). Every producer who posts while checking is costly then also posts once it is cheap, and 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. Figure 11 in Appendix A.3 draws the two pools on the average accuracy curve.

2.2 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.

[7] Appendix A.2 (Proposition A.1) further characterizes how shifts versus rotations in readers' underlying payoffs (u^+, u^-) move equilibrium participation and credibility.

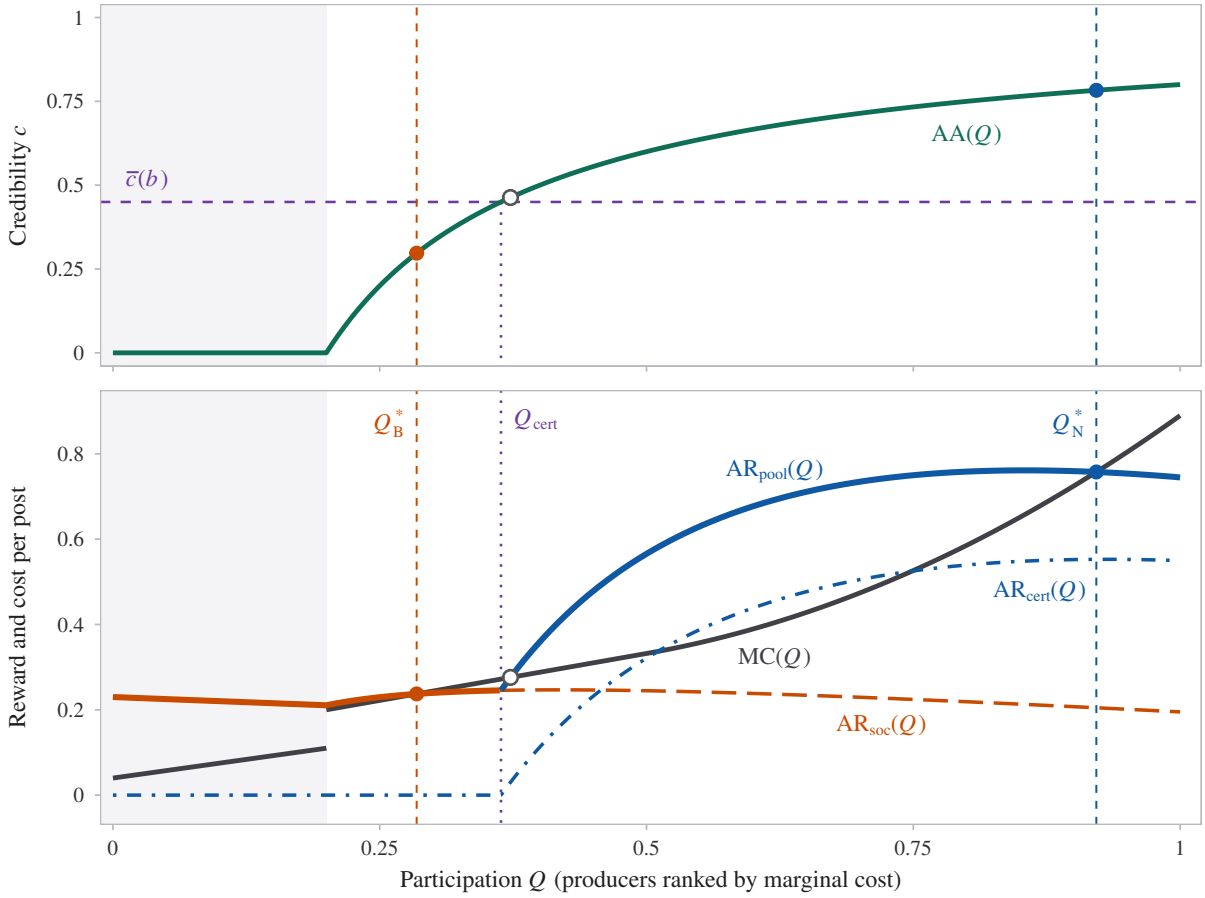


Figure 3: Two kinds of readers and two equilibria. Bottom panel: the reward from social readers, who share for appeal, the reward from certifiers, who amplify only once the pool's credibility clears their prior threshold at Q_{cert}, their sum AR_{pool}, and the marginal cost MC. The pooled reward meets 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).

Double selection is selection on production, who posts, and on amplification, who amplifies.

Figure 3 draws two kinds of readers, defined below. Social readers share for social value and baseline appeal, and their reward is nearly flat in credibility. Certifiers share a post for its consumption 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 (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

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

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 4 (Social readers and certifiers) *A cell has two kinds of readers, and the reward is the sum of their amplification,* $\mathrm{AR}_{\mathrm{pool}} = \mathrm{AR}_{\mathrm{soc}} + \mathrm{AR}_{\mathrm{cert}}$. *Social readers share for social value and baseline appeal:* $\mathrm{AR}_{\mathrm{soc}}(Q) = \tau(Q)\left[x_{0,\mathrm{soc}} + x_{1,\mathrm{soc}}\mathrm{AA}(Q)\right]$, *with* $x_{0,\mathrm{soc}} > 0$ *and* $x_{1,\mathrm{soc}} \geq 0$. *Certifiers share a post for its consumption value to a decision about the development it reports, which no other post settles (Assumption 3): they hold a common prior* $b \neq 1/2$ *of their own, while the state's prior stays one half, and* $\mathrm{AR}_{\mathrm{cert}}(Q) = \tau(Q)x_{1,\mathrm{cert}}\left[\mathrm{AA}(Q) - \bar{c}(b)\right]_+$, *with* $x_{1,\mathrm{cert}} > 0$. *Both depend on a post only through the cell's credibility.*

Social readers share mainly for appeal: $x_{1,\mathrm{soc}}$ is small, and their reward curve is nearly flat in credibility. Certifiers' reward is proportional to the decision value of Lemma 2; the proportionality is an assumption. In the baseline, certifiers stay out until average accuracy reaches $\bar{c}(b)$, at $Q_{\mathrm{cert}} = q_0/(1-\bar{c}(b))$, and the pooled reward kinks up there.

Proposition 5 (Double selection) *In the baseline with the readers of Assumption 4 and a marginal cost that satisfies Assumption 1, suppose* $Q_{\mathrm{cert}} < 1$ *and*

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

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

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

Then there are exactly three equilibria: a Belief Equilibrium $Q_B^* \in (q_0, Q_{\mathrm{cert}})$ *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. Below the certifier threshold $\bar{c}(b)$, the feed changes no certifier's decision and only social readers amplify, which removes certifiers' contribution to the credibility premium and sustains a low-credibility Belief Equilibrium: when social readers' baseline appeal is moderate ($\mathrm{MC}(q_0^-) \leq \tau(q_0)x_{0,\mathrm{soc}} \leq \mathrm{MC}(q_0^+)$, as in Lemma 1(ii)), only low-accuracy (slop) producers post ($Q_B^* = q_0, c_B^* = 0$), whereas when it exceeds $\mathrm{MC}(q_0^+)$ (condition (a) above), some low-cost investigative producers post as well ($Q_B^* \in (q_0, Q_{\mathrm{cert}})$). Above $\bar{c}(b)$, certifiers join and the pooled reward rises steeply with credibility, sustaining a high-credibility News Equilibrium when certifiers are sufficiently prevalent and responsive for condition (c) to hold. 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).

2.3 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. The cell's producer pool sets the shape of the average accuracy curve: it rises with participation where low-accuracy (slop) producers have the lowest costs, as in the baseline, and falls where accurate specialists do, as in the specialist variant of Section 2.1.4. 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 (Figure 3).

Supply: the producer pool. 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 producer pool and cost order, specified separately from the reciprocity-and-overlap demand benchmark below. Selection is adverse where low-accuracy (slop) producers have the lowest costs, as in the baseline, and advantageous where accurate specialists do, as in the specialist variant.

Demand: the reader mix. Second, two features of a producer's followers—reciprocity and overlap—enter the microfoundation of the reward and shape both reach $\tau(Q)$ and the ratio of baseline social appeal x_0 to accuracy scrutiny x_1 (Figure 4). 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.

A sharing benchmark links the network's reciprocity and overlap to demand through the N readers whom the cell's posts reach (Appendix A.4). Index the active posts in the cell by $j \in [0,Q]$. Each reader $k \in \{1,\dots,N\}$ chooses how much to share each post, $m_{kj} \geq 0$, to maximize

$$U_k = \int_0^Q m_{kj}\Big[B_k^A(c) + \sum_{k'} \omega_{kk'} m_{k'j}\Big]\,dj \;-\; \frac{1}{2}\int_0^Q m_{kj}^2\,dj \;-\; \frac{\nu}{2}\Big(\int_0^Q m_{kj}\,dj\Big)^2. \tag{15}$$

The first bracketed term is reader k's marginal return from sharing post j: it combines her stand-alone expected payoff from amplifying a post of credibility c,

$$B_k^A(c) = x_{0k} + x_{1k}c, \qquad x_{0k} = \bar{a}_k + \frac{u_k^+ + u_k^-}{2}, \qquad x_{1k} = \frac{u_k^+ - u_k^-}{2}$$

(Lemma A.1, where $\bar{a}_k$ is social appeal unrelated to accuracy and u_k^+, u_k^- are the payoffs when the post is accurate and inaccurate), with a peer complementarity $\omega_{kk'} \geq 0$ that is positive when reader k follows reader k' and zero otherwise (collected in the matrix Ω). The last two terms in equation (15) are a quadratic per-post sharing cost and a convex attention cost at rate $\nu \geq 0$ on total sharing across the Q posts in the feed. With the vector $\mathbf{B}(c)$ collecting readers' values, equilibrium sharing of each post is $[(1+\nu Q)I - \Omega]^{-1}\mathbf{B}(c)$

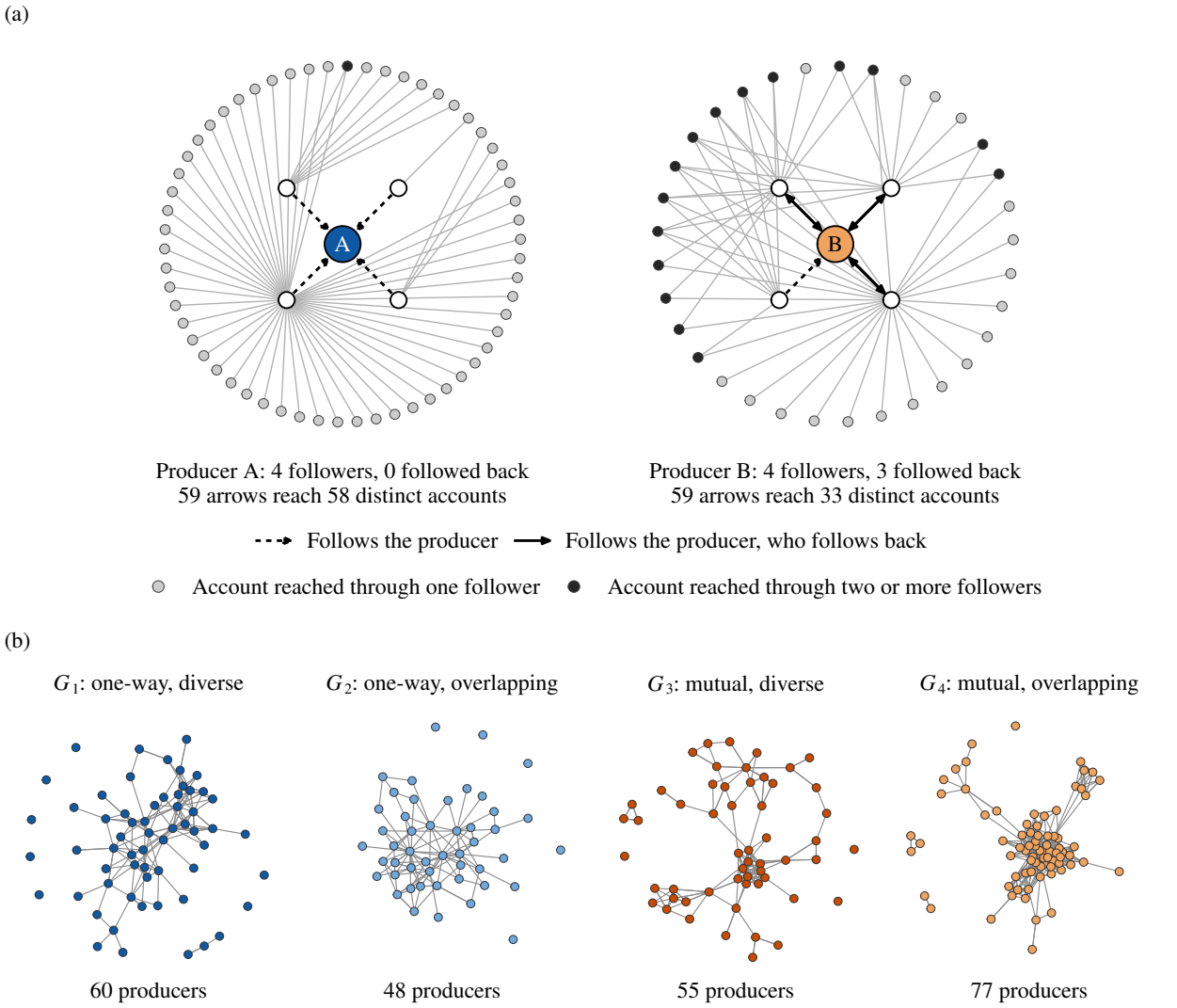


Figure 4: Reciprocity, overlap and the four network cells. (a) Two producers, each followed by four of the accounts whose following lists we collected. Producer B follows back three of her followers and producer A none, so reciprocity is high for B and low for A. The followers of producer A are followed 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 network's producers; each network has its own force-directed layout, in which position shows only who is near whom. Reciprocity and overlap are defined in equations (20) and (21).

(Lemma A.5). For a benchmark, assume each reader's sharing has the same total influence on the audience: $\mathbf{1}^\top \Omega = \omega_A \mathbf{1}^\top$, with $0 \leq \omega_A < 1$. This condition gives equal weights to readers' standalone values when their equilibrium sharing is added up. The reward is then

$$R(Q,c) = \tau(Q)\left(x_0 + x_1 c\right), \qquad \tau(Q) = \mathbf{1}^\top \left[(1+\nu Q) I - \Omega\right]^{-1} \mathbf{1}, \tag{16}$$

where x_0 and x_1 are the audience averages of x_{0k} and x_{1k} under the equal-influence condition. Here $\tau(Q)$ is the aggregate amplification multiplier of the reached audience. At $Q = 0$, baseline reach $\tau_0 = \tau(0) = \mathbf{1}^\top (I - \Omega)^{-1}\mathbf{1}$ is the sum of readers' Bonacich centralities, and feed visibility $\tau(Q)$ falls with Q because attention spent on one post raises the shadow cost $\nu \int_0^Q m_{kj'}\, dj'$ on every other post (linearized as $\tau_0 - \tau_1 Q$ in equation (5)).

Reciprocity and overlap determine the demand mix in a two-kind benchmark of this game. Fix a representative producer j with $F_j > 0$ direct followers and $E_j > 0$ second-degree arrows to $L_j = (1 - O_j)E_j > 0$ distinct further readers. All posts in the cell face this audience and reader mix. We assume reciprocal direct followers are social readers, while the other direct followers and the further readers are news readers. Their values are $B^A_{\text{soc}}(c) = x_{0,\text{soc}} + x_{1,\text{soc}}c$ and $B^A_{\text{news}}(c) = x_{0,\text{news}} + x_{1,\text{news}}c$. Social readers place relatively more weight on social value; news readers can also share for appeal.

Proposition 6 (Reciprocity, overlap and demand) *Under the equal-influence condition, suppose* $x_{0,\text{soc}} \geq x_{0,\text{news}} > 0$ *and* $0 \leq x_{1,\text{soc}} \leq x_{1,\text{news}}$*, with* $x_{1,\text{news}} > 0$ *and at least one of the two comparisons between kinds strict. The equilibrium reward has the form of equation* (16)*, with*

$$\vartheta_{\text{soc}} = \frac{F_j r_j}{F_j + E_j(1 - O_j)} = \frac{r_j}{1 + (E_j/F_j)(1 - O_j)}, \tag{17}$$

$$\tau(Q) = \frac{F_j + E_j(1 - O_j)}{1 - \omega_A + \nu Q}, \tag{18}$$

$$x_\ell = (1 - \vartheta_{\text{soc}})x_{\ell,\text{news}} + \vartheta_{\text{soc}} x_{\ell,\text{soc}}, \qquad \ell \in \{0, 1\}. \tag{19}$$

Holding participation, credibility, audience counts F_j, E_j*, peer influence* ω_A*, attention cost* ν *and reader values fixed, the proportional response to credibility,* $\partial \log R / \partial c$*, falls with reciprocity and weakly falls with overlap, strictly with overlap when* $r_j > 0$*.*

The effective social share ϑ_{soc} increases with reciprocity because more direct followers share for the relationship. It also weakly increases with overlap because fewer distinct further news readers enter the audience. At given audience counts and peer influence, demand in the one-way, diverse corner is therefore more responsive to credibility, and demand in the mutual, overlapping corner gives relatively more weight to social value. These are the two demand dimensions used to divide the empirical networks. Overlap lowers the number of further readers; the peer-influence parameter governs the strength of congestion.

The prior-threshold benchmark of Section 2.2 isolates readers' direct decision value of information. The sharing game here studies affine values and peer reinforcement; news readers may value accuracy while also sharing for appeal or in response to peers. The exact visibility in equation (18) gives the affine visibility of the baseline to first order in participation (Appendix A.4).

In the prior-threshold benchmark, the reader mix selects who amplifies, and which readers amplify depends in turn on the credibility the cell's producer pool delivers (Proposition 5). However, 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, the reader mix alone cannot reverse the sign of supply-side selection

(Proposition B.2 and Corollary B.1).[8] Within a cell, readers reward every producer alike (Lemma A.2), so a shift in reader preferences moves the reward level and slope without reordering producers by cost. Under the monotone-risk condition, opposite signs of supply selection across networks therefore require different accuracy profiles along their producer cost orders.

Section 5 takes these predictions to the four networks G_1 to G_4, holding a producer's public signal fixed through profile controls or within-producer comparisons. First, Section 5.1 tests the supply-selection margin (Proposition 4) by comparing the accuracy of the producer pool that posts during the Breaking Window, while verification is costly, against producers of the same network who enter only afterward. Second, Section 5.2 tests the amplifier-selection margin (Proposition 5) by comparing the subsequent accuracy of Breaking-Window versus later amplifiers, the accuracy elsewhere of accounts that amplify misleading or false posts, and the stance of quote tweets and replies across networks. Third, Section 5.3 examines how the equilibrium accuracy share of amplification and the across-window progression and co-movement of producer reach with network accuracy (Proposition 3) vary from G_1 (one-way, diverse) to G_4 (mutual, overlapping). Finally, Section 5.4 tests the within-producer information-asymmetry benchmark (Proposition B.6) by comparing the reach of a producer's false or premature posts and accurate-but-misleading posts relative to her own accurate posts, before and after conditioning on visible packaging cues.

3 Data

We study one news cycle: the 2022 mpox outbreak on Twitter.[9] 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 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.

We divide the cycle into four 14-day windows, W1 to W4, starting on 7 May 2022 (Table 1). 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

[8] In textbook insurance markets, 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 is non-monotone across ranks, the reader mix can also affect the sign by shifting where along the cost order the equilibria fall (Appendix B.5; see Einav and Finkelstein 2011).

[9] 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 2022b). The keyword search that defines our data uses the 2022 names.

Table 1: 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).

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.

Posts. The core data are 1,730 viral posts about the outbreak by 958 accounts, which we call producers. A post enters if it is an original post (not a retweet), matches the outbreak's keywords in English or another language or names a monkeypox vaccine, and exceeds 1,000 retweets or 100 quotes or replies by distinct users; each record carries its retweet, quote, and reply counts. We scraped these posts and cross-checked our collection against the public 2022 mpox Twitter datasets of Edinger et al. (2023) and Movahedi Nia et al. (2023), adding missing viral posts and repairing corrupted or incomplete records (Appendix D.1). To ensure that a producer's measured accuracy rests on more than its viral hits alone, we also scraped 1,079 posts on the outbreak by 197 of the same producers that fall below the virality cutoff between 7 May and 1 July 2022 (Appendix D.1). Together with the viral posts, our labeled sample comprises 2,809 posts (Section 4).

Amplification records. Across the 1,995 (1,037 viral) non-N/A labeled posts analyzed in Section 5, per-post counts total 1.21 million retweets and quote tweets, which form our post-level reach outcome. Three sets of public records capture individual amplifications and reactions on these posts (Appendix D.8). First, we collected 31,048 quote tweets and 51,107 replies, and the three language models label whether each quote tweet or reply endorses the original post or calls it out (Section 5.2). Second, a 2022 engagement panel recorded, with timestamps, every like, retweet, quote, and reply by a 20% sample of the followers of tracked accounts (drawn by user identifier): 88,354 records by 57,237 sampled followers across 56 high-engagement media posts (Appendix D.1). Third, we collected post-level retweeter lists covering 615,291 retweets of 1,833 (1,009 viral) posts by 302,634 distinct accounts (Appendix C.5); the amplifier regressions in Section 5.2 combine quote tweets, the 2022 panel's retweets, and retweets from amplifiers' timelines into one record per amplifier and post.

The follow graph and four networks. To measure audience structure, we collected the following lists of 575 influencers responsible for the most viral posts (2.00 million follow edges to 1.46 million distinct accounts; Appendix D.6). These producers are major influencers, with a median of about 183,000 followers among the 414 with known counts, and the 5,969 directed follows among 474 of the listed accounts form the producer follow graph. We place each producer in one of four networks using the two empirical proxies motivated in Section 2.3 (Figure 4). Let F_j denote the number of accounts whose collected following lists

include producer j. Reciprocity is the share of those followers that j follows back:

$$r_j = \frac{\text{followers of } j \text{ that } j \text{ follows back}}{F_j}. \tag{20}$$

A producer is one-way if $r_j < 0.30$ and mutual if $r_j \geq 0.30$; the cutoff 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:

$$O_j = 1 - \frac{\text{distinct second-degree accounts}}{\text{second-degree arrows}}. \tag{21}$$

We classify each producer relative to peers of similar degree F_j: 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; 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 yields four networks: G_1 (one-way and diverse), G_2 (one-way and overlapping), G_3 (mutual and diverse), and G_4 (mutual and overlapping).

The empirical tests in Section 5 use the 1,995 (1,037 viral) non-N/A labeled posts by 402 of these 448 producers (Appendix D.1; Appendix C.6 reports robustness to including the 403 viral posts by the 332 producers outside the follow-graph frame). Of these producers, 240 have $F_j \geq 3$ and belong to one of the four networks (60 in G_1, 48 in G_2, 55 in G_3, and 77 in G_4), while 162 have $F_j < 3$. The four networks differ sharply in their broader topology: the median producer in G_1 is followed by 13.5 listed accounts and follows 4, whereas producers in G_4 form a dense cluster in which 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 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 2 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

Table 2: 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.

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. Four posts from the corpus show the four blocks at work (Figure 5); 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.

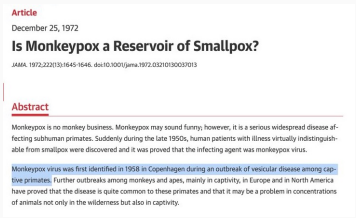
The last two posts in Figure 5 are not misinformation. Case 3, posted 23 June 2022, is an accurate post packaged as alarm: it frames an accurate chart of the record seven-day case average as a clear and present danger and an emergency (Truth: *accurate*; Appeal: *sensational*; Belief: not misleading; 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.

Case 1: The claim is never stated *Misinformation*

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'. . . "



1. Truth: *accurate* (1958 identification and polio work hold)
2. Appeal: plain to *sensational* (labelers split)
3. Belief: *misleading* (ellipsis implies vaccine work created virus)
4. Position: *fringe*, not slanted (laboratory-origin claim)

Case 2: Every sentence true, yet misleading *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?"

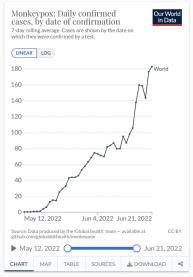


1. Truth: *accurate* (April vaccine tender and May case count hold)
2. Appeal: plain
3. Belief: *misleading* (juxtaposed dates imply foreknowledge)
4. Position: *fringe*, not slanted (planned-outbreak claim)

Case 3: Accurate post, packaged as alarm *Not misinformation*

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* (chart matches seven-day case average)
2. Appeal: *sensational* (emergency and danger framing)
3. Belief: not misleading (no untrue takeaway)
4. Position: *mainstream*, not slanted

Case 4: A partisan side, not a falsehood *Not misinformation*

Posted 23 May 2022

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



1. Truth: *no checkable fact* (opinion and poll jab)
2. Appeal: plain
3. Belief: not misleading (no untrue takeaway)
4. Position: *mainstream*, *slanted* (explicit partisan side)

Figure 5: Four posts from the corpus labeled on the four blocks. Each panel frames the post's date, text, and attached image, followed by its labels on Truth, Appeal, Belief, and Position (Table 2); a post is misinformation when it is *false* or *premature* on Truth, or *misleading* on Belief, while Appeal and Position are coded separately.

The posting date. 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 (Appendix D.4).

Human first, then language models. 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. Sixth, we tested the final rules on a holdout of 50 viral posts labeled blind by the rater. Figure 12 in Appendix D.2 draws the six steps.

Three language models from different families label the posts under the same rules, blind to one another and to any earlier labels: Claude Opus 5.5, OpenAI's gpt-6.1-sol and Google's Gemini (gemini-4-argon-high). For a viral post, the labelers read the post's text, see its images or the first frame of a video or animated image, see the producer's handle, and read the text and images of any quoted post that remains available; posts below the viral cut are labeled from their text alone. 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, and every N/A post is left out of every estimate. Misinformation is the majority verdict of a post's labelers, and the category is the one that most of that majority chose (Appendix D.3).

The validation tests show high alignment with human judgment and close agreement across the three model families (Appendix D.4). Development alignment, the share of settled answers matched by all three labelers, reached 133 of 157 (85%) in the last of four blind runs on the development set. On a holdout of 50 viral posts, drawn at random from those that are not N/A and that no development set contains and labeled blind by the human rater under the final rules with the dated record, the labelers' majority matches the rater on 84% of answers. On the paper's own labels, all three labelers give the same answer on 90.7% of answers and the same misinformation verdict on 89.4% of the viral posts that are not N/A. The remaining split posts feature genuine disagreements that also divide human raters due to differences in priors and a dated record

that was unclear on the posting date; for these posts, we rule based on the two-of-three majority.[10] Agreement among the models shows that the rules are applied consistently, not that the labels are right, because language models' errors are correlated (Dell and Rambachan 2026).

Misinformation and amplification. Misinformation, and misleading posts above all, draws more than its share of amplification. Among the 1,037 viral posts in the tests, misinformation accounts for 22% of the posts but 32% of their retweets and quote tweets, while accurate posts account for 73% of posts against 65% of amplification (the remaining 5% of posts state no checkable fact). Misleading posts, which no false sentence betrays, make up 9% of posts and 14% of amplification, alongside the 12% of posts and 17% of amplification that are false or premature.

Two features of the labels carry into the tests of Section 5. Appeal is coded apart from truth because in the model appeal raises sharing even at zero credibility (Section 2). Misinformation splits by what a reader needs to catch it: a misleading post's steering is visible in its wording, whereas catching a false or premature post requires checking outside sources, and 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 (or network cell). Section 5.1 asks how accurate the producers who post while checking is costly are, as a pool, compared with the producers of the same network who join after. Section 5.2 asks who amplifies misleading and false posts and who amplifies in the Breaking Window. Section 5.3 asks what the posts of each network are amplified for and how a producer's reach co-moves with its network's accuracy across windows. Section 5.4 asks how a post's reach changes, within producer, when the post is misleading or false. Readers pool on the network 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 5.4). Each subsection opens with the implication of the model it tests (Section 2.3) 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: 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 5.1 and the amplifiers' accuracy in Section 5.2, but no estimate for a network. Of the labeled posts,

[10] On differences in priors, a post of 20 May 2022 accurately reported Sweden classifying mpox as dangerous to public health and ended with "Here we go again…," while another of 20 May summarized the March 2021 Munich tabletop simulation of an engineered mpox pandemic without comment: all three models coded the stated facts as *accurate*, but split two to one on Belief over whether the framing planted the untrue claim that COVID-style lockdowns were returning or that the outbreak was planned (*misleading*), rather than expressing political skepticism or describing a past exercise (not misleading). On an unclear dated record, after the CDC briefly advised travelers on 6 June to wear masks and withdrew the line the next day, posts mocking the agency because mpox "isn't even airborne" (or citing a laboratory study to warn that it is "likely aerosol airborne") split the models two to one between *accurate* and *premature* or *misleading*, because the record showed close-contact spread and no long-range airborne transmission while leaving short-range aerosols under study; likewise, on 18 May early European wire summaries of national case counts mixed clinically detected suspected cases with laboratory-confirmed ones, dividing labelers two to one between *premature* and *accurate*.

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. A post's reach is $\log(1+\mathrm{RT}+\mathrm{QT})$, where RT and QT are its retweets and quote tweets.

5.1 Who posts while checking is costly

In the model, 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 (Proposition 4). The first question is which of the two holds in each network. The population is the 239 producers in a network with a checkable post, one row per producer. The indicator $\mathrm{Breaking}_j$ 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 $\mathrm{Breaking}_j = 1$. We estimate one regression per network,

$$\mathrm{Acc}_j = a_G + \beta_G \mathrm{Breaking}_j + \gamma_G' X_j + u_j, \tag{22}$$

where Acc_j is the share of producer j's checkable labeled posts that are accurate and not misleading, a_G is the network'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 network. Because each network has its own regression, β_G compares producers who post in the Breaking Window with later entrants of the same network and the same observable profile. Standard errors are heteroskedasticity-robust (HC1). Row (2) of Appendix Table 4 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. 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:

$$R_{jt} = \nu_{W(t)} + Z_j'\pi_0 + \left(\pi_c + Z_j'\pi_1\right) A_{jt} + e_{jt}, \tag{23}$$

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 (22). 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 the maintained cost-shift interpretation, equation (22) compares the pool that posts while checking is costly with the producers who join once it is cheaper, as in Proposition 4. This

interpretation holds readers' reward schedule fixed and treats later entrants as additions to a posting pool with stable within-type cost ranks. 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 network (Figure 6; Appendix Table 4, row 1). By equation (22), 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 networks 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 networks: 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 (23) gives $+0.17$ and -0.21 (row 3), and the realized-reach controls give $+0.18$ and -0.23 (row 4). Across the four networks, the signs follow reciprocity rather than overlap: β_G is positive in both one-way networks and negative in both mutual networks. 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, 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 10). Adding the producer's role, coded from its profile, to equation (22) 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 networks is -0.38 (SE 0.10). A second collection, a 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 network'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.

5.2 Who amplifies

In the certifier benchmark, who amplifies differs between the Belief Equilibrium, in which only social readers amplify, and the News Equilibrium, in which certifiers amplify too (Figure 3; Proposition 5). The second question is whether the accounts that amplify misleading and false posts, and those active in the Breaking Window, differ across networks. The amplification records are the retweets and quote tweets of the labeled posts, one record per amplifier and post, 53,247 in all. On posts by producers in a network there are 37,618 records by 30,638 amplifiers. Each quote tweet and reply on misinformation in a network 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 networks that have none count as unresolved. Appendix C.5 compares the stance labels with human raters' on 90 items.

Most quote tweets on misinformation endorse it in every network but G_1 (Figure 7(a); Appendix Table 5,

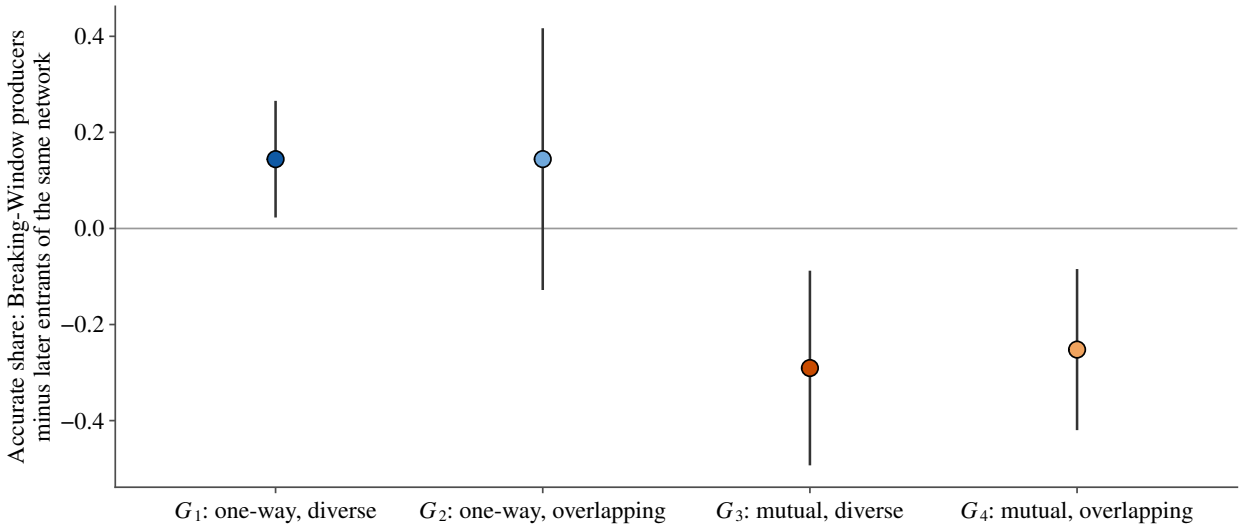


Figure 6: Who posts while checking is costly, by network. The coefficient β_G of equation (22): the difference in the share of checkable posts that are accurate and not misleading, on a scale from zero to one, between the producers whose first post falls in the Breaking Window and the later entrants of the same network, holding the producer's follower count and profile fixed. One producer-level regression per network, on the 239 producers in a network 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 network's number of producers minus one degrees of freedom. Appendix Table 4 gives the estimates and the other rows.

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 network,

$$\text{Acc}_{k,-j} = a_G + \phi_G^M \,\text{Misl}_{jt} + \phi_G^F \,\text{False}_{jt} + u_{kjt}, \tag{24}$$

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 (29), 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 network (Figure 7(b); Appendix Table 5, Panel B). False or premature posts draw less accurate amplifiers only in G_4. By equation (24), 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 network G in W2 to W4 we estimate

$$\text{Acc}_k^{\text{late}} = a_G + \beta_G^{\text{amp}} \text{Breaking}_k + \gamma_G' X_k + u_k, \tag{25}$$

where $\text{Acc}_k^{\text{late}}$ is the share of accurate posts among the checkable posts by producers in G that k amplifies in W2 to W4, 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, an indicator for a missing count, and the log of one plus its number of records in the network 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 networks. 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 5, Panel C). By equation (25), 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, 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 11, Panel A), the misleading coefficient of equation (24) 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 (25) is -0.02 in G_1 (SE 0.01) and -0.10 in G_4 (SE 0.03).

5.3 What each network amplifies and how reach co-moves across windows

In the model, a network's producer pool, reader mix and equilibrium jointly determine amplification (Figure 3). Proposition 3 relates changes in reach and credibility across equilibria. The third question is whether the posts that collect amplification, and the progression of reach and accuracy across the four two-week windows W_1 to W_4, differ across networks. For composition, the population is the producers in a network 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 network G and category n we compute the share of the network's amplification that goes to posts of that category:

$$s_{nG} = \frac{\sum_{i \in \mathscr{P}_G} w_i \mathbf{1}\{n_i = n\}}{\sum_{i \in \mathscr{P}_G} w_i}, \qquad w_i = \min\left(\text{RT}_i + \text{QT}_i,\ \bar{w}\right) + 1, \tag{26}$$

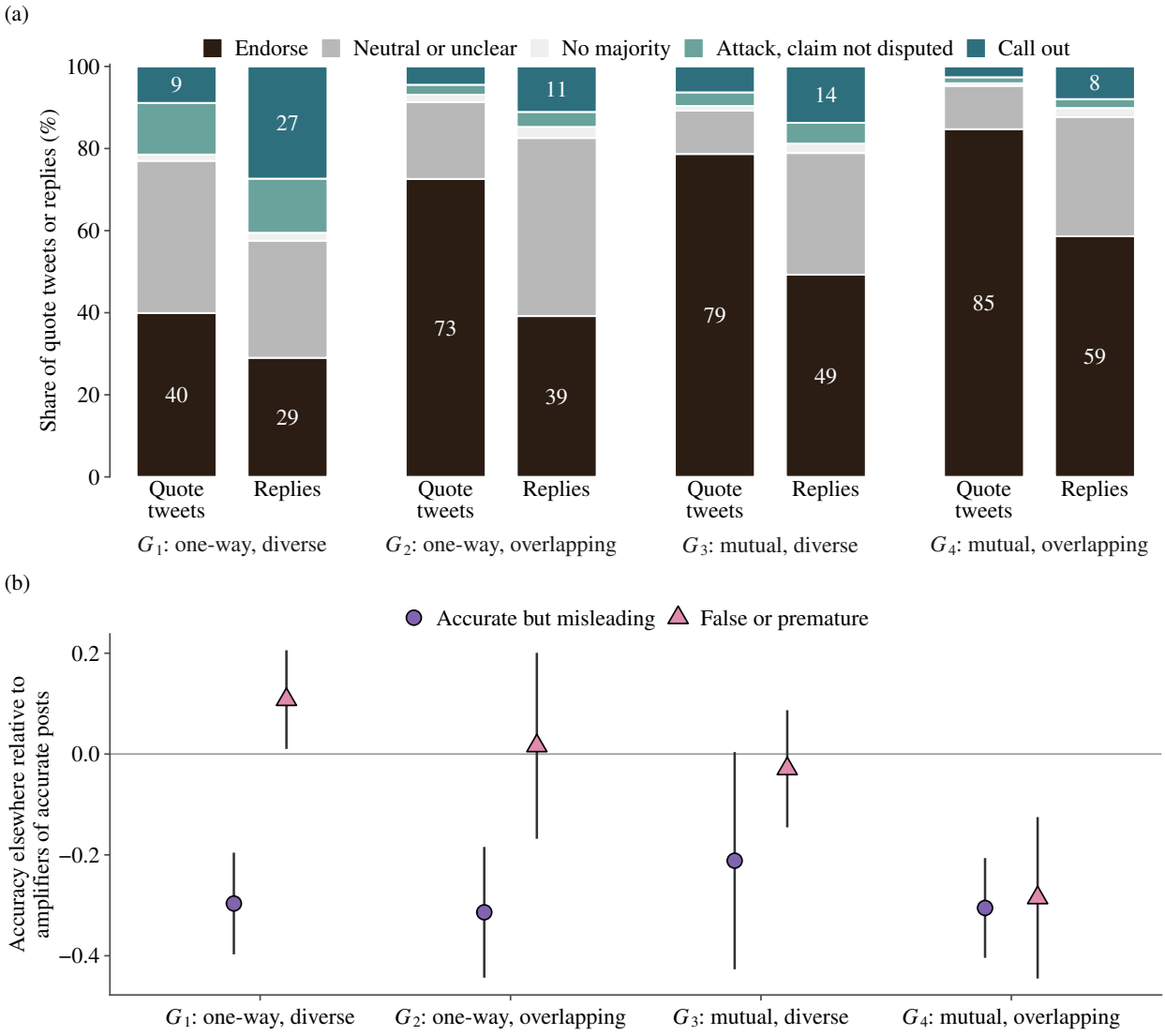


Figure 7: Who amplifies, by network. (a) The stance of the quote tweets (left bar) and replies (right bar) on misinformation posts by producers in each network, 7,822 items in all: the shares that endorse the post, call it out, attack it without disputing its claim, are neutral or unclear, or have no 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 (24): 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 network's number of producers minus one degrees of freedom. Appendix Table 5 gives the shares and estimates.

where $\mathscr{P}_G$ is the set of these posts whose producer is in network 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 2.1.2), 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 networks and is not a measure of c. To attach a standard error to the gap between networks and to hold a producer's audience size fixed, we also estimate, on the same posts and weights,

$$\mathbf{1}\{n_i = \text{accurate}\} = a + \sum_{G \neq G_1} b_G D_{iG} + f \log_{10} \text{Fol}_{j(i)} + e_i, \tag{27}$$

where D_{iG} indicates that the producer $j(i)$ of post i is in network G, Fol_j is producer j's follower count on Twitter, 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 networks are not assigned, so b_G describes how the networks' equilibria differ, not what would happen to a producer moved from one network to another.

The network predicts what is amplified (Figure 8; Appendix Table 6, 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 networks lie between: 79% in G_2 and 59% in G_3. By equation (27), the accurate share of amplification in G_4 differs from that in G_1 by -61.3 percentage points (SE 8.3; Appendix Table 6, 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).

To describe reach and accuracy across the four two-week windows W_1 to W_4, let $c_{G,W}$ be the average over network G's active producers of their share of accurate checkable posts in window W, and let $\bar{R}_{G,W}$ be the mean reach $\log(1+\text{RT}+\text{QT})$ per checkable post in that window. Figure 9(a) plots the four window points $(c_{G,W},\ \bar{R}_{G,W})$ for each network. In the one-way diverse network G_1, the Breaking Window W_1 combines the highest accuracy with the highest reach; in the mutual overlapping network G_4, W_1 combines the lowest accuracy with the highest reach, and accuracy rises as reach falls across W_2 to W_4. To test this co-movement within producer, we take the checkable labeled posts by producers in a network with at least two checkable posts (1,193 posts by 140 producers) and estimate one regression per network,

$$R_{jt} = \mu_j + \theta_G c_{G,W(t)} + u_{jt}, \tag{28}$$

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 on the network's window accuracy, and u_{jt} is the error, with standard errors clustered by producer. Within a network, $c_{G,W}$ takes four values, one per window, so any change in reach common to all producers of the network over the cycle loads onto θ_G. By equation (28), θ_G reverses sign across networks (Figure 9(b); Appendix Table 8, Panel A): it 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), positive in the one-way diverse network and negative in both mutual networks.

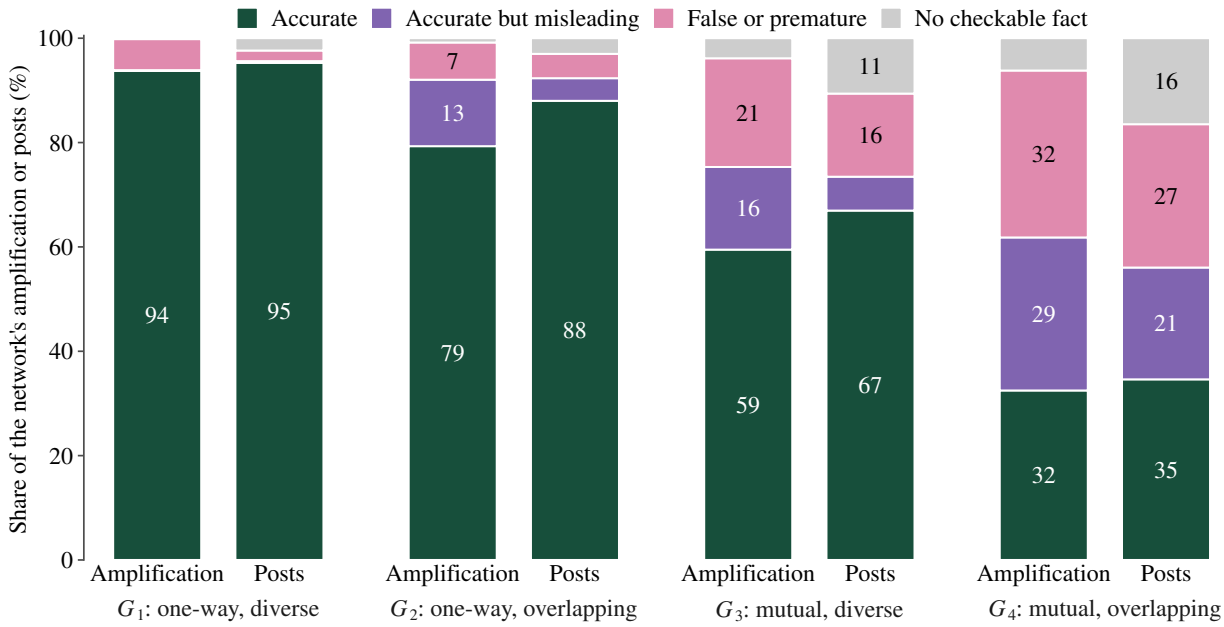


Figure 8: What each network posts and what it amplifies. For each network, the share of amplification (left bar) and the share of posts (right bar) that go to accurate posts, accurate but misleading posts, false or premature posts, and posts with no checkable fact, over the W1 to W4 posts of the producers in a network 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 (26); posts are weighted equally. Labels give the larger shares in whole percent. The shares are descriptive, so no interval is drawn; Appendix Table 6 gives the shares and the gaps between networks, with standard errors clustered by producer.

Robustness. Holding the producer's follower count fixed in equation (27) narrows the difference between G_4 and G_1 to -47.8 percentage points (SE 10.4; Appendix Table 6, 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 9, Panel A). News outlets are concentrated in the one-way networks, 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 9, Panel B).

5.4 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 (Proposition B.6). The fourth 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 network with at least two checkable posts:

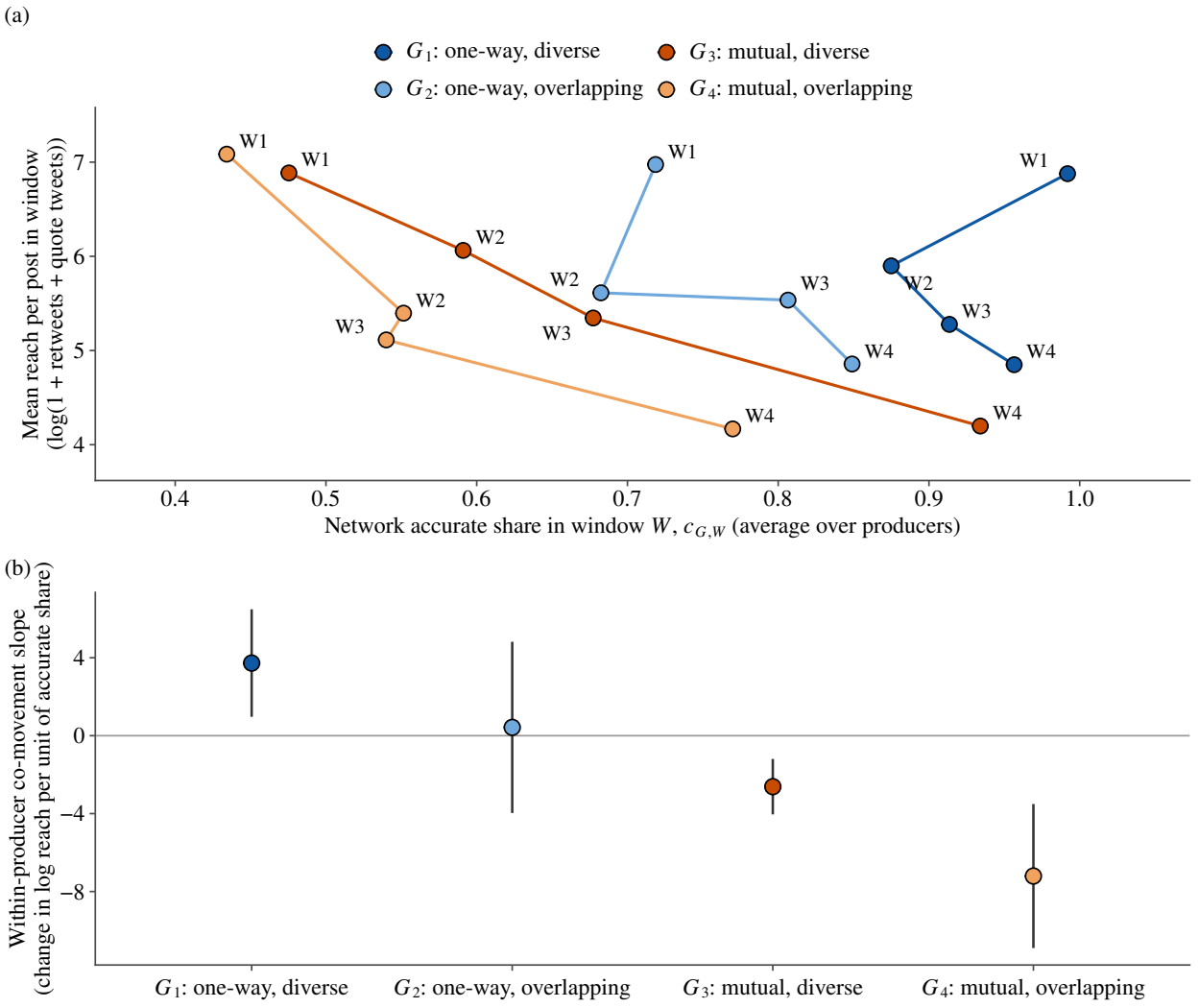


Figure 9: Reach and accuracy across windows and within-producer co-movement, by network. (a) For each network G and two-week window $W \in \{W_1, W_2, W_3, W_4\}$, the network's accurate share $c_{G,W}$ (the average over active producers in the network of their share of accurate checkable posts in window W) plotted against mean reach $\bar{R}_{G,W}$ ($\log(1 + \text{RT} + \text{QT})$ per checkable post in window W), connected across windows. (b) The within-producer co-movement slope θ_G of equation (28): one regression per network of post reach R_{jt} on $c_{G,W(t)}$ with producer fixed effects, on the 1,193 checkable labeled posts by the 140 producers in a network with at least two checkable posts. The intervals are 95 percent confidence intervals from standard errors clustered by producer and the t distribution with the network's number of producers minus one degrees of freedom. Appendix Table 8, Panel A, gives the estimates.

1,193 posts by 140 producers. We estimate one regression per network,

$$R_{jt} = \mu_j + \lambda_{d(t)} + \delta_G^M \, \text{Misl}_{jt} + \delta_G^F \, \text{False}_{jt} + u_{jt}, \tag{29}$$

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 misleading takeaway alone.

In the mutual overlapping network, misleading posts reach more than the same producer's accurate posts (Figure 10; Appendix Table 7). In G_4, by equation (29), 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 networks 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 false-or-premature coefficient is compared with the model's equal-amplification benchmark for posts with the same public record; the misleading comparison allows visible wording to affect reach (Proposition B.6).

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 (29) 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 8, Panel B).

Adding the post's sensational and slanted labels to equation (29) holds fixed cues a reader sees in the post itself (Appendix Table 8, Panel C). In G_4, 40 of the 55 false or premature posts are labeled sensational. With 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.

What the estimates do not identify. The networks are not assigned, so each estimate describes a network's equilibrium, and a difference between networks 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

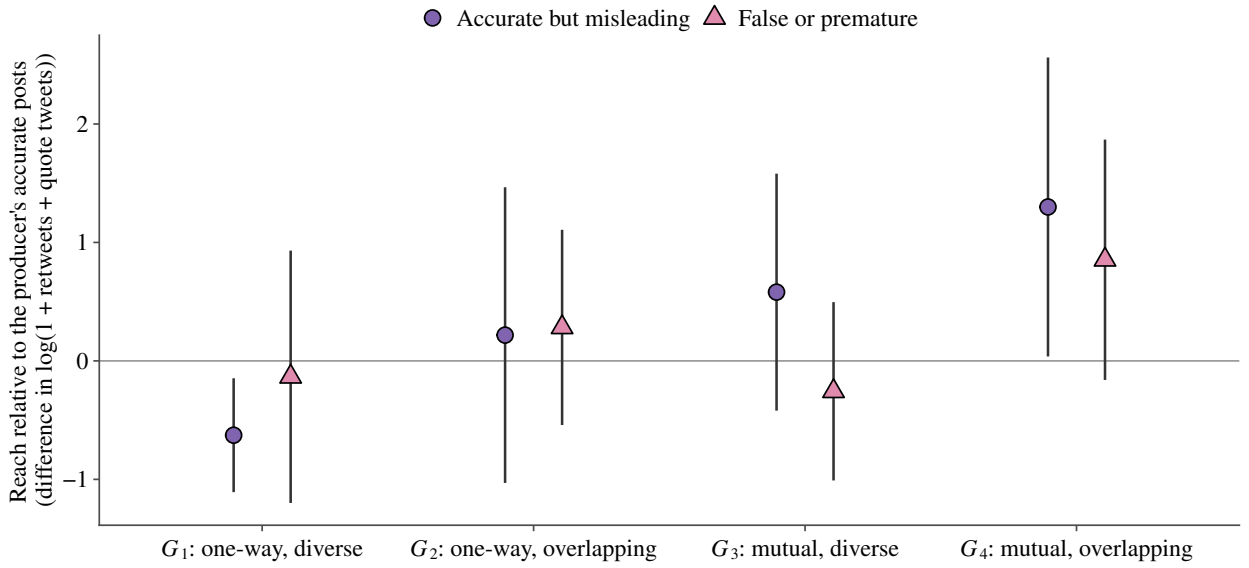


Figure 10: Reach within producer, by network. The coefficients δ_G^M on accurate but misleading posts and δ_G^F on false or premature posts in equation (29), relative to the same producer's accurate posts and net of the date of posting; the outcome is the log of one plus a post's retweets and quote tweets. One regression per network, on the checkable labeled posts, W1 to W4, by producers in the network 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 network's number of producers minus one degrees of freedom. In G_1 the coefficient on misleading posts rests on one misleading post. Appendix Table 7 gives the estimates.

two roles of the network in the model give two readings of a difference between networks, 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 5.1). 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 5.4 and 5.2). A network 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 2.3), and no test above divides it among them.

6 Related literature

Our model builds on the economics of selection markets to endogenize both the supply and the amplification of misinformation. Following Allcott and Gentzkow (2017), who note that fake news arises in equilibrium because accuracy is costly to produce and hard for consumers to infer, we cast the market for social media posts as a lemons problem (Akerlof 1970) using the demand-and-cost framework of Einav and Finkelstein (2011). Existing formal models of misinformation and verification take the set of content suppliers as given or treat false news as exogenous. In Kranton and McAdams (2024), a fixed number of suppliers paid by visibility choose whether to pay an investigation cost for a high-quality story while misinformation originates from exogenous third-party agents; in Andreottola and De Moragas (2025), competing outlets holding an

unverified rumor decide whether to publish immediately or fact-check first, with preemption pressure reducing verification and fabrication set aside; and Papanastasiou (2020) models reader sharing and fact-checking when fake news is generated exogenously. Because barriers to entry in online media are low (Allcott and Gentzkow 2017), we allow heterogeneous investigative and slop producers to enter endogenously against a congested feed, and we combine producer entry with reader heterogeneity between certifiers and social readers to generate double selection on who posts and who amplifies (Section 2).

A related literature in media economics studies how audience beliefs and incentives shape how outlets report the news. In Gentzkow and Shapiro (2006), consumers infer higher quality when reports conform to their priors, so reputation-seeking firms slant toward audience beliefs unless ex post feedback about the truth is likely. Mullainathan and Shleifer (2005) show that competing newspapers slant toward readers' beliefs by selectively omitting facts rather than fabricating them, and Gentzkow and Shapiro (2010) estimate that consumer ideology accounts for 22.3% of within-state variation in U.S. newspaper slant compared with 3.8% for owner ideology. On social platforms, Moldoveanu and Baum (2026) argue that speakers craft public assertions for third-party observers who reward engagement rather than verifiable warrant. Whereas classic media bias models hold the set of outlets fixed and study how each outlet slants its reporting, readers in our framework reward credibility pooled across producers, so audience preferences also select which producers enter the market. Our empirical labels preserve the distinction between omission and partisan positioning by coding accurate posts that lead a general reader to an untrue claim as misleading, separately from whether a post is slanted (Section 4).

We also contribute to research on how network structure and platform incentives shape content production and diffusion. In Acemoglu et al. (2024), users who share primarily with like-minded peers face a lower likelihood of being called out for spreading misinformation, which encourages low-reliability sharing and leads an engagement-maximizing platform to foster filter bubbles; empirically, Bakshy et al. (2015) find that users' friend networks and own choices restrict cross-cutting exposure on Facebook more than the ranking algorithm does. In the industrial organization of social media, Beknazar-Yuzbashev et al. (2026) study how platforms monetize user attention and engagement through advertising loads without modeling content veracity, while Filippas et al. (2025) show that users follow peers to induce follow-backs and reciprocal engagement (attention bartering), consistent with the 22.1% reciprocal tie share documented alongside one-way news subscriptions on Twitter by Kwak et al. (2010). We connect these network features to selection primitives through a linear-quadratic network sharing game (Ballester et al. 2006) in which reciprocity captures mutual attention bartering and two-hop overlap, a directed inverse analog of reach efficiency (Borgatti 1997), limits out-of-network reach. In the equal-influence benchmark, holding follower and second-degree arrow counts fixed, higher reciprocity and overlap shift effective demand toward social value relative to accuracy (Section 2.3).

On the measurement side, empirical work on misinformation has relied largely on ex post fact-checker lists or reader belief surveys. Vosoughi et al. (2018) track Twitter rumor cascades classified by six fact-checking organizations, and Allcott and Gentzkow (2017) and Pennycook and Rand (2019) draw false news items from Snopes, PolitiFact, and BuzzFeed. As Altay et al. (2023) and Allen et al. (2024) emphasize, professionally fact-checked URLs represent a small fraction of contested news and miss unflagged content that misleads

without stating a falsehood. Moreover, during breaking events, misinformation surges when demand for information outstrips credible counter-evidence (Glockner et al. 2022), and moderating a misleading post built from true facts removes accurate information from the sender's signal (Kominers and Shapiro 2024). Unlike prior analyses of the 2022 mpox outbreak that classified tweet clusters through ex post qualitative review (Edinger et al. 2023), we evaluate each post against the dated record available on its posting date and separate false, premature, and accurate-but-misleading claims (Section 4). To scale this protocol while meeting the validation standards required for text-derived and language-model measures (Gentzkow et al. 2019; Ludwig et al. 2025), we combine a human-first rulebook with three frontier language models, building on evidence that zero-shot models can outperform crowd workers (Gilardi et al. 2023) and reach fair agreement with human coders (Ziems et al. 2024), and benchmark the models' majority labels against a blind human holdout in a domain where three professional fact-checkers all agreed on 42% of rated articles (Allen et al. 2021).

Finally, our empirical strategy brings selection-market tests to a literature that has focused primarily on demand-side sharing interventions (Aridor et al. 2024). Field and survey experiments show that prompting users to consider accuracy improves the quality of the news they share (Pennycook et al. 2021), whereas offering fact-checks can reduce the sharing of true posts alongside false ones (Guriev et al. 2026). Rather than intervening on individual readers, we adapt the cost-curve test of Einav et al. (2010) to test for selection on both sides of the market. Under the maintained assumption that verification became cheaper after the Breaking Window, we compare the accuracy of producers and amplifiers active in that window with those active only afterward, by network (Section 5).

7 Conclusion

This paper studies how information asymmetry between content producers and readers drives selection in the production and amplification of social media posts, and how network structure shapes that selection. In our model, because readers cannot directly observe whether a post has been verified, they reward the pooled credibility of the producers in a network, while heterogeneity between certifiers and social readers creates a double selection margin on both who posts and who amplifies. Below the certifier threshold, only social readers amplify, which removes certifiers' contribution to the credibility premium and sustains a low-credibility Belief Equilibrium; above the threshold, certifiers amplify as well, which can sustain a high-credibility News Equilibrium when certifiers are sufficiently prevalent and responsive. To test these predictions during the 2022 mpox outbreak on Twitter, we develop a date-anchored, human-validated language-model fact-checking protocol that separates false, premature, and accurate-but-misleading posts. Across producer-influencer networks classified by follower reciprocity and two-hop overlap, the evidence is consistent with advantageous supply selection in one-way diverse networks (G_1), where amplification goes predominantly to accurate posts and producers active during the costly Breaking Window are more accurate than later entrants, and adverse supply selection in mutual overlapping networks (G_4), where amplification goes mostly to misinformation and early producers and amplifiers are less accurate. Because these contrasting patterns coexist on the same platform during the same news event, platform-wide averages miss where misinformation thrives and where verification is rewarded.

These findings carry three implications for platform design and misinformation policy. First, because amplification is concentrated on accurate posts in one-way diverse networks and on misinformation in mutual overlapping networks, engagement alone is not a measure of credibility. Incorporating source and network structure into algorithmic rankings (Lazer et al. 2018) can help separate certified reach from reciprocal echo, whereas ranking feeds or training artificial intelligence systems on raw engagement inherits the equilibrium of the networks those posts come from. Second, demand-side interventions interact with the supply selection regime: in the Belief Equilibrium, nudges that leave readers below the certifier threshold do not activate certification. Holding producer costs fixed, a marginal increase in posting rewards at a stable interior equilibrium raises credibility where supply selection is adverse and lowers it where selection is advantageous (Proposition 3(ii)). Third, generative AI tools that lower the cost of writing unverified posts may expand the share of slop producers. A larger slop share lowers pooled credibility at a given participation wherever investigative producers also post (equation (3)).

Our empirical evidence characterizes cross-network equilibria rather than the causal effect of moving a producer from one network to another. Because network membership is endogenous, differences across networks reflect the combined influence of producer pools, reader preferences, and equilibrium selection, and cross-sectional differences alone do not show that a single network admits multiple equilibria (Proposition B.8). Moreover, our Breaking Window comparisons rest on the maintained assumption that checking facts was costlier before public health agencies issued clinical guidance than afterward. Three extensions would sharpen identification in future work: exploiting exogenous shifts in the cost of fact-checking, tracking networks over time as they transition between equilibria, and endogenizing the platform's ranking algorithm alongside producer and amplifier choices.

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.

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.

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).

— (Nov. 28, 2022b). *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.