Modeling Science Communication on Weibo: Causal Inference and Network Dynamics (2015–2025)

Jingxin Yang
jingxinyang@stanford.edu

Management Science & Engineering
Stanford University

November 4, 2025

TL;DR — Set-up & Today's Asks

Set-up. Two Weibo cases with clear hot-search shocks: (i) **Zhang Xuefeng** education controversies (2023–2025); (ii) **Jiang Ping** math-competition saga (Jun & Nov 2024).
Hypotheses.

- *Algorithmic exposure (AE)* raises visibility but not proportionally *organic* engagement (authority–influence decoupling).
- *Cold diffusion:* neutral/technical takes spread comparably to emotional ones.

Design freeze. Windows fixed as: Zhang — Dec 2023 (majors/journalism controversy; public comments on humanities and journalism majors triggered wide discussion and debate), Jan 2025 (banking/deposits controversy; public remarks that "wherever my daughter works at a bank, the company's large long-term deposits will be placed in that bank" were interpreted as using financial resources to secure her job, sparking strong backlash), Sep 2025 (account restrictions; reports and user feedback that his accounts faced feature limits / throttling, used as a window on platform-level "soft governance" and visibility changes);
Jiang — Jun 12–20, 2024 (rise; initial viral phase with high visibility), Nov 2–6, 2024 (reversal; organizer / official clarifications and investigations leading to a reversal and correction of the public narrative).
AE. Post-level $\mathrm{AE}^{\mathrm{bin}}$ (primary, Top-50); topic-level $\mathrm{AE}^{\mathrm{prob}\,\mathsf{topic}}$ in appendix for dose–response.
Decisions today. Bandwidth ± 10 main (± 20 robustness); lock keywords.

Roadmap for Today

- **Research question & story**: algorithm vs. users in two high-salience cases.
- **Data**: event windows, sampling & API limits (no random sampling claim).
- **Measurement**: outcomes, AE (post-level vs topic-level), sentiment.
- **Identification**: Top-50 cutoff as a natural experiment (trending threshold; donut, balance, McCrary).
- **Decisions**: bandwidth, AE primary, finalized windows/keywords.

Research Question & Motivation

Core question: On a semi-censored, algorithm-driven platform (Weibo), is content diffusion mainly *algorithm-driven* or *user-driven*?

Why it matters:

- Algorithmic boosts vs organic sharing — impact on public debates (education/science) under censorship.
- Tests whether Western virality drivers (emotion, authority) hold in China's sociotechnical context, or are fundamentally reshaped.

Our approach: Use the Top-50 trending threshold as a natural experiment— formally, a regression discontinuity design (RDD)—that compares posts just above vs. just below the cutoff to isolate algorithmic exposure effects, and then compare them to user-driven diffusion patterns (sentiment, identity, grassroots vs experts).

Key Events: What They Are and Why They Matter

996.ICU campaign (Mar–Apr 2019).
Grassroots campaign launched on GitHub to protest the "996" overwork culture (9am–9pm, 6 days/week) in China's tech sector. It quickly spilled over onto Weibo and became a focal point for online labor-rights debates.

Gene-editing babies (2018–2025).
Chinese scientist He Jiankui announced CRISPR-edited babies, triggering four distinct waves of public outrage and policy discussion (revelation, sentencing, prison release, and visa/policy controversies). This gives us a multi-wave science/ethics case.

COVID-19 discussion waves (2020, 2022).
Weibo debates around key pandemic moments: Wuhan lockdown and confirmation of human-to-human transmission; Dr. Li Wenliang's death and censorship concerns; the 2022 Shanghai lockdown; and the sudden Zero-COVID reversal. These are high-salience, high-censorship public health episodes.

Zhang Xuefeng controversies (2023–2025).
A popular education influencer repeatedly sparked Hot Search spikes with comments on university majors, journalism, banking/deposits, and alleged account throttling. These are sharp, opinion-heavy waves in education and economic topics.

Jiang Ping math-competition saga (2024).
A high-school student was initially celebrated online as an international math-competition champion, followed by an official clarification and partial reversal. We observe two clear waves – the "rise" and the "reversal" – which are useful for pre/post contrasts in sentiment and algorithmic exposure.

Master Event Summary — Waves & RDD Fit

Zhang Xuefeng (2023–2025)

- *Waves:* W1: 2023-12 (majors/journalism); W2: 2025-01 (banking/deposits); W3: 2025-09 (account throttling).
- *Anchor $t^\star$:* first Hot Search Top-10 entry per wave.
- *Window:* $[t^\star - 3,\ t^\star + 5]$ days; extend to ± 6 if under-sampled (log change).
- *RDD:* local-linear (triangular); **donut excludes $\pm$1–2 ranks**; **bandwidth $\pm$10 (main), $\pm$20 (robustness)**; placebo cutoffs at 40/60.

Jiang Ping / AGMC (2024)

- *Waves:* WA: 2024-06-12–06-20 (rise); WB: 2024-11-02–11-06 (clarification/reversal).
- *Anchor $t^\star$:* first Hot Search Top-10 per wave. *Window:* same rule.
- *RDD:* clean two-wave contrast; estimate within-wave; report balance & McCrary per wave.

996.ICU (2019) — comparator

- *Wave:* Mar–Apr 2019 burst.
- *Anchor $t^\star$:* first Top-10 entry. *Window:* same rule.
- *RDD:* sharp bursts; donut excludes $\pm$1–2; bandwidths as above.

Gene-editing babies (2018–2025) — legacy

- *Waves:* A: 2018-11-25–11-29; B: 2019-12-30–2020-01-05; C: 2022-04-05–04-12; D: 2023-02-20–02-24.
- *Anchor $t^\star$:* first Top-10 per wave. *Window:* same rule.
- *RDD:* within-wave; placebo cutoffs at 40/60.

COVID-19 (2020, 2022) — legacy

- *Waves:* A: 2020-01-20–01-23; B: 2020-02-07–02-10; C: 2022-03-28–06-01; D: 2022-12-07–12-26.
- *Window:* same rule; *RDD:* A/B/D sharp; **C is long $\Rightarrow$ split sub-windows or use DiD**.

Frozen rule. Let $t^\star$ be the first time the relevant topic/hashtag enters **Hot Search Top-10** within a wave. Fix collection at $\boxed{[t^\star - 3,\ t^\star + 5] \text{ days}}$; extend to ± 6 if under-sampled (log change). All diagnostics (covariate balance, McCrary) are computed *within wave*; SEs clustered by account.

Event Windows (Design Freeze)

Cases and waves.

- **Zhang Xuefeng (Education)** W1: *Majors/Journalism* (2023-12); W2: *Banking/Deposits* (2025-01); W3: *Account throttling* (2025-09).
- **Jiang Ping (Math competition)** WA: 2024-06-12–06-20 (*rise*); WB: 2024-11-02–11-06 (*clarification/reversal*).

Window rule (frozen). Let $t^\star$ be the date when the relevant topic/hashtag first enters Hot Search Top-10 within the wave. Fix the collection interval at $\boxed{[t^\star - 3,\ t^\star + 5] \text{ days}}$ (default). If the wave is multi-peaked, use the first Top-10 entry as $t^\star$; if sample size is small, extend to ± 6 days (log the change).

Keywords/hashtags (examples).
Zhang: #ZhangXuefeng#, #Journalism#, #MajorChoice#, #Banking#, #Deposits#, #AccountThrottling#;
Jiang: #JiangPing#, #MathCompetition#, #MathOlympiad#. (China Standard Time; de-dup by `post_id`.)

RDD set-up (intuition).

- **Running variable:** distance to the Top-50 cutoff (positive = above the threshold).
- **Treatment:** "hot/on-list/badge" or entering Top-50 is coded as `AE`=1.
- **Estimator:** local linear (triangular kernel); *donut* excludes 1–2 ranks around the cutoff.
- **Bandwidths robustness:** ± 10 (main), ± 20 (robustness); placebo cutoffs at 40/60.
- **Diagnostics:** covariate balance and McCrary density *within each wave*; SEs clustered by account.

See Appendix "RDD Technical Details" (jump).

Data at a Glance

Unit of analysis. Event-centered windows on Weibo (2015–2025) covering COVID-19, gene editing, and education/labor controversies.

Collection strategy (not random sampling). Within each fixed event window, we *aim for near-exhaustive coverage*, not a random sample:

- use dense time slices (1–3h) and multiple keyword panels;
- query each slice repeatedly until hitting the API cap;
- deduplicate by `post_id`.

Standard search endpoints return at most ~ 500 *most recent* posts per query; without slicing this would bias us toward late posts.

Typical coverage. Per event, we collect $N \approx 5{,}000$–$30{,}000$ posts within windows, with estimated archival retention $\hat{r}_{e,t} \gtrsim 0.8$ in most months (see Appendix timeline).

Observed fields. Post text and timestamp; author identity signals (verification, org type); engagement counts (reposts, comments, likes); platform flags (`is_hot`, `rank_index`, `icon_hot`) used as AE labels.

Data Integrity & Sampling (Truth-in-Methods)

No "random sampling" claim. Within each fixed event window, we exhaustively retrieve posts when feasible; for high-volume bursts we use uniform time-slice queries (1–3h buckets, looped) to mitigate search caps.
API caps & truncation. Standard search endpoints return only the top-N (e.g., $\sim$500) per query-time slice; without slicing this induces *latest-first* bias. We therefore:

- partition windows into fine time buckets;
- repeat queries with rotated keyword panels / pagination;
- log fill rates per bucket for retention weighting.

Missingness & robustness. For each event e and time bucket t, we estimate a *retention rate* $\hat{r}_{e,t}$ = (observed posts)/(expected posts in archives). We then apply inverse-probability weights $1/\hat{r}_{e,t}$ so that under-represented buckets count more; an *archive-only* subset (no deletions) is reported as a sensitivity check.
Transparency. A monthly/daily count timeline (Appendix) visualizes spikes and gaps; all timestamps in CST; de-dup by `post_id`.

Data: Collection and Coverage

Sources & scope. Official Weibo search/topic endpoints and archival snapshots (2015–2025). Core causal cases: Zhang Xuefeng education controversies (2023–2025) and the Jiang Ping / AGMC math-competition saga (2024). Legacy / comparator events: COVID-19 discussion waves and gene-editing babies, plus a labor benchmark (996.ICU campaign).

Sampling procedure. Within each pre-defined event window (see Event Windows and Master Event Summary slides), we query in fine time buckets (1–3h) using event-specific keyword/hashtag panels with pagination to mitigate search caps; when feasible, retrieval is near-exhaustive. All posts are deduplicated by `post_id`. We also compile unique authors and their metadata (verification type, domain tags, bio keywords, follower counts) for classification and controls. *We do not claim random sampling.*

Query design. Event-specific keyword and hashtag panels; deduplication by `post_id`; timestamps normalized to CST. Examples: COVID-19: "COVID-19", "coronavirus"; Gene editing: "gene editing", "CRISPR", "He Jiankui"; 996.ICU: "996", "996.ICU"; Zhang Xuefeng: the name plus controversy-specific tags around majors/journalism, banking/deposits, and account throttling; Jiang Ping: the name plus tags such as "math competition", "Olympiad". See Appendix keyword table (p. 33).

Event windows (current design). For Zhang and Jiang waves, let $t^\star$ be the first time the relevant topic/hashtag enters Hot Search Top-10; we collect $[t^\star - 3,\ t^\star + 5]$ days (extend to ± 6 if under-sampled). For legacy/comparator events (COVID-19, gene editing, 996.ICU), we follow the wave-specific windows listed on the Master Event Summary slide (jump).

Data Limitations and Deletion Bias

API limits (search truncation). Standard search endpoints return at most ~ 500 *most recent* posts per query. During peak hours this can truncate older posts. We mitigate this by:

- shrinking queries into 1–3h time slices and looping over slices;
- rotating keyword panels and using pagination;
- logging the fill rate of each event–time cell.

Deletions and archival gaps. Pre-2019 and during sensitive moments (e.g. early COVID-19), posts may be removed or never archived. This creates missingness that is not purely random.

Retention estimates and IPW. For each event–time cell (e, t) we estimate a retention rate $\hat{r}_{e,t}$ (archived / expected posts) and apply inverse-probability weights $w_{e,t} = 1/\hat{r}_{e,t}$ in regressions. We also report an *archive-only* subset as a robustness check.

Diagnostics. An appendix timeline plots monthly/daily post counts by event. Visible dips (e.g. Feb 2020 during early COVID censorship) are highlighted and discussed as potential deletion-driven gaps.

Data Snapshot I — Volume & AE (pilot)

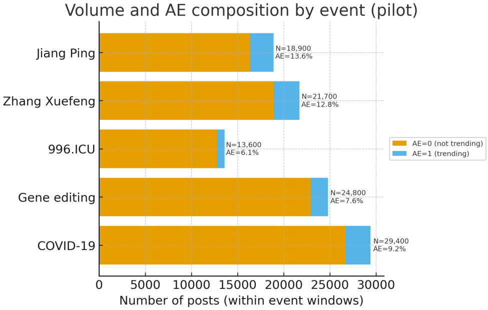

Fig. A. Posts per event (**our sample within collection windows**); not platform-wide totals.

Summary (pilot).

Event	Posts	AE=1 (%)
COVID-19	29,400	9.2
Gene editing	24,800	7.6
996.ICU	13,600	6.1
Zhang Xuefeng	21,700	12.8
Jiang Ping	18,900	13.6

Note: AE=1 by $\mathrm{AE}^{\mathrm{bin}}$. Counts reflect short Hot-Search-centered windows ($[t^\star - 3,\ t^\star + 5]$ days) with keyword filters and de-duplication, not platform-wide totals; figures are pilot-scale and will be updated with final counts.

Data Snapshot II — Collection windows by event (pilot)

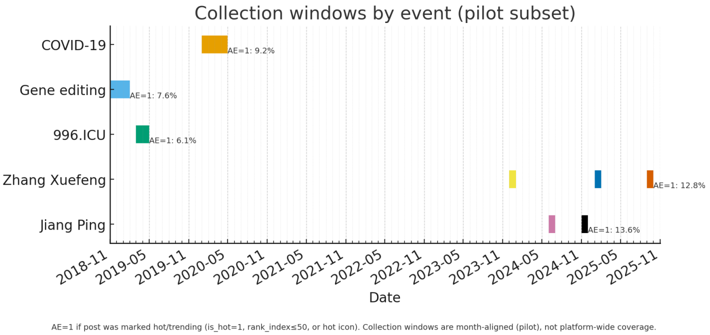

Fig. B. Collection windows by event (pilot subset). $\mathrm{AE}^{\mathrm{bin}} = 1$ if hot/trending (`is_hot`=1, `rank_index`$\leq$ 50, or hot icon). Collection windows are month-aligned around key waves, not platform-wide coverage.

Key anchors (context only): COVID-19 — Wuhan lockdown (2020-01-23); Gene editing — He Jiankui announcement (2018-11-25); 996.ICU — GitHub surge (2019-03–04); Zhang Xuefeng — majors/journalism (2023-12), banking/deposits (2025-01), account throttling (2025-09); Jiang Ping — "rise" (2024-06-12–20), "reversal" (2024-11-02–06).

Background and Motivation

Motivation. Weibo is a large, semi-censored, algorithmically mediated public sphere. Classic "emotion/authority-driven virality" findings from Western platforms may invert due to platform governance and recommendation design.

We distinguish **soft** (algorithmic/operational visibility shaping) from **hard** (takedowns/bans) censorship; our AE focuses on the former.

This study fills a gap by causally identifying how **algorithmic (soft) interventions** *shape the diffusion of science content on Weibo, a question under-tested in prior work.*

Why this matters.

- Tests whether diffusion theories travel across sociotechnical contexts.
- Offers design/measurement for *algorithmic exposure* (AE) on semi-censored platforms.
- Policy relevance: how platform governance shapes public uses of science.

From Western Findings to Our Hypotheses

What we know from Western platforms.

- Emotional content (especially anger / moral outrage) tends to spread farther and faster than neutral information. Berger & Milkman 2012; Brady et al. 2017; Vosoughi et al. 2018
- Messages from elites / authorities typically enjoy higher baseline visibility and trust.
- Political and science debates often polarize into echo chambers / ideological bubbles. Sunstein 2017; Del Vicario et al. 2016

Why Weibo might behave differently.

- Algorithmic curation and content moderation re-weight what is visible and when.
- For sensitive topics (public health, labor, education), "safer" neutral/technical frames may be favored.
- Grassroots voices can still trigger large cascades, but under stronger governance constraints.

Our next step.
Building on these anchors, we formulate four working hypotheses about diffusion on Weibo: *cold diffusion, authority–influence decoupling, grassroots amplification, and ideological silos* (next slide).

Hypotheses (tested on Weibo)

Prior work: On Western social media, emotional content and source authority often boost virality (e.g., Berger & Milkman 2012; Brady et al. 2017; Vosoughi, Roy & Aral 2018; Goel et al. 2016). On Weibo (semi-censored, algorithmically curated), these patterns may be reshaped by governance and recommendation design. We therefore test four hypotheses:

1. **Cold diffusion (neutral/technical content spreads comparably).**
   In Western settings, emotion fuels sharing. On Weibo, we hypothesize that neutral/technical posts (rational explainers, fact-checks) can spread just as widely as emotional posts, especially on sensitive or science topics.
2. **Authority–influence decoupling.**
   Trending badges and algorithmic boosts give authorities (experts, official media) extra *visibility*, but user engagement may not scale proportionally. Being pushed onto the Hot Search list may yield limited additional organic reshares and discussion.
3. **Grassroots amplification (small accounts triggering big cascades).**
   Following Goel et al. (2016), we expect ordinary users (grassroots) sometimes trigger huge cascades. On Weibo's science/public-issue debates, we hypothesize that grassroots posts have a *fatter tail* in cascade sizes than expert/official posts: small accounts occasionally create very large "blow-up" threads.
4. **Ideological silos.**
   Algorithmic recommendations and community clustering may keep discussions within their own circles, producing echo chambers. We test whether science/public-issue cascades have limited cross-cluster reach, with each camp mostly "talking to itself".

Core Research Questions and Contributions

Core RQs (mechanisms).

- **Cold diffusion:** do neutral/technical posts travel farther?
- **Authority–influence decoupling:** does AE yield visibility without proportional *organic* engagement?
- **Grassroots amplification:** do non-experts trigger heavier-tailed cascades?
- **Ideological silos:** do cascades remain within communities?

Contributions.

- Causal identification + heterogeneous information networks (HIN).
- Explicit AE operationalization ($\mathrm{AE}^{\mathrm{bin}}$ label; supervised $\mathrm{AE}^{\mathrm{prob}}$; $\mathrm{AE}^{\mathrm{pc}}$ as sensitivity).
- Cross-event design (public health/bioethics + education/labor: COVID-19, gene-editing babies, 996.ICU, Zhang Xuefeng, Jiang Ping).

Operational: Account Identity Typology

Categories.

- **Expert:** verified person with domain tag (med/sci/pop-sci) or scientist/clinician in bio; optional whitelist.
- **Org/Media:** blue-V institutions (univ/hospital/institute), media, NGO, gov.
- **Grassroots:** unverified or verified w/o domain tag; not org/media.

Signals.

- Verification type, domain tags, bio keywords, URL domain.
- Follower counts used only as controls.

Implementation & QA.

- Rule order: Org/Media $\rightarrow$ Expert $\rightarrow$ Grassroots.
- Ambiguity: whitelist/blacklist first; no tag $\Rightarrow$ Grassroots; log unresolved.
- Validation: double-code 200–300; Cohen's $\kappa \geq 0.80$; confusion matrix in appendix.

Key Variables: Algorithmic Exposure (AE)

Treatment (binary). Algorithmic Exposure (AE) indicator:

$$\mathrm{AE}_i^{\mathrm{bin}} = \nVdash \{\texttt{is\_hot} = 1 \ \vee \ \texttt{rank\_index}$$

5

0 $\vee$ $\texttt{icon\_hot} = 1\}$.

Interpretation: $\mathrm{AE}^{\mathrm{bin}} = 1$ flags posts boosted by trending/hot badges — a proxy for *soft* algorithmic promotion (distinct from removals/bans).

Score (supervised). Train a logistic model on pre-exposure features $\mathbf{Z}_i$ (author signals, media, early-lag engagement, topic/time FEs) to predict $\Pr\big(\mathrm{AE}_i^{\mathrm{bin}} = 1 \mid \mathbf{Z}_i\big)$. Use predicted probability $\mathrm{AE}^{\mathrm{prob}}$ for heterogeneity and dose-response diagnostics. Report AUC/PR and calibration on a holdout.

Sensitivity. $\mathrm{AE}^{\mathrm{pc}}$ = PC1 of standardized platform flags (robustness only).

Example. A post entering the Top-50 trending ($\texttt{rank\_index}$

5

0) is tagged $\mathrm{AE}^{\mathrm{bin}} = 1$ and is likely surfaced beyond followers via Hot/Discovery feeds.

Operational: Sentiment Measurement & Validation

Labeling target. Binary emotionality: emotional vs neutral/technical.
Primary method (lexicon). Dictionary-based Chinese sentiment scoring with intensity; bin into two classes. A *technical-term* filter (e.g., domain-specific terms such as "acute", "viral load") prevents misclassifying scientific jargon as emotional.
Cross-check (BERT). A pretrained Chinese BERT classifier (zero-/few-shot) produces an alternative label for each post; we compare agreement with the lexicon-based labels.
Human validation (plan).

- Manually code a stratified sample of ∼2,000 posts across events and AE strata.
- Report accuracy/precision/recall, ROC-AUC/PR-AUC; target Cohen's $\kappa \geq 0.75$.
- Resolve discrepancies between lexicon and BERT via adjudication.

Current status. Preliminary checks show high agreement between lexicon and BERT labels; final metrics will be reported after manual coding is complete.
Audit of "Neutral". Randomly inspect $n \approx 200$ neutral posts per wave (stratified by AE/identity); code topical type (science/technical explainer, news recap, counseling/advice, other) and report shares.
Use in analysis. Sentiment enters as (i) binary emotionality and (ii) continuous score in robustness. All regressions control for sentiment; "Cold diffusion" tests use both schemes.
Error analysis & robustness. Lexicon swap/perturbation, topic-wise relabeling placebos, and confusion matrices are in the Appendix.

Outcomes and Controls

Primary outcomes (diffusion).

- Reshares (count) — main outcome (NB2; report IRR).
- Comments, Likes (counts) as supplementary outcomes.
- Tail metrics: CCDF slope; reproduction proxy $\mathcal{R}$ (appendix).

Controls / fixed effects (used in all regressions).

- Topic & hour-of-day fixed effects; event FEs; SEs clustered by account.
- *Text/content:* length, media dummies (image/video), sentiment bins/scores, technicality, hashtag count.
- *Author/account:* account type {Expert, Organization, Grassroots}; $\log(\textbf{followers} + 1)$ **(baseline audience size)**; verification dummies.

Identification hooks (overview).

- Trending-cutoff RDD (local-linear, triangular kernel; donut; density & balance checks).
- Retention IPW ($1/\hat{r}_{e,t}$) for deletion/archival sensitivity.

Note (confounding). We explicitly control for author follower count (baseline audience) and other confounders (content type, timing, identity) in all regressions.

Methods and Identification

Counts (main). Negative binomial (NB2); report IRR with topic/hour fixed effects; SEs clustered by account.

$$\log \mu_i = \alpha + \beta\,\mathrm{Cold}_i + \boldsymbol{X}_i\boldsymbol{\gamma} + \eta_{a(i)} + \tau_{t(i)}.$$

where μ_i denotes the expected number of reshares for post i.
RDD at trending cutoff (identification). We exploit the sharp change in algorithmic exposure at the Hot Search Top-50 threshold. Let R_i denote distance to rank 50 (positive = just above the cutoff). We estimate a local-linear RDD with a triangular kernel and a *donut* (dropping $\pm$1–2 ranks):
We first show a "first stage": the probability of algorithmic exposure $\Pr\left(\mathrm{AE}_i^{\mathrm{bin}} = 1\right)$ jumps discontinuously at $R_i = 0$. Then we use this jump as a quasi-experiment (fuzzy RDD / 2SLS) to estimate the local average treatment effect on diffusion.
Policy shocks (DiD / event-study). Governance changes to Hot/Trending as quasi-exogenous shocks; Sun–Abraham estimator; long leads show no pre-trends; heterogeneity by identity.
Exposure vs influence. Exposure: AE flags and non-follower first-hop ratio. Influence: non-follower share ratio, depth > 2, Hawkes reproduction $\mathcal{R}$ (estimated post-latency; *details moved to appendix*).
Deletion bias. Snapshot retention $\hat{r}_{e,t}$; IPW $1/\hat{r}_{e,t}$; archive-only subset as robustness.

Principle: start simple. Main results rely on pre-registered NB2/RDD/DiD; dynamics (Hawkes/ABM) are used only as mechanism checks in the appendix.
Specification note. All NB2 specifications include topic & hour fixed effects; log(followers + 1), content controls, account-type/verification dummies, and clustered SEs.

Model Specs (for figures & captions)

Counts (main, NB2). Outcomes $Y_i \in \{\text{reposts}, \text{comments}, \text{likes}\}$; we model the *expected number of reshares* via:

$$\log \mathbb{E}[Y_i] = \alpha + \beta\,\text{Neutral}_i + \boldsymbol{X}_i\boldsymbol{\gamma} + \eta_{\text{topic}(i)} + \tau_{\text{hour}(i)}, \quad \text{SEs clustered by account.}$$

Here $\beta > 0$ (IRR> 1) means neutral posts are shared more than emotional posts, holding all controls fixed.
RDD at Hot Search Top-50. Running variable $R_i =$ distance to rank 50 ($R_i > 0 =$ above the cutoff, $R_i < 0 =$ below the cutoff). Treatment

$$\mathrm{AE}_i^{\mathrm{bin}} = \not\Vdash\{\texttt{is\_hot} = 1 \ \vee \ \texttt{rank\_index}$$

5

0 $\vee$ $\texttt{icon\_hot} = 1\}$. Local-linear (triangular); donut excludes $\pm$1–2 ranks; bandwidths $\pm$10 (main), $\pm$20 (robustness). Fuzzy RDD: use the first-stage jump in $\Pr\left(\mathrm{AE}^{\mathrm{bin}} = 1\right)$ at R_i=0 and 2SLS to estimate the LATE on diffusion outcomes.
Exposure vs Influence.

- *Exposure:* $\mathrm{AE}^{\mathrm{bin}}$ and non-follower first-hop ratio.
- *Influence:* non-follower interaction share; prob(depth$>$2); tail metrics (CCDF slope).

Early Signals — Sentiment ⇒ Diffusion

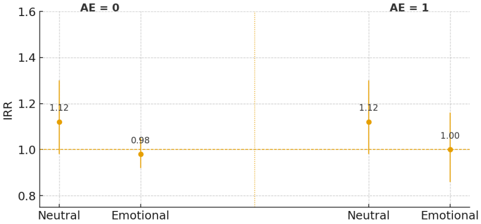

Fig. 1A. Sentiment → diffusion (IRR), single-axis faceting by AE (pilot).

Fig. Effect of sentiment on diffusion (IRR), split by AE=0/1. Each bar shows the incidence-rate ratio (IRR) of **neutral vs emotional** posts on reshare counts; the vertical lines are 95% confidence intervals. *Model:* NB2 with topic & hour fixed effects; controls for text length, media (image/video), sentiment score/bin, technicality, hashtag count, and log(followers + 1); SEs clustered by account. *Result:* IRR ≈ 1.0 in both AE groups and not statistically significant, indicating **no strong emotion advantage** in reshares.

Takeaways (pilot-scale):

- Bars are essentially centered at IRR ≈ 1 with overlapping 95% CIs ⇒ we **do not** see an emotion advantage in reshares once we control for content, timing, and baseline audience.
- This is consistent with "cold diffusion": neutral (rational/technical) posts spread at rates comparable to emotional posts (IRR≈1, n.s.).
- Controls include timing, content, and baseline audience size (log(followers + 1)); SEs clustered by account.
- Robustness next: alternative sentiment bins; BERT cross-check; within-topic relabeling placebos (see appendix plan).

Early Signals — RDD at Trending Cutoff

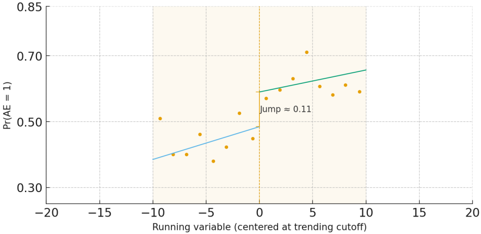

Fig. Local-linear RDD at the Hot Search Top-50 cutoff. Horizontal axis: distance in ranks to the Top-50 threshold (R_i); $R_i = 0$ is the cutoff. Vertical axis: probability that a post is algorithmically exposed ($\mathrm{AE}^{\mathrm{bin}} = 1$). Points are binned averages; lines are local-linear fits on each side (triangular kernel, donut excluding ± 1–2 ranks). *First stage:* crossing from just below to just above the cutoff raises $\Pr\left(\mathrm{AE}^{\mathrm{bin}}{=}1\right)$ by ≈ 11 percentage points (95% CI [0.07, 0.16]). *Diagnostics:* McCrary density test finds no manipulation of ranks; pre-exposure covariates are balanced across the cutoff.

Takeaways (pilot-scale):

- Intuition: posts just above and just below rank 50 are similar in content and author type, but those just above get a discrete jump in algorithmic promotion. We treat this as a quasi-experimental shock to exposure.
- Clear discontinuity in $\Pr\left(\mathrm{AE}^{\mathrm{bin}}{=}1\right)$ at the Top-50 threshold $\Rightarrow$ validates $\mathrm{AE}^{\mathrm{bin}}$ as an algorithmic-promotion proxy.
- Main bandwidth ± 10 (triangular kernel); robustness with ± 20 yields similar jumps.
- No density manipulation; pre-exposure covariates are balanced (McCrary + balance tests).
- Use fuzzy-RDD/2SLS for LATE on diffusion outcomes (NB2), reported with account-clustered SEs.

Early Signals — Heavy-tailed Cascades (Grassroots)

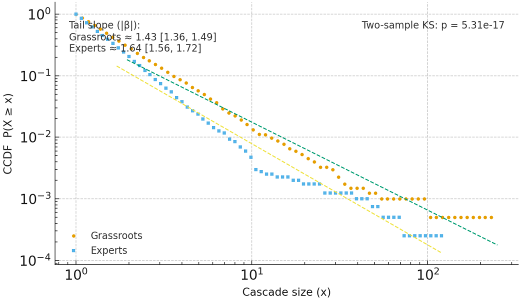

Fig. 3. CCDF of cascade sizes (log–log) for grassroots vs experts. Horizontal axis: cascade reshare count (log scale); vertical axis: fraction of posts with cascades at least that large. The step-like shape comes from discrete cascade sizes and the cumulative tail definition. Dashed lines are power-law fits on the upper tail (top 30%); estimated exponents $|\beta| = 1.43$ [1.36, 1.49] (grassroots) vs 1.64 [1.56, 1.72] (experts); KS $p = 5.31 \times 10^{-17}$. A smaller exponent means a heavier tail, so grassroots posts are more likely to trigger very large cascades; this gap remains after controlling for follower count (Appendix).

Takeaways (pilot-scale):

- **Heavier tails for grassroots** $\Rightarrow$ more frequent extreme cascades from non-experts.
- Statistical gap is large (power-law exponent gap; KS test highly significant).
- This heavier tail is **not just because experts have more followers on average**: when we compare accounts within the same follower bins, grassroots still produce fatter tails (see Appendix).
- Next: stratified fits by follower bins and archive-only subset (Appendix).

Next Steps and What I Need

Immediate plan (2–3 weeks).

- Finalize pilot for three events (COVID/gene-editing/Jasic), target 15k–30k posts; manual labels $\approx$2k.
- Publish preregistration; finalize AE logit + predicted score; tune Constructiveness weights.
- Run CS-DiD / RDD for identification; keep Hawkes in Appendix (diagnostics); finalize archival/IPW sensitivity dashboards.

Decisions for Today

Pick one per row (proposed in bold).

- **RDD bandwidth: $\pm$10 (main)**; $\pm$20 (robustness)
- **AE primary measure: $\mathrm{AE}^{\mathrm{bin}}$ (primary)**; $\mathrm{AE}^{\mathrm{pc}}$ (sensitivity)
- **Event set (heterogeneity): Zhang Xuefeng + Jiang Ping + 996.ICU**; (alt: add COVID-19 / gene editing as legacy comparators)

Why these defaults?

- $\pm$10 gives tighter local fit; $\pm$20 reported as robustness.
- $\mathrm{AE}^{\mathrm{bin}}$ is interpretable, auditable; $\mathrm{AE}^{\mathrm{pc}}$ reserved for sensitivity.
- Zhang & Jiang: high-salience education/science cases; 996.ICU: labor benchmark with a clean topical boundary.

Open science.
Code + aggregated outputs to OSF/GitHub; IRB in progress; PII de-identified.

Appendix: RDD Technical Details

Cutoff: Top-50 trending; local-linear, triangular kernel; donut $\pm$1–2 ranks.
Bandwidths: $\pm$10 (main), $\pm$20 (robustness); placebo cutoffs 40/60.
Checks: McCrary density; covariate balance within wave; SEs clustered by account.
Fuzzy RDD: first-stage jump in $\Pr\big(\mathrm{AE}^{\mathrm{bin}}{=}1\big) \Rightarrow$ 2SLS LATE on diffusion outcomes.

Appendix: Data Volume by Month (Diagnostics)

Post volume over time. Monthly counts by event window indicating coverage and gaps.

- Clear spikes around major events (e.g., Wuhan lockdown, Dr. Li Wenliang's death, Shanghai 2022 lockdown, Zero-COVID pivot).
- A pronounced dip in Feb 2020, despite intense public attention, likely reflects censorship-driven deletions and incomplete archival coverage.
- An archive-only subset is also plotted (not shown here) as a robustness check for deletion bias.
- Time-sliced queries mitigate API caps; residual truncation is noted during the sharpest peaks.

(Figure placeholder: insert monthly line plot per event when ready.)

Note: counts reflect our collection windows (pilot), not platform-wide totals.

Governance: Soft vs. Hard Censorship

Soft censorship. Algorithmic / operational shaping of visibility (de-ranking, delayed push, limited-audience flags).

Hard censorship. Takedowns, account bans, legal/administrative removals.

Relevance: AE proxies capture "soft" visibility shifts; archive-only checks address potential hard takedowns.

Appendix: Sentiment Labeling Diagnostics

Confusion matrices. Shown for (i) lexicon vs human and (ii) BERT vs human, with precision/recall/F1 by class.

Disagreement audit. Most errors arise from sarcasm/irony and domain-specific terms; adding a technical-term filter reduces false positives.

Stress-tests.

- Lexicon perturbation/swap (alternative dictionaries); results stable within CIs.
- Topic-wise within-event relabeling (placebo): coefficients remain stable.
- Threshold sensitivity for continuous scores: IRR patterns unchanged.

Appendix: Technical Checkpoints & Dynamics (Q&A)

Diagnostics & Validation.

- **RDD:** McCrary density and covariate-balance checks.
- **AE logit:** training/holdout AUC, PR, F1; calibration plot.
- **Label validation:** confusion matrix, Cohen's κ targets.
- **Deletion bias:** retention $\hat{r}_{e,t}$ estimates; IPW sensitivity tables.

Dynamics (moved from main).

- Hawkes reproduction $\mathcal{R} = \alpha/\beta$ (hourly resolution caveat).
- ABM used only for mislabel stress-tests (not for identification).
- *Why appendix?* Story-first in main talk; details on request.

Appendix: Keyword Panels (for replication)

Purpose. Exact search terms (EN/pinyin transliterations) used for data collection. This page is for replication/Q&A.

Event	**Query terms (OR-panel; case-insensitive; de-dup by post ID)**
COVID-19 (2020–2022)	`COVID`, `COVID-19`, `coronavirus`, `SARS-CoV-2`, `epidemic`, `pandemic`, `Wuhan pneumonia`
Gene editing (2018–2019)	`gene editing`, `genetic editing`, `CRISPR`, `CRISPR-Cas9`, `gene therapy`, `He Jiankui`
Jasic (2018-07–2018-08)	`Jasic`, `Jasic workers`, `Shenzhen Jasic`, `worker rights`, `labor protest`
996.ICU (2019-03–2019-05)	`996`, `996ICU`, `996.ICU`, `overtime culture`, `996 schedule`, `tech overtime`
Foxconn (various episodes)	`Foxconn`, `Zhengzhou iPhone`, `Foxconn overtime`, `Foxconn strike`
Yue Yuen (2014–2015; retrospectives)	`Yue Yuen`, `Dongguan Yue Yuen`, `shoe factory strike`

Notes. Panels expanded during the pilot via frequent-hashtag snowballing; timestamps normalized to CST; exact time windows will be documented in the preregistration. Obvious spam/ad terms excluded; posts de-duplicated by canonical post ID.