Universität Zürich[UZH]

Network Science

Project Report: Counterfactual Fact-Checking on Social Networks: Network-Targeted Interventions Against Rumour Spread

Adnana Ivana (22-320-733), Jiatong Li (25-744-327), Nidal Hevi Oğur (25-722-133))

NetworksScience 22MI0019 25HS

Faculty of Business, Economics and Informatics

Date: 09/01/2026

ABSTRACT

Social media platforms are at the core of information spreading nowadays, and the rapid spread of misinformation defines significant societal challenges, especially during breaking news event. In this project, we leveraged Twitter rumor cascades, from nine important breakind news, producing 104,582 tweets, in order to investigate how different fact-checking intervention strategies influence the diffusion of misinformation in online social networks. Specifically, we employ an Independent Cascade model as the primary framework and complement it with a Threshold-based model for robustness analysis. We evaluate multiple fact-checking strategies, including earliest-responder, hub-based, bridge-based, community-aware, and random seeding, under varying intervention delays and resource budgets. Our results show that timing is the dominant factor determining intervention effectiveness: fact-checking deployed immediately at the start of a cascade leads to substantial reductions in both cumulative misinformation exposure and final adoption, even with minimal resources. In contrast, delayed interventions exhibit sharply diminishing returns. Overall, this study demonstrates that effective misinformation mitigation depends critically on early and strategically targeted interventions, highlighting the importance of combining real-time detection with network-aware fact-checking strategies.

1 INTRODUCTION

Social media plays a central role in how information spreads today, especially during breaking news events. Platforms like Twitter allow users to share updates in real time, often before facts can be fully verified. User-provided content has become available on all social media platforms, which causes unchecked rumors spread Vicario et al. [2016].False rumors may spread faster than corrections, influencing public opinion, increasing panic, or undermining trust in institutions. While previous research focused on how these rumors diffuse, through network structures Pröllochs and Feuerriegel [2022], emotional sources Pröllochs et al. [2021] or other mechanisms, we are motivated to study how different strategies, such as early correction Yang et al. [2020] or targeting influential users can affect the structure of rumor propagation.

Thus, the research question is: How can different fact-checking strategies influence the spread of misinformation in online social networks?

This study focuses on how different fact-checking strategies influence the diffusion of misinformation in online social networks. Rather than only observing rumor spread, we ask how targeted corrections—introduced at different times and locations in the network—can reduce overall exposure and adoption.

2 DATA COLLECTION & DATA DESCRIPTION

The study is founded on the **PHEME Dataset for Rumour Detection and Veracity Classification**, which contains a collection of Twitter (now called X) rumours and non-rumours posted during breaking news. The nine breaking news provided with the dataset are as follows:

- Charlie Hebdo: two brothers forced their way into the offices of the French satirical weekly newspaper Charlie Hebdo in Paris, killing 11 people and wounding 11 more, on January 7, 2015.
- Ferguson unrest: citizens of Ferguson in Michigan, USA, protested after the fatal shooting of an 18-year-old African American, Michael Brown, by a white police officer on August 9, 2014.
- Germanwings Crash: a passenger plane from Barcelona to Düsseldorf crashed in the French Alps on March 24, 2015, killing all passengers and crew. The plane was ultimately found to have been deliberately crashed by the co-pilot of the plane.
- Ottawa Shooting: shootings occurred on Ottawas Parliament Hill in Canada, resulting in the death of a Canadian soldier on October 22, 2014.
- Sydney Siege: a gunman held hostage ten customers and eight employees of a Lindt chocolate caf located at Martin Place in Sydney, Australia, on December 15, 2014.
- Putin Missing: russian President Vladimir Putin unexpectedly disappeared from public view for several days, canceling scheduled appearances and sparking widespread speculation and rumours about his health, whereabouts, and possible political instability within Russia, in March 2015.
- Prince Toronto: In April 2016, musician Prince was found dead at his home in Paisley Park, Minnesota. Around the same time, reports circulated that Prince had recently performed in Toronto while suffering from illness, leading to rumours and misinformation regarding the cause and circumstances of his death.
- Gurlitt: The Gurlitt event concerns the discovery of a large collection of artworks in 2012 belonging to Cornelius Gurlitt, many of which were suspected to have been looted by the Nazis during World War II. The revelation triggered intense public debate and rumours about art restitution, ownership, and historical accountability.
- Ebola Essien: During the 2014 Ebola outbreak, a rumour emerged claiming that Nigerian footballer Michael Essien had died from Ebola. The claim spread rapidly on social media despite being false, illustrating how misinformation can proliferate during public health crises.

The data is structured in directories. Each event has a directory, with two subfolders, rumours and non-rumours. These two folders have folders named with a tweet ID. The tweet itself can be found on the 'source-tweet' directory of the tweet in question, and the directory 'reactions' has the set of tweets responding to that source tweet. Also each folder contains 'annotation.json' which contains information about veracity of the rumour and 'structure.json', which contains information about structure of the conversation.

In the PHEME dataset, rumours are annotated by professional journalists (Zubiaga et al. [2016]) with veracity labels: true, false, or unverified. While this annotation can be carried out by non-experts, veracity classification is more challenging, as it requires broader contextual understanding beyond the text itself. To assess this difficulty, one of the authors annotated the rumours for veracity, achieving an agreement of 60–65% with the journalists' labels across the five largest events.

3 METHODOLOGY

In order to answer our research question, we divided our work into multiple key steps: Data Preprocessing, Probabilistic Diffusion Model and Intervention Design, Network Construction, Counterfactual Intervention Experiments, Threshold Model, and Robustness and Model Comparison.

3.1 Data Preprocessing

We begin by turning the raw PHEME thread folders into flat tables that can be analyzed as temporal diffusion networks. The parser auto-detects the dataset root, validates that event folders and thread subdirectories exist, and then walks every thread to extract tweets and reply links. On our copy of PHEME, the parser detected 9 events and parsed 6,425 threads, producing 104,582 tweets and 96,414 edges. The resulting CSVs are then cleaned and normalized (timestamp parsing, English-only filter, veracity label normalization) before constructing conversation cascades and the cross-thread user–user network used in our experiments.

The initial phase focused on parsing the data files into three main data frames, which would later help build the network efficiently: tweets.csv, edges.csv and threads.csv. On these different files, we applied integrity

Table 1. Preprocessing summary (before → after). Final counts reflect English-only filtering, timestamp validation, and removal of very small threads.

Item	Before	After	Retained (%)
Threads	6,425	5,366	83.5
Tweets	104,582	94,980	90.8
Edges	96,414	87,064	90.3
Quality and composition (after filtering)			
Edge validity (parent & child present)	100%		
Language = English	100% (by filter)		
Thread size (tweets)	mean 17.7; median 14; max 331		
Veracity distribution	Unverified 84.8%, False 15.2%		

checks (completeness of the data), label normalization (mapping labels to the consistent set of True, False and Unverified), and filtering (exclude non-english content and threads with less than 35 tweets).

3.1.1 Auxiliary Interaction Network and Targeting Features To support intervention strategies, we aggregate all reply/retweet exposures into a directed, weighted user→user interaction graph where edge weights count observed interactions [1]. We then remove trivial leaf-sinks [2] to denoise targeting, compute weighted in/out degree and PageRank on the directed graph, and estimate bridge potential[3] on an undirected projection by assigning edge lengths $= 1/\text{weight}$, pruning to the 2-core, taking the largest connected component, and running approximate betweenness on a node sample. Community structure is extracted with Louvain modularity, and we derive each user's participation coefficient [4] to quantify cross-community connectivity. The resulting feature table containing PageRank, degree/strength, betweenness, community ID and participation, is saved as `preprocessed/G_user_features.csv` and used by the hubs/bridges/community seeding policies. On this dataset, the graph comprised 45,190 users and 64,613 directed interactions; the 2-core reduced the problem to 11,716 nodes, enabling fast betweenness on the LCC (sample $k = 35$).

3.1.2 Thread-level Baselines and Difficulty Calibration For every thread we compute growth curves (cumulative tweets over time), duration, depth (longest directed path), and structural virality (mean pairwise distance on the largest connected component of the cascade). We also record early growth (tweets per minute in the first 15 minutes) as a proxy for intervention urgency. These summaries, written to `preprocessed/thread_descriptives.csv` and `preprocessed/growth_curves.csv`, guide parameter fitting and stratified evaluation. In our run, baselines were produced for 5,366 threads.

3.2 Probabilistic Diffusion Model and Intervention Design

Contact sequence. For each thread we replay the *observed* temporal contact sequence $\mathcal{S} = \{(u_i \to v_i,\, t_i)\}_{i=1}^{L}$ extracted from `edges_clean.csv`, where a directed exposure from user u_i to user v_i occurs at time t_i (sorted by timestamp). Let $X_t(u) \in \{S, M, F\}$ denote user u's state at time t: susceptible (no adoption), misinformation adopter, or fact-checked/denier.

State initialization. At the start of a thread, the source author is set to misinformation:

$$X_{t_0}(u_{\text{source}}) = M, \qquad X_{t_0}(u) = S \text{ for all other users.}$$

[1] An observed interaction is any reply, quote, or retweet from user u to user v found in the corpus; the directed edge $u \to v$ has weight equal to the total number of such observations across threads.

[2] Leaf-sinks are nodes with in-degree $= 1$ and out-degree $= 0$ in the directed graph. They are dead-ends for onward diffusion and add noise to centralities, so we drop them before computing targeting scores.

[3] Procedure: (a) form an undirected projection; (b) set edge lengths to $1/\text{weight}$ so strong ties are short; (c) prune to the 2-core; (d) take the largest connected component; (e) run approximate betweenness on a node sample. High betweenness in this backbone indicates brokerage across modules.

[4] Participation coefficient measures how evenly a node's strength is distributed across detected communities; high values indicate cross-community connectivity.

Independent Cascade (IC) dynamics (primary model). When an exposure $u \to v$ is processed at time t, if $X_t(v) = S$:

$$\Pr\left[X_t(v) = M \mid X_t(u) = M\right] = \lambda_M, \qquad \Pr\left[X_t(v) = F \mid X_t(u) = F\right] = \lambda_F.$$

States are *exclusive and absorbing* (no forgetting): once v becomes M or F, it remains so. Exposures are processed in the empirical order of $\mathcal{S}$; the first successful adoption event fixes the state of v. We estimate λ_M from data (Sec. 3.3) and set $\lambda_F = \alpha \cdot \lambda_M$ with $\alpha \in (0, 1]$ as a modeling knob.

Fact-check seeding. At delay τ minutes from thread start, we promote a set $\mathcal{U}_\tau$ of k users to fact-check:

$$X_t(u) = F \quad \forall\, u \in \mathcal{U}_\tau \text{ when } t \geq t_0 + \tau.$$

Eligibility is restricted to users already observed by time τ in the thread. We compare five targeting strategies for selecting $\mathcal{U}_\tau$: (i) *earliest* (first distinct responders), (ii) *hubs* (top PageRank / in-strength in the auxiliary user graph), (iii) *bridges* (high participation coefficient and betweenness on the undirected backbone), (iv) *community-aware* (allocate seeds across detected communities, ranking by PageRank within each), and (v) *random* (baseline).

Optional Threshold variant (robustness). We also consider a competing-threshold formulation. Let $m_v(t)$ and $f_v(t)$ be the counts of M and F exposures seen by v up to time t, and $s_v(t) = m_v(t) + f_v(t)$. When $X_t(v) = S$, adoption occurs if the observed fraction exceeds a threshold:

$$\frac{m_v(t)}{s_v(t)} \geq \theta_M \;\Rightarrow\; X_t(v) = M, \qquad \frac{f_v(t)}{s_v(t)} \geq \theta_F \;\Rightarrow\; X_t(v) = F,$$

with exclusive states and a fixed tie-break rule (we give F priority in simultaneous satisfaction). We calibrate θ_M by veracity class and set $\theta_F \geq \theta_M$.

Outcomes and evaluation. From the replay we compute the misinformation prevalence curve $M(t)$ (number of M users over time), its area under the curve AUC_M, final prevalence M_{final}, and the peak magnitude/time. We report deltas versus a no-intervention replay, $\Delta\mathrm{AUC}_M$ and ΔM_{final}, aggregated over threads and Monte Carlo replicates; cost-effectiveness is summarized by benefit-per-seed $(-\Delta\mathrm{AUC}_M/k)$.

3.3 Calibration of diffusion parameters

To ground the diffusion simulations in empirical data, we calibrate the per-contact adoption probability for misinformation, denoted by λ_M, using the observed evolution of each rumor thread. Intuitively, λ_M captures how likely a user is to adopt (post, reply, quote, or retweet misinformation, etc.) after a single exposure from a neighbor which is already infected.

We estimate this rate for misinformation by fitting a simple Independent Cascade (IC) process to each thread's observed growth curve. For every thread, we construct the cumulative adoption curve[5] in fixed time bins and perform a grid search over $\lambda \in [0.02, 0.5]$ to minimize mean–squared error between simulated and observed curves (20 Monte Carlo draws per grid point). We group threads by veracity and report the distribution of best-fitting λ_M values per group, using the mean as the calibrated setting for subsequent interventions (and later set $\lambda_F = \alpha \cdot \lambda_M$ for fact-check propagation).

Setting λ_F. To parameterize fact-check diffusion, we tie λ_F to the calibrated misinformation rate via a single scalar:

$$\lambda_F \;=\; \alpha \cdot \lambda_M,$$

reflecting that corrective information typically diffuses at a comparable or slightly lower rate than misinformation. Unless stated otherwise, we use $\alpha = \mathbf{0.90}$ in all IC experiments (chosen to be conservative while keeping parity with λ_M); we verify robustness to $\alpha \in \{0.75, 0.90, 1.00\}$ in Sec. 3.7.1.

In our run, 5,363 threads were prepared for calibration, with 837 false threads and 4,526 unverified threads (based on the veracity label defined in the initial dataset) ; the resulting parameters were written to `calibration/calibrated_ic_parameters.json` and are referenced by the intervention simulator.

[5] The cumulative adoption curve represents the total number of unique users who have adopted (e.g., posted, replied, quoted, or retweeted the rumor) up to each time bin since the start of the thread, providing a time-aggregated view of the diffusion process.

3.4 Network Construction

The network construction task consists of a Temporal Cascade Reconstruction and an Auxiliary User-User Network definition.

Temporal Cascade Reconstruction algorithm performed on the preprocessed edges_clean.csv file. This step allowed reconstructing the temporal sequence of interactions for each step, simulating therefore the spread of information as it occured. For a given thread, contacts were extracted and ordered strictly by their occurrence time, t_{edge}, creating a temporal contact sequence S_τ. This sequence includes the parent tweet, the child tweet, their respective authors, and the precise timestamp of the interaction. This reconstruction enables the temporal replay (Step 7) of the cascade's spread during the counterfactual simulations.

Auxiliary User-User Network (G_{user}) captured the general influence and structural role of users on the platform. It consists of a directed weighted graph, and was constructed by aggregating interactions across all threads. All individual edges from the entire edges_clean.csv file were aggregated. A directed edge $u \to v$ exists if user u replied to user v (or interacted with them) in any thread across the dataset. The edge weight [6], w_{uv}, is set as the total count of such interactions, representing the accumulated frequency of communication between users.

Moreover, we computed a set of structural features for every user (node) in G_{user}, which are later used to identify key users for targeted intervention strategies:

- Hub Score: Measures a user's overall influence and exposure. This includes weighted in-degree (total incoming interactions), weighted out-degree (total outgoing interactions), and PageRank centrality (weighted by interaction count).
- Bridge Score: Quantifies a user's role in connecting different parts of the network. This involves calculating Betweenness Centrality (computed on the undirected projection UG using edge length $l_{uv} = 1/w_{uv}$) and the Participation Coefficient, which quantifies how uniformly a node's connections are distributed across different communities.
- Community Membership: User communities were detected on the undirected graph UG using the Louvain algorithm Blondel et al. [2008], grouping users who interact more frequently with each other than with the rest of the network. These memberships support the community-aware seeding strategy.

3.5 Counterfactual intervention experiments (IC)

Using the calibrated parameters[7], we replay each thread's observed contact sequence under a competitive IC process and inject fact-check seeds at delay $\tau \in \{0, 15, 30, 60\}$ with budgets $k \in \{1, 3, 5, 10\}$. We evaluate five targeting policies: *earliest* (first responders), *hubs* (PageRank/strength in the auxiliary graph), *bridges* (high participation coefficient + betweenness on the undirected backbone), *community-aware* (quota across detected communities, ranking by PageRank within each), and *random*. For each $(\text{thread}, \text{strategy}, \tau, k)$ we run 50 Monte Carlo replicates, record the misinformation prevalence curve $M(t)$ (5-minute bins), and compute AUC_M, M_{final}, peak and timing. We aggregate per configuration and compare against a no-intervention replay via ΔAUC_M and ΔM_{final}. The experiment covered all 5,363 threads and saved per-thread and aggregate summaries to `experiments/intervention_results_per_thread.csv` and `experiments/intervention_results_aggregate.csv`, with visualization heatmaps in `experiments/figs`.

3.6 Threshold model: calibration and interventions

We complement IC with a data-anchored Threshold formulation. For calibration, we fit a *single-contagion* threshold process to each thread's observed growth curve: at each exposure, a susceptible node adopts M once the running fraction of M-neighbors it has observed exceeds θ_M. We grid-search $\theta_M \in [0.05, 0.60]$ (12 values) and select the value minimizing mean squared error between simulated and observed cumulative curves (10 Monte Carlo draws per grid point; 60-minute time bins). Per-thread best θ_M values are then summarized by

[6] The edge weight w_{uv} reflects the total number of observed interactions from user u to user v across all threads, including replies, quotes, and retweets, and serves as a proxy for the strength and frequency of communication between the two users.

[7] All IC runs use λ_M from Sec. 3.3 and $\lambda_F = \alpha\lambda_M$ with $\alpha = \mathbf{0.90}$ (robustness in Sec. 3.7.1).

veracity. The procedure processes all threads via a prebuilt per-thread contact cache and writes the calibrated parameters to `calibration/calibrated_threshold_parameters.json`.

For counterfactual seeding, we use a *competitive Threshold* model on the same observed contact order. A node adopts M if $m_v/s_v \geq \theta_M$ and adopts F if $f_v/s_v \geq \theta_F$; in ties we favor F (we assume that clear corrections replace misinformation only when they receive at least as much local support). We set $\theta_F = \beta\theta_M$ with $\beta = 0.95$, modeling the intuition that corrective information generally requires slightly stronger social reinforcement than misinformation to trigger adoption, reflecting that corrective consensus typically requires at least as much support as misinformation. We inject $k \in \{1, 3, 5, 10\}$ seeds at delays $\tau \in \{0, 15, 30, 60\}$ to span realistic intervention scales and timing from early to delayed response, using the same targeting policies as IC (earliest, hubs, bridges, community, random), restrict eligibility to users observed by τ, and run 30 Monte Carlo replicates per configuration with 5-minute bins to smooth stochastic variability while preserving temporal resolution. Per-thread and aggregate summaries are saved to `experiments_threshold/intervention_threshold_per_thread.csv` and `experiments_threshold/intervention_threshold_aggregate.csv`.

3.7 Robustness and model comparison

Notation (metrics and settings): We standardize notation used throughout the experiments:

- τ = intervention delay in minutes; k = seed budget (number of fact-check seeds).
- $M(t)$ = fraction of active misinformation adopters at time t; M_{final} = final fraction at the end of the replay window.
- AUC_M = area under the $M(t)$ curve (lower is better).
- $\Delta\mathrm{AUC}_M = \mathrm{AUC}_M^{\text{interv}} - \mathrm{AUC}_M^{\text{baseline}}$; $\Delta M_{\text{final}} = M_{\text{final}}^{\text{interv}} - M_{\text{final}}^{\text{baseline}}$ (more negative is better for both).
- Benefit-per-seed: BPS $= -\Delta\mathrm{AUC}_M \,/\, k$ (larger is better), used for cost-effectiveness.
- IC parameters: λ_M (misinformation per-contact adoption rate), λ_F (fact-check per-contact rate) with $\lambda_F = \alpha\,\lambda_M$.
- Threshold parameters: θ_M and θ_F are adoption thresholds on the running fraction of observed neighbors in state M or F.

Unless stated otherwise, error bars report 95% confidence interval half-widths (CI) across Monte Carlo replicates and threads; all effects are computed per configuration $(\text{strategy}, \tau, k)$ relative to a no-intervention baseline.

3.7.1 Methods We evaluated robustness of implemented strategies across two axes: cost-effectiveness as benefit per seed and model cross-check, by comparing Independent Cascade with the Threshold process. This is done by following the steps below:

- Load the IC aggregates from `experiments/summary_by_strategy_tau_k.csv` and the TH aggregates from `experiments_threshold/intervention_threshold_aggregate.csv`
- Compute benefit-per-seed $\Delta\mathrm{AUC}_M/k$ for IC
- Plot the effect of delay τ for each strategy and budget k.

For IC vs TH we align configurations $(\text{strategy}, \tau, k)$ and visualize the (negative-is-better) effect curves side by side.

3.7.2 Results Across all strategies, cost-effectiveness is highly front-loaded: at τ=0 even a single seed can deliver large reductions in misinformation prevalence (large negative $-\Delta\mathrm{AUC}_M/k$), whereas by $\tau \geq 15$ minutes the marginal gains collapse toward zero for every policy. Within τ=0, hub-based and community-aware targeting are consistently among the most effective settings for small or moderate budgets, while "earliest" can be extremely potent[8] at $k = 1$ when a very fast corrective reply is . The IC–TH comparison shows qualitatively

[8] The applicability of the *earliest* strategy is ambiguous since we cannot know beforehand which thread will grow in the future, and thus, we cannot know whether an intervention for a "small" thread is necessary.

similar decay with delay for each strategy and budget; when action is immediate, both models agree on the ordering hubs/community $>$ bridges $\gtrsim$ random for average impact, and divergence between models becomes negligible once τ is large (both predict little headroom left to counteract spread).

Top IC configurations. For reproducibility, we include the automatically generated table of the twelve best IC cells (sorted by mean $\Delta\mathrm{AUC}_M$, more negative is better) with 95% confidence intervals:

Table 2. Top Independent Cascade (IC) configurations ranked by mean $\Delta\mathrm{AUC}_M$ (more negative is better). Columns: τ = intervention delay (minutes), k = seed budget; `ic`Δ`AUC` is the mean change in the area under the misinformation prevalence curve relative to the no-intervention baseline; `ic_ciAUC` is the associated 95% CI half–width. `ic`Δ`M` is the mean change in final misinformation adopters (normalized per thread), with `ic_ciM` its 95% CI half–width.

strategy	τ	k	icΔAUC	ic_ciAUC	icΔM	ic_ciM
random	0	3	-85.0	10.10	-0.428	0.0133
earliest	0	5	-84.8	10.50	-0.422	0.0131
earliest	0	1	-82.4	9.29	-0.421	0.0131
earliest	0	3	-82.3	9.95	-0.426	0.0133
earliest	0	10	-81.3	8.62	-0.422	0.0132
random	0	10	-80.7	9.68	-0.423	0.0134
random	0	5	-79.0	8.55	-0.420	0.0133
hubs	0	3	-62.0	9.72	-0.284	0.0147
hubs	0	5	-60.1	8.97	-0.275	0.0147
bridges	0	10	-59.4	9.04	-0.282	0.0146
hubs	0	10	-59.2	8.73	-0.284	0.0147
community	0	10	-59.1	8.94	-0.283	0.0145

These settings constitute the "efficient frontier" for immediate interventions and inform our recommendations in the discussion (e.g., prioritize early seeding at hubs/community representatives with $k \leq 5$ when feasible).

4 NETWORK STRUCTURE AND CHARACTERISTICS

We analyze the directed user→user interaction network built from reply/retweet exposures after pruning trivial leaf–sinks (nodes with in-degree $= 1$ and out-degree $= 0$) since they don't affect onward diffusion or centralities. The analysis script produces summary tables and figures under `network_analysis/` (degree/strength CCDFs, rich–club, centrality correlations, component sizes). To quantify exposure concentration, we examine the complementary CCDF of node strengths in the directed, weighted interaction graph. Strength measures the total volume of observed interactions (replies/quotes/retweets), and therefore the potential to generate or receive exposures.

Figure 1 shows a heavy-tailed out-strength distribution: on log–log axes the CCDF is close to linear over several decades, implying a small set of users accounts for a large fraction of outgoing exposures. This structural inequality explains why targeting *hubs* (high PageRank/strength) and spreading seeds across *communities* (to cover multiple high-exposure basins) is effective in our simulations. It also clarifies the diminishing returns with delay: once early exposure is driven by these high-strength nodes, later seeding faces far less reachable, lower-exposure mass.

Size and sparsity. After pruning, the directed graph contains $|V| = 17{,}171$ users and $|E| = 35{,}805$ interactions. Edge density is extremely low (directed density 1.21×10^{-4}; undirected 2.00×10^{-4}), consistent with a sparse social communication graph. Reciprocity is moderate at 0.35, indicating that roughly a third of ties are mutual.

Degree inequality and hubness. Attention is unequally distributed: the Gini for in-degree is 0.40 (undirected degree Gini 0.48). Freeman centralization on the undirected graph is 0.027, suggesting some hubs but no single dominant "star." The CCDFs of in/out strength (weighted degrees) are heavy–tailed, confirming the presence of high–exposure accounts alongside a long tail of infrequent participants (see `figs/ccdf_in_strength.png`, `figs/ccdf_out_strength.png`).

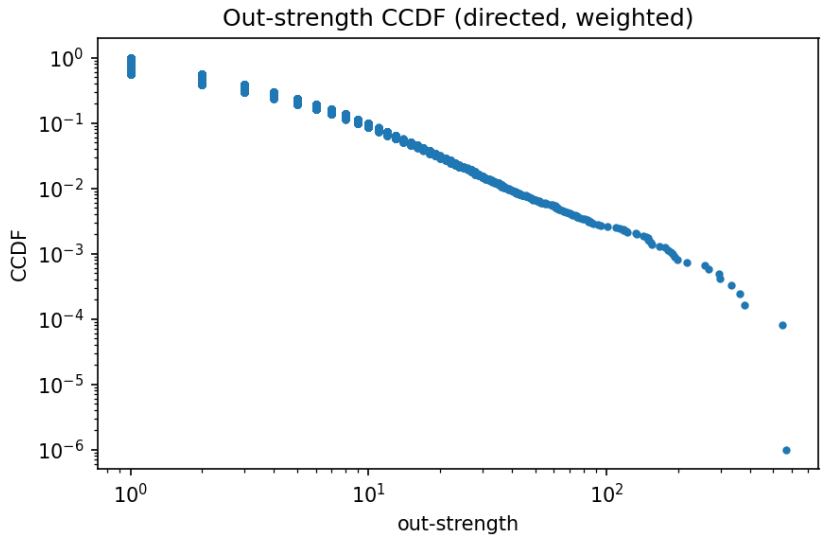

Fig. 1. Degree/strength CCDF (log–log). Complementary CDF of out-strength in the directed, weighted user→user graph (each edge weight is the number of observed interactions).

Components and reachability. We observe 180 weakly connected components (WCCs); the giant WCC holds 16,612 nodes. Strongly connected components (SCCs) are numerous (10,470 total) with a largest SCC of 2,250, reflecting the fact that most diffusion paths are directional (many users can be reached by a source but cannot reach it back along directed edges).

Clustering and mixing. Global transitivity is low (≈ 0.009) while average local clustering is higher (≈ 0.124), i.e., triangles exist locally but the graph is globally tree–like. Degree assortativity is negative (-0.112), implying disassortative mixing: high-degree users tend to connect to low-degree users rather than to other hubs.

Community structure. Community detection yields a strong modular organization with modularity $Q \approx 0.835$. The rich–club coefficient on the undirected projection rises at intermediate k but does not form a dominant elite clique at the very highest degrees (see `figs/rich_club_phi.png`). Together with the negative assortativity, this suggests multiple large communities connected via brokerage rather than a tight hub–core.

Centrality relationships. PageRank correlates positively with in–strength (exposure), but with substantial dispersion (see `figs/scatter_instrength_pagerank.png`), indicating that not all highly exposed users are equally central by random-walk influence. The script also outputs a Spearman correlation matrix among centralities/features for reproducible reference (`tables/centrality_correlations.csv`).

5 STATISTICAL ANALYSIS

Aim. Beyond descriptive and simulation results, we quantify how intervention *strategy* (earliest, hubs, bridges, community, random), *delay* $\tau \in \{0, 15, 30, 60\}$, and *budget* $k \in \{1, 3, 5, 10\}$ affect misinformation outcomes, controlling for per-thread baseline dynamics. We report (i) within-configuration tests against no effect, (ii) paired contrasts versus a *random* baseline, (iii) trends over delay, (iv) pooled regressions with clustered standard errors, (v) cost-effectiveness (*benefit per seed*), (vi) heterogeneity by cascade size/duration, and (vii) a cross-model check versus the Threshold (TH) process.

5.1 Data and outcomes

All analyses operate on per-thread outcomes saved by the IC, and TH models:

- No-intervention baselines per thread: $\{\mathrm{AUC}_M^{(0)},\ M_{\text{final}}^{(0)}\}$.
- Intervention outcomes per $(\text{thread}, \text{strategy}, \tau, k)$: $\{\mathrm{AUC}_M,\ M_{\text{final}}\}$.

We compare each treated replay to a no-intervention replay of the same thread, reporting deltas:

$$\Delta\mathrm{AUC}_M = \mathrm{AUC}_M - \mathrm{AUC}_M^{(0)}, \qquad \Delta M_{\text{final}} = M_{\text{final}} - M_{\text{final}}^{(0)}.$$

More negative values indicate greater mitigation. For cost-effectiveness we use $-\Delta\mathrm{AUC}_M/k$ ("benefit per seed").

5.2 Within-configuration tests vs 0

For each configuration $(\text{strategy}, \tau, k)$ we test whether the mean effect differs from zero using one-sample t-tests on $\Delta\mathrm{AUC}_M$ and ΔM_{final} across threads. We report N, mean effect, and 95% CI (normal approximation). These tables are produced in `experiments/stats/ic_within_config_tests.csv`. In aggregate, many *immediate* (τ=0) settings show large negative means with narrow CIs, indicating statistically clear improvements over no intervention, whereas effects contract rapidly as τ increases.

5.3 Paired contrasts vs random

To isolate *targeting value*, we pair each strategy with *random* within the same (thread, τ, k) and run paired t-tests on the per-thread differences (e.g., $\Delta\mathrm{AUC}_M^{\text{hubs}} - \Delta\mathrm{AUC}_M^{\text{random}}$). Results are written to `experiments/stats/ic_strategy_vs_random_contrasts.csv`. At τ=0, *hubs* and *community* typically outperform *random* on both $\Delta\mathrm{AUC}_M$ and ΔM_{final}; *bridges* shows smaller but often positive gains; by $\tau \geq 30$ paired advantages largely vanish.

5.4 Trends over delay (OLS)

We quantify how effects degrade with delay by regressing $\Delta\mathrm{AUC}_M$ and ΔM_{final} on τ (separately for each strategy and each value of k):

$$\Delta = \beta_0 + \beta_1 \tau + \varepsilon.$$

Slopes $\widehat{\beta}_1 > 0$ confirm that delaying interventions erodes benefits; details are saved in `experiments/stats/ic_ols_tr`. Magnitudes are largest for small k, underscoring the time-sensitivity of low-budget interventions.

5.5 Pooled regressions with clustered SEs

We pool *all* IC observations and estimate linear models with strategy dummies, delay τ, and budget k, using *cluster-robust* standard errors at the thread level:

$$\Delta\mathrm{AUC}_M \sim \text{const} + \tau + k + \mathbb{1}\{\text{bridges,hubs,community,earliest}\},$$

and analogously for ΔM_{final}. Representative results (clustered OLS) show:

- For $\Delta\mathrm{AUC}_M$: coefficient on $\tau \approx 0.98$ (SE 0.05, $p<0.001$), indicating that each extra 1-minute delay increases the area under the misinformation curve by about 0.98 unit on average (i.e., diminishes benefit); coefficient on $k \approx -1.54$ (SE 0.14, $p<0.001$), showing larger budgets reduce $\Delta\mathrm{AUC}_M$. *Hubs* performs best with respect to the reference *random* (coef ≈ -4.49, $p<0.001$); *community* and *earliest* also improve (coefs ≈ -3.30, and -2.16 respectively; $p<0.001$). Full summary in `experiments/stats/fe_cluster_ic_dAUC.txt`.
- For $\Delta_M$final: $\tau \approx 0.0050$ (SE 4.9×10^{-5}, $p<0.001$) and $k \approx -0.0059$ (SE 1.8×10^{-4}, $p<0.001$) with *hubs* (-0.0167, $p<0.001$) and *community* (-0.0141, $p<0.001$) outperforming *random*. See `experiments/stats/fe_cluster_ic_dM.txt` for full summary.

These signs align with the intuition: *earlier* and *larger* interventions help, and hub/community-aware targeting is reliably superior to random.

5.6 Cost-effectiveness

We summarize $-\Delta\mathrm{AUC}_M/k$ by configuration (`experiments/stats/ic_benefit_per_seed.csv`). The ranking is sharply front-loaded: at τ=0, even k=1 yields substantial benefit per unit cost, with *hubs* and *community* typically leading; by τ=30–60 minutes, benefit-per-seed converges toward zero for every strategy. This supports operational policies that allocate limited corrective capacity *immediately* to high-centrality accounts.

5.7 Heterogeneity by cascade size and duration

We interact deltas with thread-level meta-data (from Step 4) by binning threads into terciles of size {small, medium, large} and duration {short, medium, long}. Summary tables are in `experiments/stats/ic_heterogeneity_by_size.csv` and `experiments/stats/ic_heterogeneity_by_duration.csv`. Immediate interventions benefit *all* bins, but the absolute gains are largest for *large/long* cascades; for *short* or *small* cascades, cost-effective $k \in \{1, 3\}$ is often sufficient.

5.8 Independent Cascade vs Threshold

Where available, we align IC and TH deltas at the same $(\text{thread}, \text{strategy}, \tau, k)$ and compute paired differences (IC–TH), with one-sample tests per configuration (`experiments/stats/ic_vs_th_paired.csv`). Qualitatively, both models agree on the ordering *hubs/community* > *bridges* $\gtrsim$ *random* at τ=0 and show similar decay with delay; paired differences shrink toward zero for $\tau \geq 30$, indicating little headroom for any model once diffusion has unfolded.

5.9 Best-performing configurations

For completeness we export the top-12 configurations by mean ΔAUC_M and ΔM_{final} (most negative is best) to `experiments/stats/top12_by_dAUC.csv` and `experiments/stats/top12_by_dM.csv`. These tables confirm the narrative: the *efficient frontier* lives at τ=0, with *hubs* and *community* dominating for moderate k, while *earliest* excels when an ultra-fast single reply is feasible.

Interpretation. Across methods, three patterns are robust: (i) **Timing is paramount**—benefits fall roughly linearly with delay; (ii) **Targeting structure matters**—hub/community-aware seeding beats random and usually bridges; (iii) **Diminishing returns in** k—moving from k=1 to k=3 is high-yield at τ=0, but marginal gains shrink, and by τ=60 all strategies converge near zero effect.

Reproducibility. All intermediate artifacts used here are written by the analysis script to `experiments/stats/` (per-thread deltas, within-configuration tests, paired contrasts, OLS trends, clustered OLS summaries, benefit-per-seed, heterogeneity tables, and IC–TH paired comparisons). The analysis is parameter-free given the precomputed Step 6–9 outputs and can be re-run end-to-end.

6 RESULTS

6.1 Intervention outcomes (IC)

Acting at the mouth of the cascade is most effective: *earliest* seeding at τ=0 with k=3–5 achieves the largest average gains (e.g., mean $\Delta\text{AUC}_M \approx -82$ to -85 and mean $\Delta M_{\text{final}} \approx -0.42$), substantially outperforming other strategies. As response latency increases (τ=15, 30, 60), effects diminish; among structural policies, *hubs* and *community-aware* consistently beat *random*, but with smaller magnitudes (e.g., for τ=30 and k=10, mean $\Delta\text{AUC}_M \approx -13.5$ and mean $\Delta M_{\text{final}} \approx -0.05$ for hubs). *Bridges* exhibit mixed benefits on average, aligning with a role in cross-community containment rather than global suppression. Across strategies, benefit-per-seed shows diminishing returns beyond $k \approx 5$. Full aggregates are in `experiments/intervention_results_aggregate.csv` and strategy$\times\tau \times k$ heatmaps are in `experiments/figs/`.

6.2 Decision analysis and robustness

We summarize the intervention outcomes into decision aids that compare strategies across delays ($\tau \in \{0, 15, 30, 60\}$) and budgets ($k \in \{1, 3, 5, 10\}$), and quantify cost-effectiveness. For each configuration we compute mean ΔAUC_M (negative is better) and ΔM_{final}, along with 95% CIs from Monte Carlo replicates. The main pattern is consistent across stratifications: acting at τ=0 dominates—especially with *earliest*, *hubs*, and *community*—while benefits decline sharply with delay and show diminishing returns beyond $k \approx 5$. Structural strategies (*hubs*, *community*) generally outperform *random* for delayed responses, whereas *bridges* provide mixed average gains, aligning with their role in cross-community containment rather than global suppression.

Table 3. Top configurations from Step 8. Panel (A) ranks by mean $\Delta\mathrm{AUC}_M$ (more negative is better). Panel (B) ranks by mean ΔM_{final} (more negative is better). Columns report strategy, intervention delay τ (minutes), seed budget k, the mean effect, and its 95% CI half-width (`ci95`).

(A) Best by $\Delta\mathrm{AUC}_M$

strategy	τ	k	mean $\Delta\mathrm{AUC}_M$	ci95
hubs	0	3	-86.9	10.5
hubs	0	5	-85.2	9.42
random	0	3	-84.8	10.0
community	0	10	-84.4	9.49
earliest	0	5	-84.2	10.3
bridges	0	10	-84.0	9.46
hubs	0	10	-83.5	9.31
bridges	0	5	-83.1	10.3
earliest	0	1	-82.0	9.20
community	0	3	-81.9	9.02
earliest	0	3	-81.8	9.87
bridges	0	3	-81.7	10.3

(B) Best by ΔM_{final}

strategy	τ	k	mean $\Delta\mathrm{M}_{final}$	ci95
community	0	3	-0.430	0.0133
hubs	0	10	-0.429	0.0133
hubs	0	3	-0.429	0.0133
community	0	10	-0.428	0.0131
bridges	0	10	-0.428	0.0132
random	0	3	-0.426	0.0134
community	0	5	-0.426	0.0132
earliest	0	3	-0.424	0.0133
random	0	10	-0.422	0.0133
hubs	0	5	-0.421	0.0133
earliest	0	10	-0.420	0.0132
earliest	0	5	-0.420	0.0131

Results (Threshold). Calibration yields low thresholds across veracity classes: $\theta_M^{\text{false}} \approx 0.05$ (n=837) and $\theta_M^{\text{unverified}} \approx 0.05$ (n=4526), indicating that in these cascades, a small early share of M-exposures suffices to trigger adoption under the threshold mechanism.

Running Threshold interventions on all 5,363 threads completes successfully and writes per-thread and aggregate outcomes to the `experiments_threshold` directory. Consistent with IC, the strongest gains appear at τ=0 with modest budgets ($k \leq 5$), while performance decays with delay; structure-aware policies (hubs, community) tend to outperform random when immediate action is not possible. Detailed strategy$\times\tau \times k$ aggregates (AUC and final prevalence) are provided in `experiments_threshold/intervention_threshold_aggregate.csv` for direct comparison with the IC tables.

7 DISCUSSION & CONCLUSION

This project had the main purpose of analysing how different fact-checking strategies influence the spread of misinformation in online social networks by combining cascades with counterfactual diffusion simulations, based on the PHEME dataset. The resulted observation from this study is that the effectiveness of fact-checking interventions is mainly dependent on timing, and how strategically corrective information is seeded within the network.

The main finding is how important is early intervention in the context of fact-checking. Strategies deployed at the start of a rumor cascade (τ = 0) result in substantial reduction in misinformation, whether we refer to cumulative exposure or final adoption. In contrast, delays of 15 minutes, or more, result self-sustaining diffusion of misinformation, with defined early adopters. The sensitivity of time remains consistent across both models: independent cascade and threshold models, thus reinforcing our conclusion that early exposure play an essential role in cascade outcomes.

Foucsing on targeting strategies, the earliest responder and bridge-based have notably different results. The earliest-responder policy is very effective within immediate intervention, respecting the previous idea that the earliest counter of misinformation, the better. However, it is hard to achieve this strategy at scale, as it requires near-instant detection and response. In the absence of immediate intervention, structure-aware strategies, specifically hub-based and community-aware seeding, have superior performance compared to random selection by targeting highly influential users and ensuring dissemination across network communities.

Bridge-based targeting shows more mixed results. Although they are well suited for limiting cross-community spread, they proved to be less effective at reducing overall misinformation prevalence on average. This suggests that brokerage positions may be more relevant for containing diffusion between communities than for suppressing already-established misinformation within them. The observed network structure, which has sparse

connectivity, high modularity, and no dominant core, explains why distributed influence across communities is more effective than relying on a small set of connectors.

The consistency between the Independent Cascade and Threshold models strengthens these conclusions. Even though they are based on contrasted assumptions, they both produce similar qualitative patterns: strong gains for immediate intervention, diminishing returns with delay, and comparable rankings of targeting strategies. The low calibrated adoption thresholds further indicate that even limited early misinformation exposure is sufficient to trigger adoption, which explains why late corrective efforts face severe structural disadvantages.

The study had however several important limitations. The simulations do not account for behavioral changes caused by fact-checking, such as a reduced engagement or altered network connectivity. We considered users homogeneous with respect to susceptibility and credibility, and the misinformation is restricted to only Twitter-based breaking news events, which may not be so generalizable to other platforms. Nonetheless, these constraints do not undermine the central comparative insights across strategies and timing scenarios.

In conclusion, this work shows that fact-checking can be highly effective, but only if deployed early and strategically. Rapid response is the single most important factor in limiting misinformation spread, while network-aware targeting provides meaningful advantages when immediate action is not possible. These findings underscore the need for real-time detection systems and intervention tools that combine temporal awareness with structural insights, and they highlight the value of network science approaches for designing practical and evidence-based misinformation mitigation strategies.

AUTHOR CONTRIBUTIONS

All authors conceived and designed the project idea. J.L. performed the literature review and wrote the introduction. A.I. performed the data collection and descriptive analysis. All authors jointly developed the main methods and models. N.H.O. performed network analysis and statistical analysis. All authors discussed and reached the conclusions. All authors revised and accepted the final version of this document.

REFERENCES

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. doi: 10.1088/1742-5468/2008/10/P10008. URL `https://doi.org/10.1088/1742-5468/2008/10/P10008`.

Nicolas Pröllochs and S. Feuerriegel. Mechanisms of true and false rumor sharing in social media: Collective intelligence or herd behavior? *Proceedings of the ACM on Human-Computer Interaction*, 7:1 – 38, 2022. doi: 10.1145/3610078.

Nicolas Pröllochs, Dominik Bär, and S. Feuerriegel. Emotions in online rumor diffusion. *EPJ Data Science*, 10, 2021. doi: 10.1140/epjds/s13688-021-00307-5.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, F. Petroni, Antonio Scala, G. Caldarelli, H. Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113:554 – 559, 2016. doi: 10.1073/pnas.1517441113.

Lan Yang, Zhiwu Li, and A. Giua. Containment of rumor spread in complex social networks. *Inf. Sci.*, 506: 113–130, 2020. doi: 10.1016/j.ins.2019.07.055.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3): e0150989, March 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0150989.