When Do Propagation Models Beat Cascade Summaries? A Released-Signal Audit of Text-Free Misinformation Detection

Supplementary material

Xuan-Bach Le Phat T. Tran-Truong Son Ha Xuan Duy Nguyen

WSDM 2027, Paper 2182

Correspondence: lexuanbach@hcmut.edu.vn
Assigned paper DOI: https://doi.org/10.1145/3862322.3863898

Scope

Part A documents the implemented full diagnostic configuration: heterogeneous hypergraph construction, next-slice contrastive pretraining, the stratum-averaged adjustment head, and total-variation (TV) disagreement abstention. Part A isolates each component's contribution through matched diagnostics. The five-seed analysis finds no resolved marginal hypergraph gain, while the adjustment and abstention diagnostics add no distinct class decisions at the reported operating point. Part B collects the full cross-dataset matrix, future-test details, retweet-tree experiments, title-only reference, feature ablations, calibration metrics, subsampling stability, size-stratified splits, and tuning-parity checks. Together, the two parts connect the main findings to their protocols and saved outputs.

Part A: Audited testbed diagnostics

1 Notation and Setup

Part A concerns only the binary event-list diagnostic testbed. Part B also reports the four-class Twitter15/16 audit. In Part A, a story $s \in \mathcal{S}$ carries a binary label $y_s \in \{0, 1\}$ (1= misinformation) and a text-free cascade-summary vector $\mathbf{c}_s \in \mathbb{R}^7$, as defined in the main paper's released-signal section. The audit testbed additionally represents s as a set of hyperedges over a heterogeneous hypergraph $\mathcal{H}=(\mathcal{V}, \mathcal{E})$ built *per data split*. The node set partitions into four families

$$\mathcal{V} = \mathcal{V}_{\text{user}} \,\dot\cup\, \mathcal{V}_{\text{url}} \,\dot\cup\, \mathcal{V}_{\text{media}} \,\dot\cup\, \mathcal{V}_{\text{slice}},$$

(legacy "user" nodes, URLs, media objects, and cascade-time slices), and each hyperedge $e \in \mathcal{E}$ carries a type $\tau(e) \in \mathcal{T} = \{\text{share}, \text{reply}, \text{coexpose}, \text{burst}\}$. In the event-list releases, the first family contains event proxies rather than identified users. The implementation constructs share edges by URL and 20-minute window, approximates reply threads by URL and five-minute bin, forms coexposure edges by media identifier, and creates one burst edge per story over its non-empty slices. The two diagnostic stratification variables are early-window popularity $P \in \mathbb{Z}_{\geq 0}$ and hash-bucket concentration $E_{\text{hash}} \in [0, 1]$, discretized into $K_P{=}K_E{=}4$ strata indexed by (p, e). For four early-proxy bucket shares q_b, $E_{\text{hash}} = (\max_b q_b - 1/4)/(3/4)$. This measures bucket concentration, not communities or echo chambers. We write $H{=}\mathbf{z}_s$ for the fused story representation, f_θ^{obs} for the observational head, and f_θ^{adj} for the diagnostic stratum-averaged head.

2 Hypergraph Encoder: Construction

The hypergraph encoder combines event proxies and auxiliary objects through type-aware attention. We use a two-layer hypergraph neural network (HGNN) (hidden dimension 64, four attention heads, dropout 0.2) in the hyperedge-to-node message-passing family of [2, 6]. Let $\mathbf{x}_v^{(0)}$ be the input feature of node v. Event-proxy inputs contain log degree and an eight-bin degree indicator, while URL, media, and slice inputs contain log counts. These numeric features are padded to a common dimension, and node/hyperedge types parameterize the attention queries. Each layer alternates a node→hyperedge update with a hyperedge→node update.

Node → hyperedge. For hyperedge e with incident node set $N(e)$ and type $\tau(e)$,

$$\begin{aligned} \mathbf{m}_e^{(\ell)} &= \sum_{v \in N(e)} \alpha_{e,v}^{(\ell)} \, \mathbf{W}_{\tau(e)}^{(\ell)} \, \mathbf{x}_v^{(\ell)}, \\ \alpha_{e,v}^{(\ell)} &= \frac{\exp\big(\langle \mathbf{q}_{\tau(e)}^{(\ell)}, \, \mathbf{W}_{\tau(e)}^{(\ell)} \mathbf{x}_v^{(\ell)} \rangle\big)}{\sum_{w \in N(e)} \exp\big(\langle \mathbf{q}_{\tau(e)}^{(\ell)}, \, \mathbf{W}_{\tau(e)}^{(\ell)} \mathbf{x}_w^{(\ell)} \rangle\big)}, \end{aligned} \tag{1}$$

where $\mathbf{W}_\tau^{(\ell)}$ is a type-specific transform and $\mathbf{q}_\tau^{(\ell)}$ is a per-type query.

Hyperedge → node. For node v incident to hyperedges $N(v)$,

$$\mathbf{x}_v^{(\ell+1)} = \sigma\Big(\mathbf{x}_v^{(\ell)} + \textstyle\sum_{e \in N(v)} \beta_{v,e}^{(\ell)} \, \mathbf{m}_e^{(\ell)}\Big), \tag{2}$$

with $\beta_{v,e}^{(\ell)}$ defined symmetrically to α and σ a ReLU nonlinearity. After L=2 layers, the structural readout mean-pools the cascade-slice nodes of story s,

$$\mathbf{z}_s^{\text{struct}} = \frac{1}{|\mathcal{V}_{\text{slice}}(s)|} \sum_{v \in \mathcal{V}_{\text{slice}}(s)} \mathbf{x}_v^{(L)}, \tag{3}$$

which is fused residually with the summary branch, $\mathbf{z}_s = \mathbf{z}_s^{\text{struct}} + \psi(\mathbf{c}_s)$. Setting $\psi \equiv 0$ recovers a structure-only model, while zeroing $\mathbf{z}_s^{\text{struct}}$ recovers a summary-only model. Table S8 reports marginal changes of $+0.000/+0.007$ on the powered datasets, with neither comparison resolved.

Release-induced degeneracy. The four event-list releases contain one URL and no media object per cascade, while event proxies are nearly unique ($u \approx n$). Reply edges are URL/time-bin approximations rather than observed threads, and the current burst builder creates one edge per story. These choices leave little recurring cross-story relational signal. The unresolved marginal gain therefore concerns this release-specific construction. Recurring verified identities would enable a different test.

3 Next-Slice Contrastive Pretraining

The implementation samples 30% of cascade slices that have a successor in the same story. For an anchor slice a, the positive v^+ is the next five-minute slice from that story. The default sampler draws eight negative slices from other stories: up to half from stories in the top decile of event-proxy-set Jaccard similarity and the remainder uniformly from all slice nodes. Label-free sampling never reads the story label. A separate label-aware switch draws negatives from opposite-label stories.

A bilinear head scores the anchor against the positive and negative candidates. With candidate set $\mathcal{C}(a) = \{v^+\} \cup \mathcal{N}^-(a)$, the implemented cross-entropy loss is

$$\mathcal{L}_{\text{ssl}} = -\mathbb{E}_a \log \frac{\exp g_\phi(a, v^+)}{\sum_{v \in \mathcal{C}(a)} \exp g_\phi(a, v)}. \tag{4}$$

$$(P, E_{\text{hash}}) \to H, \qquad (P, E_{\text{hash}}) \to Y, \qquad H \to Y$$

$$f_\theta^{\text{adj}}(h) = \sum_{p,e} \hat{q}(p, e)\, f_\theta(h, p, e)$$

The adjusted head conditions on (P, E_{hash}) and averages over its empirical strata.

Figure S1: Hypothetical motivation for the adjustment diagnostic. Its arrows are not identified from the event-list data. The implemented head simply averages predictions over empirical strata.

Equation (4) trains next-slice prediction. Its negatives mix data-dependent overlap sampling with uniform sampling, so the standard marginal-negative InfoNCE mutual-information bound does not apply. In the exploratory three-seed CoAID ablation, label-free pretraining raises macro-F1 from 0.531 to 0.563. The original full arm matches the label-aware diagnostic seed for seed at 0.587, but its sampler flag was not stored. This identity supports that attribution but does not establish the mode of every historical DLF-Full run. This downstream diagnostic motivates testing the pretext at the primary five-seed depth.

4 Stratum-Averaged Adjustment Diagnostic

The implementation conditions a classifier on the fused representation $H{=}\mathbf{z}_s$ and the discretized pair (P, E_{hash}), then averages its predicted probabilities over the empirical training-stratum mass:

$$f_\theta^{\text{adj}}(h) = \sum_{p,e} \hat{q}(p, e) f_\theta(h, p, e). \tag{5}$$

An annealed Kullback-Leibler agreement penalty encourages f_θ^{adj} to remain close to f_θ^{obs}. Figure S1 records the hypothetical motivation used when the component was designed.

Equation (5) is a stratum-averaged prediction head, not an identified intervention. Both P and E_{hash} contribute to the learned representation, and neither has an established pre-treatment or confounder interpretation in these releases. A causal interpretation would require additional identification assumptions and evidence beyond this computation.

Observed behavior. At the reported operating point, permuting (P, E_{hash}) leaves the adjusted probabilities bit-identical. Replacing hash concentration with a Louvain diagnostic has the same outcome. Across adjustment-agreement weights $\lambda_2 \in [0.1, 10]$, the mean absolute probability gap is at most 0.005, while 2.8 to 4.2% of CoAID argmax decisions change. At the reported point, bit-identical probabilities imply zero Brier and ECE change. Elsewhere, the sweep only measures a small probability gap. This diagnostic neither establishes nor excludes confounding.

5 Empirical Stress Tests

The artifact retains two exploratory stress tests. A synthetic $\gamma \in [1, 3]$ perturbation injects label noise without instantiating a confounder between representation and outcome and therefore is not a Rosenbaum sensitivity analysis. A separate 1.8× popularity-stratum reweight changes macro-F1 by less than 0.10 in the recorded run. That result is an empirical summary-feature perturbation, not a total-variation risk guarantee. Neither check identifies a causal estimand or validates the adjustment head. A formal sensitivity analysis would require a defined treatment/estimand, an applicable hidden-bias model, and a complete estimator and interval.

6 TV-Disagreement Abstention Rule

The rule uses disagreement between the observational and adjusted heads:

$$d_s = \mathrm{TV}\Big(f_\theta^{\mathrm{adj}}(\mathbf{z}_s), f_\theta^{\mathrm{obs}}(\mathbf{z}_s)\Big), \qquad \text{abstain if } d_s > \kappa. \tag{6}$$

The threshold κ is selected on validation data for target coverage c=0.9 [3]. The statistic can rank difficult stories only when its values vary across stories. The current heads provide little such variation: the weight sweep changes mean disagreement from 0.0051 to 0.0041, and the reported operating point has no probability or class difference. A secondary seed-level non-degeneracy check succeeds on only 2/3 CoAID seeds and 1/3 ReCOVery seeds.

Calibrated-confidence abstention has macro-F1 0.638/0.692 on CoAID/ReCOVery at coverage 0.875/0.841, compared with 0.555/0.676 for TV disagreement at coverage 0.792/0.785. Coverage is therefore not matched. These are macro-F1 values, not selective-risk scores. This comparison evaluates the disagreement statistic in the audited configuration. General conclusions about abstention require matched-coverage tests.

7 Diagnostic Requirements

The adjustment head needs defensible covariates and predictions that change across their strata. Here, permuting (P, E_{hash}) leaves the reported outputs unchanged, and the λ_2 sweep keeps $\mathbb{E}|p_{\mathrm{adj}} - p_{\mathrm{obs}}| \le 0.005$. The abstention rule needs variation in this disagreement and must improve over calibrated confidence at matched coverage. The retained comparison uses unequal coverage and cannot establish that advantage. The hypergraph encoder needs recurring verified identities or other cross-story relations and a resolved marginal improvement over the summary branch. Its five-seed marginal changes are $+0.000/+0.007$ on ReCOVery/PolitiFact (Table S8). The audited event-list releases cannot establish these conditions.

Part B: Additional results

8 Dataset card

This section expands the main paper's released-signal table into a detailed dataset card. Every statement here is checkable against a file in the release or in `outputs/r13/signal_matrix/signal_matrix.json`.

Provenance and terms. All nine corpora are public research releases (CoAID, ReCOVery, FakeNewsNet PolitiFact/GossipCop, UPFD PolitiFact/GossipCop, Twitter15/16, Weibo). We use each under the terms published with it by its original distributor and redistribute none of them. The artifact ships loaders and split files, not corpus data. The retained files contain no recoverable corpus licence, so users must consult the original distributors for applicable terms.

Hydration. No tweet, post or account is hydrated. In the event-list regime the releases distribute tweet/retweet *ids* only. Timestamps are decoded from the Twitter snowflake encoding of the id itself, and each id is used as an event proxy. Consequently $u \approx n$, no stable user recurs across cascades, and our analysis does not attempt account re-identification.

Auxiliary objects are degenerate. Every cascade in all four event-list releases carries exactly one URL and zero media nodes: 2,275 CoAID and 1,603 ReCOVery cascades measured directly, and 677 PolitiFact and 17,977 GossipCop cascades in `outputs/r12/example`. Two of the seven summary features ($\log r$, $\log m$) are therefore constant by construction, and a third ($\log u$) duplicates $\log n$. This is a property of the releases, not a measured stability.

Truncation. The event-list loader caps each cascade at 200 events. Of the 677 PolitiFact stories, 271 sit at that cap, which bounds $\log n$ throughout that regime. No cap is applied in the retweet-tree regime.

Class balancing and subsampling. The four event-list corpora are class-balanced before splitting, and GossipCop is additionally subsampled to 1,500 stories per class. Every seed redraws *both* the subsample and the split, so the reported ±std already absorbs subsample variation. A second independent family of five draws

moves macro-F1 by at most 0.011 (§17). Prevalence in the unbalanced releases is 5/28/54/23% misinformation on CoAID/ReCOVery/PolitiFact/GossipCop. The UPFD releases and Weibo are near-balanced as distributed, and Twitter15/16 are four-class with roughly equal class sizes.

User features. Only the two UPFD corpora expose per-node user information through released 10-d aggregate profile vectors rather than raw profiles. Twitter15/16 and Weibo ship no user features at all, so node features there are purely structural (depth one-hot, log out-degree, root flag, log size). This is why the "profile" regime rests on two corpora.

Event times. None of the five retweet-tree releases, as distributed here, carries an event-time field: the UPFD raw directories hold adjacency, labels, node-feature matrices and split indices only, and the BiGCN-released RvNN tree files hold `(eid, parentIdx, childIdx, ..., bag-of-words)`. Chronological analysis therefore uses the event-list corpora.

Deletion and correction status. None of the nine releases exposes whether a source article was later deleted, corrected or retracted, or whether a sharing account was later suspended. The pipeline does not model these statuses, so its predictions cannot distinguish corrected stories from uncorrected ones on that basis.

9 Historical inference

The eight secondary GCNFN/BiGCN seed-paired tests are archived in `outputs/r11/seed_paired_gnn/seed_paired_gnn.json`; five have $p < 0.05$. They are outside the 23-test family.

The event-list multiplicity family originally combined 12 story-clustered contrasts (PropMLP+, RandomForest, and fixed 7-feature HistGBT versus PropMLP on each corpus) with 11 pooled-row contrasts (four DLF-Full, four GATv2, two DLF-Full+, and the hidden-size check). We re-estimated the four DLF-Full and two DLF-Full+ contrasts by resampling stories, keeping all seeds of a story together. Their stored per-story predictions reproduce every historical per-seed macro-F1 and the archived row-level estimates, intervals, and p-values exactly. Point estimates are unchanged, and the clustered p-values are 0.38, 0.61, 0.21, and 0.24 for DLF-Full on CoAID, ReCOVery, PolitiFact, and GossipCop, and 0.65 and 0.57 for DLF-Full+ on ReCOVery and PolitiFact. The GATv2 comparator was a separate PropMLP run with unsaved predictions, so the four GATv2 contrasts keep archived row-level values. The hidden-size check resamples five seed-level differences. Holm and BH resolved sets are identical before and after re-estimation. The standalone `paired_rf` and original rich-summary outputs also use pooled rows. The main text uses the clustered rich-summary estimates from `rev_stats`. In a cluster contrast, losses are averaged within story before resampling stories. In a row contrast, repeated story-seed observations can be sampled independently. The latter can understate uncertainty. The archived Holm/BH values are reproducible arithmetic, but they do not turn the mixed-unit family into uniformly valid clustered inference. No historical contrast was refitted or silently replaced with a newer run.

Three HistGBT configurations have distinct recorded five-seed macro-F1 means:

Configuration	CoAID	ReCOVery	PolitiFact	GossipCop
A: tuned, 15 features	.629	.646	.706	.877
B: fixed, 7 features	.571	.665	.670	.871
C: fixed, 15 features	.704	.670	.698	.878

A is the main descriptive table (`r9/stanford2`). B is the event-list family and power comparison (`r10/rev_stats`, `r10/power`). C is the rich-summary comparison and label-permutation control (`r10/ssl_tree`, `r13/label_permutation`). These results are not interchangeable. The stale GATv2 rows in the historical power file describe an earlier run and are excluded from the paper's MDE claims. The Weibo clustered file contains GCN and summaries rather than the complete graph pool. GCN still has the highest descriptive graph score in the full table. Its interval is conditional on that selected model, not selection-adjusted.

The three-seed label-free pretraining result is separate from the five-seed main-table DLF-Full configurations. The r7 full ablation has the same first three CoAID scores as the r8 label-aware diagnostic. Main-table DLF-Full uses the later r9 rich-summary outputs, whose CoAID scores differ. Those outputs do not retain a per-run sampler flag, so the table is not evidence of a uniquely label-free pretraining benefit. The artifact provides a provenance ledger and recomputes available standard deviations with ddof=0 from saved seed values. The main-table GossipCop PropMLP+ mean is retained from the transfer summary. Its SD is omitted because the matching seed vector is

Table S1: Cross-dataset transfer, macro-F1 (5 seeds). Diagonal in bold.

Model	train ↓ / test →	CoAID	ReCOVery	PolitiFact	GossipCop
HistGBT (15d)	CoAID	**0.629**	0.472	0.521	0.361
	ReCOVery	0.522	**0.646**	0.516	0.574
	PolitiFact	0.489	0.445	**0.706**	0.645
	GossipCop	0.387	0.494	0.414	**0.877**
HistGBT (7d)	CoAID	**0.587**	0.489	0.429	0.437
	ReCOVery	0.485	**0.631**	0.527	0.504
	PolitiFact	0.445	0.432	**0.651**	0.443
	GossipCop	0.425	0.557	0.399	**0.862**
RandomForest (15d)	CoAID	**0.669**	0.449	0.557	0.284
	ReCOVery	0.515	**0.689**	0.529	0.602
	PolitiFact	0.394	0.427	**0.721**	0.613
	GossipCop	0.345	0.508	0.371	**0.877**
RandomForest (7d)	CoAID	**0.613**	0.445	0.432	0.397
	ReCOVery	0.491	**0.657**	0.538	0.485
	PolitiFact	0.397	0.387	**0.645**	0.417
	GossipCop	0.414	0.577	0.385	**0.862**
PropMLP+ (15d)	CoAID	**0.610**	0.449	0.522	0.349
	ReCOVery	0.548	**0.699**	0.513	0.510
	PolitiFact	0.450	0.472	**0.717**	0.595
	GossipCop	0.362	0.525	0.523	**0.851**
PropMLP (7d)	CoAID	**0.618**	0.426	0.475	0.398
	ReCOVery	0.500	**0.659**	0.488	0.451
	PolitiFact	0.366	0.511	**0.657**	0.541
	GossipCop	0.411	0.604	0.483	**0.836**

unavailable. A later validation-selection run has the same rounded mean but a different SD, and is not substituted for that historical row.

10 Cross-dataset transfer: full matrix

Table S1 gives the complete transfer matrix for all six summary models. Panel B of the main text's generalization table shows HistGBT-15d and PropMLP+. Rows are the training corpus and columns the test corpus. Standardization uses the source training split, HistGBT tunes on the source validation split, and all cells use five seeds. Uniform random binary prediction has expected macro-F1 near 0.5, while a constant predictor scores 0.333. Transfer gaps (in-domain minus mean off-diagonal) range from 0.11 to 0.47 across the six models: 0.11 to 0.18 for ReCOVery sources, 0.14 to 0.24 for CoAID, 0.18 to 0.24 for PolitiFact, and 0.34 to 0.47 for GossipCop sources.

11 Temporal split: details

The temporal split orders stories by their first event and holds out the latest 10% (deterministic within each seed-dependent balanced sample). Priors below pool the five runs, whereas the main drift diagnostic uses seed 0. Future-test sizes are n=64 (PolitiFact, prior 0.54 misinformation), n=300 (GossipCop, 0.60), and n=92 (ReCOVery, 0.99). The ReCOVery future era is almost entirely misinformation, so macro-F1 is degenerate there (0.22 to 0.37 for every model) and the main text omits it. Importance weighting (IW) fits a logistic train-vs-future domain classifier on the features, weights training stories by the clipped density ratio $[0.1, 10]$, and re-fits the tree. Table S2 gives paired accuracy deltas. Its historical bootstrap pools story-seed rows without clustering repeated stories, so these intervals are descriptive, not population-level evidence. No interval excludes zero.

Table S2: Importance weighting for the trees on the future-test era: paired accuracy delta (IW − base) with historical pooled-row 95% bootstrap CI and p. Repeated stories are not clustered.

	ReCOVery	PolitiFact	GossipCop
HistGBT (15d)	-0.017 $[-0.048, 0.013]$, p=0.30	$+0.009$ $[-0.019, 0.038]$, p=0.58	-0.003 $[-0.012, 0.007]$, p=0.62
RandomForest (15d)	$+0.004$ $[-0.022, 0.033]$, p=0.80	$+0.013$ $[-0.016, 0.041]$, p=0.44	-0.003 $[-0.009, 0.004]$, p=0.47

Table S3: Log-transformed vs. raw counts/durations in the 7-d summary (macro-F1, 5 seeds, same splits).

	CoAID	ReCOVery	PolitiFact	GossipCop
PropMLP, log / raw	0.618 / 0.555	0.659 / 0.666	0.657 / 0.678	0.836 / 0.852
RandomForest, log / raw	0.613 / 0.605	0.657 / 0.664	0.645 / 0.654	0.862 / 0.863
HistGBT (7d, default), log / raw	0.561 / 0.561	0.618 / 0.618	0.652 / 0.649	0.841 / 0.841

12 Retweet-tree release details

Data. UPFD [1] PolitiFact has 314 trees under the official 62/31/221 split. GossipCop has 5,464 trees under the official 1,092/546/3,826 split. Both provide 10-dimensional profile attributes, and we do not treat repeated vectors as stable identities. Twitter15 has 1,470 four-class trees, Twitter16 has 808 four-class trees, and Weibo has 4,527 binary trees. These releases provide structural inputs comprising a depth one-hot vector of length 8, log out-degree, a root indicator, and log tree size. **Summary vector.** Each tree yields log size, depth, log maximum width, mean branching, leaf fraction, log root fan-out, mean depth, structural virality, and depth per node. UPFD additionally uses the mean, standard deviation, and maximum of each profile attribute, producing 39 dimensions in total. **Models.** Logistic regression, RandomForest, validation-tuned HistGBT, and a two-layer MLP form the summary family. The graph family comprises our GCN and GATv2 implementations with the UPFD readout, and BiGCN with raw root features distributed through both branches and mean pooling. We select learning rates and stopping epochs on validation data and z-score node features from training nodes. **Statistics.** Tables display five-seed mean macro-F1. Paired inference uses story-level accuracy. The script selects the graph and summary member by mean test macro-F1 before bootstrapping their 0/1-loss difference. The reported intervals and TOST results are therefore descriptive and not selection-adjusted. Across the five corpus comparisons, nominal p-values are 0.71/0.0106/0.6876/0.5596/0.2956. A post-hoc Holm correction maps the UPFD-GossipCop value to 0.053 and all others to 1.0. **Timing.** Per seed, summary models take 0.4 to 7 seconds, GCN takes 0.4 to 9, GATv2 takes 0.7 to 21, and BiGCN takes 1 to 25 seconds on CPU. UPFD-PolitiFact GCN reaches 0.783 macro-F1, consistent with the published profile-feature operating point.

13 Title-text reference point

We compare title text with cascade summaries on the same stratified splits and seeds as the event-list tables. Title-only term frequency-inverse document frequency (TF-IDF) logistic regression uses unigrams and bigrams, sublinear term frequency, and C=4. It reaches $0.790_{0.019}$ macro-F1 on PolitiFact and $0.765_{0.019}$ on GossipCop. Logistic regression on the 15-d summary reaches $0.743_{0.025}$ / $0.848_{0.009}$, and HistGBT (15d) reaches $0.706_{0.052}$ / $0.872_{0.023}$. Concatenating title TF-IDF with the summary in one logistic regression gives $0.811_{0.037}$ / $0.875_{0.008}$. Every story has a released title. Fusion improves the measured point estimates. Deletion and paraphrase robustness require separate tests.

14 Raw vs. log features

Table S3 shows that tree-model macro-F1 changes by at most 0.01 under the transform, while PropMLP moves by -0.06 on the 24-story CoAID fold and by $+0.01$ to $+0.02$ on the powered folds. This sweep uses a per-seed `random_state` and a default HistGBT, so its tree cells need not reproduce the event-list table of the main text exactly. The log-versus-raw contrast within each row is unaffected. We keep log features in the main text for comparability with the published loaders. The powered-fold changes do not alter the reported comparisons.

Table S4: 15-d vs. 15-d + 6 point-process/growth features (macro-F1, 5 seeds). Δacc: paired per-story accuracy gain of the 21-d HistGBT over the 15-d HistGBT [95% CI], p.

	CoAID	ReCOVery	PolitiFact	GossipCop
HistGBT 15d / 21d	0.629 / 0.615	0.646 / 0.675	0.706 / 0.741	0.877 / 0.882
RandomForest 15d / 21d	0.669 / 0.679	0.689 / 0.697	0.721 / 0.741	0.877 / 0.877
PropMLP+ 15d / 21d	0.610 / 0.616	0.699 / 0.710	0.717 / 0.731	0.851 / 0.857
Δacc HistGBT	−0.008	+0.028	+0.034	+0.005
95% CI	[−0.083, 0.067]	[−0.011, 0.067]	[−0.003, 0.072]	[−0.005, 0.015]
p	0.90	0.18	0.10	0.34

15 Point-process and growth-curve features

Per cascade we fit a univariate exponential-kernel Hawkes process by maximum likelihood (baseline rate μ, branching ratio α, decay β, using four restarts with L-BFGS-B and times in hours) and compute three growth-curve descriptors (log hour of peak hourly rate, peak-hour share of events, log-log slope of cumulative count vs. time over the first 24 h). Table S4 compares each model on the 15-d summary vs. the 21-d vector under identical splits and seeds. The displayed changes are unresolved on ReCOVery/PolitiFact and small on GossipCop, so the reported event-list conclusions are unchanged.

16 Calibration: AUPRC and Brier

Table S5 completes the ROC-AUC block of the main text's natural-prevalence table with AUPRC and Brier scores at each corpus's natural prevalence, from the same run (`outputs/r10/calibration`, 5 seeds). Base rates are CoAID 5%, ReCOVery 28%, PolitiFact 54%, GossipCop 23%, so the AUPRC scale differs sharply by corpus and AUPRC is comparable only *within* a column. Brier is dominated by the base rate for the same reason: CoAID's ≈0.05 reflects its 5% prevalence, not good calibration.

The ordering matches the ROC-AUC picture in the main text's natural-prevalence table. PropMLP+ has the best Brier on ReCOVery (0.186) and PolitiFact (0.184), HistGBT the best on GossipCop (0.077), and CoAID point estimates cluster at 0.051 to 0.052. Isotonic recalibration of HistGBT improves ECE on three corpora in the separate ECE run. In the ranking run reported here, Brier changes by (+0.000 CoAID, −0.004 ReCOVery, +0.015 PolitiFact, +0.002 GossipCop), while AUPRC falls on all four corpora (−0.020/ − 0.024/ − 0.072/ − 0.020): it moves probability mass without improving the ranking.

Table S5: AUPRC and Brier at natural prevalence, 5 seeds (source `outputs/r10/calibration`). Lower Brier is better. AUPRC is comparable only within a column because base rates differ.

	CoAID	ReCOVery	PolitiFact	GossipCop
AUPRC				
PropMLP	0.066	0.430	0.739	0.763
PropMLP+	0.074	0.470	0.823	0.823
RandomForest	0.116	0.445	0.815	0.817
HistGBT	0.100	0.423	0.815	0.822
HistGBT +iso	0.080	0.399	0.743	0.802
Brier				
PropMLP	0.051	0.194	0.205	0.092
PropMLP+	0.051	0.186	0.184	0.078
RandomForest	0.051	0.192	0.202	0.078
HistGBT	0.052	0.202	0.203	0.077
HistGBT +iso	0.052	0.198	0.218	0.079

17 GossipCop subsampling stability

GossipCop is capped at 1,500 stories per class. The cap is applied inside the loader, which shuffles the positive and negative pools with `Random(seed)` before cutting, and the same seed then drives the class-stratified split. Each of the five seeds therefore redraws *both* the subsample and the split, so the ±std reported in the main text's event-list table already absorbs subsample variation rather than holding the subsample fixed.

To check that the particular family of five draws is not special, we repeat the whole five-seed protocol with a seed offset of 1,000, giving a second, independent family of five subsample-plus-split draws. Macro-F1 moves by +0.011 (logistic regression $0.827_{0.012} \rightarrow 0.838_{0.015}$), +0.010 (RandomForest $0.865_{0.022} \rightarrow 0.875_{0.010}$) and +0.007 (GradientBoosting $0.859_{0.014} \rightarrow 0.866_{0.012}$). Every gap is smaller than the corresponding base standard deviation, and the model ordering is unchanged, supporting stability across these two families of draws. Source: `outputs/r9/tabular_baselines/gossipcop_stability.json`.

18 Cascade-size-stratified split

Stories are ordered by cascade size n and split deterministically 80/10/10: *small→large* trains on the smallest 80% and tests on the largest 10%, while *large→small* is the mirror image. The five seeds average model randomness only. The *random* column is a class-stratified split run inside the same script with the same models and seeds, so the Δ rows are internally consistent. The random RandomForest and HistGBT cells differ from the event-list table of the main text because here they use a per-seed `random_state` and an untuned HistGBT. Three of those four cells differ by 0.004 or less, while RandomForest on PolitiFact differs by 0.022 on a 64-story fold whose seed standard deviation is 0.052.

The protocol confounds size with prevalence. Sorting GossipCop by cascade size also sorts it by label: the largest-10% fold is 76.3% misinformation and the smallest-10% fold 93.7%, against a balanced training prior. The large GossipCop drops in Table S6 therefore combine size extrapolation with substantial prior shift. PolitiFact's test priors (51.6% and 34.4%) are closer to balance. Even there, the small-to-large split lowers PropMLP, RandomForest, and HistGBT, while PropMLP+ changes by +0.026. The mixed directions rule out a uniform size-split effect in this comparison. Separating size extrapolation from prior shift requires a new evaluation with prevalence held fixed across size strata.

Table S6: Cascade-size-stratified split (macro-F1, 5 seeds). Δ is against the random split in the same run. Test-fold misinformation prior in a separate row. Source `outputs/r11/size_split`.

		random	**small→large**	**large→small**
PolitiFact	PropMLP	0.657	0.610 (−0.047)	0.652 (−0.005)
	PropMLP+	0.717	0.743 (+0.026)	0.655 (−0.062)
	RandomForest	0.645	0.600 (−0.045)	0.631 (−0.014)
	HistGBT	0.710	0.674 (−0.036)	0.686 (−0.024)
	test prior	bal.	0.516	0.344
GossipCop	PropMLP	0.836	0.345 (−0.491)	0.495 (−0.341)
	PropMLP+	0.851	0.553 (−0.298)	0.593 (−0.258)
	RandomForest	0.862	0.503 (−0.359)	0.555 (−0.307)
	HistGBT	0.881	0.619 (−0.262)	0.606 (−0.275)
	test prior	bal.	0.763	0.937

19 GNN Tuning Parity

The shared validation-selected grid is the four (epochs, lr) points $(25, 5{\times}10^{-4})$, $(40, 10^{-3})$, $(40, 5{\times}10^{-3})$, $(80, 5{\times}10^{-3})$. PropMLP's arm adds its own default operating point $(25, 10^{-2})$ as a fifth candidate, so the flat-feature side selects from five and the graph side from four. Selection is on the validation fold only. On CoAID: GCNFN $0.586_{0.099}$, BiGCN $0.608_{0.123}$, RGCN $0.599_{0.124}$, EBGCN $0.611_{0.138}$. Under that same grid, PropMLP reaches 0.591 and PropMLP+ 0.533. On this 24-story fold matched tuning gives a higher graph mean (0.611 vs. 0.591), the only such event-list corpus. The fold's MDE (0.12 to 0.13) cannot resolve the difference. Against

Table S7: Future-test macro-F1 over five seeds under four label-free corrections. The main paper plots paired changes against **base**.

	Method	base	test-std.	quantile	imp.-wt	CORAL
Polit.	PropMLP	0.458	0.313	0.386	0.463	0.450
	PropMLP+	0.483	0.424	0.464	0.505	0.457
	RForest	0.450	0.299	0.345	0.431	0.501
Gossip	PropMLP	0.753	0.748	0.790	0.781	0.760
	PropMLP+	0.786	0.791	0.807	0.809	0.792
	RForest	0.851	0.738	0.847	0.848	0.755

PropMLP's untuned 0.618 the same grid leaves EBGCN below it, illustrating how the comparator budget changes the descriptive ordering. The EBGCN model infers per-incidence edge weights over the same hypergraph with a bidirectional (top-down/bottom-up) encoder. Its rows for ReCOVery, PolitiFact, and GossipCop in the main text's event-list table use the same grid (ReCOVery: $0.675_{0.020}$, per seed 0.666/0.664/0.714/0.656/0.673 against PropMLP 0.685/0.680/0.632/0.670/0.630, seed-paired t-test p=0.50).

Extending the matched-selection sweep to ReCOVery, the powered corpus where the fixed-budget table shows a graph model ahead, reverses the sign of that gap. Under the same grid: GCNFN $0.555_{0.065}$, BiGCN $0.561_{0.070}$, RGCN $0.663_{0.032}$, EBGCN $0.672_{0.022}$, against PropMLP 0.673 and PropMLP+ 0.693. The ordering of EBGCN and PropMLP in the main text's event-list table (0.675 vs. 0.659) is therefore sensitive to matched tuning: under the shared grid, the highest graph mean and PropMLP differ by −0.001, while PropMLP+ is higher by 0.021. (This sweep reproduced EBGCN at 0.672 against the dedicated run's 0.675. The two differ on one seed, giving an observed run-to-run difference of ≈0.003 on this cell.) The same sweep on PolitiFact gives GCNFN $0.581_{0.037}$, BiGCN $0.600_{0.051}$, RGCN $0.631_{0.045}$ and EBGCN $0.658_{0.027}$ against PropMLP 0.671 and PropMLP+ 0.721 (source `outputs/r11/gnn_parity_politifact`). Under the fixed budget EBGCN leads PropMLP there by +0.003 (0.660 vs. 0.657), an unresolved gap. Matched tuning reverses the sign to −0.013 (0.658 vs. 0.671). The reversal is a point-estimate sign change, as that pair has no paired test. In this earlier partial grid, CoAID is the sole cell whose point estimate favours the graph model, and its MDE cannot resolve that gap. The later eight-trial study in Section 31 changes this ordering. The two studies use different candidate grids. The earlier PropMLP parity run gives 0.840 on GossipCop (`outputs/r10/propmlp_parity/propmlp_parity.json`).

20 Future-Test Drift Corrections

The main paper's drift-correction figure shows the *paired* change each label-free correction makes to future-test macro-F1. Table S7 gives the underlying absolute cells. Source `outputs/r10/temporal_drift/summary_v2.json`. Deltas are computed per seed against the **base** column and the interval plotted in the figure is the seed-level 95% Student-t interval over the five paired differences (four degrees of freedom). It includes seed-dependent balancing and model randomness, and is not a story-bootstrap interval or population-level shift guarantee. Positive intervals exclude zero for PolitiFact CORAL/RandomForest, GossipCop importance-weighted PropMLP and PropMLP+, and GossipCop quantile-mapped PropMLP+. These intervals are unadjusted and descriptive.

A fifth, exploratory arm. The shipped JSON also carries a *recency* arm (linear up-weighting of later training stories) that the main paper does *not* count among its four evaluated corrections. We report it only here, as exploratory: paired deltas $-0.109/-0.029/-0.070$ on PolitiFact and $-0.004/+0.001/-0.013$ on GossipCop for PropMLP/PropMLP+/RandomForest. Five of six point estimates are nonpositive. We keep this exploratory arm separate from the four-correction comparison.

21 Structure vs. summary: DLF-Full ablation

The separate size-feature-drop results are archived in `outputs/r11/feature_drop/feature_drop.json`.

Table S8 reports the DLF-Full structure-versus-summary ablation. GossipCop is omitted because each configuration costs ∼800 s per seed. Its central changes are Δstruct +0.047/+0.000/+0.007, with a hidden-128 check. These

Table S8: Structure-versus-summary ablation: macro-F1 over five seeds, hidden size 64, 4+25 epochs. Δstruct is full minus summary-only.

Variant	CoAID	ReCOVery	PolitiFact
structure-only ($\psi \equiv 0$)	$0.538_{0.089}$	$0.556_{0.060}$	$0.582_{0.055}$
summary-only ($\mathbf{z}^{\text{struct}}$=0)	$0.607_{0.130}$	$0.663_{0.032}$	$0.671_{0.044}$
full (fusion)	$\mathbf{0.654_{0.106}}$	$\mathbf{0.663_{0.034}}$	$\mathbf{0.678_{0.041}}$
Δstruct	+0.047	+0.000	+0.007
hidden-128, 8+35 ep.: Δstruct	+0.049	+0.002 (p=0.87)	NA

Table S9: Inputs used in original evaluations and the corresponding audited setting.

Family	What it needed originally	Here
BiGCN / EBGCN	TF-IDF text on every node, root-text enhancement	no text, with structure-only trees 0.28 to 0.35 on Twitter15/16, 0.88 to 0.89 on Weibo (retweet-tree table of the main text)
UPFD	user-preference encodings + news text	profile-only: absolute test-selected GNN-summary gap $\leq$ 1.1 points
Hypergraph (HG-SL)	users recurring across stories	no stable user ID, and repeated profile attributes are not identity
Adjustment head	measurable, defensible covariates	P, E_{hash} describe event proxies, and adjusted output is permutation-invariant
TV abstention	two heads that disagree	same argmax at the reported point, with $\leq 4.2\%$ of CoAID decisions flipping over the λ_2 sweep

comparisons do not establish a marginal hypergraph gain on the powered event-list corpora.

22 Inputs Across Model Families

Table S9 summarizes differences between the inputs used in the original model families and those available in this audit. The main paper discusses the implications in prose.

23 Notation

Table S10 maps the cascade features, hypergraph objects, and prediction heads used in the main text and Part A.

24 Label-permutation negative control

Gradient-boosted trees are strong on tabular inputs and could in principle exploit a label-independent artefact of the split or the features rather than the labels themselves. Flipping 10% of training labels measures noise sensitivity. Permuting the training labels provides a separate check for label-independent signal.

Protocol. Splits, seeds, features and *test* labels are left untouched. Only the training labels are permuted, within the training fold, once per seed. HistGBT on the 15-d summary and RandomForest on the 7-d summary are then retrained and scored on the untouched test fold. Five seeds, four event-list corpora. Script `scripts/rev8_label_permutation.py`, output `outputs/r13/label_permutation`. The **real** column uses fixed hyperparameters rather than validation selection. We include the main event-list table's values to expose the resulting differences.

Decision rule fixed before this diagnostic run. The expected result was a decline toward the no-signal region between the constant-predictor macro-F1 of 0.333 and the roughly 0.50 expectation of uniform random prediction on a balanced binary fold. A corpus materially above 0.50 under permuted training labels would have been reported as a possible label-independent artifact. This rule was recorded in the local script docstring, but it was not externally preregistered.

Result. The eight permuted means in Table S11 span 0.467 to 0.529, and all five-seed intervals contain 0.50. The decline from the same script's real arm ranges from 0.12 to 0.36. Per-seed spread tracks fold size: RandomForest

Table S10: Notation used throughout the main text and this supplement.

s, y_s	**story and its label (credible / misinformation)**
n_s, u_s, r_s, m_s	**cascade size, distinct event proxies, URLs, and media objects**
t_0, t_{cut}	**first event, earlier of 10th event and t_0+1 h (early window)**
P_s, $E_{\text{hash},s}$	**early-window proxy count and normalized dominance of the largest of four deterministic hash buckets $\in [0,1]$**
spread_s	**wall-clock span from first to last event**
$\mathbf{c}_s \in \mathbb{R}^7$ / $\mathbb{R}^{15}$	**cascade-summary vector (PropMLP) / with 8 temporal-dynamics features (PropMLP+)**
$\mathcal{H}=(\mathcal{V},\mathcal{E})$	**per-split heterogeneous hypergraph of event proxies, URLs, media, and time slices**
$\mathbf{z}_s^{\text{struct}}$, $\psi(\mathbf{c}_s)$	**hypergraph readout and summary branch, fused residually**
event-list / retweet-tree	**released event proxies and timestamps without stable user identity or shipped edges / shipped retweet edges with profile or structural attributes**
$f_\theta^{\text{obs}}, f_\theta^{\text{adj}}$	**observed prediction and empirical stratum-averaged diagnostic prediction. Neither is an identified intervention**

Table S11: Label-permutation control over five seeds. Only training labels are permuted, while test labels and splits remain fixed. The real arm uses this script's fixed hyperparameters.

Corpus	Model	n_{test}	Main table	real	permuted [95% CI]
CoAID	HistGBT (15d)	24	0.629	0.704	0.467 [0.392, 0.541]
CoAID	RForest (7d)	24	0.595	0.613	0.469 [0.272, 0.666]
ReCOVery	HistGBT (15d)	92	0.646	0.670	0.506 [0.450, 0.561]
ReCOVery	RForest (7d)	92	0.651	0.657	0.483 [0.383, 0.584]
PolitiFact	HistGBT (15d)	64	0.706	0.698	0.498 [0.410, 0.587]
PolitiFact	RForest (7d)	64	0.667	0.645	0.529 [0.392, 0.665]
GossipCop	HistGBT (15d)	300	0.877	0.878	0.518 [0.453, 0.583]
GossipCop	RForest (7d)	300	0.865	0.862	0.518 [0.423, 0.614]

ranges 0.245 to 0.624 on the 24-story CoAID fold and 0.397 to 0.592 on the 300-story GossipCop fold. This negative control detects no label-independent predictive signal. A repeated-permutation null distribution and tests of other leakage paths remain separate requirements.

25 Degree-preserving edge rewire

We report the edge-rewire control for all four corpora and quantify how much attachment structure the intervention changes.

Protocol. For each cascade, the undirected tree is rewired by a connected double-edge swap (`networkx.connected_double_edge_swap`, $|E|$ swaps), re-rooted at the original root, and the parent pointers re-derived by breadth-first search. This preserves node count, edge count, connectivity and the full degree sequence. The tree therefore stays a tree, with cascade size, root fan-out, leaf fraction and mean branching held *exactly*. Attachment arrangement can change, along with depth, mean depth and structural virality. Only the `edge_index` consumed by the graph model is rewired: the summary baselines keep the intact tree, because the retweet-tree summary vector is rebuilt from the tree itself and rewiring the shared item would move both sides of the comparison. Node features, labels, splits and seeds are unchanged. The rewire is cached per story, so seeds vary only model initialisation. The cache is keyed by story ID. The four corpora have 8,056 rows, but Twitter15 and Twitter16 share 165 IDs with identical parsed parent arrays, leaving 7,891 distinct cached trees. Of these, 7,780 were rewired and 111 were too small to swap (< 4 nodes or < 2 edges). The degree sequence was verified preserved on all 7,780.

Internal control. Because only the graph model's edges changed, the four summary baselines must reproduce bit-identically between the intact and rewired runs. They do, on all four corpora and all five seeds. This is checked programmatically in `scripts/rev8_edge_rewire_analysis.py`.

Result. Table S12 gives the paired comparison. No corpus shows a resolvable change at five seeds. On UPFD-GossipCop, whose test-selected nominal interval for the graph-versus-summary gap excludes zero, the rewired model stays above the tuned tree (0.937 vs. 0.931).

Scope of the rewiring intervention. These cascades are close to stars, so a degree-preserving rewire has little arrangement to destroy. On the 3,826 UPFD-GossipCop test trees, 75.2% of nodes attach directly to the root

Table S12: Degree-preserving rewire for BiGCN over five seeds. Δ is rewired minus intact, and p is from a secondary seed-paired t-test outside the Holm/BH family.

Corpus	intact	seed std	rewired	Δ	p
UPFD-GossipCop	0.9404	0.0089	0.9366	−0.0038	0.23
UPFD-PolitiFact	0.7516	0.0249	0.7553	+0.0038	0.27
Twitter15	0.3418	0.0171	0.3243	−0.0175	0.20
Twitter16	0.3114	0.0236	0.2747	−0.0368	0.25

and mean depth is 1.20. The rewire changes structural virality on 31.6% of trees, by 0.030 on average, and maximum depth on 23.4%, by 0.28. Twitter15 is only slightly less degenerate (75.1% root-attached, mean depth 1.49, with virality changing on 51.2% of trees, by 0.27). This result tests attachment arrangement under the rewiring constraints of these releases. Profile value and broader propagation claims require separate interventions.

26 Validation-only selection: protocol

The main text reports validation-based model choice within the archived component pool. The rule reads validation scores only, but the pool declaration records that CoAID test regret was already known, so this is not a prospectively blinded comparison of pools. The rule and pool are stored with the run. For each corpus, the chosen component is the argmax of mean validation macro-F1 over seeds 0 to 4. Ties break on lower validation standard deviation and then on a declared simplest-first order. The pool is RandomForest, PropMLP, PropMLP+, HistGBT, EBGCN, GATv2, DLF-Full, and DLF-Full+ on CoAID, ReCOVery, and PolitiFact, and the first four of those on GossipCop, where the graph and DLF dumps were not complete.

Three components could not be used as published, because their published protocol reads the validation fold and therefore makes their validation score in-sample. HistGBT selected hyperparameters on validation and refitted on train+validation. It runs a five-fold stratified search inside the training fold over the same grid and fits once on train. EBGCN and GATv2 selected epochs and learning rate on validation. Each uses an inner 85/15 stratified holdout carved from train, with the winner refitted on the full training fold over the same grid and the same number of fits. Five of the eight components reproduce their published main-table cell exactly under regeneration (RandomForest, PropMLP, PropMLP+, DLF-Full, DLF-Full+). The three modified ones do not, which is the expected consequence of removing their access to the validation fold.

Validation macro-F1 by component, averaged over five seeds, places both graph components outside the top two on every corpus where they appear: on CoAID EBGCN is third and GATv2 sixth of eight. On both ReCOVery and PolitiFact, they rank seventh and eighth. A leakage audit (`verify_no_leakage.py`) recomputes the selection with every `test_*` field set to `None` and recovers the frozen manifest exactly across 140 (corpus, model, seed) dumps.

For the retweet-tree corpora the same idea is implemented directly in `scripts/rev6_graph_informative.py` behind `-selection val`. The default `-selection test` reproduces the published, exploratory protocol bit-for-bit. Under `val`, HistGBT switches to the inner-train search described above, every summary member is scored once on the untouched validation fold, and the graph and summary winners are the argmax of mean validation macro-F1. Selection opportunities remain unequal: each graph member contributes its maximum validation score over epochs and two learning rates, while each summary member contributes one score. This gives the graph family more opportunities to fit validation variation and limits selection parity.

27 One-feature and finer-horizon controls

Cascade size alone. Logistic regression on the single standardised feature $\log(1+n)$, with the same deterministic splits and seeds as every other event-list result, reaches macro-F1 $0.514_{0.057}$ on CoAID, $0.623_{0.029}$ on ReCOVery, $0.496_{0.081}$ on PolitiFact, and $0.707_{0.022}$ on GossipCop (`scripts/rev9_size_only.py`). Against PropMLP's 0.618, 0.659, 0.657, and 0.836, and a majority floor of 0.333, one feature recovers most of the ReCOVery signal and little of PolitiFact's. The single-feature control shows what the remaining summaries add beyond size.

Finer observation horizons. `scripts/rev5_censoring.py` gains a `CENSORING_FINE` switch that adds 1 h, 6 h and 3 d windows to the published 1/2/7-day sweep, recomputing the shared windows identically so the two runs can be checked against each other. On PolitiFact, random-split PropMLP macro-F1 is 0.602 at one hour, 0.599 at six hours, 0.692 at one day, 0.694 at three days, 0.661 at seven days and 0.657 uncapped. RandomForest follows the same shape. On GossipCop the curve is nearly flat: 0.822 at one hour, 0.833 at six hours, 0.828 at one day and 0.836 uncapped, so an hour of events already supports 98% of the full-cascade macro-F1. The future-test gap shrinks on PolitiFact from 0.199 uncapped to 0.089 at one hour because truncation degrades the training era as well as the future one, whereas on GossipCop it stays between 0.076 and 0.086 at every horizon. The shorter horizon preserves much of GossipCop's random-split performance while leaving its temporal gap largely intact.

28 Profile-vector permutation: design

The degree-preserving rewire above destroys the *attachment arrangement* while holding the degree sequence fixed. This control does the complementary thing: it destroys the association between a released 10-dimensional user profile vector and the graph, while leaving every edge in place. Both profile-permutation variants change only the graph model's node features. The summary baselines retain their intact inputs, and the analysis script verifies that their reported columns remain bit-identical.

Two variants are run (`scripts/rev9_profile_permute.py`). Under **within**, the non-root profile rows are permuted inside each tree: the topology, the tree's profile multiset and the news root are all preserved, and only which profile vector occupies each structural position changes. Because the summary features are the mean, standard deviation and maximum over a tree's profile rows, they are invariant under this permutation by construction. Under **global**, the non-root rows are permuted across the whole corpus, so every tree's topology and the corpus-wide attribute marginal survive but per-tree composition does not. Node 0 of a UPFD graph is the news item rather than a user and is held fixed in both variants. The permutation is applied to the train, validation and test folds alike, so the models see one consistent regime rather than a train/test mismatch.

On UPFD-PolitiFact the three-way contrast is complete. Rewiring moves the BiGCN model by +0.0038 (p=0.27). The **within** permutation moves GCN, GATv2 and BiGCN by −0.0204 (p=0.13), −0.0133 (p=0.60) and −0.0053 (p=0.61), none resolved. The **global** permutation moves the same three by −0.0376 (p=0.003), −0.0525 (p=0.021) and −0.0315 (p=0.039), all resolved at the seed level. The ordering is consistent across the three architectures: neither the attachment arrangement nor a profile vector's position inside a cascade carries resolvable signal, whereas which released profiles appear in a cascade at all does.

29 Executing the BiGCN Release

The main text reports that the authors' released BiGCN code (`external/BiGCN`, commit `5779f2c`) runs in this environment. We record the compatibility shims so readers can distinguish the executed release from our local adaptation.

Dependency substitution. The release pins `torch_geometric` 1.3.2, `torch_scatter` 1.4.0, `torch_sparse` 0.4.3 and `torch_cluster` 1.4.5. Those wheels do not build against the installed toolchain, but the released source imports only one symbol from `torch_scatter`: `scatter_mean(x,data.batch,dim=0)`, used at `model/Twitter/BiGCN_Twitter.py` lines 44 and 78 and `model/Weibo/BiGCN_Weibo.py` line 43. We install a recorded shim that forwards that call to `torch_geometric.utils.scatter` with `reduce="mean"` (`scripts/official_bigcn_shim.py`), restore the `np.Inf` alias that NumPy 2.0 removed and that `tools/earlystopping.py` still reads, and force the `fork` start method, which was the platform default when the code was published and which the driver requires because it passes `num_workers=5` to `DataLoader` without an `if __name__` guard. No line of the authors' source is edited. The file is executed with `runpy`. The authors' own preprocessor `Process/getTwittergraph.py` runs unmodified and builds 1,470 graphs from the 1,490 labelled Twitter15 events (20 trees have maximum parsed child index below two and are dropped by the loader).

Test-based stopping. The release contains no validation fold. `Process/rand5fold.py::load5foldData` returns only a train and a test list for each of the five folds, and `train_GCN` builds its evaluation loader from `x_test`, passes those metrics to `EarlyStopping`, and returns the accuracy and per-class F1 *at the early-stopping*

checkpoint. The published protocol therefore selects the reported epoch on the test fold. A validation-only rerun therefore requires a protocol change. We distinguish the executed release's test-based stopping from our validation-selected comparisons.

Source-level fidelity. We compared the archived BiGCN commit with our local implementations. The artifact's `IMPLEMENTATION-FIDELITY.md` binds each inspected file by hash. The released Twitter model uses the first convolution's root embedding after the second convolution. Our tree adaptation uses the raw root features at both concatenations. The released source applies feature dropout at the framework default of 0.5 and drops 20% of edges in each direction. The local tree source uses feature dropout 0.2 and no edge dropping. Its optimizer searches learning rates 0.001 and 0.005 with one parameter group, whereas the released source sets 0.0005 and a five-times smaller rate for the bottom-up convolutions. Both use weight decay 0.0001. These are inspected source settings, not reconstructed per-trial logs. The saved UPFD validation-selection outputs independently record five seeds and a 60-epoch cap, but omit the chosen learning rate and stopping epoch for each trial.

The event-list BiGCN model is a separate adaptation: its branches act on constructed edge-type subsets, not a tree and its reversal. Our GCNFN, UPFD-prop, and EBGCN implementations also have local operators and readouts. Pinned-source comparisons against GNN-FakeNews (GCNFN and UPFD) and the EBGCN release document operator, readout, input and optimizer differences. The UPFD release has no propagation-only variant. Source comparison does not supply matched-input official accuracy runs or recover missing historical configurations.

Run status. Execution verifies compatibility, text inputs, and test-based stopping. The as-released five-fold Twitter15 run reached the third fold within the recorded compute budget and did not produce the final aggregate. The partial log therefore supports a protocol audit, not an official accuracy replication. Official UPFD, GCNFN, and EBGCN runs remain future work.

Input confound. The released model is `Net(5000,64,64)` over 5000-dimensional TF-IDF/bag-of-words node vectors read from `data.TD_RvNN.vol_5000.txt`. Our audited implementation is text-free by construction. An as-released official run therefore differs from our local adaptation in implementation *and* in input, and any gap between them is not attributable to implementation alone.

30 Reproducibility Artifacts

The code, stored outputs, and reproduction instructions are public at https://github.com/lexuanbach/wsdm2027-propagation-audit.

Checksums. `scripts/rev9_dataset_card.py` writes a SHA-256 card over every released input the loaders open: 52 files under `data/raw`, 7 under `external/BiGCN/data`, and 20 under `external/upfd`. Caches the pipeline regenerates (the `*graph/` directories built by the BiGCN preprocessor, PyG `processed/` folders) are counted, not hashed, because they are outputs rather than released inputs. The card records provenance and hashes. Corpus terms remain those of the original distributors.

Split manifests. Explicit story lists make seed-generated splits directly comparable. `scripts/rev9_split_manifest.py` materialises the ordered story-id list for every (corpus, seed, fold) together with a SHA-256 over that list, importing the corpus configurations from `scripts/rev2_power.py` so the manifest cannot drift from the evaluated splits. The UPFD corpora are excluded because they ship official index files that the loader consumes verbatim.

Per-story predictions. The retweet-tree runner emits, for every corpus, the per-seed story ids, gold labels and per-model predictions, matching the per-(corpus, model, seed) dumps that already existed for the event-list corpora.

31 Equal-trial tuning, declared sampler, and DLF ablation

After acceptance we ran a new study under a protocol frozen before any training output (`revision/a4-pa9-pa11-protocol.json`). This protocol and the directory below are in the companion `camera-ready-analysis.zip`, available in the public repository. The study reused the stored event-list splits, checked story by story, and the official UPFD splits, with five seeds. All 2,340 trials completed. Two amendments, both recorded with file hashes,

reduced the worker count and dropped GossipCop tuning of the six event-list graph baselines before any of those jobs started.

Declared sampler. Every DLF run declares its negative sampler. These are new runs and do not recover the flags of the historical table runs. With label-free negatives the rerun reproduced all 40 historical DLF-Full and DLF-Full+ per-seed scores to four decimals, and with label-aware negatives it reproduced 23. Because the historical code revision is unrecorded, this agreement supports but does not establish the historical flag. Switching the sampler moved accuracy by at most 0.042, and no difference survived Holm correction.

DLF-Full ablation. With the same data, seeds, and budget, removing pretraining, adjustment, or the hypergraph changed accuracy by at most 0.050, and none of 12 contrasts survived Holm correction. Validation selection preferred the variant without pretraining on ReCOVery, PolitiFact, and GossipCop and the summary-only variant on CoAID. At matched 90% coverage, disagreement abstention never beat confidence abstention from the same model, and it had the largest risk-coverage area on every corpus. DLF-Full+ was not ablated.

Equal-trial tuning. Every included model received exactly eight validation-scored configurations per corpus and seed, a factorial grid around its own default. Models were fitted on training data only, and test data were scored only after selection. Tuning lifted the fixed-budget graph baselines markedly, for example GATv2 on ReCOVery from 0.338 to 0.656 macro-F1. The 0.338 is this study's default, distinct from the main table's previously tuned 0.597. No tuned graph baseline establishes an accuracy advantage over tuned PropMLP on these corpora. CoAID macro-F1 point estimates nevertheless exceed tuned PropMLP (0.578): GCNFN 0.618, BiGCN 0.622, RGCN 0.598, and GATv2 0.602; the corresponding accuracy contrasts remain unresolved. UPFD-prop stayed at the majority floor in all 120 trials. Eleven of 34 declared orderings reversed between default and tuned settings, seven of them on the 24-story CoAID fold, and all reversed pairs remained unresolved. On GossipCop, the RandomForest advantage over PropMLP fell from 0.025 (Holm p=0.012) to 0.003 after tuning, and tuned HistGBT led tuned PropMLP+ by 0.008. Both tuned gaps are equivalent within 0.03. On UPFD-GossipCop, validation again selected BiGCN and HistGBT, and the graph gain was 0.017 accuracy (95% CI 0.010 to 0.024), also inside the 0.03 margin. Equal budget here means an equal number of trials rather than equal compute. Twitter15/16 and Weibo require fresh test populations after their earlier test-based selection and are excluded here. Results, every trial, and the amendments are in `revision/a4-pa9-pa11`.

32 Limitations and future work

The comparisons establish predictive value within the released signals and declared selection protocols. The 0.03 accuracy margin represents three errors per hundred stories, not an established deployment tolerance. For the main paper's validation-selected fits, UPFD-GossipCop's story-clustered 90% interval $[0.0065, 0.0186]$ lies within margins 0.02, 0.03, and 0.05 but not 0.01. UPFD-PolitiFact's $[-0.0226, 0.0570]$ lies within none. The GossipCop 95% interval excludes zero: a detectable gain and equivalence within a declared margin can coexist. These retrospective margin checks change no selected fit. Equal-trial tuning covers four event-list corpora and UPFD (Section 31), but GossipCop graph baselines, the test-selected corpora, and equal-compute budgets remain open. The next comparisons are matched-input official UPFD/GCNFN/EBGCN runs and an untouched structure-only evaluation. A factorial study of real identities, profiles and timing, complete split provenance, and independent retraining are future work. These extensions require observed edges and identities with documented provenance. Seed-level drift intervals summarize five saved runs. A later training-era selector and additional powered temporal and publisher-disjoint tests remain needed. Re-estimating the four GATv2 contrasts with story clustering requires the unsaved PropMLP comparator predictions. Future pretraining comparisons must retain per-run negative-sampler flags. The missing historical flags cannot be inferred from a current default.

33 Historical test-selected tree matrix

This is the original test-selected comparison, retained separately from the validation-selected UPFD headline in the main paper. Model columns report macro-F1, whereas the paired gap reports accuracy. HistGBT's earlier GossipCop fit yields macro-F1 0.931, while its later inner-training-fold fit yields 0.928. The comparisons must not be combined cell by cell. Selection used test macro-F1, so the displayed intervals and equivalence verdicts are descriptive and do not account for selection. In particular, a zero-containing interval does not establish equivalence.

Table S13: Test-selected summary and graph comparisons on released retweet trees over five seeds. Bold marks the highest macro-F1 per row. 'Within 0.03' denotes TOST equivalence of the accuracy gap at margin 0.03.

(a) Summary-model macro-F1

Dataset	Node inputs	Test/seed	LogReg	RForest	HistGBT	MLP
UPFD-PolitiFact	profile	221	0.787	0.765	0.792	**0.793**
UPFD-GossipCop	profile	3,826	0.906	0.914	0.931	0.926
Twitter15	structural	149	0.325	0.329	0.284	**0.351**
Twitter16	structural	82	0.294	0.292	0.291	**0.336**
Weibo	structural	453	0.889	0.884	0.877	0.888

(b) Graph macro-F1 and selected graph-versus-summary accuracy

Dataset	GCN	GATv2	BiGCN	Accuracy gap [95% CI]	p	Within 0.03
UPFD-PolitiFact	0.783	0.776	0.752	-0.011 $[-0.067, 0.046]$	0.71	unresolved
UPFD-GossipCop	0.902	0.917	**0.940**	$+0.009$ $[+0.002, +0.017]$	0.011	yes
Twitter15	0.307	0.328	0.342	-0.008 $[-0.044, 0.028]$	0.69	unresolved
Twitter16	0.314	0.327	0.311	-0.015 $[-0.059, 0.030]$	0.56	unresolved
Weibo	**0.894**	0.890	0.891	$+0.005$ $[-0.004, 0.015]$	0.30	yes

The validation-selected main-paper comparison and the eight-trial study in Section 31 use different fits.

34 Running-example details

Running example. We use one FakeNewsNet PolitiFact story before balancing. Its identifier is `politifact:01e21af5a8d19ce7`, and its label is misinformation. Its 103 events span 46.6 days, and its tenth event occurs after 19.6 h. The resulting vector is $(4.64, 4.64, 0.69, 0, 1.39, 1.00, 15.21)$. The first two entries coincide because we treat each event identifier as a distinct proxy. The value E_{hash}=1 arises because all three proxies observed in the first hour fall into one hash bucket. Our balanced analysis uses 628 of the 677 usable raw stories. The loader caps each cascade at 200 events. A total of 271 raw stories reach this cap, while the example does not.

35 Temporal diagnostics and correction details

In the separate seed-0 drift diagnostic, temporal spread on PolitiFact has KS 0.49, size or proxy count has KS 0.19, and the class prior moves from 0.45 to 0.55. On GossipCop, spread and early popularity have KS 0.21 and 0.19, while the prior moves from 0.49 to 0.64 (supplement Part B reports priors pooled over seed-dependent balanced samples). Panel (a) of the main future-test feature-shift figure plots only the four features that can move. The $g(r)$ and $g(m)$ features have KS exactly 0 because each cascade in these releases has one URL and no media. The event-proxy representation also makes $g(u)$ duplicate $g(n)$. These zero values reflect degenerate release fields, not temporal stability.

Capping cascades to fixed one-day, two-day, and seven-day windows leaves a PolitiFact loss from 0.19 to 0.26. Right censoring alone does not explain the temporal gap. At shorter horizons, one hour of events gives PropMLP macro-F1 0.602 on PolitiFact and 0.822 on GossipCop, compared with 0.657 and 0.836 without a cap. The PolitiFact future-test gap narrows to 0.089 at one hour because both random and future-test performance decline.

Mitigation. Panels (b,c) of that figure use circles for PropMLP, triangles for PropMLP+, and squares for RandomForest. We test per-era standardization, per-era quantile transformation, importance weighting through a train-versus-test domain classifier [4], and correlation alignment (CORAL) [5]. On PolitiFact, CORAL/RandomForest gains 0.051 with a five-seed interval $[0.001, 0.102]$. RandomForest standardization and quantile mapping lose 0.151 and 0.105. On GossipCop, positive intervals occur for importance-weighted PropMLP and PropMLP+ ($+0.028$/ $+0.023$) and quantile-mapped PropMLP+ ($+0.021$). These unadjusted seed intervals are descriptive, not evidence of general recovery. Tree importance weighting remains unresolved. GossipCop CORAL changes RandomForest from 0.851 to 0.755.

Selecting a correction by validation-era F1 chooses CORAL for GossipCop RandomForest, although it performs poorly on the future period. The validation fold is sampled from the training era and does not represent

the transformed future distribution. Evaluating a shift-aware selector requires a later training-era slice that approximates the target shift, so we reserve that test for future work. The PolitiFact loss persists across rolling-origin test sizes (macro-F1 0.458, 0.346, and 0.185 for the latest 10%, 20%, and 30%), and Saerens EM prior correction changes 0.458 to 0.352.

References

[1] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2051–2055. doi:10.1145/3404835.3462990

[2] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3558–3565. doi:10.1609/aaai.v33i01.33013558

[3] Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*. 4878–4887. https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[4] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* 8 (2007), 985–1005. https://www.jmlr.org/papers/v8/sugiyama07a.html

[5] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30. 2058–2065. doi:10.1609/aaai.v30i1.10306

[6] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. 2019. HyperGCN: A new method for training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems 32*. https://proceedings.neurips.cc/paper/2019/hash/1efa39bcaec6f3900149160693694536-Abstract.html