

PHANTOM: Benchmark Validity and Causal Limits in Coordinated Listening

Nischit Girish Kumar

NEXUS Research Initiative — Independent Research

NEXUS Research Initiative · Independent Technical Report · October 2026

Abstract

High scores in coordination detection can reward benchmark artifacts while concealing unreliable operating thresholds. PHANTOM rejected attractive listening benchmarks when origin, catalog and duplicate shortcuts could explain performance; even perfect raw ranking failed threshold transfer. We develop structure-preserving controlled interventions and a relational detector normalized against a local conditional permutation reference. The population detector targets a declared recurring coordination pattern that the intervention deliberately instantiates. Controlled transfer across separate ListenBrainz populations culminates in a prospectively committed, pre-access holdout evaluation: 378 of 378 treatments and none of 378 matched controls alert across 42 source groups, with AUC 1.000. These labels denote controlled interventions on observed histories, and repeated worlds share background. The holdout tests a new population under the same intervention family, rather than universal coordination detection. Historical reconstruction fully recovers member and target sets for 154 of 168 parents, but common-stimulus and shifted-overlap tests defeat broader operating claims. Applying observational-equivalence reasoning shows why event-only evidence cannot settle intent. PHANTOM therefore supports investigation by presenting structural evidence, alternative explanations and gaps in provenance. Its contribution is the empirical separation of benchmark validity, conditional transfer and causal interpretation.

In plain English. Synchronized listening can be structurally unusual without revealing intent. Early high-scoring benchmarks were rejected when shortcuts were discovered. Across two controlled ListenBrainz populations, the frozen detector separated planted coordination from matched controls, but common-stimulus tests showed that event patterns alone cannot determine whether coordination reflects manipulation or a shared external trigger. PHANTOM therefore presents structural evidence and uncertainty to support investigation.

1 Introduction

Synchronized listening can be structurally unusual without revealing why people listened together. A shared broadcast and an organized manipulation campaign can align the same listeners, recordings and times. PHANTOM asks how a detector can provide useful structural evidence when benchmark artifacts, threshold instability and missing causal information each threaten its interpretation.

We rejected high-scoring early benchmarks when origin, catalog and duplicate shortcuts could explain their scores. Even after those shortcuts were repaired, perfect raw ranking coexisted with 22 alerts among 180 controls at the frozen threshold. We therefore asked whether a locally normalized rule could retain its operating behavior on another observed population, and where the evidence would stop being informative.

PHANTOM’s population detector targets a specific recurring coordination pattern: the selector requires supported recurring structure, and the planted intervention deliberately instantiates it. The holdout tests the fixed rule on new observed background under the same intervention family. Separate historical reconstruction and common-stimulus/overlap tests expose boundaries that population separation alone cannot resolve.

Contributions.

1. **Benchmark validity.** A benchmark development methodology shows how origin, catalog, duplicate and calibration shortcuts can create misleading coordination results. Structure-preserving counterfactual construction addresses the identified shortcuts.
2. **Conditional structural transfer.** A locally normalized relational method transfers across separate observed ListenBrainz populations under a fixed controlled mechanism. Failed raw/global threshold transfer provides the historical comparison; a separate internal holdout supplies bounded graph reconstruction evidence.
3. **Failure boundaries and interpretation.** Common-stimulus and overlap failures delimit broader claims and motivate an investigation architecture that preserves evidence, alternatives and missing provenance while withholding event-only intent judgments.

Permutation testing, MAD scaling, group detection and observational equivalence are established methods. The contribution is the connected empirical account and the restrictions it places on the resulting system.

2 Related work and comparison scope

CopyCatch detects temporal lockstep behavior [2]; CatchSync uses unusual, synchronized connectivity and can operate without timing data [7]; FRAUDAR studies dense bipartite patterns under camouflage [4]. Behavioral-trace networks [8] and dynamic coordination communities [14] broaden the structural setting. Tardelli et al. include malicious and other coordination: separating structure from intent is established context. CopyCatch, CatchSync and FRAUDAR were not run as matched empirical baselines in PHANTOM.

Temporal-network randomization defines anomalies through preserved constraints [3]. PHANTOM uses established median/MAD scaling [11] and a nonzero finite-randomization comparison [10]. Exchangeability for ordinary common-stimulus listening remains a separate assumption. The empirical question is whether the fixed combination transfers within the declared construction.

PISA addresses repeat-aware listening recommendation [16]; its dataset [15] supplies histories rather than verified fraud labels. SAGE addresses negative harvesting for streaming fraud in a different contextual and labeling regime [17]. Its WSDM Companion publication is relevant context, not a matched comparison. A comparable baseline experiment would need equivalent inputs, tuning populations and evaluation units.

Causal interpretation requires assumptions beyond observed associations [9]. Shalizi and Thomas analyze confounding among homophily, contagion and covariate effects in observational networks [13]. We apply observational-equivalence reasoning to the PHANTOM event-only setting. Guidance on communicating scope and fallibility informs the investigation interface [1]; analyst effectiveness remains unmeasured.

Recent temporal and multiplex coordination work [6] emphasizes representation choice. BAG [5] compares anomaly detection at node and edge levels, whereas PHANTOM evaluates constructed worlds and

original parents. Graph Evidential Learning [18] models graph and reconstruction uncertainty; PHANTOM’s retained causal alternatives are not learned uncertainty distributions. Selection-aware recommendation evaluation [12] addresses a related evaluation concern, although PHANTOM does not estimate inverse-propensity effects.

3 Data, observation and provenance

Evaluation units. A *world* is one detector input: observed background plus a controlled treatment or matched control assignment. A *source group* (or block) contains the background listeners reused across several worlds. A *parent* is an original planted campaign whose members and targets reconstruction attempts to recover. Control alert rate is the fraction of control worlds with an alert, not a verified real-world false-positive rate.

3.1 Deezer: source scale and experimental scale

The audit covers 704,728,520 events from 3,444,027 users across 184 UTC dates, March through August 2022. Event files contain 43,803 observed track identifiers; the separate embedding file contains 50,000 rows. There are 500 event parquet files and one embedding file. Integer user, track, session and timestamp fields have no reported nulls in the audited representation. Source scale describes ingestion and auditing, not the number of events passed simultaneously through the frozen graph detector.

All 70,472,852 source sessions contain exactly ten records. This regularity prevents assuming that the release’s sessions represent naturally delimited listening episodes. Same-second events have no established within-second order. Embedding temporal provenance is unresolved, limiting claims about prospective embedding-based evaluation. These limitations concern the release representation; they are not behavioral conclusions about listeners.

The audit identifies 2,435,880 excess exact duplicate rows, about 0.3456% of events, affecting 496,504 users. Actor-level distortion can be substantial even when the global duplicate rate is small. Ingestion retains duplicates for auditability; reconstruction uses unique events and reports multiplicity separately. Retaining the source records documents the release; deduplication prevents repeated identical rows from manufacturing graph evidence.

The bounded development sample contains 671,410 rows, 3,252 complete release-relative histories and 19,380 tracks. Completeness means within the available release, not lifetime listening. Full-source auditing used disk-backed DuckDB processing, a 3 GB memory cap, four threads and 64 buckets. It did not load all events into a single pandas frame or NetworkX graph. The graph experiment later uses much smaller source blocks.

3.2 ListenBrainz external population

Both ListenBrainz evaluations use snapshot 2723, dated 15 October 2026. A snapshot publication date differs from the dates of the events it contains: this snapshot supplies both the May 2022 and September 2026 windows. External validation is restricted to 1 May 2022 inclusive through 29 May exclusive. The source-window audit records 16,006,621 events and 16,783 listeners. Requiring at least ten events and two distinct items leaves 16,185 eligible listeners. Predeclared selection retains 4,032 listeners in 42 disjoint groups of 96, with 4,004,802 original events. Histories are complete within this source window only.

The selected projection has 1,391,257 recording-MSID keys. This is an identifier count, not a count of verified unique audio recordings. Identities are source-specific; equal integers across datasets would not imply the same listener or recording. May acquisition retained three required event columns from nine relevant archive partitions via byte ranges; the full archive remained remote and September final listener histories were not accessed. The external audit reports embedded CC0 documentation for its source. That does not license unrelated Deezer records or every project artifact.

3.3 Labels, rights and missing context

Real histories supply observations. The treatment, protected parent labels and matched control are constructed for research. No ListenBrainz participant is identified as a real fraud actor. Neither source supplies the device, exposure, IP, geography, referral, organizer or reliable listening-duration evidence needed to settle many causal questions.

The visible Deezer release page did not resolve redistribution rights during this review. A repository's code license would not settle dataset rights. The publication package therefore uses aggregate evidence and synthetic demonstration material; raw source records are excluded from the public repository. Any additional data redistribution requires separate rights clearance. Local source maps preserve identities for internal review, but their absolute paths are not automatically public-release content.

4 Benchmark validity before detector performance

4.1 Origin, catalog and session shortcuts

The early simulator combined six campaign families, three difficulty levels and four popularity regimes. Generated actors and observed background differed in ways that revealed origin independently of coordination. A logistic baseline looked useful on campaign detection, but synthetic-neutral actors alerted much more often than observed background. Session-sensitive features were especially problematic because source sessions were fixed-size blocks.

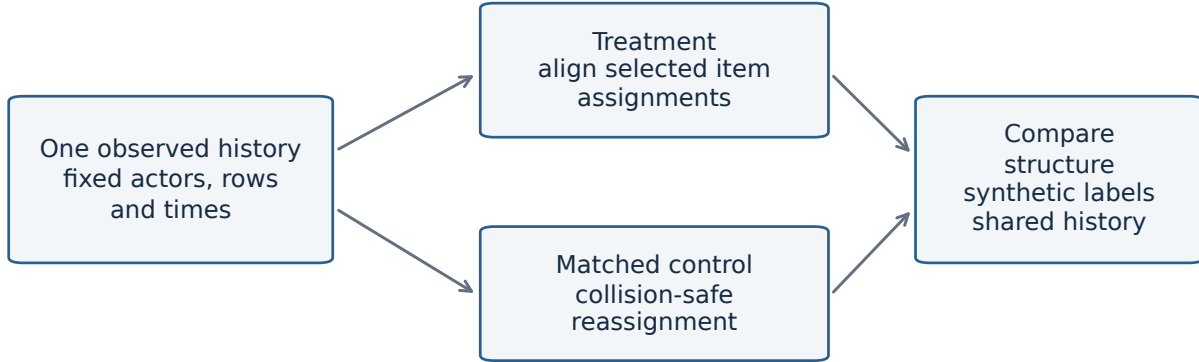
Session removal was necessary but insufficient: neutral alerts remained 12.08% against 4.79% observed-background alerts. Observed background was unlabeled, so this is a benchmark-shift diagnostic, not real-world specificity. Catalog and calendar probes exposed further generated/observed differences. All 72 relational candidates in one realism stage failed augmented-origin gates; their detection performance was not accepted as scientific evidence.

A feature can improve AUC by identifying a simulator instead of the intended target. Testing only generated positives against observed background would not expose that problem. Explicit neutral and origin probes therefore precede performance interpretation. Passing those probes rules out identified shortcuts, not every possible confounder. This distinction motivates documenting rejected constructions rather than reporting only the best-looking accepted score.

4.2 Counterfactual pairs and collision-safe reassignment

An additive intervention changed actor marginals even when paired with a sham addition. A subsequent structure-preserving reassignment held per-actor item and timestamp multisets fixed, but changed semantic duplicate counts in 20/180 pairs. That processing defect invalidated the relational comparison before detector performance could be interpreted.

Collision-safe construction repairs this defect. Treatment and control preserve actor identity, row counts, item multisets, timestamp multisets and declared duplicate invariants, while changing cross-actor item alignment. The controlled recurring schedule has 32 slots and eight observed items repeated four times; difficulty changes synchronized slots at 32, 24 and 16. This is a specific relational intervention, not a universal model of real coordination.



Preserve each actor's item multiset and timestamp multiset; check semantic duplicates.

Figure 1: Matched construction changes relational assignment while preserving declared actor marginals. Paired worlds share histories. Synthetic labels describe the intervention and do not label real participants' intent.

For event multiset $E = \{(r, u, i, t)\}$, let M_u and T_u denote selected item and timestamp multisets. Admissible pairs require

$$M_u^T = M_u^C, \quad T_u^T = T_u^C, \quad |E_u^T| = |E_u^C|. \quad (1)$$

Background rows remain fixed. Duplicate invariants are additional checks, not consequences of these equalities. Item-time association can change: this is exactly the intervention. Thus preserving marginals does not imply preserving every imaginable joint per-actor feature. Claims about reduced leakage are supported by specified probes, not a proof that every actor statistic is identical.

Whole source blocks and campaigns define splits. Shared histories within pairs, repeated difficulties and construction seeds induce dependence. Many worlds are not equally many independent populations. Development, previously used historical validation and external evaluation remain separate.

5 A population-level relational-surprise detector

5.1 Coactivity and its opportunities

Let Q be the set of unique (u, i, b) triples, with $b = \lfloor t/60 \rfloor$. Let n_{ib} count actors on item i in bin b , $N_b = \sum_i n_{ib}$ count actor-item presences, and d_{ub} count actor u 's distinct items in that bin. The inherited population coactivity statistic is

$$C(E) = \frac{\sum_{i,b} n_{ib}(n_{ib} - 1)}{\max\{1, \sum_b [N_b(N_b - 1) - \sum_u d_{ub}(d_{ub} - 1)]\}}. \quad (2)$$

The numerator counts ordered same-item cross-actor coincidences. The denominator counts ordered cross-actor opportunities, including different items while excluding within-actor pairs. The optimized null implementation uses fixed-background sufficient counts to compute this statistic exactly. It is a population-level structural score, not an individual risk or fraud probability.

5.2 A restrictive visible schedule

The frozen selector requires exact timestamp groups with at least eight distinct users, each present once. The same actor set must occur exactly 32 times, occupy 32 distinct minute bins and span at least 27 days. Exactly one qualifying actor-set group must exist; otherwise the method fails closed as unsupported. This narrow selector is part of the method and must accompany any transfer claim.

For an eligible world, each selected actor's sorted item multiset is independently permuted across selected slots, using 4,095 draws and seed 41101. Times, identities and background remain fixed. The null asks what coactivity would arise under exchangeable selected item assignments conditional on the preserved schedule. Common external stimuli can violate this assumption. Rejecting it can reveal unusual coordination without identifying why it occurred.

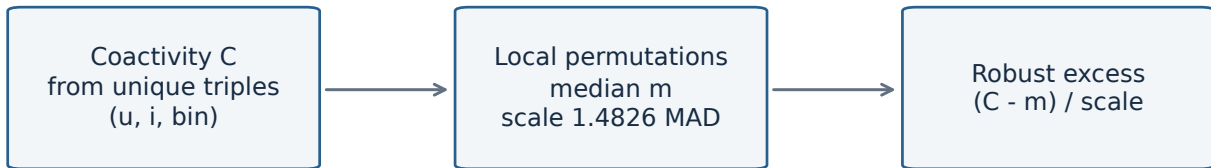
5.3 Normalization and threshold

For null values $C^{(1)}, \dots, C^{(B)}$, define

$$m = \text{median}_r C^{(r)}, \quad s = 1.4826 \text{ median}_r |C^{(r)} - m|, \quad (3)$$

$$Z(E) = (C(E) - m)/s \quad \text{when } s > 0. \quad (4)$$

If $s = 0$, the implementation returns zero when $C(E) = m$, 10^{12} when larger and -10^{12} when smaller. These sentinels preserve order; they are not extreme Gaussian measurements. The frozen alert rule is $Z \geq 3.0$. The 1.4826 factor is a normal-reference MAD convention, not a guarantee that Z has normal tails.



4095 draws; alert at excess ≥ 3 . Zero-scale rule is explicit.

Definition schematic: no fitted density, Gaussian tail probability or invented observations.

Figure 2: Conditional local normalization. This schematic contains no fitted density or invented observations. A robust-excess threshold is not a causal probability or Gaussian significance level.

Perfect ranking is distinct from transferable operating behavior. Table 1 reports the failed raw/global calibration results alongside conditional normalization and its later population transfers.

Table 1: Historical method comparison. Every row has perfect ranking; operating thresholds differ. Global and local development procedures use the same eight blocks but different fit/calibration designs; validation and transfer rows use other populations. This is not one head-to-head test. AUC is mean block (also 1.000 pooled where recorded); the raw row reports pooled and all four block AUCs. Control counts and pooled rates use worlds; worst-block rates retain source grouping. n/r: not recorded in the cited aggregate. Global procedures and the local empirical tail fail the worst-block requirement; local robust excess passes the declared development gates.

Method	Evaluation population	AUC	Control alerts (pooled)	Worst block
Raw fitted-score threshold	Historical validation (4 blocks)	1.000	22/180 (12.22%)	n/r
Global pooled 95th percentile	Development (8 blocks)	1.000	23/180 (12.78%)	41.67%
Global maximum block 95th percentile	Development (8 blocks)	1.000	22/180 (12.22%)	38.89%
Global maximum calibration control	Development (8 blocks)	1.000	12/180 (6.67%)	29.63%
Local empirical tail	Development (8 blocks)	1.000	6/180 (3.33%)	11.11%
Local robust excess	Development (8 blocks)	1.000	0/180 (0%)	0%
Frozen local robust excess	External May 2022 (42 groups)	1.000	0/378 (0%)	0%
Frozen local robust excess	Pre-access September 2026 (42 groups)	1.000	0/378 (0%)	0%

6 Controlled external-population validation

The external test freezes the selector, null, 4,095 draws and threshold before use on ListenBrainz. Sampling uses seed 41201; construction uses seeds 41211–41213. Forty-two groups, three seeds and three difficulties yield $42 \times 3 \times 3 = 378$ treatment/control pairs and 756 worlds. Pairs share histories, and multiple pairs share a source group.

Frozen outputs record AUC 1.000, treatment detection 100%, control alerts 0/378 and worst-group control alert rate 0%. AUC describes ranking across the constructed labels; detection and alert counts describe behavior at the fixed threshold. These are external-population results under controlled planted coordination, not real fraud classification. Changing the population while retaining the intervention family tests a useful but bounded form of transfer.

Whole-group resampling yields degenerate summaries because every observed group has perfect outcomes. The final-evaluation section explains their uncertainty limitation; the supplement retains the historical bootstrap details.

The external result supports controlled population transfer of the frozen detector. Reconstruction and investigation retain their separate evaluation populations.

7 Campaign reconstruction from semantic presence

7.1 Representation and burst support

A population alert does not identify a usable investigation object. Reconstruction needs members, targets, periods and supporting relationships. The accepted reconstruction method restricts events to 1–28 May

2022 UTC in the Deezer listening-events dataset and reduces repeated (u, i, t) rows to one unique event (a semantic atom), retaining multiplicity descriptively. Exact duplicate rows consequently cannot manufacture additional presence-based evidence. Canonical actor order uses item/time fingerprints; ambiguous fingerprints trigger abstention rather than arbitrary identity-based tie breaking.

The burst selector groups sorted atoms when consecutive gaps are at most 600 seconds. A qualifying burst needs at least eight distinct users and at least eight atoms from users appearing once in that burst. Each selected actor must occur in at least eight qualifying bursts. This actor filter is applied once, after which retained groups must contain at least two selected atoms. Multi-atom actor cells are not converted into randomized singleton observations. The procedure encodes a recurring burst support assumption; it is not a graph over every audited event.

7.2 A fixed catalog candidate family

Connect active actors when their selected catalogs share at least three items and have Jaccard similarity at least 0.10. Connected components with at least three actors form candidate pools. Within-actor item permutations preserve this catalog graph, keeping the candidate family fixed under the null. Connectivity alone does not establish an alert.

For actor u at selected burst slot s , write item i_{us} and time t_{us} . Pair evidence at bin width h is

$$w_{uv}^{(h)} = \sum_{s \in S_u \cap S_v} \mathbf{1}[i_{us} = i_{vs}] \mathbf{1}[\lfloor t_{us}/h \rfloor = \lfloor t_{vs}/h \rfloor]. \quad (5)$$

Consensus takes the smaller 60- and 300-second pair counts. Pooling averages over all unordered pairs, including zero contributions:

$$T(G) = \binom{|G|}{2}^{-1} \sum_{\{u,v\} \subseteq G} \min\{w_{uv}^{(60)}, w_{uv}^{(300)}\}. \quad (6)$$

The complete-pair denominator prevents a few strong edges from portraying a large sparse component as uniformly coordinated. Since the two fixed bin systems share an origin and are nested, agreement is not independent evidence of arbitrary time-shift robustness.

7.3 Conditional null and fragment association

The accepted null uses 511 within-actor selected-item permutations and seed 55101, preserving catalogs, presence and times. Observed and null pooled statistics receive the same median/MAD standardization. Let Z_g denote the observed normalized score and $Z_g^{(r)}$ its null counterpart. With $M^{(r)} = \max_g Z_g^{(r)}$ over the fixed family,

$$\hat{p}_g = \frac{1 + \sum_{r=1}^{511} \mathbf{1}[M^{(r)} \geq Z_g]}{512}. \quad (7)$$

The frozen rule requires $Z_g \geq 3$ and $\hat{p}_g \leq 0.01$. Resolution is $1/512$. Although the code names this a familywise comparison, interpretation depends on randomization validity, fixed candidates and estimated standardization. A maximum alone does not prove error control for organic listening.

Admitted components link when target Jaccard is at least 0.25 and period overlap is at least half the shorter interval. Targets are catalog items shared by at least two selected members, some without direct temporal support. That distinction is reported. A merged object retains constituent component statistics; it receives no invented joint p -value or organizer claim. Linking fragments is structural bookkeeping, not evidence of one causal authority.

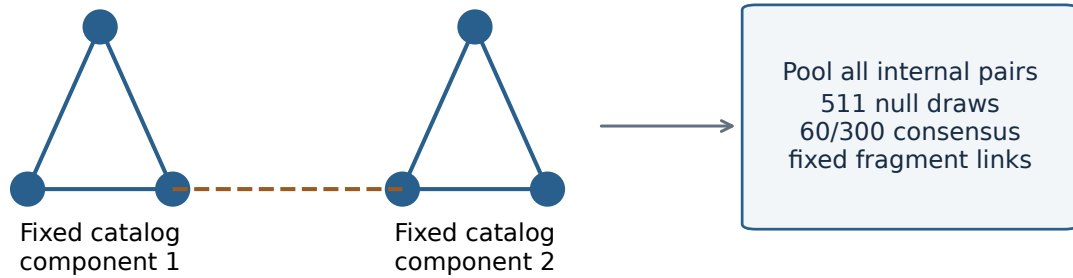


Illustration only. Shared catalog geometry does not itself establish an alert or cause.

Figure 3: Illustrative graph workflow; nodes do not represent observed people. Fixed catalog pools are tested for temporal surprise, then associated using target and period overlap. No joint merged significance is invented.

8 Reconstruction evaluation and rejected alternatives

Development uses six Deezer source blocks, 759 observed users and 50,523 original events. The external ListenBrainz numerical population was excluded from reconstruction tuning. Seven regimes, three levels, two seeds and six blocks produce 252 pairs, with two treatment parents per world. The selected method reaches member F1 0.9067 and 2/252 control alerts. The selected method combines burst-conditioned support, consensus timing and fragment links; other investigated graph variants have different evidence.

The historical internal holdout contains two reserved Deezer blocks, 84 treatment/control pairs and 168 original treatment parents (two planted campaigns per treatment world). The saved candidate objects exactly recover member *and* target sets for 154/168 parents (91.67%); there are no partial member/target recoveries, and 14 parents are missed and unmatched. There are no unmatched predicted candidates. Each parent contributes either one or zero to member precision, recall, F1 and target F1, explaining their common mean $154/168 = 11/12 = 0.91\bar{6}$.

At the world level, 76 treatments recover both parents, two recover one, and six recover neither. Recovery is uneven: all 14 missed parents occur in the hardest level-3 reduced-participation and gradual-evolution cells, where recovery falls to 1/8 in each; every other regime-difficulty cell recovers 8/8. The strong pooled average masks severe degradation in these two hard cases (Table 2). The two holdout blocks recover 78/84 and 76/84 parents respectively. The designated fragmented regime recovers both protected halves in the same matched object for all 24/24 parents. This fragment criterion differs from general parent recovery.

Structural ambiguity is not separately coded in the frozen holdout payload, so a missed parent cannot be assigned an ambiguity-abstention cause. Every recovered object explicitly withholds fraud attribution; that causal abstention is distinct from structural recovery. Control alerts are 0/84 and original-only alerts 0/2. Exact duplicate perturbations preserve output. The supplementary parent table gives the full regime-difficulty breakdown and checks exact-set recovery against the saved world metrics.

The evaluation uses one-to-one parent/object assignment by member F1, with unmatched parents scored zero. For matched membership sets P, T , precision is $|P \cap T|/|P|$, recall $|P \cap T|/|T|$, and F1 $2|P \cap T|/(|P| + |T|)$. Aggregation follows per-parent computation. One must not recompute aggregate F1 from unrelated aggregate precision and recall. Targets and fragment recovery use separate definitions and denominators.

Table 2: Exact member-and-target recovery by regime and difficulty in the internal holdout, from the existing frozen per-parent outputs. Bold cells account for all 14 misses (seven each). The supplement retains the full detailed record.

Regime	Level 1	Level 2	Level 3
Broader targets	8/8	8/8	8/8
Distributed	8/8	8/8	8/8
Fragmented	8/8	8/8	8/8
Gradual evolution	8/8	8/8	1/8
Low signal	8/8	8/8	8/8
Reduced participation	8/8	8/8	1/8
Weak temporal	8/8	8/8	8/8
Overall: 154/168 recovered (91.67%)			

Table 3: Primary evidence by stage. Populations, tasks and denominators are distinct; no overall accuracy score is implied. Final metrics are confirmed from frozen outputs, with the original verifier failure preserved.

Evidence stage	Population	Task	Result	Interpretation limit
External validation	May 2022 ListenBrainz; 4,032 listeners; 42 groups; 378 pairs	External controlled detection	AUC 1.000; 378/378 treatments; 0/378 controls	Synthetic intervention labels; already-used external validation.
Internal reconstruction holdout	Two historical holdout blocks; 84 pairs; 168 parents	Graph reconstruction	154/168 parents fully recovered; 0/84 controls	Small internal holdout; no external graph validation.
Historical stress tests	Six reused development blocks; 12 moderate worlds per shock family	Hard-negative robustness	Popularity/release shocks: 12/12 alerts each; shifted-overlap F1 0.3636	Overall moderate operating region failed.
Investigation development	321 synthetic worlds; 480 candidates; 477 parents	Investigation development	416/480 candidates reviewable; 327/477 parents covered	Coverage differs by denominator; analyst utility unmeasured.
Pre-access holdout	September 2026 ListenBrainz; 4,032 listeners; 42 groups; 378 pairs	Pre-access controlled detection	Block/pooled AUC 1.000; 378/378 treatments; 0/378 controls	Same intervention family; new background.

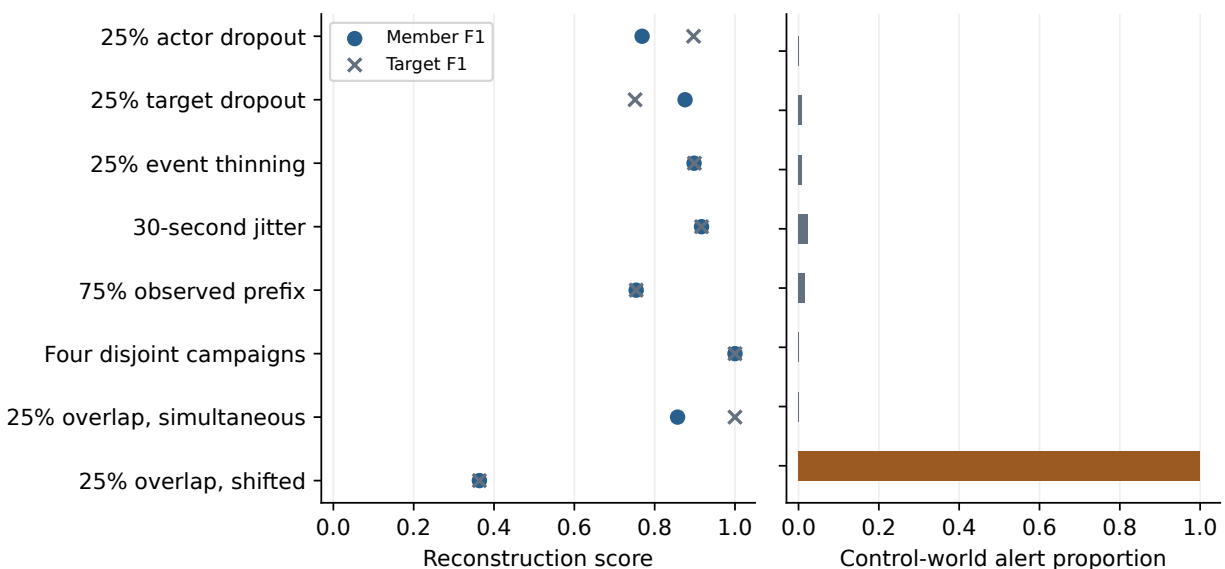
An appealing simple catalog method alerted on every control. Initial graph nulls also produced large control alert rates, while duplicate-sensitive singleton logic allowed representation artifacts to change behavior. These failures motivated semantic presence, better conditioning and pooled candidates. Comparing only against a weak implementation would obscure why the accepted method was selected.

Post-veto member F1 scores parents removed by a veto as zero. A hybrid behavioral veto reduced this measure from roughly 0.9067 to 0.5357, erasing structurally supported cases even as control alert rate fell. Non-veto augmentation preserved the core but produced zero added within-block ranking gain across 210 comparisons and failed other stability/cost gates. Both were rejected. The final system has no successful behavioral fusion claim, GNN contribution or independently validated learned risk score.

9 Robustness includes the failure region

Stress testing reuses six development blocks and imports the accepted core unchanged. Primary scores retain original member/target truth even when actors or targets disappear from view. Secondary surviving-truth analyses score only the portion of the original truth still visible after perturbation, distinguishing lost evidence from residual reconstruction failure. This denominator choice matters: removing half the actors while retaining the original recovery target necessarily changes attainable recovery.

Separate moderate settings yield member F1 0.7687 for 25% actor loss, 0.8757 for 25% target loss, 0.8981 for 25% event thinning, 0.9167 for 30-second jitter and 0.7540 for a 75% prefix. A bounded four-disjoint-campaign panel yields F1 1.000. These results apply to sampled separate perturbations, not their combinations or interpolated severities. Nonmonotonic outcomes are retained rather than smoothed into an invented degradation curve. The identical-looking jitter mean is a different fraction: 231/252 parents fully recovered, versus 154/168 in the reconstruction holdout. The jitter panel uses six development blocks; its equality to 11/12 is numerical, not repeated recovery of the same parents.



Historical development only; perturbations tested separately.
Popularity/release shocks additionally alerted 12/12 each.

Figure 4: Historical stress results transcribed from the frozen moderate-gate table. Successful individual perturbations do not establish a safe joint operating region. Reconstruction F1 and control-world alert rate are different quantities.

Time-shifted 25% overlap yields member and target F1 0.3636 and alerts on every control in its panel; simultaneous overlap behaves differently. Popularity-shock and release-event controls each alert in 12/12 moderate worlds, with background absorption 0.2544 (the fraction of background actors included in predicted groups) and stress absorption 0.3376 (the corresponding rate within designated stress cohorts). Routine listening alerts in 2/12 worlds. Severe missingness yields member F1 0.5365 at 50% actor loss, target F1 0.5572 at 50% target loss, and member F1 0.2913 at a 25% prefix. The overall predeclared moderate operating region failed.

A shared broadcast can create real alignment that violates the conditional null. Such an alert may be a valid structural observation while failing a benchmark gate intended to suppress common-cause controls. Relabeling the output after the fact does not make that gate pass. The appropriate response is to preserve the failure and distinguish structural evidence from cause and actionability.

10 Observational equivalence and causal attribution

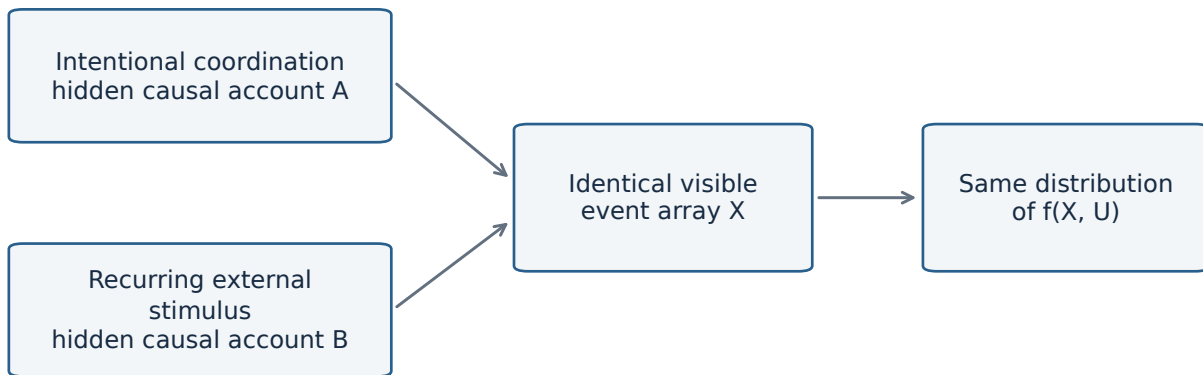
Following observational-identification precedent [9, 13], we apply observational-equivalence reasoning to the PHANTOM event-only setting.

Conditioning away common-cause structure suppressed shocks but also erased shifted coordinated parents. Partial routine and overlap repairs did not satisfy the combined gate. The resulting question was informational: if hidden causal accounts produce exactly the same visible events, what could an event-only method distinguish?

Let H_A mean intentional coordination and H_B a recurring external stimulus. Suppose both produce $X = x$. A deterministic method receives the same input. For a randomized method $f(X, U)$ whose random variable U has a law independent of the hidden account,

$$\mathcal{L}(f(X, U) \mid H_A, X = x) = \mathcal{L}(f(X, U) \mid H_B, X = x). \quad (8)$$

Here X contains all permitted detector inputs; the comparison fixes the same visible array and gives U the same distribution under both accounts. For any output set A , integrating $\mathbf{1}\{f(x, u) \in A\}$ against that common law yields the same probability. This is an identical-input implication, not a claim that arbitrary observational distributions identify a cause. Feature transformations of x cannot manufacture the missing distinction. Six constructed causal-twin fixtures demonstrate the construction; they are evaluator cases, not observed real-world causal examples. They do not show that all real cases are twins, or that richer provenance cannot distinguish causes.



The distinction requires additional provenance or assumptions, not a stronger score on the same X .

Figure 5: Different causal explanations can give identical detector-visible events. The limit concerns the supplied information, not every possible investigation or every fraud-detection task.

A gate conflict makes this precise. If a reconstruction score lies in $[0, 1]$ and no alert gives score zero, then $\mathbb{E}[F1] \leq \Pr(\text{alert})$. Requiring mean F1 at least 0.7 for one twin implies alert probability at least 0.7.

Requiring at most 0.07 alerts for its identical-input common-cause twin is incompatible. These constants belong to the constructed gate conflict; they are not estimated natural-population rates.

Robust excess is therefore evidence against a specified assignment null, not a posterior probability of intent. A small Monte Carlo comparison does not identify an organizer. Reconstruction creates a supported structural object, not an unobserved causal label. Exposure histories, referral evidence or independent coordination provenance could change the information setting, but no validated combined causal estimator is reported here.

11 An investigation system that preserves ambiguity

11.1 Support and missing evidence

The contract distinguishes SUPPORTED, SUPPORTED_WITH_CAUTION, AMBIGUOUS_COMMON_CAUSE and OUT_OF_SUPPORT. These describe applicability and limitations of structural evidence, not increasing fraud probabilities. Supported structure still withholds event-only causal attribution. Filling every evidence field is insufficient for an unqualified supported classification.

Evidence packages retain the source object, members, targets, period, temporal support, null description, support checks, alternatives, missing provenance and lineage. Missing data remains explicitly null. Diagnostic objects stay accessible when interpretation is out of support. This prevents a display or aggregate from silently strengthening the source claim.

11.2 Equivalence instead of unjustified scalar priority

Defensive tests exposed unstable parent association under bridges and overlap, plus ranking changes caused by lexical identifiers. A scalar priority could prefer one unresolved explanation without scientific justification. Remediation improved invariance, but sorting could not resolve shared-response ambiguity.

The accepted design uses sets of candidates equivalent under the stored evidence signature with equal scientific priority. Display order is nonsemantic; complete membership survives filtering and exports. A set does not assert one causal campaign or enumerate every possible explanation. Equivalence depends on the evidence signature, whereas the general causal limit applies beyond a particular implementation.



SUPPORTED | SUPPORTED_WITH_CAUTION
 AMBIGUOUS_COMMON_CAUSE | OUT_OF_SUPPORT

Support is about evidence applicability. All event-only causal attribution remains withheld.

Figure 6: Complete evidence, alternatives and equal-priority sets support review. Support concerns applicability; all event-only causal attribution remains withheld.

11.3 Coverage and unresolved disagreements

Combined development contains 480 candidates in 321 synthetic worlds and 477 original parents. Of the candidates, 416 are reviewable; 327 original parents have a reviewable match. Thus 416/480 measures candidate coverage and 327/477 original-parent coverage. Substituting the first for the second inflates recovery. Expected out-of-support recall remains 40/44, leaving four disagreements.

All 230 development equivalence classes require provenance, including singleton classes. A singleton means no competing member under this representation, not causal certainty. Later unseen synthetic fixtures exercise familiar mechanisms rather than fresh real populations. The contribution is an inspectable and testable specification for reporting evidence.

12 Inspectable research interface

A read-only dashboard exposes support, constituent evidence, alternatives, lineage and complete equal-priority sets on wholly synthetic fixtures. Missing provenance stays explicit and display order carries no scientific meaning. The supplement retains the dashboard screenshot and engineering receipts. The dashboard implements the evidence-reporting specification; analyst utility remains unmeasured.

13 September population holdout

After May external validation, we fixed the September population-detector holdout protocol. Reconstruction and investigation development then continued with the detector fixed and without accessing or using the September histories. The pre-final manuscript was frozen before final population access. A pre-access amendment clarified an underspecified randomization mapping, and the completed execution implementation was locked before the September population was materialized and scored. The holdout was not used to tune the detector; no scientific tuning occurred after the final outcomes were seen. The resulting outputs were then verified as described below. The protocol applied only to the population detector.

The September histories already existed in ListenBrainz snapshot 2723, dated 15 October 2026. The protocol, exclusions, seeds, metrics and gates were fixed before the September population was accessed or scored. “Pre-access” describes the evaluation boundary; it does not mean that the September events were collected after model freeze. The supplement summarizes the ordering and identifies the supporting protocol, output and verification records; exact execution timestamps remain in the internal provenance archive.

May and September each contain $42 \times 96 = 4,032$ listeners because the final reused the same predeclared sampling design. The September selection uses a separate 1 September inclusive–29 September exclusive 2026 UTC window and excludes all 16,783 identities in the May source window. Identical sample counts therefore do not mean the same listeners were reused. May contributes 4,004,802 selected observed events; September contributes 3,714,102.

Three construction seeds and three difficulties per group give 378 treatment/control pairs. Mean block and pooled AUC are 1.000, with 378/378 treatment alerts and 0/378 control alerts. These are controlled intervention labels on observed histories. The result extends population transfer under the same mechanism; it does not extend reconstruction evidence to September.

The weakest treatment has $Z = 36.5189$, a margin of 33.5189 above the frozen threshold of three. The closest control has $Z = 2.9228$, only 0.0772 below it. Thus perfect empirical decisions coexist with unequal margins, including a near-threshold control. No zero-scale sentinel occurs among the 756 final scores. The

Table 4: Descriptive stratification of the predeclared September difficulties, using the existing frozen scores. Each row contains 126 worlds across the same 42 groups. Signed margin is $Z - 3$; positive values alert. No score has zero null scale and no sentinel is displayed. Medians/ranges are descriptive, not confidence intervals. Exact quartiles and unrounded values accompany the supplement.

Sync. slots	Arm	Alerts / count	Median Z	Z range	Signed-margin range
32	Treatment	126/126	248.213	[75.447, 323.035]	[72.447, 320.035]
32	Control	0/126	-0.284	[-2.505, 2.923]	[-5.505, -0.077]
24	Treatment	126/126	185.967	[55.790, 242.956]	[52.790, 239.956]
24	Control	0/126	-0.096	[-2.280, 2.248]	[-5.280, -0.752]
16	Treatment	126/126	123.978	[36.519, 164.785]	[33.519, 161.785]
16	Control	0/126	-0.095	[-2.861, 2.120]	[-5.861, -0.880]

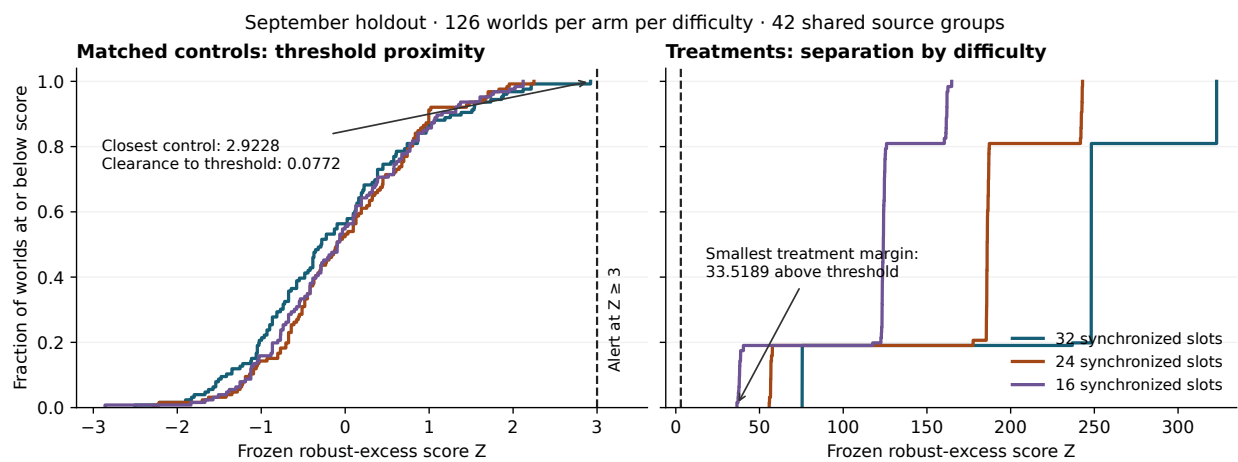


Figure 7: Frozen September score ECDFs by predeclared difficulty. Each line represents 126 dependent worlds per arm across 42 source groups. The control panel uses a narrower horizontal scale to expose threshold proximity; the treatment panel preserves the full score range. The dashed line is the unchanged alert threshold $Z = 3$. No zero-scale sentinels occur, no observations are capped, and no smoothing or new predictions are used.

existing score distributions show strong treatment–control separation and decreasing treatment scores as difficulty increases.

Uncertainty. Resampling observed groups with uniformly perfect outcomes necessarily reproduces perfect summaries. The original conditional bootstrap therefore supplies no information about unseen failure mechanisms; its degenerate summaries are retained in the supplement, not presented as population bounds. If the 42 groups were independent Bernoulli units, zero observed group failures would give an approximate one-sided 95% upper event probability of $1 - 0.05^{1/42} = 6.9\%$. Group independence is not established, so this is an illustrative reference calculation, not a valid population bound for PHANTOM.

Execution and verification. The final scientific execution ran once, and deterministic replay reproduced its outputs. The original verifier failed while generating a receipt after computation; later verification confirmed that the outputs were unchanged. No scientific tuning occurred. The supplement explains the pre-access lock replacement, original failure and post-run verification; detailed timing and incident

Table 5: Failure/remediation summary. A later scope change does not retroactively pass an earlier failed gate.

Problem	Failure	Implication	Change
Synthetic origin	Generated neutrals separated	Detector could identify simulator	Origin and neutral gates
Sessions	All source groups had ten rows	Unvalidated behavioral meaning	Session-free designs
Catalog shortcuts	Catalog exposed source origin	Graph gains were confounded	Matched actor marginals
Duplicates	20/180 pairs changed keys	Preprocessing changed support	Collision-safe construction
Raw calibration	22/180 controls alerted	AUC did not rescue threshold	Local median/MAD scaling
Graph null	High control alert rate	Reconstruction could be spurious	Presence, conditioning, pooling
Hybrid	Post-veto member F1 fell to 0.5357	Veto removed useful evidence	Retain core
Common stimulus	Every shock control alerted	Structure did not identify intent	Alternatives and provenance
Shifted overlap	F1 0.3636	Parent association failed	Explicit support boundary
Scalar priority	IDs and ambiguity changed rank	Ordering implied preference	Equal-priority sets

records remain in the internal provenance archive. Protocol commitments were recorded locally; verification used separate processes, without institutional attestation.

Future method changes or comparator experiments require new evidence under a separately fixed protocol.

14 Failures as design constraints

Some failures have representation fixes; others demand a different statistical question. Duplicate presence is a representation issue. Null conditioning refines what departures count as unusual. Causal twins expose missing information that a transformation of the same events cannot restore. A responsible research system must distinguish these categories rather than turning every unresolved requirement into a stronger score.

The chronology supplement records objectives, hypotheses, methods, gates, decisions and resulting changes for every stage. It retains rejected approaches and the conditions under which later work proceeded. The main narrative is organized by scientific problem so that readers can assess the accepted method without navigating internal phase numbers.

15 Reproducibility

The supplement retains protocol chronology, selectors, seeds, implementation identities and the verification incidents. Both manuscripts and companion reports use the same canonical metrics file, claim ledger and bibliography identity ledger; figure and reference numbers may differ. Descriptive stratification reads only frozen artifacts and creates no new predictions or null draws. Earlier publications and their source

records remain available. Full scientific reproduction requires authorized source access; public examples explain the methods with synthetic data.

16 Limitations

Construct and label validity. Labels encode controlled assignment interventions on observed histories, not verified manipulation. Common-stimulus controls can contain real synchronization. Event-only evidence lacks exposure, referral and organizer provenance; structural surprise is not a calibrated intent probability.

Support and coverage. The selector and intervention share a restrictive recurring pattern. The final assesses population transfer under this mechanism, not general coordination difficulty. Eligibility coverage on unrestricted streams and exchangeability under natural listening are unknown.

Population and generalization. Repeated worlds share histories, source groups and construction choices. The final has 42 groups rather than 756 independent populations. Independence across groups is unestablished, and conditional resampling cannot supply uncertainty over unseen mechanisms.

Reconstruction. Internal reconstruction evidence covers two historical blocks; later perturbations reuse development histories. Member/target recovery leaves period accuracy and cause distinct. The available holdout payload does not identify structural-ambiguity abstention reasons for missed parents. External detector results do not validate an end-to-end reconstruction or investigation system.

Comparative evidence. Historical calibration rows differ in population or fitting design. No matched run of published methods supports superiority. The retained failures reject the tested procedures, not every possible alternative.

Analyst and deployment utility. No analyst study, real-world queue precision estimate, enforcement outcome or production evaluation is available. Reviewable candidates do not guarantee original-parent coverage. The dashboard implements an evidence contract whose benefit remains to be measured.

Reproducibility and rights. Public reproduction covers aggregate reporting and wholly synthetic teaching examples. Full experiments require authorized data access; raw histories, protected truth and uncleared research code remain excluded. Local commitments are an audit trail rather than external timestamp attestation. PHANTOM's scientific system remains frozen; September cannot serve as unseen evidence for a revised method.

17 Conclusion

PHANTOM's central result is the distinction between benchmark validity, ranking, operating-threshold transfer and causal interpretation. Apparently strong benchmarks failed shortcut checks, and perfect ranking did not ensure a transferable global threshold. Local conditional normalization subsequently transferred across controlled populations under a declared recurring mechanism. Reconstruction recovered

some parents and missed others; common-stimulus and overlap tests ruled out broader claims. These findings support a system that presents structural evidence and unresolved alternatives for investigation. Broader mechanisms, matched comparative experiments, contextual evidence and analyst studies remain separate scientific work.

Authorship and AI assistance

Nischit Girish Kumar is the sole human author. NEXUS Research Initiative denotes independent research, not institutional sponsorship. **AI Assistance Disclosure:** AI-assisted tools were used extensively for implementation, debugging, analysis scripting, and verification, with minor assistance in manuscript editing. Nischit Girish Kumar originated and directed the research, designed the system architecture and experimental methodology, set the evaluation criteria, interpreted the results, and made the final scientific decisions. The manuscript was written and revised by Nischit Girish Kumar, who assumes responsibility for the work and its conclusions.

References

- [1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for Human-AI Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2019. doi: 10.1145/3290605.3300233. URL <https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/>.
- [2] Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. CopyCatch: Stopping Group Attacks by Spotting Lockstep Behavior in Social Networks. In *Proceedings of the 22nd International Conference on World Wide Web*, 2013. doi: 10.1145/2488388.2488400. URL https://alexbeutel.com/papers/www2013_copycatch.pdf.
- [3] Laetitia Gauvin, Mathieu Géniois, Márton Karsai, Mikko Kivelä, Taro Takaguchi, Eugenio Valdano, and Christian L. Vestergaard. Randomized Reference Models for Temporal Networks. *SIAM Review*, 64(4):763–830, 2022. doi: 10.1137/19M1242252. URL <https://arxiv.org/abs/1806.04032>.
- [4] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. FRAUDAR: Bounding Graph Fraud in the Face of Camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939747. URL https://bhooi.github.io/papers/fraudar_kdd16.pdf.
- [5] Fengrui Hua, Yiyan Qi, Zikai Wei, Yuxing Tian, Chengjin Xu, Xiaojun Wu, Jia Li, and Jian Guo. BAG: Benchmarking anomaly detection on dynamic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14892–14900, 2026. doi: 10.1609/aaai.v40i17.38510. URL <https://ojs.aaai.org/index.php/AAAI/article/view/38510>.
- [6] Letizia Iannucci, Elisa Muratore, Antonis Matakos, and Mikko Kivelä. Detecting Coordinated Activities Through Temporal, Multiplex, and Collaborative Analysis. arXiv preprint, 2025. URL <https://arxiv.org/html/2512.19677v1>.
- [7] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. CatchSync: Catching synchronized behavior in large directed graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014. doi: 10.1145/2623330.2623632. URL <https://www.cs.cmu.edu/~christos/PUBLICATIONS/14-kdd-catchsync.pdf>.
- [8] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):455–466, 2021. doi: 10.1609/icwsm.v15i1.18075. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/18075>.

- [9] Judea Pearl. Causal Inference in Statistics: An Overview. *Statistics Surveys*, 3:96–146, 2009. doi: 10.1214/09-S057. URL <https://web.cs.ucla.edu/~kaoru/SS-2009-57-Sup.pdf>.
- [10] Belinda Phipson and Gordon K. Smyth. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010. doi: 10.2202/1544-6115.1585. URL <https://gksmyth.github.io/pubs/PermPValuesPreprint.pdf>. Article 39.
- [11] Peter J. Rousseeuw and Christophe Croux. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993. doi: 10.1080/01621459.1993.10476408. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408>.
- [12] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as Treatments: Debiasing Learning and Evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1670–1679, 2016. URL <https://proceedings.mlr.press/v48/schnabel16.html>.
- [13] Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40:211–239, 2011. doi: 10.1177/0049124111404820. URL <https://arxiv.org/abs/1004.4704>.
- [14] Serena Tardelli, Leonardo Nizzoli, Maurizio Tesconi, Mauro Conti, Preslav Nakov, Giovanni Da San Martino, and Stefano Cresci. Temporal dynamics of coordinated online behavior: Stability, archetypes, and influence. *Proceedings of the National Academy of Sciences*, 121(20):e2307038121, 2024. doi: 10.1073/pnas.2307038121. URL <https://arxiv.org/abs/2301.06774>.
- [15] Viet Anh Tran. Deezer listening events dataset. Zenodo, version 1, 2024. URL <https://zenodo.org/records/13890194>.
- [16] Viet-Anh Tran, Guillaume Salha-Galvan, Bruno Sguerra, and Romain Hennequin. Transformers Meet ACT-R: Repeat-Aware and Sequential Listening Session Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024. doi: 10.1145/3640457.3688139. URL <https://arxiv.org/html/2408.16578v1>.
- [17] Sudheer Tubati and Amit Goyal. SAGE: Scalable Automatic Gating Ensemble for Confident Negative Harvesting in Fraud Detection. In *Companion Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, 2026. doi: 10.1145/3779211.3793166. URL <https://www.amazon.science/publications/sage-scalable-automatic-gating-ensemble-for-confident-negative-harvesting-in-fraud-detection>.
- [18] Chunyu Wei, Wenji Hu, Xingjia Hao, Yunhai Wang, Yueguo Chen, Bing Bai, and Fei Wang. Graph evidential learning for anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2025. doi: 10.1145/3711896.3736989. URL https://www.yunhaiwang.net/KDD2025/kdd_2025_GEL_Web/GEL_KDD25_camera.pdf.