

PHANTOM: Benchmark Validity and Causal Limits in Coordinated Listening

Nischit Girish Kumar

NEXUS Research Initiative — Independent Research

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.

Keywords

coordinated behavior, listening events, conditional randomization, campaign reconstruction, observational equivalence, investigation systems

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 [5]; FRAUDAR studies dense bipartite patterns under camouflage [4]. Behavioral-trace networks [6] and dynamic coordination communities [11] 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 [9] and a nonzero finite-randomization comparison [8]. 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 [13]; its dataset [12] supplies histories rather than verified fraud labels. SAGE addresses negative harvesting for streaming fraud in a different contextual and labeling regime [14]. 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 [7]. Shalizi and Thomas analyze confounding among homophily, contagion and covariate effects in observational networks [10]. 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.

3 Data and controlled benchmark

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 Data scale is not detector scale

The Deezer audit covers 704,728,520 events, 3,444,027 users and 43,803 observed tracks across March–August 2022. There are 500 event parquet files and one 50,000-row embedding file. All 70,472,852 source sessions have exactly ten records, so session boundaries lack validated behavioral meaning. The source contains 2,435,880 excess exact duplicates, about 0.3456%; individual effects can be larger than the global rate suggests. Same-second ordering and embedding temporal provenance remain unresolved. The release is associated with the PISA dataset [12].

The development sample has 671,410 rows and 3,252 complete release-relative histories. Full-source auditing is disk-backed; frozen detection and reconstruction use bounded samples. The graph system was not evaluated simultaneously on 704.7 million events. Neither source supplies verified fraud labels or sufficient duration, device, exposure, referral and organizer context for intent attribution. Deezer redistribution rights were not resolved from the visible release page, and raw data is excluded from the public package.

Both ListenBrainz evaluations use snapshot 2723, dated 15 October 2026; snapshot publication is distinct from event time. For external validation it is restricted to 1 May 2022 inclusive through 29 May exclusive. After eligibility and predeclared sampling, 4,032 listeners in 42 disjoint groups contribute 4,004,802 original events. The representation contains 1,391,257 recording-MSID keys, not verified distinct audio recordings. Selected histories are complete only within that window. Source observations are real; interventions and labels described next are synthetic.

4 Making a relational benchmark interpretable

4.1 Rejecting high scores with the wrong explanation

An early simulator created multiple campaign families and difficulty regimes around observed background. Generated actors differed in origin, catalog and session properties. Removing session features did not remove the mismatch: synthetic-neutral alerts were 12.08% versus 4.79% for observed background. Because the latter is unlabeled, this is a distribution-shift diagnostic

rather than a specificity estimate. Further origin probes rejected relational candidates before their downstream scores could be interpreted.

A paired additive intervention still altered actor marginals. A stronger reassignment preserved per-actor item and timestamp multisets, but initially changed semantic duplicate counts in 20/180 pairs. That preprocessing difference invalidated the benchmark even if a detector separated labels. Collision-safe reassignment then repaired the duplicate invariant. Intermediate performance gains were accepted only after the benchmark checks passed.

For selected actor u , let M_u be its item multiset and T_u its timestamp multiset. Treatment/control worlds preserve M_u , T_u , row counts, identity and declared duplicate properties while changing cross-actor item assignment; background remains fixed. Joint item-time association is allowed to change, so preserving marginals is not a claim that every possible actor feature is invariant. The intervention uses 32 recurring slots and eight observed items repeated four times, with difficulty varying synchronized slots at 32, 24 and 16.

Whole source blocks and campaigns define evaluation splits. A matched pair is dependent, and several seeds and difficulty levels share a source group. A larger number of constructed worlds is not the same as a larger number of independent populations. Validation has already been used and is therefore no longer unseen.

4.2 Perfect ranking and failed threshold transfer

AUC can remain perfect when both treatment and control scores shift above a threshold. Table 1 records the historical calibration comparisons, keeping each population and fitting design explicit.

The accepted response normalizes each eligible world against its own constrained assignment null. This improves transfer in the tested construction by accounting for local baseline and dispersion. It does not remove the need to state the null or show that natural listening satisfies its exchangeability assumptions.

5 Conditional detection and reconstruction

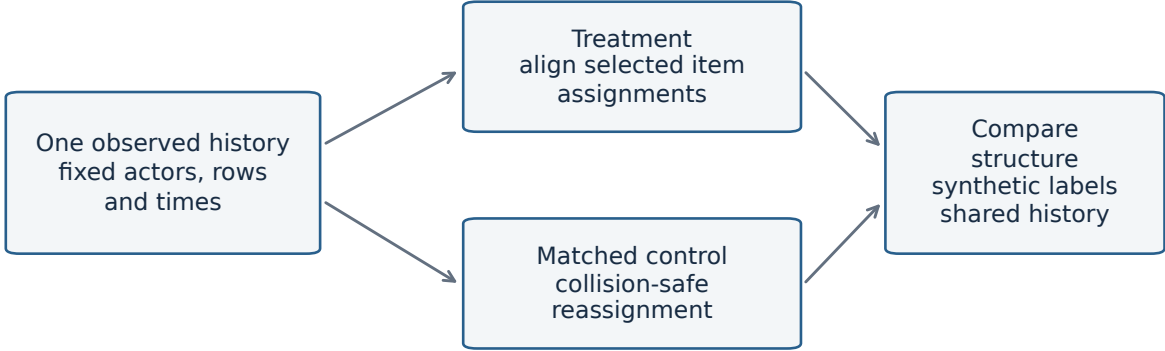
5.1 Population coactivity and a local reference

For minute bins $b = \lfloor t/60 \rfloor$, form unique user–item–minute presences $Q = \{(u, i, b)\}$. Let n_{ib} count actors per item/bin, N_b all actor-item presences per bin, and d_{ub} actor u ’s distinct items in bin b . Population coactivity is

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

Numerator and denominator count ordered cross-actor coincidences and opportunities. The implementation evaluates the full statistic through fixed-background sufficient counts, avoiding repeated materialization of complete null worlds.

The visible selector is narrow: an exact timestamp group has at least eight distinct users each appearing once; the same actor set must recur exactly 32 times in distinct minute bins spanning at least 27 days. Exactly one eligible group is required, otherwise the method is unsupported. This explicit support restriction is essential to interpreting the reported transfer.



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

Figure 1: Matched controlled worlds preserve declared actor marginals and share source histories. Labels describe the planted intervention, not real participant intent.

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%

Each actor’s selected sorted item multiset is independently permuted across its fixed slots, with 4,095 draws and seed 41101. Index randomization draws by r . Let m be the median null coactivity and $s = 1.4826 \text{ median } |C^{(r)} - m|$. Robust excess is $Z = (C - m)/s$ when $s > 0$, with the frozen alert rule $Z \geq 3$. If $s = 0$, equality maps to zero and positive/negative deviations to signed 10^{12} sentinels. These sentinels are ordering conventions; the score is neither a Gaussian tail probability nor an intent probability.

The null asks whether selected item assignments are exchangeable over the preserved schedule. A shared external stimulus can violate that assumption without intentional manipulation. Detecting a departure is consequently a structural result conditional on the reference model.

5.2 Duplicate-invariant campaign objects

Reconstruction starts from unique user–item–timestamp events (semantic atoms) in the Deezer listening-events dataset’s May 2022 window and retains multiplicity descriptively. Canonical ordering uses actor item/time fingerprints; ambiguous fingerprints cause abstention. Selected bursts join consecutive atoms whose gaps are at most 600 seconds. A burst needs at least eight distinct users and eight singleton-user atoms; an actor must participate in at least eight qualifying bursts. After this one-time actor filter, retained groups need two atoms. The procedure does not silently randomize multi-atom cells as if they were singleton observations.

Catalog overlap defines a fixed candidate graph: connect actors sharing at least three selected items and catalog Jaccard at least 0.10, then retain components with at least three actors. Within-actor permutations preserve these catalogs and candidate components. Catalog connectivity alone is not an alert.

For actors u, v , count shared selected slots with the same item and same fixed bin at width h , giving $w_{uv}^{(h)}$. The component statistic is

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

The denominator includes all unordered pairs, including zeros. The 60/300 bins share an origin and are nested: their agreement must not be described as arbitrary time-jitter robustness.

The graph null uses 511 within-actor selected-item permutations, seed 55101. Observed and null component scores receive the same median/MAD normalization. For observed Z_g and null $Z_g^{(r)}$, define $M^{(r)} = \max_g Z_g^{(r)}$ and

$$\hat{p}_g = \frac{1 + \sum_{r=1}^{511} \mathbf{1}[M^{(r)} \geq Z_g]}{512}. \quad (3)$$

A candidate is admitted as a reported structural object only when $Z_g \geq 3$ and $\hat{p}_g \leq 0.01$. Resolution is $1/512$. The maximum compares a fixed candidate family, but universal familywise calibration for organic listening is not established merely by this construction.

Admitted components link when target Jaccard is at least 0.25 and period overlap at least half the shorter interval. Targets are items shared by two selected member catalogs and may lack direct temporal support. Merged objects retain constituent scores, without an invented merged p -value or common-organizer claim. The accepted method is the frozen burst-schedule, consensus, pooled-fragment variant; alternative graph methods are not interchangeable with it.

5.3 What is evaluated

AUC describes ranking of constructed labels. Detection and control alert rate describe the fixed threshold at the world level. Reconstruction uses one-to-one parent/object assignment by member F1; unmatched parents receive zero. Member precision and recall describe matched sets, while target F1 measures target-set recovery. The supplement defines the specialized fragmented-parent metric. Aggregate F1 follows per-parent computation, not the harmonic mean of independently averaged precision and recall. Robustness primary scores retain original truth even when evidence is removed, with secondary analyses against only the remaining visible truth reported in the historical record.

Output semantics. A supported non-alert, an unsupported input and withheld attribution are distinct outcomes. Unsupported inputs are not negative classifications. Merged objects preserve constituent evidence; no jointly calibrated merged significance or common-organizer claim is introduced.

6 Evaluation logic

Admissibility precedes performance: origin, neutral and duplicate checks determine whether a score answers the intended relational question. Ranking, threshold behavior, reconstruction and reviewability then have separate gates. Passing selected perturbations cannot compensate for a failed conjunctive operating requirement. Duplicate defects, a misspecified natural-process null

and missing causal information are different failure mechanisms; they do not admit one generic scoring repair.

7 Evidence and failure boundaries

7.1 External population: strong separation, bounded transfer

The external test freezes the relational detector before applying it to ListenBrainz. Sampling seed 41201 and construction seeds 41211–41213 produce $42 \times 3 \times 3 = 378$ pairs across groups, seeds and difficulties, or 756 worlds. AUC is 1.000, all 378 treatments alert, no matched controls alert, and worst-group control alert rate is zero. These results concern the *controlled planted-coordination benchmark*; real listeners are not assigned real fraud labels.

The historical uncertainty procedure uses 2,000 whole-group bootstrap replicates, seed 41102. Perfect outcomes in every observed group yield degenerate percentile intervals. They describe this empirical distribution, not uncertainty about unseen mechanisms. Disjoint users do not establish independent groups when source periods and construction choices are shared. A zero observed alert count must therefore remain a count, not a claim of zero population error.

External transfer covers the population detector only. It does not validate graph reconstruction, support classification, analyst performance, queue precision or deployment. The intervention remains narrow, and support eligibility is not measured as coverage of an unrestricted production stream.

7.2 Reconstruction and why added complexity failed

Graph development uses six Deezer blocks and 252 treatment/control pairs, yielding member F1 0.9067 and 2/252 control alerts for the selected core. The external ListenBrainz numerical population was excluded from reconstruction tuning. The historical internal reconstruction holdout contains two reserved Deezer blocks, 84 pairs and 168 original treatment parents. The saved objects exactly recover both member and target sets for 154/168 parents (91.67%); 14 are missed and unmatched, with no partial recoveries or unmatched predicted objects. Each parent’s four headline metrics is either zero or one, so their common mean is $154/168 = 11/12 = 0.91\bar{6}$.

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 3). Control alerts are 0/84. The supplement distinguishes structural misses from withheld causal attribution and records the complete breakdown.

Earlier graph variants were unacceptable: catalog structure could give attractive F1 while alerting on every control, and initial nulls retained high burden. A behavioral hybrid later reduced post-veto member F1 to 0.5357 after veto decisions, with vetoed parents scored zero. Non-veto augmentation preserved the core but added zero within-block ranking gain across 210 comparisons. These failures support retaining the simpler accepted core.

Table 2: 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.

Table 3: 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%)			

7.3 Separate perturbations and a failed overall envelope

Historical stress tests reuse the development blocks. Member F1 is 0.7687 under 25% actor loss, 0.8981 under 25% event thinning, 0.9167 under 30-second jitter and 0.7540 with a 75% prefix. A four-disjoint-campaign panel reaches F1 1.000. These are sampled separate settings, not a jointly validated robust operating region. The 30-second jitter mean is independently $231/252 = 11/12$ on development parents, not the holdout outcomes.

Shifted 25% overlap yields member/target F1 0.3636 and full control-world alert rate, unlike simultaneous overlap. Popularity-shock and release-event controls each alert in all 12 moderate worlds. In those settings, 0.2544 of the evaluated background actors enter predicted groups (background absorption). Routine-listening controls also fail the intended moderate gate. The overall operating-envelope claim is therefore rejected despite several successful perturbations.

A shared stimulus may generate unusual synchronization under the assignment null. Such an alert can be structurally meaningful while failing a manipulation-control requirement. Renaming the

output cannot make the original gate pass. We therefore report the failure and narrow the claim.

8 From failed attribution to an investigation system

8.1 Why event-only scores cannot settle cause

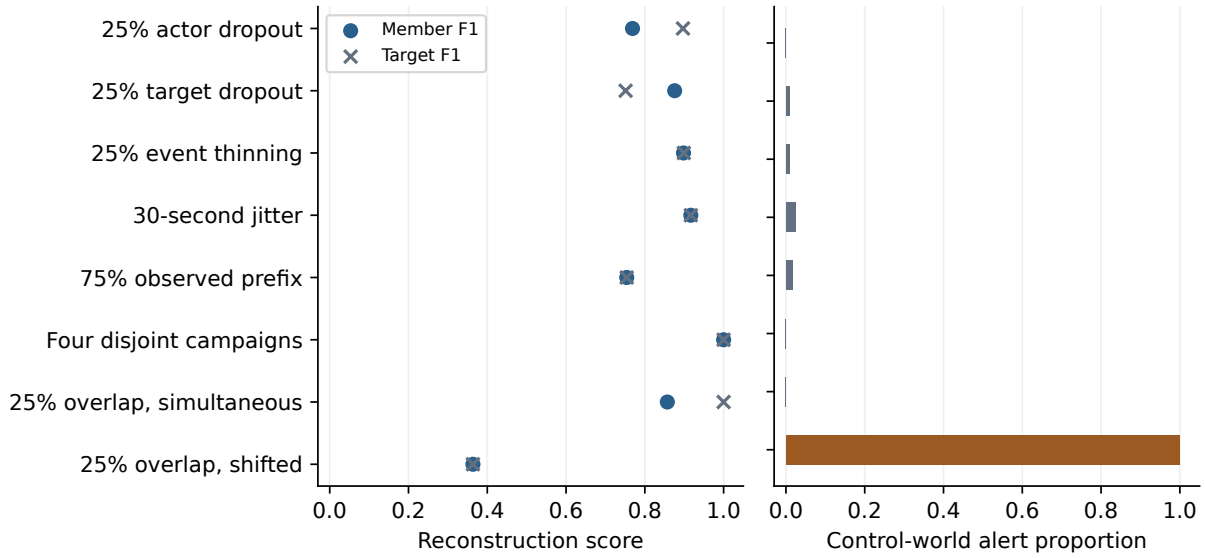
Observational identification already distinguishes evidence from causal assumptions [7, 10]. We apply observational-equivalence reasoning to the PHANTOM event-only setting. Let intentional coordination H_A and a recurring external stimulus H_B supply the same permitted input $X = x$. A deterministic detector returns the same output; if its randomness U has the same law independent of the hidden account, then

$$\mathcal{L}(f(x, U) \mid H_A) = \mathcal{L}(f(x, U) \mid H_B). \quad (4)$$

Six constructed causal-twin fixtures instantiate different hidden accounts with identical visible events; they are evaluator cases, not observed real-world causal examples. They establish neither the prevalence of such cases nor their relative probability. With F1 bounded by one and no-alert F1 zero, $\mathbb{E}[F1] \leq \Pr(\text{alert})$. A required mean F1 of 0.7 on one twin is incompatible with at most 0.07 alert probability on the other. These are constructed gates, not estimated population rates. The practical consequence is to retain structural evidence and request contextual provenance; richer transformations of identical event inputs cannot supply the missing causal distinction.

8.2 Support, alternatives and equal-priority sets

PHANTOM consequently separates supported structure from causal interpretation. The operating contract names supported, supported-with-caution, ambiguous-common-cause and out-of-support states. These concern evidence applicability, not fraud probabilities. Event-only attribution remains withheld even for a well-supported structural object. Missing provenance stays explicitly null, and diagnostic objects remain available when interpretation is withheld.



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

Figure 2: Both moderate successes and a failure region on reused development histories. Plotted aggregates come from frozen reports; no smoothing, new trials or independent-world interpretation is introduced. The additional shock failures prevent a global safe-envelope claim.

Later defensive testing found identifier-sensitive ranking and unresolved alternatives under bridges and shared responses. A repaired scalar ranking still did not justify a causal preference. The accepted development layer instead uses sets of candidates equivalent under the stored evidence signature, preserving complete membership and equal scientific priority through display, filtering and export. Display order is nonsemantic. A singleton is not causal certainty, and a set is not proof of a common organizer.

Combined development has 480 candidates across 321 synthetic worlds and 477 original parents. Candidate reviewability is 416/480, whereas original-parent reviewable coverage is 327/477. Expected out-of-support recall remains 40/44. These different denominators must not be substituted. All 230 equivalence classes require provenance. Unseen synthetic fixtures exercise familiar mechanisms, without providing fresh real-population or analyst validation.

8.3 Inspectable implementation, unmeasured utility

A read-only dashboard exposes support, constituent evidence, alternatives, lineage and complete equivalence sets on wholly synthetic fixtures. Missing provenance remains explicit, and display order has no scientific meaning. It performs no scientific recomputation or enforcement. The accompanying report describes fidelity checks; analyst effectiveness remains unmeasured.

9 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 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 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.

10 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.

11 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.

Acknowledgments

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. Nischit Girish Kumar is the sole human author. NEXUS Research Initiative denotes independent research, not institutional sponsorship.

References

- [1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fournery, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3290605.3300233

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]

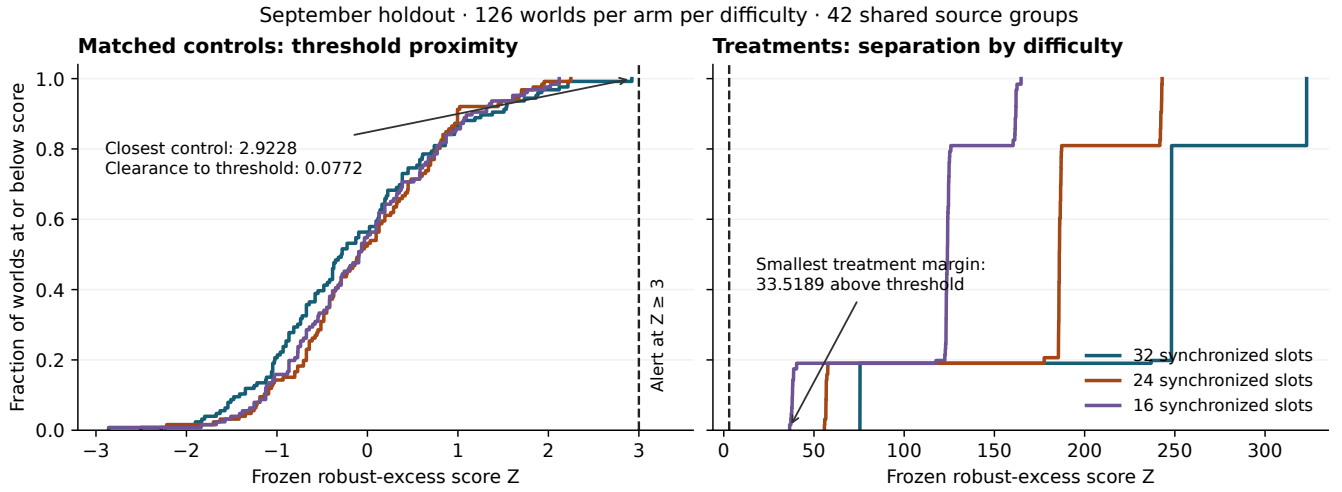


Figure 3: Frozen September score ECDFs by predeclared difficulty. Each line represents 126 dependent worlds per arm across 42 source groups. The control panel uses a narrower horizontal scale to expose threshold proximity; the treatment panel preserves the full score range. The dashed line is the unchanged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

- [2] Alex Beutel, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. 2013. CopyCatch: Stopping Group Attacks by Spotting Lock-step Behavior in Social Networks. In *Proceedings of the 22nd International Conference on World Wide Web*. doi:10.1145/2488388.2488400
- [3] Laetitia Gauvin, Mathieu Génois, Márton Karsai, Mikko Kivelä, Taro Takaguchi, Eugenio Valdano, and Christian L. Vestergaard. 2022. Randomized Reference Models for Temporal Networks. *SIAM Rev.* 64, 4 (2022), 763–830. doi:10.1137/19M1242252
- [4] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. 2016. FRAUDAR: Bounding Graph Fraud in the Face of Camouflage. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:10.1145/2939672.2939747
- [5] Meng Jiang, Peng Cui, Alex Beutel, Christos Faloutsos, and Shiqiang Yang. 2014. CatchSync: Catching Synchronized Behavior in Large Directed Graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:10.1145/2623330.2623632
- [6] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (2021), 455–466. doi:10.1609/icwsm.v15i1.18075
- [7] Judea Pearl. 2009. Causal Inference in Statistics: An Overview. *Statistics Surveys* 3 (2009), 96–146. doi:10.1214/09-SS057
- [8] Belinda Phipson and Gordon K. Smyth. 2010. 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 Article 39.
- [9] Peter J. Rousseeuw and Christophe Croux. 1993. Alternatives to the Median Absolute Deviation. *J. Amer. Statist. Assoc.* 88, 424 (1993), 1273–1283. doi:10.1080/01621459.1993.10476408
- [10] Cosma Rohilla Shalizi and Andrew C. Thomas. 2011. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research* 40 (2011), 211–239. doi:10.1177/0049124111404820
- [11] Serena Tardelli, Leonardo Nizzoli, Maurizio Tesconi, Mauro Conti, Preslav Nakov, Giovanni Da San Martino, and Stefano Cresci. 2024. Temporal Dynamics of Coordinated Online Behavior: Stability, Archetypes, and Influence. *Proceedings of the National Academy of Sciences* 121, 20 (2024), e2307038121. doi:10.1073/pnas.2307038121
- [12] Viet Anh Tran. 2024. Deezer listening events dataset. Zenodo, version 1. doi:10.5281/zenodo.13890194
- [13] Viet-Anh Tran, Guillaume Salha-Galvan, Bruno Sguerra, and Romain Hennequin. 2024. Transformers Meet ACT-R: Repeat-Aware and Sequential Listening Session Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*. doi:10.1145/3640457.3688139
- [14] Sudheer Tubati and Amit Goyal. 2026. 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*. doi:10.1145/3779211.3793166