# Nodewise E-Values Under Graph Interference: Causal Abuse Filtering for Agent Platforms

June Kim
Independent Researcher
june@june.kim
ORCID 0009-0005-3153-9396

July 6, 2026

## Abstract

A misbehaving AI agent on an email platform sends well-formed, DKIM-signed, paid messages while laundering malicious output through its earned reputation. Content filters pass every message, and the agent's interaction graph is indistinguishable from an honest subcontractor's. The discriminating question is causal: what would happen downstream if the platform throttled this node? We compose four published pieces into a filter that answers it with finite-sample guarantees: an exposure-integrated estimand under a Bernoulli friction design, doubly robust nodewise scores, betting e-values, and e-BH for false discovery rate control under the dependence a shared graph induces. A prior-art search, refreshed at this draft, finds each piece separately and the composition nowhere. On a synthetic agent platform (400 agents, 20 relays that deposit harm in their neighborhoods while active, 20 degree-matched honest subcontractors with identical edge structure), the filter flags 91% of relays after 100 randomized decision windows and 98% after 200 at target FDR 0.10, with realized FDR 0.000 and zero subcontractor false flags across 68,000 subcontractor decisions. The guarantees are exactly as conditional as the design. Replace randomized friction with reactive throttling, with the analyst still assuming the design, and at the finer materiality bar ($\tau$ = 0.25) realized FDR reaches 0.691 where the matched randomized control holds 0.000. Cut the friction rate from 50% to 5% and the e-values stay valid while detecting nothing at any tested horizon. Randomization is the precondition, and the experiment prices it.

## 1 The Relay Problem

Spam filtering is a predicate: each message passes a test or fails it. Agent abuse defeats predicates. The motivating attack is the relay, developed narratively in Return to Sender: an agent builds a clean record of delivered, paid work, then quietly forwards incoming tasks to an unvetted agent outside the trust topology, signs the tainted results with its own DKIM key, and ships them under its earned reputation. Every message it sends is well-formed. Payments settle. Volume is normal. The abuse lives in what the node does to its neighborhood, and no per-message test sees it.

Graph structure does not see it either. A relay and an honest subcontractor have the same edges: both accept tasks, both forward work, both return results. The difference is counterfactual. Throttle the relay and the downstream harm stops; throttle the subcontractor and nothing changes but latency. Distinguishing them requires intervening, because the quantity that separates them is defined by an intervention. The platform must inject randomized friction (throttle windows, send delays, fanout caps) and measure what changes. That turns abuse detection into a problem of causal inference on a graph: for each node v, test whether its interventional effect $\Delta_v$ exceeds a harm threshold $\tau$, with false discovery rate control across all nodes tested.

Return to Sender specified this filter as a composition of four published pieces and observed that no paper composes them. Here we state the construction and run it. The contributions:

1. The five-step composition, stated as a procedure with its validity conditions: design-integrated estimand, nodewise doubly robust scores, betting supermartingales, terminal e-values, e-BH.
2. A synthetic validation in which the composition runs end to end and separates relays from degree-matched honest subcontractors at controlled FDR, the discrimination that motivates the construction.
3. Measured failure modes. The two stated preconditions (designed randomization, adequate friction) are each ablated, and each ablation produces the predicted failure with a number attached: FDR 0.691 under confounded throttling, zero power under rare friction.

## 2 The Construction

The setting is a platform observing an interaction graph over `n` agents across repeated decision windows `t = 1, 2, ….` In each window the platform assigns each node a binary treatment `Z_{v,t}` (1 = throttled) by independent Bernoulli(`p`) draws, the reference design. After the window it observes `Y_{v,t}`, the harm recorded in `v`'s closed 1-hop neighborhood, clipped at a known cap `C` so the outcome is bounded: abuse reports, bounced attestations, flagged deliverables. Clipping is the analyst's choice, made before analysis; the tested estimand is the clipped-harm contrast.

Step 1: estimand. For each node, the effect of its own throttle on its own neighborhood's harm, with everyone else's treatments integrated over the reference design:

$$\Delta_v = E_\pi[\, Y_v \mid Z_v = 0 \,] - E_\pi[\, Y_v \mid Z_v = 1 \,]$$

where the expectation averages over the Bernoulli design $\pi$ for all other nodes. This is the marginal-contrast member of the exposure-mapping family of Aronow and Samii (2017). Integrating neighbors out, rather than conditioning on an exposure cell such as "all neighbors untreated," keeps positivity manageable on an overlapping graph: the estimand only requires that `v` itself has positive probability of each arm, which the design guarantees. The price is that $\Delta_v$ answers the marginal question, the average effect of throttling `v` under business-as-usual randomization of everyone else. For abuse filtering that is the operative question: the platform throttles one node at a time against the ambient traffic distribution.

Step 2: nodewise scores. The Horvitz-Thompson score

$$S_{v,t} = Y_{v,t} \cdot ((1 - Z_{v,t})/(1 - p) - Z_{v,t}/p)$$

is conditionally unbiased for $\Delta_v$ given the past, by design randomization alone. The $\pm Y/p$ swing dominates its variance, so we augment it with an outcome model in the standard doubly robust (AIPW) form of Robins, Rotnitzky and Zhao (1994): per-node, per-arm running means over past windows only. Because the model is fit on the past, the augmented score remains conditionally unbiased for $\Delta_v$ no matter how wrong the means are, and the residual terms shrink as they converge. This substitution alone lifts power at 100 windows from 0.52 to 0.91 (§3).

Step 3: accumulate. Email is a repeated game, so each window contributes a fresh score. For each node we run a betting supermartingale against the null $H_v : \Delta_v \le \tau$, in the style of Waudby-Smith and Ramdas (2023): wealth multiplies by $1 + \lambda(S_{v,t} - \tau)$ each window, with the bet $\lambda$ capped at the score's realizable floor so wealth stays nonnegative. A uniform mixture over a small grid of bet sizes stands in for bet tuning. Under `H_v` each factor has conditional mean at most 1, so wealth is a nonnegative supermartingale with initial value 1.

Step 4: e-values. By the optional stopping theorem, the wealth at any stopping time is an e-value: a nonnegative statistic with expectation at most 1 under the null (Grünwald, de Heide and Koolen, 2024). No correction for repeated looking, no minimum sample size; the platform reads the running wealth whenever it wants a verdict.

Step 5: e-BH. Rank the `n` terminal e-values, flag the largest `k` such that $e_{(k)} \ge n/(\mathrm{qk})$. Wang and Ramdas (2022) prove this controls FDR at level `q` under arbitrary dependence among the e-values, which is the whole point. On a shared graph, neighboring nodes' e-values are dependent through every edge they share, and no independence assumption holds.

The composition is finite-sample valid: each node's e-value is valid marginally by the supermartingale argument, and e-BH tolerates the dependence between them. What validity does not survive is loss of the

design. If treatments are assigned by anything other than the assumed randomization, step 2's conditional unbiasedness fails and everything downstream inherits the bias. §4 measures exactly this.

One structural remark from Return to Sender carries over: single-shot nodewise testing is impossible in general. With one treatment realization and one outcome per node, and no homogeneity assumption across nodes, two potential-outcome systems can produce identical observed data with opposite truth values for a given node's null. Repetition under fresh randomization makes the per-node question answerable at all.

## 3 Synthetic Validation

World. An Erdős–Rényi graph over `n = 400` agents with mean degree 8. Twenty relays and twenty subcontractors are drawn from the same mid-degree band, so the two groups match in degree distribution and differ in nothing observable from the graph. In each window, a relay that is not throttled deposits $\beta = 3$ units of harm into its closed neighborhood; a subcontractor deposits nothing, ever. Background harm arrives as per-neighbor Bernoulli noise at rate 0.05, so every neighborhood shows nonzero harm and honest nodes near relays sit in contaminated neighborhoods. Harm is clipped at `C = 6`. The design throttles each node independently at `p = 0.5` per window. We test $H_v : \Delta_v \leq \tau$ with $\tau$ `= 1` (one clipped-harm unit per window is the materiality bar) and run e-BH at `q = 0.10`. Every cell below is 200 replications with fresh graphs; with `n = 400` and `q = 0.10`, flagging all 20 relays requires the 20th-ranked e-value to reach 200.

Main result. Realized FDR and power (fraction of relays flagged), by number of randomized windows:

| Windows | HT power | AIPW power | FDR (both) |
|---|---|---|---|
| 10 | 0.000 | 0.000 | 0.000 |
| 25 | 0.000 | 0.000 | 0.000 |
| 50 | 0.000 | 0.000 | 0.000 |
| 100 | 0.521 | 0.909 | 0.000 |
| 200 | 0.778 | 0.977 | 0.000 |

Three readings. First, the filter works: at 100 windows the doubly robust variant flags 91% of relays, at 200 windows 98%, and across every designed cell (3,400 platform runs, 68,000 subcontractor decisions) it flags zero subcontractors. The degree-matched agent whose edges look exactly like a relay's is never touched, because its neighborhood's harm does not respond to its throttle. That is the discrimination no predicate or graph-structural filter can make. Second, variance reduction is worth more than horizon: the outcome model buys more power at 100 windows than doubling the horizon buys the raw HT score. Third, realized FDR sits at 0.000 against a target of 0.10 because every null in this world has slack (`Δ_v = 0` against $\tau$ `= 1`) and e-BH is conservative. The guarantee spent is an upper bound, and the ablation below shows what it costs to lose it.

Evidence is anytime-valid by the supermartingale construction, a property inherited from the cited theorems rather than exercised by this experiment, which reports fixed horizons. A platform would read wealth continuously. At hourly decision windows, 100 windows is roughly four days of ambient randomized friction. The power step between 50 and 100 windows is the e-BH bar: individual relay e-values grow exponentially but need to clear `n/(qk)` $\approx$ `200`, and at 50 windows the median relay's wealth is still climbing toward it.

## 4 Ablations: Pricing the Preconditions

The construction states two preconditions: treatments must follow the assumed randomization, and the design must put adequate probability on both arms. Each ablation deletes one and keeps everything else fixed. Both use the doubly robust score. Flag counts below are totals across the 200 replications of a cell (4,000 subcontractor decisions per cell), never single-run tallies.

Confounded throttling. The platform behaves the way an untutored abuse team would: throttle whoever's neighborhood was hot last window. The analyst, unaware, still computes weights as if Bernoulli(0.5) held. Reactive throttling synchronizes an honest node's treatment with its abusive neighbor's activity cycle (relay active, neighborhood hot, both throttled next window; relay throttled, neighborhood cool, both released), which manufactures positive drift in null nodes' scores. Measured over all honest nodes, the phantom drift lands near 0.5 harm units per window (0.477 to 0.507 across cells, against 0.000 under the design; the `null_drift` field in the result log). At $\tau$ `= 1` the materiality bar happens to sit above the drift and realized FDR stays under target at 100 windows (0.054), which is luck. The 17 subcontractor false flags there are the first casualties. At $\tau$ `= 0.25`, the bar an analyst hunting subtler abuse would set:

| Windows | FDR, confounded | FDR, randomized control | Subcontractor false flags, total (200 runs) |
|---|---|---|---|
| 50 | 0.046 | 0.000 | 9 vs 0 |
| 100 | 0.691 | 0.000 | 555 vs 0 |

Over two thirds of everything flagged is innocent. Most of the false flags are ordinary bystander nodes, and the degree-matched subcontractors, untouched in every designed run, are flagged 555 times. The matched control at the same $\tau$, same world, same horizon, holds FDR at 0.000 with power 0.989. The analyst cannot know the phantom-drift magnitude in advance, so raising $\tau$ can hide the drift in one regime but never repairs the design that produced it; only randomization does. This is the sense in which the e-values are meaningless without designed randomization: they remain numbers, they stop being evidence.

Rare friction. Keep the randomized design but cut the throttle rate to `p = 0.05`, the budget a platform reluctant to inconvenience agents might tolerate. The score floor scales as `C/p`, so the safe bet sizes shrink by an order of magnitude, and the throttled arm is observed a twentieth of the time. Result: power 0.000 at every horizon up to 200 windows, FDR 0.000, nothing flagged. Strict positivity still holds (every node has probability 0.05 of each arm); what collapses is its effective version, the failure mode Return to Sender called positivity collapse. The e-values stay valid and say nothing. Friction is the measurement instrument, and a platform that will not spend friction cannot buy evidence. The practical lever is concentrating the budget (higher `p` inside targeted audit cohorts) rather than diluting it platform-wide.

The third stated failure mode, harm propagating beyond the assumed neighborhood, is not ablated here. Under Bernoulli randomization the marginal estimand stays unbiased regardless of propagation radius, because neighbors are integrated out, never conditioned on. The radius re-enters for exposure-conditioned estimands and for interpreting which neighborhood's harm a flagged node is responsible for, and it stands as a live limit for any deployment that needs attribution rather than detection.

## 5 Limits

The world is synthetic and built to be favorable. Honest nodes have exactly zero effect by construction, so every null has slack against $\tau$ and no near-boundary honest node exists. A world with mildly harmful-but-tolerable agents would test the threshold, where this one tests the machinery. Relay effects are homogeneous and stationary ($\beta = 3$ forever); real attackers adapt, and an adapting relay that senses throttle windows and pauses is a different, adversarial game. The graph is static across windows and Erdős–Rényi rather than the heavy-tailed topology a real platform grows. Harm is observed noiselessly and attributed to the right neighborhoods; real harm signals arrive late, mislabeled, and gameable. The design is known exactly; a real platform's friction interacts with retries and queueing in ways that smear the assigned treatment. And the parameters that make the numbers land (cap `C`, bar $\tau$, budget `p`, window length) are product judgment applied to a specific network, which is the part Return to Sender argued no paper supplies. What the validation establishes is narrower and worth having: the composition is implementable, its guarantees hold where its preconditions hold, and each precondition, when deleted, fails in the predicted direction at measured size.

## 6 Related Work

The prior-art table from Return to Sender, plus one neighbor a fresh sweep at this draft (July 2026) surfaced. Each paper supplies a piece; none compose all four.

| Paper | Supplies | Missing |
|---|---|---|
| Puelz, Basse, Feller & Toulis (2022) | Finite-sample randomization tests under interference via biclique decomposition | P-values, one hypothesis at a time, no FDR |
| Athey, Eckles & Imbens (2018) | Exact p-values for non-sharp nulls under network interference | Single-shot, single hypothesis |
| Wang, Dandapanthula & Ramdas (2025) | Stopped e-BH for sequential FDR at arbitrary stopping times | No interference |
| He & Song (2024) | Nodewise doubly robust estimation under network dependence | Asymptotic; estimation, no testing |

| Paper | Supplies | Missing |
|---|---|---|
| Huang, Li & Toulis (2025) | Finite-sample tests for monotone spillovers via sub-network partitioning | Randomization p-values, no e-values, no FDR |
| Dalal et al. (2024) | Anytime-valid causal inference via confidence sequences and DML | No interference |
| Leung (2024) | Finite-sample conditional tests via random graph null models | Tests existence of interference, no nodewise effects |
| Viviano et al. (2024) | Platform experiments under approximate interference networks | Total effects, no per-node testing |
| Ogburn, Shpitser & Lee (2020) | Identification on networks via structural causal models | Identification, no finite-sample testing |
| Cortez et al. (2022) | Randomization inference of heterogeneous effects under interference | Homogeneity hypotheses, no nodewise nulls, no FDR |
| Gauthier, Bach & Jordan (2026) | Per-agent test supermartingales with BH-type FDR across agents, anytime-valid, in repeated games | Observational monitoring against equilibrium benchmarks; no interventions, no causal estimand, no interference |

The four composed pieces are themselves textbook or near-textbook: exposure mappings (Aronow & Samii 2017), AIPW (Robins, Rotnitzky & Zhao 1994), betting e-processes (Waudby-Smith & Ramdas 2023; Grünwald, de Heide & Koolen 2024), and e-BH (Wang & Ramdas 2022). The contribution here is the wiring and the evidence that the wired system does what the wiring diagram promises.

## 7 Conclusion

A relay defeats every per-message filter because its abuse is a property of its causal role, and the filter that catches it must therefore be causal: randomize friction, score neighborhoods, bet against innocence, control discoveries. The composition needs no new theorem, only four published pieces wired in the right order. Wired, it separates a relay from a subcontractor whose graph is identical, at FDR 0.000 against a 0.10 budget, in about a hundred randomized windows. The preconditions are priced rather than assumed: reactive throttling turns over two thirds of discoveries false, and starving the friction budget buys validity with no evidence. The standing claim is the composition and its measured behavior on the synthetic platform; the open work is the adversarial version, where the relay knows it is being measured.

## 8 Availability

Simulation code, tests, and the result log are archived on Zenodo at 10.5281/zenodo.21216212 (return-to-sender, AGPL-3.0): under 400 lines of NumPy, six property tests (score unbiasedness for both estimators, supermartingale validity under the null, e-BH correctness), and `run_experiment.py`, which regenerates every number above from fixed seeds in under half a minute on a laptop. This paper is CC BY-SA 4.0: fork it, formalize it, ship it, and keep it open.

## LLM collaboration disclosure

The construction, the estimand, the failure-mode predictions, and the prior-art table are the author's, specified in Return to Sender before any code existed. The simulation harness and this prose were drafted with Anthropic's Claude (Fable 5) from that specification and the experiment logs; the parameter choices and the decision of what to claim are the author's. No LLM decided what to publish.

## Funding

This work was conducted independently, with no external, institutional, or commercial funding. All compute costs were borne by the author.