WHEN RANDOMIZATION IS NOT ENOUGH: AUDITING EXPOSURE IN NETWORK EXPERIMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized network experiments observe assignments and outcomes, but the physical exposure that defines a causal effect may be generated by an unknown mechanism. The same assignment–outcome law can then support exposure effects of opposite signs. We formulate active partial identification: bridge experiments measure physical exposure under selected assignments to narrow the union of candidate-implied causal targets. Under candidate completeness, outcome sufficiency, and calibrated invariant probes, candidate intervals and a structural e-process give a causal envelope valid throughout adaptive querying and stopping. Packing and feasible splits bound intrinsic audit complexity—the worst-case depth of a decision tree over causal targets—by causal-target geometry rather than by the number of mechanisms. For candidates with a common pre-audit law, heterogeneous-channel lower bounds separate search and confirmation costs; supplied ordered target cells with feasible cuts and a common binary-symmetric channel attain these scales up to universal constants. Experiments separate acquisition, calibration, and information-cost effects and quantify bridge sensitivity: target-aware auditing pays off when target and model geometries differ; public-data analysis illustrates the missing-anchor boundary left by unmeasured exposure.

1 INTRODUCTION

Randomized network interventions assign whom to target, but the scientific question often concerns what people actually receive. In a public-health messaging experiment, direct delivery, resharing, and feed ranking can produce different receipt events under the same assignment. Whether targeting works and whether receipt helps are different causal questions; the latter matters when choosing how to deliver the next intervention. Exposure mappings connect assignments to causal comparisons (Manski, 2013; Ugander et al., 2013; Aronow & Samii, 2017), but the mapping may be unknown or misspecified (Sävje, 2024). This is an identification problem. Two mechanisms can induce the same assignment–outcome law yet imply opposite-sign physical-exposure effects (Theorem 5.1). Even unlimited observations from that experiment cannot distinguish them. The question is what additional information would resolve the causal conclusion.

Learning a representation and measuring receipt provide different information. Learned exposure representations use network and outcome information to estimate peer effects (Adhikari et al., 2026; Huber et al., 2026); sensitivity analyses bound effects under specified mapping uncertainty (Schröder et al., 2026). Outcome-based specification tests can detect some departures, but not unrestricted alternatives uniformly (Gao et al., 2026). Active causal design acquires new observations to narrow a target's uncertainty (Ailer et al., 2024). We study a distinct measurement choice: which physical-exposure label to acquire next while keeping the main outcome sample fixed. In the message example, a feasible probe sets a local sender pattern and records receipt. Probes should distinguish mechanisms that change the causal decision, rather than merely separate many mechanisms. A disagreement affecting one receiver may determine whether the effect interval crosses zero, whereas many other disagreements leave the target nearly unchanged.

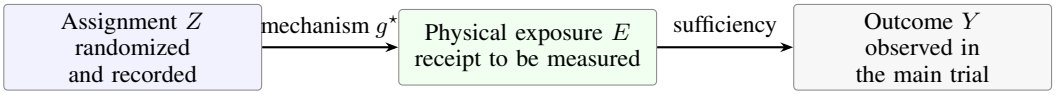
We formulate *active partial identification*: enlarge the information underlying a partially identified target (Manski, 2003) through targeted exposure measurements. The analyst supplies candidate mechanisms for one prespecified physical event. Main-experiment outcomes give each candidate an effect interval; bridge labels test its exposure predictions. Target-Aware Bridge Auditing (TAB)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

selects probes by their potential to narrow the retained intervals' hull, adjusted for reliability and cost. A structural e-process—a nonnegative evidence process calibrated for sequential testing—supports adaptive queries and stopping (Howard et al., 2021; Emmenegger et al., 2023). The guarantee requires a correct candidate, outcome-sufficient exposure, and invariant probes with calibrated labels. Under these conditions, the envelope covers the physical effect throughout auditing with probability at least $1 - \alpha - \delta$. Here α and δ bound main-data and structural errors. Thus outcomes determine candidate precision, while exposure measurements resolve structural ambiguity. An independently learned map can supply a candidate if it predicts the same event and is frozen before evaluation.

Our contributions are threefold.

- **Identification and active formulation.** We characterize sign nonidentification and the sharp union of causal targets compatible with the main experiment and exact audits. This specifies what new exposure measurements can identify (Theorems 5.1 and 5.2).
- **Anytime-valid causal reporting.** Candidate intervals and nested structural retention yield coverage under outcome-adaptive querying, continuous reporting, and stopping (Theorem 5.3). Sensitivity bounds separate mechanism and outcome errors (Theorem 5.7) from bridge-transport errors (Theorem D.4).
- **Target-aware information complexity.** Packing and feasible splits bound ideal audit depth. Under a common pre-audit law, search and confirmation have distinct information-per-cost lower bounds; supplied ordered target cells with feasible cuts attain both scales under a common binary-symmetric channel (Theorems 5.4 to 5.6). The construction adapts classical noisy search and sequential testing (Feige et al., 1994; Naghshvar & Javidi, 2013).



Bridge labels test the first link; causal interpretation uses the second.

Figure 1: **Active exposure anchoring.** Main randomization gives candidate-specific effect intervals. Bridge probes measure the physical event and update which intervals enter the causal envelope.

2 Related Work

Appendix A gives a fuller review and compares information sources and inferential outputs (Table 2).

Exposure representations. Interference designs distinguish direct, indirect, total, and overall effects (Hudgens & Halloran, 2008). Exposure mappings define comparisons (Manski, 2013; Ugander et al., 2013; Aronow & Samii, 2017); Sävje (2024) separates their estimands from sufficiency assumptions. Schröder et al. (2026) derive sharp direct/spillover bounds under exposure-propensity sensitivity restrictions and estimate them with orthogonal scores. Their information set is fixed. Bridge audits instead acquire exposure labels to update the candidate set and its causal report.

Structural validation. Outcome-based tests assess exposure restrictions (Hoshino & Yanagi, 2026; Tiwari & Basu, 2026; Huber et al., 2026). Against larger exposure-mapping models, worst-case type-I and type-II errors sum to at least one (Gao et al., 2026). Additional outcome restrictions can restore testability. Validation data correct specified network-membership errors (Chao et al., 2025). Bridge labels test the assignment-to-exposure mechanism.

Target-directed design. Sequential indirect design acquires treatment–outcome observations to narrow causal bounds (Ailer et al., 2024). ACI adapts assignments for Gaussian-process estimation with supplied neighborhood summaries (Zhu et al., 2024); ACE reduces posterior causal variance by selecting units or treatments (Song et al., 2023). Network bandits use supplied exposures (Zhang & Wang, 2025; Wang et al., 2025). In contrast, the audits studied here use exposure-only labels to inform a time-uniform causal envelope. Appendix A.3 specifies a shared indirect-design protocol.

Sequential inference and information. Likelihood ratios and confidence sequences support adaptive sampling (Ville, 1939; Howard et al., 2021; Emmenegger et al., 2023), including causal targets (Cook et al., 2024; Dalal et al., 2024); bounded-mean betting supplies fixed-cohort

bands (Waudby-Smith & Ramdas, 2024). Search/confirmation information (Burnashev, 1976; 1977; Naghshvar & Javidi, 2013) and nonuniform testing costs (Nitinawarat & Veeravalli, 2015; Vershinin et al., 2025) are classical. Our bounds reduce causal precision to target-separated testing; ordered search adapts Feige et al. (1994) and likelihood-ratio tests. Table 2 compares information sources and outputs.

3 Causal Model and Auditable Information

The setting has three objects: a physical event, its effect on outcomes, and the mechanism generating it. We define these separately, then connect the two information sources—randomized outcomes and new exposure labels—to the causal target.

3.1 Notation

For units $[n] = \{1, \dots, n\}$, an assignment $Z \in \mathcal{Z}_n$ follows the known design p_n. Potential outcomes $Y_i(z)$ are fixed; (Z, Y^{obs}), with $Y_i^{\text{obs}} = Y_i(Z)$, has law P_0. We use natural logarithms. Appendix A lists the symbols, their meanings, and where they are introduced.

3.2 Physical exposure and causal target

Exposure mappings ordinarily summarize assignments (Aronow & Samii, 2017). Here we first specify the event to be measured, separating its definition from outcome restrictions (Sävje, 2024).

Definition 3.1 (Physical exposure)**.** *For each unit i and assignment z, $E_i(z) \in \{0, 1\}$ denotes an event with a prespecified physical meaning, such as receipt of an experiment-induced message.*

Receipt within a fixed time window is distinct from reading or understanding. To attribute outcomes to the specified event, we impose exposure sufficiency, as in design-based interference models (Aronow & Samii, 2017; Sävje, 2024).

Assumption 3.2 (Outcome exposure sufficiency)**.** *There are fixed exposure-level potential outcomes $Y_i^E(0), Y_i^E(1)$ such that*

$$Y_i(z) = Y_i^E\{E_i(z)\} \quad \textit{for every } (i, z). \tag{1}$$

This restricts $E \to Y$: equal exposure must imply equal outcomes for a unit, giving the receipt contrast its causal meaning. Sender identity or dose must therefore act through E; a richer exposure or bounded residual effect relaxes this restriction (Theorem 5.7). The physical contrast is $\tau_n^E = n^{-1} \sum_i \{Y_i^E(1) - Y_i^E(0)\}$.

3.3 Candidate mechanisms and union targets

Because $z \mapsto E_i(z)$ is unknown, the analyst supplies a finite class $\mathcal{G}_n = \{g^{(1)}, \dots, g^{(M_n)}\}$ with $g_i : \mathcal{Z}_n \to \{0, 1\}$. Each candidate predicts the same physical event. Hop radii and receipt thresholds give interpretable families; learned maps may propose candidates (Adhikari et al., 2026). Class selection and positivity screening precede the independent evaluation assignment. We first analyze correct specification of this class, extending the supplied-map premise (Aronow & Samii, 2017):

Assumption 3.3 (Exact exposure mechanism)**.** *Some $g_n^\star \in \mathcal{G}_n$ satisfies $E_i(z) = g_{n,i}^\star(z)$ for every (i, z).*

This restricts $Z \to E$: one candidate must describe physical receipt so that its interval can carry the causal target. Delivery rules and independent receipt data guide the class. A certified approximation buffer permits an omitted exact mechanism (Corollary C.4); nonrejection alone cannot certify completeness (Example A.1).

Even an incorrect candidate defines a design-based contrast (Sävje, 2024). Its exposure probability is $\pi_{ie}(g) = \mathbb{P}_{p_n}\{g_i(Z) = e\} > 0$. The corresponding outcome mean is $\mu_{ie}(g; P_0) = \mathbb{E}_{P_0}[\mathbf{1}\{g_i(Z) = e\} Y_i^{\text{obs}}]/\pi_{ie}(g)$. Their average contrast is

$$\theta_n(g; P_0) = n^{-1} \sum_i \{\mu_{i1}(g; P_0) - \mu_{i0}(g; P_0)\}. \tag{2}$$

Under outcome sufficiency, $\theta_n(g_n^\star; P_0) = \tau_n^E$; other candidates generally give pseudo-targets. Let $\mathfrak{M}_g$ be the submodel with mechanism g and sufficient exposure. Their union is $\mathfrak{M}_n = \bigcup_g \mathfrak{M}_g$. The compatible candidates are $\mathcal{G}_n(P_0) = \{g : \mathfrak{M}_g \text{ can induce } P_0\}$. The unaudited identified set is $\Theta_{n,0}(P_0) = \{\theta_n(g; P_0) : g \in \mathcal{G}_n(P_0)\}$. This is a population identification set. The algorithm uses the supplied class $\mathcal{G}_n$ and finite-data intervals; it need not compute compatibility with the unknown law P_0.

3.4 Main randomization and bridge information

Candidate-wise inference. Main randomization supplies an effect estimate and interval for each mechanism. Bridge audits keep the estimates fixed and decide which candidate intervals remain plausible. Using design-based inverse-probability estimation (Aronow & Samii, 2017), let $\mathcal{H}_n^{\mathrm{aug}}$ be an external sigma-field independent of the evaluation assignment, and let m_{ie} be measurable with respect to it. These adjustments may be pilot predictions or zero. Define

$$\psi_i(g) = m_{i1} - m_{i0} + \frac{\mathbf{1}\{g_i(Z) = 1\}}{\pi_{i1}(g)}(Y_i^{\mathrm{obs}} - m_{i1}) - \frac{\mathbf{1}\{g_i(Z) = 0\}}{\pi_{i0}(g)}(Y_i^{\mathrm{obs}} - m_{i0}).$$

Set $\widehat{\tau}_n(g) = n^{-1}\sum_i \psi_i(g)$. Intervals require control of exposure support, residual size, and shared randomized inputs. A dependency graph encodes the last condition (Janson, 1988); write $\underline{\pi}_g = \min_{i,e} \pi_{ie}(g)$ and $\underline{\pi} = \min_g \underline{\pi}_g$.

Assumption 3.4 (Support and local dependence)**.** *Every candidate has $\underline{\pi}_g > 0$. Conditional on the augmentation data, $|Y_i(z) - m_{ie}| \le R_g < \infty$ whenever $g_i(z) = e$ and $p_n(z) > 0$, with $R_g > 0$. Under this conditional law, the scores for every candidate share a valid dependency graph $H_n^\cup$, with chromatic number $\chi_n^\cup$ and maximum degree $D_n^\cup$.*

This is a sampling condition on $Z \mapsto \psi_i(g)$, not a further restriction on either causal arrow. Positivity and bounded dependence make finite-sample intervals possible. Disjoint vertex sets without connecting edges index independent scores. Thus shared outcome inputs must also enter the graph. A conservative graph or another valid interval construction can replace the local bound. For weights $w_g > 0$, $\sum_g w_g = 1$, fixed before evaluation, Theorem C.2 gives $I_n(g; \alpha w_g) = [\ell_n(g), u_n(g)]$ centered at $\widehat{\tau}_n(g)$ with radius (24). Any valid coloring suffices. Fixed-cohort modelwise bands provide an alternative (Proposition C.5 and lemmas C.6 and C.7).

Bridge measurements. A query q in the finite feasible menu $\mathcal{Q}_n$ specifies a receiver $i(q)$ and assignment $z(q)$; candidate g predicts $a_q(g) = g_{i(q)}\{z(q)\}$. For local candidates, setting their common dependency footprint suffices, with arbitrary completion elsewhere. Global candidates require correspondingly feasible control (Appendix A.1). Write A_t for the recorded label and include an independent random seed U in the initial information. The query Q_t is *predictable*: it is selected before A_t, using

$$\mathcal{F}_{t-1} = \sigma(\mathcal{H}_n^{\mathrm{aug}}, U, Z, Y^{\mathrm{obs}}, Q_1, A_1, \ldots, Q_{t-1}, A_{t-1}).$$

Sequential likelihood inference needs a conditional observation law (Naghshvar & Javidi, 2013; Emmenegger et al., 2023). Here the probe must also preserve the main mechanism.

Assumption 3.5 (Bridge invariance and calibrated channel)**.** *The probe preserves the main assignment-to-exposure mechanism. A channel K_q and charged cost $c_q > 0$ are specified before auditing, with*

$$\mathbb{P}(A_t = a \mid \mathcal{F}_{t-1}, Q_t = q, g_n^\star = g) = K_q\{a \mid a_q(g)\}.$$

A binary-symmetric channel (BSC) flips its input with probability $\eta_q \in [0, 1/2)$; $\eta_q = 0$ gives an exact label.

This preserves $Z \to E$ across protocols and calibrates the measurement $E \to A$, so audit evidence applies to the main mechanism. A selected-query error bound can replace the exact channel (Proposition D.3); transport and calibration errors have separate allowances (Theorem D.4 and proposition A.3). Matched reference/probe receipts test a necessary transfer implication (Appendix A.2). The candidate likelihood is $L_t(g) = \prod_{r=1}^t K_{Q_r}\{A_r \mid a_{Q_r}(g)\}$, with $L_0(g) = 1$.

With exact labels, let $H_s = \{(q_r, b_r) : 1 \le r \le s\}$ record queries and noiseless answers. Then

$$\Theta_{n,s}(P_0, H_s) = \{\theta_n(g; P_0) : g \in \mathcal{G}_n(P_0),\ a_{q_r}(g) = b_r,\ r \le s\}. \tag{3}$$

Exact answers remove inconsistent predictions; noisy answers enter the retention rule in Section 4.

3.5 POPULATION AND INFERENTIAL TARGET GEOMETRY

Audit design separates unresolved targets from sampling uncertainty. For nonempty $V \subseteq \mathcal{G}_n$, the population diameter is $\Delta_n(V) = \max_{g\in V} \theta_n(g; P_0) - \min_{g\in V} \theta_n(g; P_0)$. The finite-data hull is $\mathcal{I}_n(V) = [\min_{g\in V} \ell_n(g), \max_{g\in V} u_n(g)]$, with width $W_n(V) = \max_{g\in V} u_n(g) - \min_{g\in V} \ell_n(g)$. The first measures intrinsic ambiguity; the second determines reportable precision.

An exact answer leaves $V_{q,a} = \{g \in V : a_q(g) = a\}$. Section 4 compares the resulting interval widths when choosing a query. Section 5.3 asks how many exact queries can resolve the population targets. These are different questions because finite-data intervals also contain sampling error. Table 3 summarizes the conditions and validation requirements.

4 NESTED TARGET-AWARE BRIDGE AUDITING

An audit has two tasks: decide which mechanisms the labels still support, and choose the next useful measurement. The retention rule supports coverage; query choice affects how much the envelope contracts.

Evidence and permanent rejection. Recall that $L_t(g)$ is candidate g's likelihood for the labels collected so far. For a candidate h, compare it with the fixed mixture over the original class: $m_t = \sum_{g\in\mathcal{G}_n} w_g L_t(g)$ and $\mathsf{E}_t(h) = m_t/L_t(h)$. A large $\mathsf{E}_t(h)$ means that the observed labels favor the mixture over h. It is evidence against h, not a posterior probability. The weights and mixture class remain fixed after rejections. Here E_t denotes evidence; $E_i(z)$ denotes physical exposure. The running maximum $\mathsf{E}_t^{\max}(h) = \max_{0\le s\le t} \mathsf{E}_s(h)$ remembers the strongest evidence so far. At $1/\delta$, h is permanently rejected. Later evidence can fall without restoring it (Example A.2).

Set $\mathsf{E}_t(h) = \infty$ when $L_t(h) = 0$, a null event under h. Under h and any predictable query policy, $\mathsf{E}_t(h)$ is a nonnegative supermartingale starting at one (a martingale for full-support channels). Ville's inequality applies to this process, not to its running maximum, and bounds the probability of ever rejecting a true h by δ (Ville, 1939; Howard et al., 2021). Thus

$$\mathcal{V}_t(\delta) = \{h : \mathsf{E}_t^{\max}(h) < 1/\delta\} \tag{4}$$

retains the true candidate at every audit time with probability at least $1 - \delta$. For nonempty $\mathcal{V}_t$, retain its main-experiment intervals and report their hull:

$$\mathcal{C}_{n,t}(\alpha, \delta) = [\min_{g\in\mathcal{V}_t} \ell_n(g),\ \max_{g\in\mathcal{V}_t} u_n(g)]. \tag{5}$$

The intervals stay fixed; labels change their membership in the report. An empty $\mathcal{V}_t$ reports model failure, with $\mathcal{C}_{n,t} = \varnothing$. Nonemptiness records nonrejection within the supplied class (Example A.1).

Information-priced target gain. For binary entropy $h(u) = -u\log u - (1-u)\log(1-u)$, with $0\log 0 := 0$, a BSC(η_q) has

$$C_q = \log 2 - h(\eta_q), \qquad \kappa_q = (1-2\eta_q)\log\frac{1-\eta_q}{\eta_q}, \quad 0 < \eta_q < 1/2. \tag{6}$$

C_q is channel capacity; κ_q is the mean log-likelihood gain between opposite predictions. Exact channels have $C_q = \log 2$ and $\kappa_q = \infty$. Fix a tolerance $\varepsilon > 0$ and resolution levels $\{\beta_s\}_{s\ge1} \subset (0, 1)$. The batch-size proxy is

$$r_q(\beta_{t+1}) = \begin{cases} 1, & \eta_q = 0, \\ \max\{1, \lceil \log(1/\beta_{t+1})/\kappa_q \rceil\}, & 0 < \eta_q < 1/2. \end{cases} \tag{7}$$

Recall that $W_n(V)$ is the current interval width. For a query q, let $\Gamma_n^I(q; V) = \max_a W_n(V_{q,a})$ be the worst width after an exact answer, omitting empty cells. For $V = \mathcal{V}_t$, define

$$\mathrm{Score}_t(q) = \frac{\log\{W_n(V) \vee 2\varepsilon\} - \log\{\Gamma_n^I(q; V) \vee 2\varepsilon\}}{c_q r_q(\beta_{t+1})}. \tag{8}$$

The numerator rewards a query only for the width reduction guaranteed by either exact answer; the denominator approximates its resolution cost. Truncation at 2ε stops rewarding resolution beyond

the requested precision. This score affects efficiency, while Theorem 5.3 preserves coverage for any predictable rule. Each raw label updates (4); β_{t+1} is a pricing parameter, not an error budget.

Operational loop. TAB selects queries, updates retention, and reports the hull (Algorithm 1). It stops at a precision or decision threshold, an exhausted budget, zero one-step gain, or model failure. Continuing through a zero-gain plateau is also valid. Example A.2 works through one receipt-query update.

5 Identification, Honest Inference, and Audit Complexity

The theory follows the information available to the analyst. The main experiment first determines which causal targets remain possible (Theorems 5.1 and 5.2). Noisy exposure labels can then narrow a finite-data envelope while preserving coverage at every audit time (Theorem 5.3). With validity established, target geometry determines the distinctions to resolve and channel information limits their cost (Theorems 5.4 and 5.5). Ordered queries attain these information scales in a structured class (Theorem 5.6); approximation bounds handle departures from exact specification (Theorem 5.7). Figure 2 links the results to TAB; Appendices B–E give proofs.

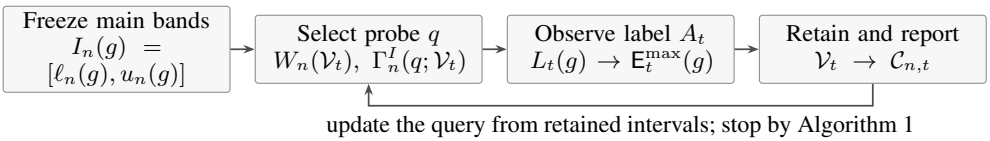


Figure 2: **From identification to adaptive reporting.** Theorem 5.2 defines the targets; Theorem 5.3 validates retention and reporting; Theorems 5.4 to 5.6 bound query complexity and cost.

5.1 What randomization cannot identify

Even perfect knowledge of the assignment–outcome law can leave the exposure effect unresolved. This is a lack of identification, rather than sampling precision.

Theorem 5.1 (Sign nonidentification without exposure anchoring)**.** *There exist a two-unit Bernoulli randomized experiment and two models satisfying Assumptions 3.2 and 3.3 that induce the same complete law of (Z, Y^{obs}) while their physical-exposure contrasts are $+1$ and -1. No procedure based only on the main randomized law can therefore identify the sign uniformly over the union.*

For independent Bernoulli$(1/2)$ assignments, take $(E_1, E_2) = (Z_2, Z_1)$ with $Y_i^E(e) = e$, or $(E_1, E_2) = (1 - Z_2, 1 - Z_1)$ with $Y_i^E(e) = 1 - e$. Both give $(Y_1, Y_2) = (Z_2, Z_1)$, but the effects are $+1$ and -1. The same observable response therefore supports opposite receipt effects. The appropriate object is the set of all compatible targets, characterized next.

Theorem 5.2 (Union-model identified set)**.** *If every compatible candidate submodel point-identifies its physical-exposure contrast, then*

$$\{\tau_n^E(m) : m \in \mathfrak{M}_n,\ P_m = P_0\} = \{\theta_n(g; P_0) : g \in \mathcal{G}_n(P_0)\}. \tag{9}$$

If, in addition, exact bridge answers satisfy Assumption 3.5, then after history H_s the sharp set is $\Theta_{n,s}(P_0, H_s)$ in (3)*.*

Sharpness means that every value in the identified set is attainable under the observed law and assumptions. Thus a positive sign requires all compatible targets to be positive, while equal targets give point identification even if several mechanisms remain. Exact audits can resolve sign uniformly precisely when each feasible-query signature class is sign-homogeneous (Corollary B.1).

5.2 Candidate inference and anytime causal coverage

The population set uses the main law and exact audit answers. The reported envelope $\mathcal{C}_{n,t}$ instead contains the intervals of retained candidates. With finite data and noisy labels, two errors matter: a candidate interval can miss its target, or auditing can discard the true mechanism. Allocate α to the bands of Proposition C.1 and Theorem C.2 and δ to structural rejection. True-candidate marginal coverage also suffices (Proposition C.5).

Theorem 5.3 (Anytime-valid active partial identification)**.** *Let $\alpha, \delta \in (0,1)$ with $\alpha + \delta < 1$. Under Assumptions 3.2 to 3.5, for arbitrary predictable queries,*

$$\mathbb{P}\{g_n^\star \in \mathcal{V}_t(\delta)\ \forall t\} \geq 1 - \delta, \tag{10}$$

$$\mathbb{P}\{\tau_n^E \in \mathcal{C}_{n,t}(\alpha, \delta)\ \forall t\} \geq 1 - \alpha - \delta. \tag{11}$$

An empty set counts as a coverage failure. The second statement also holds at every almost surely finite stopping time.

On the joint good event, one retained interval contains the physical target at every audit time; their hull must contain it too. A union bound combines the two errors without independence. Query choice and stopping can therefore adapt to the outcomes and previous labels without a new error allocation at each look. Calibrated bridge deviations are handled separately by Theorem D.4.

5.3 How many target-relevant audits are needed?

Recall that $\Delta_n(V)$ is the range of population targets, whereas $W_n(V)$ is a finite-data interval width. Let $D_\varepsilon^\star(V, \mathcal{Q}_n)$ be the smallest worst-case exact-query depth achieving $\Delta_n \leq 2\varepsilon$, or infinity if impossible. A leaf may retain several mechanisms with close targets, unlike exact hypothesis recovery (Nowak, 2011). The packing number $N_\varepsilon(V)$ counts the largest collection of targets more than 2ε apart (Tsybakov, 2009). A node is (ρ, ε) target-splittable when some query has $\Gamma_n^\theta(q; V) := \max_a \Delta_n(V_{q,a}) \leq (1-\rho)\Delta_n(V)$, omitting empty cells. Packing counts necessary distinctions; splittability describes the available measurements.

Theorem 5.4 (Target-complexity sandwich)**.** *For every nonempty $V \subseteq \mathcal{G}_n$, $\lceil \log_2 N_\varepsilon(V) \rceil \leq D_\varepsilon^\star(V, \mathcal{Q}_n)$. If $0 < \rho < 1$ and every reachable nonterminal node is (ρ, ε) target-splittable, then*

$$D_\varepsilon^\star(V, \mathcal{Q}_n) \leq d_\rho(V) := \begin{cases} 0, & \Delta_n(V) \leq 2\varepsilon, \\ \left\lceil \dfrac{\log\{\Delta_n(V)/(2\varepsilon)\}}{-\log(1-\rho)} \right\rceil, & \Delta_n(V) > 2\varepsilon. \end{cases}$$

If no valid tree exists, $D_\varepsilon^\star = \infty$.

The lower bound counts leaves needed for separated effects; the upper bound constructs a tree by repeatedly contracting target diameter. Many mechanisms can share one leaf. If their target diameter is already at most 2ε, no audit is needed, however large the class. Under simultaneous bands, $W_n(V) \geq \Delta_n(V)$, so stopping at $W_n(V) \leq 2\varepsilon$ certifies the target resolution. Modelwise bands retain physical-target coverage but not this simultaneous statement about all pseudo-targets (Proposition C.5). Theorem D.9 gives a corresponding hard-decoding contraction.

5.4 Information limits with heterogeneous audit costs

Exact-query depth ignores telemetry noise and resource prices. We now lower-bound their cost for target-separated candidates indistinguishable before auditing. For a finite-output channel K_q with binary input, define its capacity $C_q = \sup_{p \in [0,1]} I_p(X; A \mid q)$ and

$$\mathfrak{C} = \sup_{q \in \mathcal{Q}_n} \frac{C_q}{c_q}, \qquad \mathfrak{K}(g,h) = \sup_{q \in \mathcal{Q}_n} \frac{\mathrm{KL}\{K_q(\cdot \mid a_q(g)), K_q(\cdot \mid a_q(h))\}}{c_q}. \tag{12}$$

$\mathfrak{C}$ is the best information rate per cost for searching the class; $\mathfrak{K}(g,h)$ is the best rate for distinguishing a specific pair. Queries on which g and h agree contribute zero. Write $\mathsf{Cost}_\mathsf{T} = \sum_{t=1}^{\mathsf{T}} c_{Q_t}$ for the accumulated cost.

Theorem 5.5 (Information per cost within a common pre-audit law)**.** *Let $V = \{g_1, \ldots, g_K\}$ be compatible hypotheses with pairwise target separation greater than 2ε, $K \geq 2$. All information available before auditing, including augmentation data, has the same joint law under these hypotheses. The predictable queries use the calibrated channels in Assumption 3.5. Let $0 < \delta < 1/2$. A procedure stops almost surely at T and returns a transcript-measurable interval of diameter at most 2ε. Suppose its coverage is at least $1 - \delta$ under each hypothesis. For $F_K(\delta) = \log K - h(\delta) - \delta \log(K-1)$ and $0 < \mathfrak{C} < \infty$,*

$$\frac{1}{K} \sum_{j=1}^{K} \mathbb{E}_j \mathsf{Cost}_\mathsf{T} \geq \frac{F_K(\delta)}{\mathfrak{C}} \geq \frac{\{(1-\delta)\log K - \log 2\}_+}{\mathfrak{C}}. \tag{13}$$

For any separated pair with $0 < \mathfrak{K}(g_j, g_k) < \infty$*,*

$$\mathbb{E}_j \mathsf{Cost}_\mathsf{T} \geq \frac{\mathrm{kl}(1-\delta, \delta)}{\mathfrak{K}(g_j, g_k)}. \tag{14}$$

The reverse bound uses $\mathfrak{K}(g_k, g_j)$ *and* $\mathbb{E}_k \mathsf{Cost}_\mathsf{T}$*. If all available queries have zero information for a separated pair, no such uniformly valid procedure exists.*

The search bound grows with the number of separated targets; confirmation grows with the confidence required against a single alternative. They are separate necessary bounds, not an additive lower bound. A global packing applies only within a common pre-audit law; informative initial data reduce the required audit information (Appendix B.5). Corollary B.2 specializes both rates to a unit-cost BSC.

5.5 Ordered search at the two information scales

A lower bound does not supply a successful policy. Ordered targets permit separate search and confirmation stages. Recall that C_η and κ_η measure search information and pairwise evidence per label. Suppose the population targets are supplied in nondecreasing order and every cut $a_q(g_k) = \mathbf{1}\{k \leq q\}$ is feasible. Partition them greedily into maximal consecutive blocks $B_1, \ldots, B_K$ of diameter at most 2ε. Then $K = N_\varepsilon(\mathcal{G}_n)$ (Corollary B.3).

Theorem 5.6 (Search and confirmation for ordered target cells)**.** *In this ordered setting, let* $K \geq 2$*,* $0 < \delta < 1/2$*, and suppose every cut supplies fresh unit-cost BSC*(η) *labels,* $0 < \eta < 1/2$*. There is a procedure that stops almost surely and returns the true block's target hull with probability at least* $1 - \delta$*, with*

$$\sup_{g \in \mathcal{G}_n} \mathbb{E}_g \mathsf{T} \leq A_0 \frac{\log K}{C_\eta} + A_1 \frac{\log(1/\delta)}{\kappa_\eta}, \tag{15}$$

where A_0, A_1 *are universal constants. If* $K = 1$*, no label is needed; exact channels require at most* $\lceil \log_2 K \rceil$ *labels.*

A backtracking walk locates a plausible block; fresh likelihood-ratio tests verify its boundaries and restart after rejection. Search pays for locating the target, confirmation for making the error probability small, including failed attempts. Adapting noisy search (Feige et al., 1994) gives explicit constants $A_0 = 1200$, $A_1 = 4$ (Appendix D). These rates concern supplied population cells and feasible cuts; TAB instead scores estimated intervals. Fixed-budget and heterogeneous per-cut alternatives appear in Corollary B.3 and proposition D.6.

5.6 Mechanism and outcome-sufficiency robustness

Finally, accurate labels may still support only an approximate mechanism. The next bound separates errors in predicting exposure from failures of exposure to summarize outcomes. For an approximate candidate g°, set $\gamma_{ie} = \mathbb{P}\{g_i^\circ(Z) = e, E_i(Z) \neq e\}$, suppose $\sup_z |Y_i(z) - Y_i^E\{E_i(z)\}| \leq \varepsilon_i$, and let $B_i = |Y_i^E(1) - Y_i^E(0)|$. Here γ_{ie} measures exposure mismatch in arm e under the main design, ε_i bounds departure from outcome sufficiency, and B_i is the unit-level effect magnitude.

Theorem 5.7 (Mechanism and outcome-sufficiency bias)**.** *Under candidate positivity,*

$$|\theta_n(g^\circ; P_0) - \tau_n^E| \leq b_n(g^\circ), \tag{16}$$

$$b_n(g^\circ) = n^{-1} \sum_i \left[2\varepsilon_i + B_i \left\{ \frac{\gamma_{i1}}{\pi_{i1}(g^\circ)} + \frac{\gamma_{i0}}{\pi_{i0}(g^\circ)} \right\} \right]. \tag{17}$$

The first term prices outcome-sufficiency error; the second prices exposure mismatch, amplified by poor overlap. With certified buffers and retention of g°, expanding the candidate intervals by $b_n(g)$ gives robust coverage; a calibration or buffer failure budget ζ changes it to $1-\alpha-\delta-\zeta$ (Corollary C.3). Independent gold-label validation can certify an average omission buffer (Corollary C.4); selected-query transport and outcome-sufficiency bounds require separate evidence.

The theory separates three precision limits: candidate-band width from the main randomization, structural ambiguity reduced by audits, and bias from model approximation. Each calls for a different source of information.

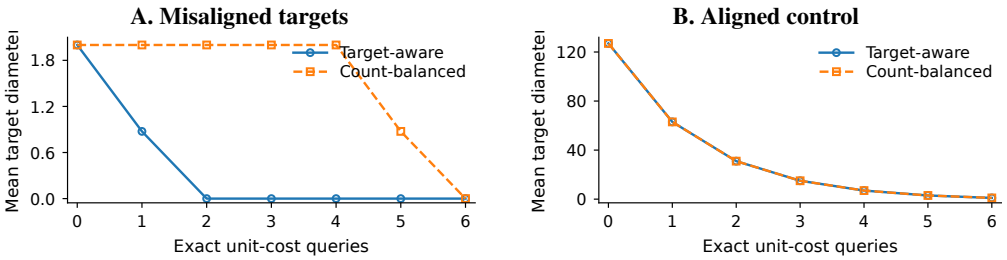


Figure 3: **Target geometry determines useful queries.** Mean target diameter over all 128 truths. Policies differ in A and coincide in B; vertical scales differ. Figure 4 gives reporting and cost diagnostics.

6 Experiments

Three experiment families test retention, acquisition, and real-outcome ambiguity; Appendix F gives designs and full results.

1. Structural validity and target geometry. With fixed target intervals, nominal coverage $1-\delta = .95$, and BSC$(.35)$ labels, union full-path coverage is 0.9766. Selecting one map gives path coverage 0.3899 despite terminal coverage 0.9849. Target and count balancing need 1.875 versus 5.875 exact queries when geometries differ, and 6 each when aligned (Figure 3). Network-embedded ordered search separates search and confirmation costs (Appendix F.2).

2. Network acquisition and non-oracle precision. With $B = 96$ independently randomized 70-node components, modelwise betting reduces terminal width by 28.1% relative to simultaneous betting, with path coverage 0.9917. Count-based querying is narrower at $B = 16$. On one connected cycle, the certified half-width is $\underline{\pi}_g^{-1}\sqrt{\log(2/\alpha)/J}$. Here $J = n/b$ counts randomized blocks. Few blocks can leave the full effect range unresolved (Figure 8). At 30 labels and $n = 4096, b = 1$, target-anytime width is 0.161, versus 0.591 without audits; fixed-horizon controls appear in Appendix F.5. The 70-node oracle benchmark reaches 0.554, versus 35.97 for its analytic replay (Appendix F.3).

At transport rate .30, valid multiplicative correction uses all 80 labels and leaves width 1.603 out of 2; bounded-error betting reaches .294 at 79.43 labels under the same error envelope. Targets are supplied in this structural diagnostic (Appendix F.3.3).

3. Real-outcome boundary and exclusion sensitivity. Weighted cells from 1,975 Twitter accounts (Coppock et al., 2016) give an assigned-opportunity contrast of .0195, but a binary-IV proxy-effect region of $[-.1405, .4823]$, widening to $[-.1473, .5450]$ under 5% exclusion violations. Receipt remains latent. Appendix F.6 specifies the proxy assumptions; Appendix F.7 examines overlap under a published Facebook design.

7 Discussion

Active partial identification asks which measurement would change a causal decision. Main randomization sets candidate precision; exposure labels reduce structural ambiguity. Once all retained intervals lie on the same side of a fixed decision threshold, further mechanism identification is unnecessary for that decision (Corollary B.1). For message delivery, this means deciding whether receipt helps without reconstructing every routing rule. Appendix C.5 relates sample size, tolerance, and budget.

8 Limitations

Applications require feasible probes, conditional sensor/transport calibration, and a justified candidate class and outcome model, or certified approximation buffers. Calibration contributes to total cost. The evidence covers synthetic audits and public-data ambiguity. Precomputed scoring and updates cost $O(|\mathcal{Q}_n|M_n)$ and $O(M_n)$; design integration can dominate. Ordered search requires supplied population cells and feasible cuts. Corollary B.4 gives necessary budget and precision checks.

AI USE STATEMENT

Generative AI was used to aid or polish writing, including clarifying exposition, editing grammar, and preparing LaTeX. The authors have manually reviewed and verified the AI-assisted content. The authors are responsible for the final text, claims, code, and results.

ETHICS STATEMENT

The Twitter study is a secondary analysis of an unrestricted public CC0 replication archive. This work collected no new participant data, initiated no new participant contact or intervention, reports only aggregate quantities, and made no attempt to reidentify individuals. Future bridge experiments involving delivery telemetry would require context-specific review of participant welfare, privacy, platform governance, and the risks of deliberately manipulating network exposure before deployment.

REPRODUCIBILITY STATEMENT

The supplementary artifact contains executable simulations, fixed configurations, observed inputs, audit paths, and independent verification tools for every primary numerical result. The public-data calculations start from the supplied weighted response cells and condition on those estimates. Appendix F specifies the designs, estimands, baselines, and uncertainty summaries. Appendices B–E give the proofs. At the artifact root, `reproduce.sh` runs the experiments and `README.md` maps results to their inputs and checks.

REFERENCES

Shishir Adhikari, Sourav Medya, and Elena Zheleva. Learning exposure mapping functions for inferring heterogeneous peer effects. In *International Conference on Learning Representations*, 2026.

Elisabeth Ailer, Niclas Dern, Jason Hartford, and Niki Kilbertus. Targeted sequential indirect experiment design. In *Advances in Neural Information Processing Systems*, volume 37, pp. 122029–122053. Curran Associates, Inc., 2024. doi: 10.52202/079017-3877.

Sina Akbari, Jalal Etesami, and Negar Kiyavash. Optimal experiment design for causal effect identification. *Journal of Machine Learning Research*, 26(28):1–56, 2025.

Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017. doi: 10.1214/16-AOAS1005.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 519–528, 2012. doi: 10.1145/2187836.2187907.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. doi: 10.1080/01621459.1997.10474074.

Marat V. Burnashev. Data transmission over a discrete channel with feedback. random transmission time. *Problems of Information Transmission*, 12(4):250–265, 1976.

Marat V. Burnashev. Note on the article "data transmission over a discrete channel with feedback. random transmission time". *Problems of Information Transmission*, 13(1):80, 1977.

Cong Cao. Causal inference under interference with learned exposure mappings. *arXiv preprint arXiv:2608.19224*, 2026. doi: 10.48550/arXiv.2608.19224.

Ariel Chao, Donna Spiegelman, Ashley Buchanan, and Laura Forastiere. Estimation and inference for causal spillover effects in egocentric-network randomized trials in the presence of network membership misclassification. *Biostatistics*, 26(1):kxaf009, 2025. doi: 10.1093/biostatistics/kxaf009.

Thomas Cook, Alan Mishler, and Aaditya Ramdas. Semiparametric efficient inference in adaptive experiments. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 1033–1064. PMLR, 2024.

Alexander Coppock, Andrew Guess, and John Ternovski. Replication data for: When treatments are tweets: A network mobilization experiment over Twitter. Harvard Dataverse, doi: 10.7910/DVN/29548, 2015.

Alexander Coppock, Andrew Guess, and John Ternovski. When treatments are tweets: A network mobilization experiment over Twitter. *Political Behavior*, 38(1):105–128, 2016. doi: 10.1007/s11109-015-9308-6.

Abhinandan Dalal, Patrick Blöbaum, Shiva Kasiviswanathan, and Aaditya Ramdas. Anytime-valid inference for double/debiased machine learning of causal parameters. *arXiv preprint arXiv:2408.09598*, 2024. doi: 10.48550/arXiv.2408.09598.

Nicolas Emmenegger, Mojmír Mutný, and Andreas Krause. Likelihood ratio confidence sets for sequential decision making. In *Advances in Neural Information Processing Systems*, volume 36, pp. 26686–26698. Curran Associates, Inc., 2023. doi: 10.52202/075280-1159.

Uriel Feige, Prabhakar Raghavan, David Peleg, and Eli Upfal. Computing with noisy information. *SIAM Journal on Computing*, 23(5):1001–1018, 1994. doi: 10.1137/S0097539791195877.

Chao Gao, Christopher Harshaw, Fredrik Sävje, and Yitan Wang. On the impossibility of specification testing of interference models based on exposure mappings. *arXiv preprint arXiv:2605.09726*, 2026. doi: 10.48550/arXiv.2605.09726.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal Bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems*, volume 23, pp. 766–774. Curran Associates, Inc., 2010.

Yihui He and Eric J. Tchetgen Tchetgen. A general exposure-mapping-agnostic framework for causal inference under interference. *arXiv preprint arXiv:2607.04644*, 2026. doi: 10.48550/arXiv.2607.04644.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Tadao Hoshino and Takahide Yanagi. Conditional randomization tests for the specification of interference structure. *Journal of Applied Econometrics*, 2026. doi: 10.1002/jae.70076.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Martin Huber, Jannis Kueck, and Mara Mattes. Learning and testing exposure mappings of interference using graph convolutional autoencoders. In *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pp. 656–675. PMLR, 2026.

Michael G. Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008. doi: 10.1198/016214508000000292.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. doi: 10.2307/2951620.

Svante Janson. Normal convergence by higher semiinvariants with applications to sums of dependent random variables and random graphs. *The Annals of Probability*, 16(1):305–312, 1988. doi: 10.1214/aop/1176991903.

Shervin Javdani, Yuxin Chen, Amin Karbasi, Andreas Krause, Drew Bagnell, and Siddhartha Srinivasa. Near optimal Bayesian active learning for decision making. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 430–438, 2014.

Jason J. Jones, Robert M. Bond, Eytan Bakshy, Dean Eckles, and James H. Fowler. Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election. *PLOS ONE*, 12(4):e0173851, 2017. doi: 10.1371/journal.pone.0173851.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Charles F. Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013. doi: 10.1111/j.1368-423X.2012.00368.x.

Tobias Maringgele and Jalal Etesami. Optimal experiments for partial causal effect identification. *arXiv preprint arXiv:2605.06993*, 2026. doi: 10.48550/arXiv.2605.06993.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013. doi: 10.1214/13-AOS1144.

Sirin Nitinawarat and Venugopal V. Veeravalli. Controlled sensing for sequential multihypothesis testing with controlled markovian observations and non-uniform control cost. *Sequential Analysis*, 34(1):1–24, 2015. doi: 10.1080/07474946.2014.961864.

Robert D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011. doi: 10.1109/TIT.2011.2169298.

Matvey Ortyashov and AmirEmad Ghassami. A sensitivity analysis framework for causal inference under interference. *arXiv preprint arXiv:2511.21534*, 2025. doi: 10.48550/arXiv.2511.21534.

Fredrik Sävje. Causal inference with misspecified exposure mappings: Separating definitions and assumptions. *Biometrika*, 111(1):1–15, 2024. doi: 10.1093/biomet/asad019.

Maresa Schröder, Miruna Oprescu, Stefan Feuerriegel, and Nathan Kallus. Causal inference on networks under misspecified exposure mappings: A partial identification framework. *arXiv preprint arXiv:2602.03459*, 2026. doi: 10.48550/arXiv.2602.03459.

Difan Song, Simon Mak, and C. F. Jeff Wu. ACE: Active learning for causal inference with expensive experiments. *arXiv preprint arXiv:2306.07480*, 2023. doi: 10.48550/arXiv.2306.07480.

Vydhourie Thiyageswaran, Tyler H. McCormick, and Jennifer Brennan. Data-adaptive exposure thresholds under network interference. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131360–131393. Curran Associates, Inc., 2025. doi: 10.52202/085713-4373.

Supriya Tiwari and Pallavi Basu. Randomization tests for model specification in causal inference under network interference. *arXiv preprint arXiv:2608.22890*, 2026. doi: 10.48550/arXiv.2608.22890.

Christopher Tosh and Sanjoy Dasgupta. Diameter-based active learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3444–3452, 2017.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 329–337, 2013. doi: 10.1145/2487575.2487695.

George Vershinin, Asaf Cohen, and Omer Gurewitz. Active sequential hypothesis testing with non-homogeneous costs. *arXiv preprint arXiv:2509.11632*, 2025. doi: 10.48550/arXiv.2509.11632.

Jean Ville. *Étude Critique de la Notion de Collectif*. Gauthier-Villars, Paris, 1939.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945. doi: 10.1214/aoms/1177731118.

Zichen Wang, Haoyang Hong, Chuanhao Li, Haoxuan Li, Zhiheng Zhang, and Huazheng Wang. Design-based bandits under network interference: Trade-off between regret and statistical inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3982.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Bar Weinstein and Daniel Nevo. Causal inference with misspecified network interference structure. *Biometrics*, 82(1):ujag023, 2026. doi: 10.1093/biomtc/ujag023.

Steven Wilkins-Reeves, Shane Lubold, Arun Chandrasekhar, and Tyler McCormick. Single-network asymptotics for causal inference with partial network data. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 7341–7369. PMLR, 2026.

Anpeng Wu, Haiyi Qiu, Zhengming Chen, Zijian Li, Ruoxuan Xiong, Fei Wu, and Kun Zhang. Causal graph transformer for treatment effect estimation under unknown interference. In *International Conference on Learning Representations*, 2025.

Zhiheng Zhang and Zichen Wang. Online experimental design with estimation-regret trade-off under network interference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0538.

Banghua Zhu, Ziao Wang, Nadim Ghaddar, Jiantao Jiao, and Lele Wang. On the optimal bounds for noisy computing. *arXiv preprint arXiv:2306.11951*, 2023. doi: 10.48550/arXiv.2306.11951.

Hongtao Zhu, Sizhe Zhang, Yang Su, Zhenyu Zhao, and Nan Chen. Integrating active learning in causal inference with interference: A novel approach in online experiments. *arXiv preprint arXiv:2402.12710*, 2024. doi: 10.48550/arXiv.2402.12710.

Appendices

When Randomization Is Not Enough: Auditing Exposure in Network Experiments

Contents

A Notation, Positioning, and Auxiliary Facts . . . **15**
A.1 Feasible probes and external calibration . . . 17
A.2 A necessary transfer diagnostic . . . 18
A.3 Interface to sequential indirect design . . . 18

B Proofs for Identification, Anytime Validity, and Information Limits . . . **20**
B.1 Proof of Theorem 5.1 . . . 20
B.2 Proof of Theorem 5.2 . . . 20
B.3 Proof of Theorem 5.3 . . . 20
B.4 Proof of Theorem 5.4 . . . 21
B.5 Proof of Theorem 5.5 . . . 21
B.6 Ordered target cells and fixed-budget search . . . 22
B.7 A pre-audit feasibility screen . . . 23

C Candidate Inference and Robustness Proofs . . . **23**
C.1 Proof of Proposition C.1 . . . 24
C.2 Proof of Theorem C.2 . . . 24
C.3 Proof of Theorem 5.7 . . . 25
C.4 Selecting a finite class and bounding omission bias . . . 25
C.5 Precision, sample size, and audit budget . . . 26
C.6 Modelwise bands and fixed-cohort calibration . . . 26

D Design Guarantees, Composite Channels, and Failure Semantics . . . **28**
D.1 Version-set semantics . . . 28
D.2 Safe inference under bounded channel error . . . 29
D.3 Sensitivity to bridge-mechanism changes . . . 30
D.4 Sequential pair resolution . . . 31
D.5 Search at capacity scale and sequential boundary confirmation . . . 32
D.6 Hard-decoding contraction benchmark . . . 34
D.7 Aligned geometry, computation, and audit records . . . 35

E Proof of the Dependency-Graph Central Limit Theorem . . . **35**
E.1 Proof of Theorem E.2 . . . 36

F Experiments and Reproducibility . . . **37**
F.1 Structural validity, target geometry, and ordered search . . . 37
F.2 Separating search and confirmation cost in a network construction . . . 38
F.3 End-to-end synthetic-network audit and ablations . . . 39
F.4 Non-oracle inference on one connected network . . . 47
F.5 Matched-budget passive and fixed-horizon comparisons . . . 48
F.6 Real-outcome boundary audit: Twitter . . . 49
F.7 Published-design overlap and data requirements . . . 51

A Notation, Positioning, and Auxiliary Facts

We write $[n] = \{1, \dots, n\}$, $Y_i^{\mathrm{obs}} = Y_i(Z)$, and use natural logarithms. For probabilities $p, q \in [0,1]$, $\mathrm{kl}(p,q) = p\log(p/q) + (1-p)\log\{(1-p)/(1-q)\}$, with the usual limiting conventions; $\mathrm{KL}(P\|Q)$, also written $\mathrm{KL}(P,Q)$, denotes relative entropy. The symbols $x \vee y$, x_+, $\mathbf{1}(A)$, and $\mathrm{hull}(S)$ mean $\max(x,y)$, $\max(x,0)$, the indicator of A, and the smallest interval containing S, respectively. For a common BSC error η, C_η and κ_η are the values of C_q and κ_q in (6) with $\eta_q = \eta$. In finite-class experiments $M = |\mathcal{G}_n|$.

Table 1: Notation index. The audit index t, cohort size B, and audit budget B_{aud} refer to different resources.

Symbol	Meaning	First definition or result
$n, [n], i$	Population size, unit index set, and focal unit	Section 3.1
$\mathcal{Z}_n, Z, z, p_n$	Assignment space, random assignment, its realization, and known design	Section 3.1
$Y_i(z), Y_i^{\mathrm{obs}}, P_0$	Fixed potential outcome, observed outcome, and main-data law	Section 3.1
$E_i(z), Y_i^E(e), \tau_n^E$	Physical event, exposure-level outcome, and physical causal contrast	Section 3.2
$\mathcal{G}_n, M_n, g_n^\star$	Candidate class, size, and exact mechanism	Section 3.3
$\mathfrak{M}_g, \mathfrak{M}_n$	Candidate causal submodel and their union	Section 3.3
$\mathcal{G}_n(P_0), \Theta_{n,0}$	Candidates compatible with the complete main law and their target set	Section 3.3
$\pi_{ie}(g), \mu_{ie}(g; P_0)$	Exposure probability and candidate-weighted mean outcome	Section 3.3
$\theta_n(g; P_0)$	Candidate-implied contrast; physical target under the correct mechanism	(2)
$\mathcal{H}_n^{\mathrm{aug}}, m_{ie}$	External data and fixed outcome adjustment	Section 3.4
$\psi_i(g), \widehat{\tau}_n(g)$	Unit score and its population average estimator	Section 3.4
$\underline{\pi}_g, \underline{\pi}, R_g$	Minimum exposure probabilities and residual bound	Assumption 3.4
$H_n^\cup, \chi_n^\cup, D_n^\cup$	Common score-dependency graph, chromatic number, and maximum degree	Assumption 3.4
$w_g, I_n(g), r_n(g; a)$	Weights, interval, and radius at error a	(24)
$\ell_n(g), u_n(g)$	Lower and upper candidate-interval endpoints	Section 3.4
$\mathcal{Q}_n, q, i(q), z(q)$	Feasible probe menu, probe, receiver, and controlled assignment	Section 3.4
$a_q(g), Q_t, A_t$	Predicted label, query, and noisy label	Section 3.4
$U, \mathcal{F}_{t-1}, t$	Seed, pre-label information, audit time	Section 3.4
K_q, c_q, η_q	Conditional observation channel, charged cost, and BSC error	Assumption 3.5
$H_s, (q_r, b_r), \Theta_{n,s}$	Exact history, query/answer, identified set	(3)
$L_t(g), m_t$	Candidate likelihood and its fixed-weight mixture	Section 4
$\mathsf{E}_t(g), \mathsf{E}_t^{\max}(g)$	Current evidence and running maximum	Section 4
$\mathcal{V}_t(\delta), \mathcal{C}_{n,t}$	Nested retained set and reported causal envelope	(4)–(5)
α, δ, ζ	Main-interval, structural-retention, and external-certification error budgets	Theorem 5.3 and proposition A.3
$\varepsilon, \Delta_n(V), W_n(V)$	Target half-tolerance, population diameter, and finite-data envelope width	Section 3.5
$\mathcal{I}_n(V), V_{q,a}$	Interval hull and exact-answer child set	Section 3.5
$\Gamma_n^\theta, \Gamma_n^I$	Worst-answer population and inferential widths	Sections 5.3 and 4
$N_\varepsilon, D_\varepsilon^\star, \rho$	Target packing, minimum worst-case query depth, and split contraction	Section 5.3
$C_q, \kappa_q, C_\eta, \kappa_\eta$	Channel capacity and pairwise divergence; common-BSC versions	(6)
$\beta_t, r_q(\beta_t), \mathrm{Score}_t(q)$	Pricing level, label-count proxy, query score	Section 4
$\mathfrak{C}, \mathfrak{K}(g,h)$	Best search and pairwise information per cost	(12)
$\mathsf{T}, \mathsf{Cost}_\mathsf{T}$	Stopping time and total charged audit cost	Theorem 5.5
$K, B_1, \dots, B_K$	Target-cell count and ordered blocks	Theorem 5.6
$g^\circ, \gamma_{ie}, \varepsilon_i, B_i, b_n(g)$	Approximate mechanism, mismatch, outcome error, effect magnitude, and bias buffer	Theorem 5.7

Table 1 continued

Symbol	Meaning	First definition or result
$\rho_q^{\mathrm{br}}, \xi_t, \nu_q$	Physical transport bound, channel-law deviation, and total error bound	Theorem D.4 and proposition D.3
$B, N = 70B$	Independent components and total units	Appendix F.3.1
$b, J = n/b, B_{\mathrm{aud}}$	Connected-design block size, number of randomized blocks, and audit budget	Appendix F.4; Algorithm 1

Information sources and antecedents. Partial identification studies the parameters compatible with observed data and maintained assumptions (Manski, 2003). Bridge experiments enlarge the information set by measuring exposure, rather than only restricting the initial ambiguity. Learned exposure mappings and inference with partial network data address uncertainty in the exposure representation and network structure (Cao, 2026; Wilkins-Reeves et al., 2026). Bridge audits add direct measurements of the physical event. Intervention design can minimize identification cost (Akbari et al., 2025) or optimize a partial-identification region under a budget (Maringgele & Etesami, 2026). Active learning can target equivalence classes, decisions, or version-space diameter (Golovin et al., 2010; Javdani et al., 2014; Tosh & Dasgupta, 2017). Generalized binary search balances hypotheses to recover a binary-valued function (Nowak, 2011). The target complexity in Theorem 5.4 instead permits unresolved mechanisms whose causal targets are sufficiently close.

Table 2: Positioning by added information and inferential output. Each row specifies a source of structural information and the object it supports.

Paradigm	Structural information	Output	Distinction
Known-map design inference	Fixed exposure map	Estimator / interval	Recovered when the retained set is a singleton
Fixed-ambiguity analysis	Specified uncertainty region	Bias bound / identified set	Propagates structural uncertainty to the target
Outcome-based map testing	Main assignment and outcomes	Specification decision	Tests outcome sufficiency or nested maps; no new physical-exposure label
Passive validation	Gold-standard and surrogate network measurements	Measurement-error correction	Collects validation data under a fixed protocol
Targeted experiment design	Feasible interventions or tests	Target / decision resolution	General antecedent for width- and decision-aware querying
This paper	Adaptive noisy physical-exposure labels plus main randomization	Anytime causal union interval	Couples active exposure anchoring, randomization bands, and structural e-process retention

Further exposure-representation methods. Bias and sensitivity bounds quantify network or mapping error (Weinstein & Nevo, 2026; Ortyashov & Ghassami, 2025). Mapping-agnostic designs (He & Tchetgen Tchetgen, 2026), data-adaptive thresholds (Thiyageswaran et al., 2025), graph-transformer representations (Wu et al., 2025), and learned exposure functions (Adhikari et al., 2026; Huber et al., 2026) use different structural restrictions or information from the main study. The exposure labels here are new measurements from a feasible intervention protocol.

Candidate-class scope. A finite class specifies the mechanisms against which each label is tested. Threshold grids, hop radii, and a small collection of delivery rules give tractable families; unrestricted truth-table enumeration is exponential. A growing sequence $\mathcal{G}_n$ is fixed independently of each evaluation assignment. With bounded R_g and uniform weights, simultaneous radii vanish when $\chi_n^{\cup} \log(M_n/\alpha)/(n\underline{\pi}^2) \to 0$. Modelwise bands avoid the $\log M_n$ calibration penalty for the physical target, but not enumeration cost or approximation bias. Theorem 5.7 requires a certified approximation; the omitted-truth example below shows why nonrejection is insufficient.

Example A.1 (Nonempty retention under an omitted truth). Let Z_1, Z_2 be independent Bernoulli$(1/2)$ assignments, with physical exposure $E_i(z) = z_1$ and $Y_i^E(e) = e$ for both units. Then $\tau_n^E = 1$. Supply only $h_i(z) = z_2$. The complete main law implies $\theta_n(h; P_0) = \mathbb{E}(Y_i \mid Z_2 = 1) - \mathbb{E}(Y_i \mid Z_2 = 0) = 0$, so the oracle candidate interval $I_n(h) = \{0\}$ covers its own target exactly. For any full-support audit channel, the singleton mixture satisfies $m_t = L_t(h) > 0$ and $\mathsf{E}_t(h) = 1$ on every label path. Hence $\mathcal{V}_t = \{h\}$ forever, although its causal report misses the physical effect. This is failure of candidate completeness, not of candidate-target calibration. The artifact checks all four main assignments and all 2^{12} label paths for a BSC$(.1)$ and repeated query $(1, 0)$.

Table 3: Assumptions, diagnostics, and responses to failure.

Condition	Why it is needed	Diagnostic	Response to failure
Candidate completeness	Ensures one candidate carries the physical target	Rejection can reveal failure; nonrejection is inconclusive	Specify a revised inference protocol or certified approximation bounds
Outcome sufficiency	Gives the exposure contrast its causal meaning	Not testable from exposure labels alone	Richer outcomes, repeated assignments, or sensitivity buffer
Bridge invariance	Transfers audit labels to the main mechanism	Replicated protocols and transport checks	Certified conditional transport bound (Theorem D.4) or the no-audit union
Channel calibration	Makes likelihood evidence valid	Gold-standard calibration and drift monitoring	Bounded-error safe e-process
Positivity	Defines every candidate score and interval	Exact design calculation	Prespecified pruning or target redesign
Predictability	Protects optional-stopping validity	Time-stamped query log	Specify the query before acquiring its label
Common dependence bound	Controls simultaneous randomization bands	Overlap of score-determining assignments	Denser graph bound or alternative design

Example A.2 (One public-health bridge update). Suppose the main randomized message trial leaves three scientifically prespecified delivery mechanisms: direct delivery D, one-hop receipt R, and feed-mediated receipt F. Their simultaneous candidate intervals are $I_n(D) = [-.06, .10]$, $I_n(R) = [.04, .18]$, and $I_n(F) = [.12, .25]$, so the no-audit envelope is $[-.06, .25]$ and still crosses zero. For candidates depending only on the stated delivery neighborhood, a feasible bridge query sets that neighborhood to a pattern for which $(a_q(D), a_q(R), a_q(F)) = (0, 1, 1)$. Take uniform mixture weights, $\delta = .05$, and independent BSC$(.10)$ telemetry. If the first two labels are both one, then

$$(L_2(D), L_2(R), L_2(F)) = (.01, .81, .81), \qquad \mathsf{E}_2(D) = \frac{(.01 + .81 + .81)/3}{.01} = 54\tfrac{1}{3} > 20,$$

whereas the running e-values for R and F never exceed one. Thus $\mathcal{V}_2(.05) = \{R, F\}$ and the reported envelope becomes $[.04, .25]$. If the third label is zero, the likelihoods become $(.009, .081, .081)$, giving $\mathsf{E}_3(D) = 19/3 < 20$. Yet $\mathsf{E}_3^{\max}(D) = 163/3 > 20$, so D remains excluded. The current evidence can fall; the rejection record cannot. Telemetry changes which mechanisms are retained, while the main-trial interval endpoints stay fixed.

A.1 Feasible probes and external calibration

A local intervention is enough when the mechanism is local. For a query at unit i, suppose each candidate has the form $g_i(z) = f_{g,i}(z_{S_q})$ for a known common footprint S_q. The protocol sets $z_{S_q} = x_q$, measures the prespecified physical event at i, and records its sensor output. Every completion $z(q)$ with this restriction has the same prediction $a_q(g) = f_{g,i}(x_q)$, for every g. If the true mechanism is in the class and bridge invariance holds, this also fixes its physical event. Thus a complete assignment in the notation need not be replayed over the whole network. A delivery probe can set the focal receiver's eligible sender pattern, hold the interface and exposure window fixed, and timestamp the resulting receipt event. Carryover and changes in receiver behavior enter the bridge-transport bound. Candidate mechanisms with global dependencies require global control or a separately calibrated randomized-completion law; only implementable protocols enter $\mathcal{Q}_n$.

Calibration precedes outcome-adaptive selection. The charged cost c_q is the resource price assigned to a probe. Its channel requires measurements: a gold-label calibration sample compares the recorded sensor output with independently verified physical receipt. Suppose S prespecified channel strata have stable error probabilities η_s and m_s independent calibration pairs, with empirical errors $\widehat{\eta}_s$. For $\zeta \in (0, 1)$, set

$$\overline{\eta}_s = \min\left\{1, \widehat{\eta}_s + \sqrt{\frac{\log(S/\zeta)}{2m_s}}\right\}.$$

Hoeffding's inequality and a union bound give $\mathbb{P}\{\eta_s \le \overline{\eta}_s \text{ for all } s\} \ge 1 - \zeta$. Using these bounds for future probes requires the conditional error law of each selected query to remain within its stratum bound; an average over the unselected menu does not provide that condition. When a bound is at least $1/2$, its stratum supplies no informative bounded-error sign test.

Proposition A.3 (Coverage with externally estimated channel bounds)**.** *Let independent pre-audit calibration data have a good event of probability at least* $1-\zeta$*. For each calibration realization on that event, suppose the future selected-query error laws satisfy the bounds supplied to Proposition D.3, conditional on the calibration data and the past. If the true mechanism is represented, outcome sufficiency holds, and its main-data interval has marginal coverage at least* $1-\alpha$*, the resulting all-time causal envelope has coverage at least* $1-\alpha-\delta-\zeta$*.*

Proof. For every good calibration realization, the conditional-expectation argument in Proposition D.3 and Ville's inequality bound true-candidate rejection by δ. Integrating over the calibration data adds at most ζ for the bad event. On joint retention and true-candidate band coverage, the envelope contains the physical target at every audit time. A union bound adds the band failure probability α. □

Sensor calibration bounds A relative to the bridge's physical event; transport calibration relates that event to the main mechanism. The latter requires matched protocols or a stated sensitivity range ρ_q^{br}. Under conditionally independent BSC telemetry it gives $\nu_q = \eta_q + (1-2\eta_q)\rho_q^{\mathrm{br}}$. Neither the main outcomes nor unverified sensor labels identify this decomposition. The mismatch and outcome-sufficiency inputs in Theorem 5.7 have distinct validation requirements. The numerical studies evaluate supplied bounds and their cost.

A.2 A NECESSARY TRANSFER DIAGNOSTIC

Fix S assignment–context strata and compare reference and probe protocols with the same distribution of units, assignments, and pre-exposure state. Let p_{sd} be the probability of independently verified receipt in stratum s and protocol $d \in \{0,1\}$. Bridge invariance implies $p_{s0} = p_{s1}$. With prespecified independent sample sizes m_{sd}, empirical means $\widehat{p}_{sd}$, and $\xi \in (0,1)$, define

$$u_{sd} = \sqrt{\frac{\log(4S/\xi)}{2m_{sd}}}, \qquad L_s = \left\{ |\widehat{p}_{s1} - \widehat{p}_{s0}| - u_{s1} - u_{s0} \right\}_+.$$

Simultaneously over strata, $|p_{s1} - p_{s0}| \geq L_s$ with probability at least $1-\xi$: apply two-sided Hoeffding bounds to the $2S$ means and the triangle inequality. Rejecting when any $L_s > 0$ is therefore a level-ξ check of this necessary implication. Matching is essential; differences in context composition can otherwise explain rejection. Repeated looks require a time-uniform replacement for these fixed-sample bounds.

Equal marginal receipt rates do not establish equality of individual mechanisms or a selected-query transport bound: opposite changes can cancel. Gold receipts separate this check from sensor calibration. With only sensor outputs, the comparison additionally requires equal known channels or channel-error allowances. Thus rejection can challenge transfer, while nonrejection leaves unmeasured strata and conditional deviations unresolved.

A.3 INTERFACE TO SEQUENTIAL INDIRECT DESIGN

Ailer et al. (2024) acquire (Z, X, Y) under selected instrument policies and narrow a functional $Q[f]$ of an outcome response. Our audit observes physical exposure while the main outcome sample is fixed. Their kernel criterion and TAB can be compared on a common, richer protocol: repeated randomized encouragements with observed binary receipt $X = E$, bounded additive outcomes $Y = f(E) + U$, $Z \perp U$, and no direct $Z \to Y$ path. Choose $Q[f] = f(1) - f(0)$, a bounded linear functional for the corresponding RKHS; it equals the physical contrast in the constant-effect submodel. Both policies must receive the same initial data, feasible encouragement menu, gold receipt and outcome observations, and charged cost per observation. Discretizing the indirect-design policy to that menu is an explicit adaptation. Compare both at a fixed horizon with the same target, then evaluate any anytime claim under its own valid reporting rule. The exposure-only budget used here instead supports the controls in Appendix F.5.

ACI learns direct and spillover response functions using a supplied neighborhood summary (Zhu et al., 2024); ACE chooses units or treatments to reduce posterior variance (Song et al., 2023). Learned exposure functions can also propose candidates (Adhikari et al., 2026). To enter the present

inference, each proposed map must predict the same prespecified physical event, be frozen before the independent evaluation assignment, and pass design-based positivity screening. Predictive fit and a receipt measurement provide different evidence about that map.

Lemma A.4 (Adaptive likelihood-ratio supermartingale)**.** *Fix $g, h \in \mathcal{G}_n$. Under the law in which h is true,*

$$M_t^{g,h} = L_t(g)/L_t(h)$$

is a nonnegative supermartingale with respect to $\{\mathcal{F}_t\}$; it is a martingale if every $K_q(\cdot \mid e)$ has full support.

Proof. Conditional on $\mathcal{F}_{t-1}$, $Q_t = q$ is fixed. Put $p_g(a \mid q) = K_q\{a \mid a_q(g)\}$. On the support of the true law,

$$\mathbb{E}_h\left[\frac{p_g(A_t \mid q)}{p_h(A_t \mid q)}\middle|\mathcal{F}_{t-1}\right] = \sum_{a:p_h(a|q)>0} p_g(a \mid q) \le 1.$$

The sum equals one under full support. Multiplication by the nonnegative $M_{t-1}^{g,h}$ proves the claim. Histories with $L_t(h) = 0$ are null under h. □

Lemma A.5 (Mixture e-process and nested confidence set)**.** *Under candidate h, $\mathsf{E}_t(h) = \sum_g w_g L_t(g)/L_t(h)$ is a nonnegative supermartingale starting at one. Consequently,*

$$\mathbb{P}_h\{\sup_t \mathsf{E}_t(h) \ge 1/\delta\} \le \delta,$$

and $\mathcal{V}_t(\delta)$ is nested.

Proof. Take the fixed convex combination in Lemma A.4; Ville's inequality gives the crossing probability. The running maximum in (4) is nondecreasing, so membership can only be lost. □

Lemma A.6 (Majority error)**.** *For independent BSC(η) errors $B_1, \dots, B_r$ with $\eta < 1/2$ and odd r,*

$$\mathbb{P}\left(\sum_{s=1}^{r} B_s > r/2\right) \le \exp\{-r(1-2\eta)^2/2\}.$$

Proof. Hoeffding's inequality (Hoeffding, 1963) at deviation $1/2-\eta$ gives $\exp\{-2r(1/2-\eta)^2\}$. □

Lemma A.7 (Variance under a dependency graph)**.** *If centered $X_1, \dots, X_n$ admit a dependency graph of maximum degree D and $\mathrm{Var}(X_i) \le \sigma^2$, then*

$$\mathrm{Var}(\textstyle\sum_i X_i) \le n(D+1)\sigma^2.$$

Proof. Nonadjacent variables have zero covariance. Cauchy–Schwarz bounds each remaining covariance by σ^2, and the graph has at most $nD/2$ edges. □

Lemma A.8 (Hoeffding inequality under graph coloring)**.** *If $X_i \in [a_i, b_i]$ have a dependency graph of chromatic number χ, then for $t > 0$,*

$$\mathbb{P}\left(\sum_i (X_i - \mathbb{E}X_i) \ge t\right) \le \exp\left\{-\frac{2t^2}{\chi \sum_i (b_i - a_i)^2}\right\},$$

and likewise for the lower tail.

Proof. Color the graph by $C_1, \dots, C_\chi$. Variables within each color are jointly independent. Generalized Hölder followed by Hoeffding's lemma yields

$$\log \mathbb{E}\exp\{\lambda \sum_i (X_i - \mathbb{E}X_i)\} \le \chi\lambda^2 \sum_i (b_i - a_i)^2/8.$$

Chernoff optimization proves the upper tail; apply it to $-X_i$ for the lower tail. □

B Proofs for Identification, Anytime Validity, and Information Limits

B.1 Proof of Theorem 5.1

Proof. Let Z_1, Z_2 be independent Bernoulli$(1/2)$. In model M_0, set

$$E_1(z) = g_1^{(0)}(z) = z_2, \qquad E_2(z) = g_2^{(0)}(z) = z_1,$$

and $Y_i^E(0) = 0, Y_i^E(1) = 1$. Then $(Y_1^{\text{obs}}, Y_2^{\text{obs}}) = (Z_2, Z_1)$ and $\tau^{E,(0)} = 1$.

In model M_1, set

$$E_1(z) = g_1^{(1)}(z) = 1 - z_2, \qquad E_2(z) = g_2^{(1)}(z) = 1 - z_1,$$

and $Y_i^E(0) = 1, Y_i^E(1) = 0$. The observed outcome vector is again (Z_2, Z_1) for every assignment, but $\tau^{E,(1)} = -1$. Hence the complete observed-data laws coincide. Any procedure based on that law has the same distribution under the two models and cannot identify two opposite signs uniformly. □

B.2 Proof of Theorem 5.2

Proof. If $m \in \mathfrak{M}_n$ induces P_0, then $m \in \mathfrak{M}_g$ for some $g \in \mathcal{G}_n(P_0)$. Candidate-specific identification gives $\tau_n^E(m) = \theta_n(g; P_0)$, proving the left-hand side of (9) is contained in the right. Conversely, compatibility of $g \in \mathcal{G}_n(P_0)$ supplies an $m_g \in \mathfrak{M}_g$ inducing P_0, and identification gives $\tau_n^E(m_g) = \theta_n(g; P_0)$. Every value on the right is attained. Under Assumption 3.5, exact bridge answers are predictions of the same mechanism and therefore replace $\mathcal{G}_n(P_0)$ by precisely the subset satisfying all observed predictions. □

Corollary B.1 (Sign and point identification from exposure information)**.** *Under Theorem 5.2, let the compatible target set after an exact history be nonempty. Its positive sign is identified if and only if its minimum is positive; the negative sign is identified if and only if its maximum is negative. Point identification holds if and only if it contains one target value, even when several mechanisms remain.*

For the compatible class, let $g \sim h$ when $a_q(g) = a_q(h)$ for every feasible q. A finite exact-audit procedure identifies sign uniformly if and only if every equivalence class has a single value of $\operatorname{sign}\theta_n$*. When this condition holds, querying each member of the finite menu once is sufficient. In finite-sample inference with a nonempty envelope, declaring positive effect only when* $\inf \mathcal{C}_{n,t} > 0$ *(and negative only when* $\sup \mathcal{C}_{n,t} < 0$*) has probability at most $\alpha + \delta$ of ever making a wrong-sign declaration under Theorem 5.3.*

Proof. Sharpness makes every compatible target attainable. All have positive sign exactly when the minimum is positive, and the negative and singleton claims follow in the same way. Exact answers from the full menu reveal precisely the signature class of the true candidate. They identify sign when that class is sign-homogeneous. Conversely, candidates in a common signature class produce identical answers to every adaptive exact-query policy; if their signs differ, their common pre-audit law and identical transcripts preclude uniform sign identification. Finally, on the all-time coverage event the true effect belongs to every reported envelope, so neither wrong-sign declaration can occur. The event's complement has probability at most $\alpha + \delta$. □

B.3 Proof of Theorem 5.3

Proof. Take $h = g_n^\star$. By Lemma A.5 and Ville's inequality,

$$\mathbb{P}_h\{\exists t : \mathsf{E}_t^{\max}(h) \geq 1/\delta\} \leq \delta.$$

Thus $h \in \mathcal{V}_t(\delta)$ for every t outside an event of probability δ.

Let $\mathcal{E}_{\text{band}}$ be the simultaneous event from Theorem C.2; it has probability at least $1 - \alpha$. On $\mathcal{E}_{\text{band}}$ and the retention event,

$$\tau_n^E = \theta_n(h; P_0) \in I_n(h) \subseteq \mathcal{C}_{n,t}$$

for every t; the retention event also makes every $\mathcal{V}_t$ nonempty. A union bound gives (11). Because the inclusion holds simultaneously for all t, evaluating it at an almost surely finite stopping time spends no additional error probability. □

B.4 Proof of Theorem 5.4

Proof. Let S be a 2ε-packing of size $N_\varepsilon(V)$. A leaf with population diameter at most 2ε cannot contain two elements of S. A binary tree of depth d has at most 2^d leaves, so $2^d \geq N_\varepsilon(V)$, proving the lower bound.

If $\Delta_n(V) \leq 2\varepsilon$, the root is already a valid leaf. Otherwise, at each reachable nonterminal node select the guaranteed split. Along a path of length d, induction gives population diameter at most $(1-\rho)^d \Delta_n(V)$. This is at most 2ε once

$$d \geq \frac{\log\{\Delta_n(V)/(2\varepsilon)\}}{-\log(1-\rho)}.$$

Taking the ceiling constructs a valid tree of depth $d_\rho(V)$. □

B.5 Proof of Theorem 5.5

Proof. Let J be uniform on the K hypotheses. Let O contain all pre-audit observations and let U be an independent algorithmic random seed. The common initial law gives $I(J; O, U) = 0$. The query policy and stopping rule are the same measurable functions of these data under every hypothesis.

Pad the stopped history with an absorbing symbol: $X_t = (Q_t, A_t)$ on $\{\mathsf{T} \geq t\}$ and $X_t = \dagger$ otherwise. Put $H^{(m)} = (O, U, X_1, \ldots, X_m)$ and $H^\infty = (O, U, X_1, X_2, \ldots)$. Continuing and choosing a query are measurable from the preceding history. Thus neither operation supplies additional conditional information about J. Conditional on continuing with $Q_t = q$, the input $a_q(g_J)$ is binary and the new label follows K_q. Data processing and the definition of capacity give

$$I(J; A_t \mid H^{(t-1)}, Q_t = q, \mathsf{T} \geq t) \leq C_q \leq \mathfrak{C} c_q.$$

Here the conditional expression is interpreted at each realized history; averaging over histories and applying the chain rule yields

$$I(J; H^{(m)}) \leq \mathfrak{C}\, \mathbb{E}\left[\sum_{t=1}^{\mathsf{T}\wedge m} c_{Q_t}\right].$$

Relative entropy increases to its full value on increasing sigma-fields. Since the costs are positive, monotone convergence on both sides gives

$$I(J; H^\infty) \leq \mathfrak{C}\, \mathbb{E}\mathsf{Cost}_\mathsf{T} = \frac{\mathfrak{C}}{K} \sum_{j=1}^{K} \mathbb{E}_j \mathsf{Cost}_\mathsf{T}.$$

This remains valid when the right-hand side is infinite. The almost surely finite stopped output is measurable from H^∞. Its diameter allows at most one packed target. Decode that index when present, and use any fixed index otherwise. Uniform interval coverage bounds the average decoding error by δ. Let $p_{\mathrm{err}} \leq \delta$ denote this decoding error. Fano's inequality (Tsybakov, 2009) gives

$$I(J; H^\infty) \geq \log K - h(p_{\mathrm{err}}) - p_{\mathrm{err}} \log(K-1) \geq F_K(\delta).$$

The function $h(p) + p\log(K-1)$ is increasing for $0 \leq p \leq 1 - 1/K$, which contains $[0, \delta]$. Also $F_K(\delta) > 0$ for $\delta < 1/2$ and $K \geq 2$, and $F_K(\delta) \geq (1-\delta)\log K - \log 2$. These inequalities prove both bounds in (13).

For a separated pair g_j, g_k, let $P_j^{(m)}, P_k^{(m)}$ be the laws of $H^{(m)}$. They have the same initial-data distribution. At each continuing history the policy selects the same query as a function of that history, so the conditional divergence comes only from the label:

$$\mathrm{KL}(P_j^{(m)}, P_k^{(m)}) = \mathbb{E}_j \sum_{t=1}^{\mathsf{T}\wedge m} \mathrm{KL}\{K_{Q_t}(\cdot \mid a_{Q_t}(g_j)), K_{Q_t}(\cdot \mid a_{Q_t}(g_k))\} \leq \mathfrak{K}(g_j, g_k)\mathbb{E}_j \mathsf{Cost}_{\mathsf{T}\wedge m}.$$

For finite $\mathfrak{K}(g_j, g_k)$ the finite-history conditional divergences are well defined and the equality follows from the KL chain rule. Increasing sigma-fields and positive costs give

$$\mathrm{KL}(P_j^\infty, P_k^\infty) \leq \mathfrak{K}(g_j, g_k)\mathbb{E}_j \mathsf{Cost}_\mathsf{T}.$$

Let D be the event that the reported interval contains θ_j. Its coverage and diameter imply $P_j^\infty(D) \geq 1-\delta$ and $P_k^\infty(D) \leq \delta$. Binary data processing therefore gives

$$\mathfrak{K}(g_j, g_k)\mathbb{E}_j\mathsf{Cost}_\mathsf{T} \geq \mathrm{kl}\{P_j^\infty(D), P_k^\infty(D)\} \geq \mathrm{kl}(1-\delta, \delta).$$

The last inequality follows because binary KL increases in its first argument and decreases in its second on $p \geq 1-\delta > \delta \geq q$. Reverse j, k to obtain the other directional bound. If every query has zero divergence for this pair, every finite transcript has the same law under both hypotheses. So does the infinite padded transcript. It cannot support an event with probabilities at least $1-\delta$ and at most δ, establishing impossibility for $\delta < 1/2$. It applies regardless of the expected cost. □

Informative initial observations. The same argument permits different initial-data laws for K worlds with fixed, pairwise 2ε-separated targets. Let $H_0 = (O, U)$, $I_0 = I(J; H_0)$ for uniform J, and $d_{jk}^0 = \mathrm{KL}(P_j^{H_0}, P_k^{H_0}) < \infty$. Under the remaining channel, coverage, and stopping conditions of Theorem 5.5,

$$\frac{1}{K}\sum_j \mathbb{E}_j\mathsf{Cost}_\mathsf{T} \geq \frac{\{F_K(\delta) - I_0\}_+}{\mathfrak{C}},$$

$$\mathbb{E}_j\mathsf{Cost}_\mathsf{T} \geq \frac{\{\mathrm{kl}(1-\delta, \delta) - d_{jk}^0\}_+}{\mathfrak{K}(g_j, g_k)}.$$

Indeed, the information chain rules now add I_0 and d_{jk}^0 to the respective upper bounds for the stopped transcript; the decoding and data-processing lower bounds are unchanged. Subtract and use nonnegative costs. These refinements can be vacuous when the initial data already separate the targets. Applying them as numerical budget screens requires certified upper bounds on the initial information.

Corollary B.2 (Homogeneous binary-symmetric channels)**.** *Under the hypotheses of Theorem 5.5, let every audit have unit cost and pass through BSC(η) with $0 < \eta < 1/2$. Then*

$$\frac{1}{K}\sum_{j=1}^K \mathbb{E}_j\mathsf{T} \geq \frac{F_K(\delta)}{C_\eta} \geq \frac{\{(1-\delta)\log K - \log 2\}_+}{C_\eta}, \tag{18}$$

and, for every separated pair,

$$\min\{\mathbb{E}_j\mathsf{T}, \mathbb{E}_k\mathsf{T}\} \geq \frac{\mathrm{kl}(1-\delta, \delta)}{\kappa_\eta}. \tag{19}$$

Proof. The capacity is $C_\eta = \log 2 - h(\eta)$. A query with opposite predictions has divergence $(1-2\eta)\log\{(1-\eta)/\eta\} = \kappa_\eta$ in both directions; other queries contribute zero. Also $\mathsf{Cost}_\mathsf{T} = \mathsf{T}$. Substitution in Theorem 5.5 proves both bounds. If a pair has no informative query, the impossibility clause applies. □

B.6 Ordered target cells and fixed-budget search

Corollary B.3 (Certified noisy search for ordered target cells)**.** *Fix $\delta \in (0,1)$ and suppose $g_n^\star \in \mathcal{G}_n$. The population target values are supplied to the procedure. Index the candidates so that $\theta_k := \theta_n(g_k; P_0)$ obeys $\theta_1 \leq \cdots \leq \theta_M$, and suppose every cut $a_q(g_k) = \mathbf{1}\{k \leq q\}$ is feasible. Greedily partition the candidates into maximal consecutive blocks $B_1, \ldots, B_K$ of target diameter at most 2ε; then $K = N_\varepsilon(\mathcal{G}_n)$. If $K = 1$, no query is needed. Otherwise let $d = \lceil \log_2 K \rceil$ and use fresh conditionally memoryless BSC(η) labels. For $0 < \eta < 1/2$, balanced binary search over the blocks, resolving each cut by the majority of*

$$r = \left\lceil \frac{2\log(d/\delta)}{(1-2\eta)^2} \right\rceil_{\mathrm{odd}} \tag{20}$$

labels, returns the block containing $g_n^\star$ with probability at least $1-\delta$ using at most dr labels. Here $\lceil x \rceil_{\mathrm{odd}}$ is the least odd integer not smaller than $\max\{1, x\}$. If $\eta = 0$, at most d labels suffice.

Proof. Start at g_1 and form a maximal consecutive block whose target diameter is at most 2ε; repeat from the first unassigned candidate. Let b_s be the index of the first candidate in block B_s. Maximality of B_s implies $\theta_{b_{s+1}} - \theta_{b_s} > 2\varepsilon$, so $\{g_{b_1}, \dots, g_{b_K}\}$ is a 2ε-packing and $K \le N_\varepsilon(\mathcal{G}_n)$. Conversely, each block has diameter at most 2ε, so a 2ε-packing contains at most one member of each block; hence $N_\varepsilon(\mathcal{G}_n) \le K$. Thus $K = N_\varepsilon(\mathcal{G}_n)$.

If $K = 1$, the sole block already has diameter at most 2ε and no query is required. Hence assume $K \ge 2$ and put $d = \lceil \log_2 K \rceil$. Every boundary between consecutive blocks is an available cut. Balanced binary search on the ordered block indices therefore has depth at most d. At a visited cut and $0 < \eta < 1/2$, use r fresh independent BSC labels and take their majority. By Lemma A.6 and the definition of r,

$$\mathbb{P}(\text{wrong majority at this cut} \mid \text{past}) \le e^{-r(1-2\eta)^2/2} \le \delta/d.$$

There are at most d visited cuts, so a conditional union bound gives probability at most δ that any decision is wrong. On the complementary event, binary search returns the block containing the true candidate; its target hull has diameter at most 2ε. The label count is deterministically at most dr. When $\eta = 0$, each cut is resolved by one label, so at most d labels suffice. □

B.7 A PRE-AUDIT FEASIBILITY SCREEN

Corollary B.4 (Frozen-band and information-cost budget screens)**.** *Fix the main-data intervals and let V^+ be the candidates with nonempty intervals. If $V^+ \neq \varnothing$, define $w_{\mathrm{floor}} = \min_{g \in V^+}\{u_n(g) - \ell_n(g)\}$. Every nonempty envelope formed solely by retaining a subset of these intervals has width at least w_{floor}. Thus $w_{\mathrm{floor}} > 2\varepsilon$ rules out the requested precision from exposure audits alone, irrespective of the labels' accuracy or the audit budget.*

Separately, suppose a supplied common-law witness satisfies Theorem 5.5, with certified pairwise 2ε target separation, $K \ge 2$, and known finite positive information rates. Any uniformly valid procedure that always reports width at most 2ε within deterministic budget B_{aud} must satisfy

$$B_{\mathrm{aud}} \;\ge\; B_{\min} := \max\left\{ \frac{F_K(\delta)}{\mathfrak{C}},\; \max_{g \neq h} \frac{\mathrm{kl}(1-\delta, \delta)}{\mathfrak{K}(g,h)} \right\}.$$

If a target-separated pair has no informative query, no finite budget suffices. These are necessary conditions, not a sufficient spending rule.

Proof. A nonempty retained hull contains at least one of the original nonempty intervals and is at least as wide as that interval. Minimizing over all candidates proves the first claim. A deterministic cost cap bounds each hypothesis-specific expected cost and their uniform average by B_{aud}. Applying both bounds in Theorem 5.5 and taking the maximum gives the second claim. Its zero-information case is the final clause of that theorem. □

The first screen is computable directly from the observed intervals. The second is computable only after a target-separated common-law witness has been justified, as in the constructed ordered-network experiment. Estimated targets alone cannot certify that witness: main outcomes may already distinguish some candidates, and population target order may be unknown. The maximum rather than the sum is required because the lower-bound theorem bounds search and confirmation separately. The constructive ordered-search upper bound, in contrast, is additive. Neither bound prices the cost of obtaining external calibration data unless that cost is included in c_q.

C CANDIDATE INFERENCE AND ROBUSTNESS PROOFS

Proposition C.1 (Candidate expectation and variance)**.** *Under Assumption 3.4 and the external-augmentation construction in Section 3, conditional on the augmentation data,*

$$\mathbb{E}\{\widehat{\tau}_n(g) \mid \mathcal{H}_n^{\mathrm{aug}}\} = \theta_n(g; P_0) \quad \textit{for every } g \in \mathcal{G}_n. \tag{21}$$

For every candidate,

$$\mathrm{Var}\{\widehat{\tau}_n(g) \mid \mathcal{H}_n^{\mathrm{aug}}\} \le \frac{R_g^2(D_n^{\cup} + 1)}{n\underline{\pi}_g^2}. \tag{22}$$

If Assumptions 3.2 and 3.3 also hold, then $\theta_n(g_n^\star; P_0) = \tau_n^E$.

Theorem C.2 (Simultaneous candidate band)**.** *Under Assumption 3.4 and external augmentation, for each g and $x > 0$,*

$$\mathbb{P}\{|\widehat{\tau}_n(g) - \theta_n(g; P_0)| \geq x \mid \mathcal{H}_n^{\mathrm{aug}}\} \leq 2\exp\left\{-\frac{n\underline{\pi}_g^2 x^2}{2\chi_n^{\cup} R_g^2}\right\}. \tag{23}$$

Thus the half-width

$$r_n(g; \alpha w_g) = \frac{R_g}{\underline{\pi}_g}\sqrt{\frac{2\chi_n^{\cup}\log\{2/(\alpha w_g)\}}{n}} \tag{24}$$

gives simultaneous coverage of all candidate-implied targets with probability at least $1-\alpha$. The weights are fixed independently of the evaluation assignment. They may instead be measurable with respect to the independent augmentation data, in which case the argument is conditional on those data.

The common graph controls dependence for every candidate using the same main assignment.

C.1 Proof of Proposition C.1

Proof. Fix $g \in \mathcal{G}_n$ and condition on the external augmentation data. By the definitions of μ_{ie} and π_{ie},

$$\begin{aligned}\mathbb{E}\{\psi_i(g)\} &= m_{i1} - m_{i0} + \{\mu_{i1}(g; P_0) - m_{i1}\} - \{\mu_{i0}(g; P_0) - m_{i0}\} \\ &= \mu_{i1}(g; P_0) - \mu_{i0}(g; P_0).\end{aligned}$$

Averaging proves (21) for every operational candidate and also proves augmentation invariance. For $g = g_n^\star$, Assumption 3.3 and outcome sufficiency imply $\mu_{ie}(g_n^\star; P_0) = Y_i^E(e)$, hence $\theta_n(g_n^\star; P_0) = \tau_n^E$.

For the variance claim, fix any g and define its random part

$$V_i = \frac{\mathbf{1}\{g_i(Z)=1\}}{\pi_{i1}}(Y_i^{\mathrm{obs}} - m_{i1}) - \frac{\mathbf{1}\{g_i(Z)=0\}}{\pi_{i0}}(Y_i^{\mathrm{obs}} - m_{i0}).$$

Only one summand is nonzero, so

$$\mathrm{Var}(V_i) \leq \mathbb{E}V_i^2 \leq R_g^2\left(\frac{1}{\pi_{i1}} + \frac{1}{\pi_{i0}}\right) = \frac{R_g^2}{\pi_{i1}\pi_{i0}} \leq \frac{R_g^2}{\underline{\pi}_g^2}.$$

The constant augmentation term has no variance. Apply Lemma A.7 to the centered scores and divide by n^2 to obtain (22). □

C.2 Proof of Theorem C.2

Proof. Fix g. Across assignments in the support of p_n, the random residual part of $\psi_i(g)$ lies between $-R_g/\underline{\pi}_g$ and $R_g/\underline{\pi}_g$; its range is at most $2R_g/\underline{\pi}_g$. The augmentation difference is constant. Applying Lemma A.8 to the common dependency graph with $t = nx$ gives each one-sided probability at most

$$\exp\left\{-\frac{n\underline{\pi}_g^2 x^2}{2\chi_n^{\cup} R_g^2}\right\}.$$

Adding the tails proves (23). Substitution of (24) makes the failure probability at most αw_g. A union bound over g, using $\sum_g w_g = 1$, proves simultaneous coverage. □

Corollary C.3 (Robust causal envelope)**.** *For $\alpha, \delta \in (0,1)$ with $\alpha + \delta < 1$, fix a candidate g° satisfying $|\theta_n(g^\circ; P_0) - \tau_n^E| \leq b_n(g^\circ)$. Let $\mathcal{R}_t$ contain candidates with certified deterministic bounds $b_n(g)$ and satisfy $\mathbb{P}\{g^\circ \in \mathcal{R}_t\ \forall t\} \geq 1-\delta$. For $\mathcal{R}_t \neq \varnothing$, define $\mathcal{C}_{n,t}^{\mathrm{rob}} = [\min_{g\in\mathcal{R}_t}\{\ell_n(g) - b_n(g)\}, \max_{g\in\mathcal{R}_t}\{u_n(g) + b_n(g)\}]$; set it to $\varnothing$ otherwise. If the candidate bands are simultaneous with probability $1-\alpha$, then $\mathbb{P}\{\tau_n^E \in \mathcal{C}_{n,t}^{\mathrm{rob}}\ \forall t\} \geq 1-\alpha-\delta$. Retention can use* (4) *when g° specifies the actual conditional label law, or Proposition D.3 when its conditional error bound holds. If certified bias bounds have a simultaneous failure probability at most $\zeta \in [0, 1-\alpha-\delta)$, the guarantee becomes $1-\alpha-\delta-\zeta$.*

C.3 Proof of Theorem 5.7

Proof. Write $g = g^\circ$, $G_i = g_i(Z)$, $E_i = E_i(Z)$, and

$$\Delta_i(Z) = Y_i(Z) - Y_i^E\{E_i(Z)\}, \qquad |\Delta_i(Z)| \le \varepsilon_i.$$

By the direct definition in (2), the unit-i contribution to $\theta_n(g; P_0)$ is

$$\frac{\mathbb{E}\{\mathbf{1}(G_i = 1)Y_i(Z)\}}{\pi_{i1}(g)} - \frac{\mathbb{E}\{\mathbf{1}(G_i = 0)Y_i(Z)\}}{\pi_{i0}(g)}.$$

The contribution of Δ_i to these two arms has absolute value at most

$$\frac{\mathbb{E}[\mathbf{1}(G_i = 1)|\Delta_i|]}{\pi_{i1}(g)} + \frac{\mathbb{E}[\mathbf{1}(G_i = 0)|\Delta_i|]}{\pi_{i0}(g)} \le 2\varepsilon_i.$$

For the exposure-level component and $e \in \{0, 1\}$, the candidate-weighted mean agrees with $Y_i^E(e)$ on $\{G_i = e, E_i = e\}$. On $\{G_i = e, E_i \neq e\}$ its discrepancy is B_i, hence

$$\left| \frac{\mathbb{E}[\mathbf{1}(G_i = e)Y_i^E(E_i)]}{\pi_{ie}(g)} - Y_i^E(e) \right| \le B_i \frac{\gamma_{ie}}{\pi_{ie}(g)}.$$

Combining both arms and the outcome residual, then averaging over i, gives (16)–(17).

For Corollary C.3, let $\mathcal{E}_{\text{band}}$ be the simultaneous candidate-band event and $\mathcal{E}_{\text{ret}} = \{g^\circ \in \mathcal{R}_t \ \forall t\}$. By assumption, their failure probabilities are at most α and δ. On $\mathcal{E}_{\text{band}} \cap \mathcal{E}_{\text{ret}}$, every retained set is nonempty and its robust hull contains

$$[\ell_n(g^\circ) - b_n(g^\circ), u_n(g^\circ) + b_n(g^\circ)] \ni \tau_n^E,$$

where the inclusion follows from the candidate band and (16). Thus the asserted simultaneous coverage follows by a union bound, with no independence requirement. If the certified bias bounds are random, intersect also with their simultaneous validity event, whose failure probability is at most ζ. □

C.4 Selecting a finite class and bounding omission bias

Candidate construction can use physical delivery rules or independent training data. Freeze the family before validation and evaluation; assess positivity using the known assignment law. An independent gold-receipt sample can then measure approximation quality even when the exact mechanism is absent. The validation design must match the main assignment law rather than the adaptively selected audit menu.

Corollary C.4 (Gold-label certificate for an approximate candidate)**.** *Let $\zeta \in (0, 1)$ and suppose a fixed class of size M_n has $\pi_{ie}(g) \ge \pi_0 > 0$. Take m independent validation draws with I_j uniform on $[n]$, $Z_j \sim p_n$ independent of I_j, and gold label $E_{I_j}(Z_j)$. The draws are independent of the main evaluation assignment. Define*

$$\widehat{d}(g) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{g_{I_j}(Z_j) \neq E_{I_j}(Z_j)\}, \qquad \overline{d}(g) = \min\left\{1, \widehat{d}(g) + \sqrt{\frac{\log(M_n/\zeta)}{2m}}\right\}.$$

If $B_i \le B_{\max}$ and $n^{-1}\sum_i \varepsilon_i \le \overline{\varepsilon}$, then with probability at least $1 - \zeta$, simultaneously for all g,

$$|\theta_n(g; P_0) - \tau_n^E| \le 2\overline{\varepsilon} + \frac{B_{\max}}{\pi_0}\overline{d}(g).$$

A candidate chosen from these bounds may be used in Corollary C.3 with this additional failure budget, provided its selected-query label law has a separately valid retention guarantee.

Proof. Let $d(g) = n^{-1}\sum_i \mathbb{P}_{p_n}\{g_i(Z) \neq E_i(Z)\}$. The validation indicators have mean $d(g)$. One-sided Hoeffding and a union bound give $d(g) \le \overline{d}(g)$ for every g with probability at least $1 - \zeta$. In Theorem 5.7, $\gamma_{i1} + \gamma_{i0} = \mathbb{P}\{g_i(Z) \neq E_i(Z)\}$ and each denominator is at least π_0, so $b_n(g) \le 2\overline{\varepsilon} + B_{\max}d(g)/\pi_0$. The uniform event permits selection using the validation sample. □

This certificate controls average main-design mismatch, not completeness or the error of a selected bridge query. Correlated validation observations need their own dependence bound. Without gold receipt or a substantive approximation bound, expanding the class offers sensitivity analysis rather than a controlled probability of including the truth.

C.5 Precision, sample size, and audit budget

With uniform weights, $R_g \le R$, and $\underline{\pi}_g \ge \pi_0$, (24) gives $r_{\max} \le (R/\pi_0)\sqrt{2\chi_n^{\cup}\log(2M_n/\alpha)/n}$. On simultaneous candidate coverage, every retained set obeys $W_n(V) \le \Delta_n(V) + 4r_{\max}$: each estimated center is within its radius of its population target. Thus $\Delta_n(V) \le \varepsilon$ and $r_{\max} \le \varepsilon/4$ suffice for $W_n(V) \le 2\varepsilon$. One explicit sufficient main-design requirement is

$$n \ge \frac{32R^2\chi_n^{\cup}}{\pi_0^2\varepsilon^2}\log\frac{2M_n}{\alpha}.$$

The remaining geometric objective is target diameter at most ε, hence tolerance $\varepsilon/2$ in the exact-query problem. In the supplied-order, feasible-cut setting, Theorem 5.6 achieves this population objective at scales $\log N_{\varepsilon/2}/C_\eta$ and $\log(1/\delta)/\kappa_\eta$. Its successful output is a target cell, not the raw-label retained set $\mathcal{V}_t$. For finite-data reporting from that cell, take the hull of its simultaneous candidate bands. On correct cell selection and simultaneous band coverage, this hull covers the physical target and has width at most 2ε when $r_{\max} \le \varepsilon/4$. A union bound gives probability at least $1-\alpha-\delta$. These query scales do not give a contraction rate for the operational TAB score. Certified bias buffers add $2\max_g b_n(g)$ to the width bound. This separates sampling, structural resolution, and approximation rather than treating a larger audit budget as a substitute for main-design overlap.

C.6 Modelwise bands and fixed-cohort calibration

Simultaneous bands cover every candidate-implied target on one event. The causal union needs less: an interval covering the physical target under the fixed mechanism that generates the data. This distinction changes the band calibration, not the structural test.

Proposition C.5 (Modelwise coverage and audit-time retention)**.** *Let $\mathfrak{M} = \bigcup_{g\in\mathcal{G}}\mathfrak{M}_g$ be a class of full causal models, and let $\tau(P)$ be the physical target under $P \in \mathfrak{M}$. Here P specifies the exposure mechanism as well as the observed-data law P_0. Suppose that, for every g and every $P \in \mathfrak{M}_g$, an interval $I_g(O)$ computed from the main data satisfies*

$$P\{\tau(P) \in I_g(O)\} \ge 1-\alpha, \qquad P\{g \in V_t \text{ for all } t\} \ge 1-\delta. \tag{25}$$

Then, with $C_t = \operatorname{hull}\bigcup_{g\in V_t} I_g(O)$ and $\operatorname{hull}\varnothing = \varnothing$,

$$\inf_{P\in\mathfrak{M}} P\{\tau(P) \in C_t \text{ for all } t\} \ge 1-\alpha-\delta. \tag{26}$$

The two events in (25) *need not be independent. Queries and stopping may depend on the main data and the preceding audit history.*

Proof. Fix $P \in \mathfrak{M}$ and choose one index g_P for which $P \in \mathfrak{M}_{g_P}$. This index depends on the data-generating model, not on the realized observations. The band event for g_P and its all-time retention event fail with probabilities at most α and δ. On their intersection,

$$\tau(P) \in I_{g_P}(O) \subseteq C_t \qquad \text{for every } t.$$

The union bound proves the result for this P; taking the infimum proves (26). Empty reports are failures on the same event definition. □

The simultaneous construction in Theorem C.2 remains sufficient and still supports statements about all pseudo-targets, including $W_n(V) \ge \Delta_n(V)$. Marginal error α suffices for Proposition C.5, but does not provide that simultaneous geometric certificate or validate an outcome-selected single-map interval.

Independent components and fixed population size. Consider B independently randomized components with 70 units each, no between-component interference, and one shared mechanism. Let

$$X_{kg} = \frac{1}{70}\sum_{i\in\text{component } k}\psi_i(g), \qquad \theta_N(g) = \frac{1}{B}\sum_{k=1}^{B}\mathbb{E}X_{kg}, \qquad N = 70B.$$

All potential outcomes and augmentations are fixed, or we condition on independent augmentation data. Component means may differ. The cohort size B, candidate class, support bounds, and clipping rule are chosen before the main assignments and outcomes. Under the true mechanism, Proposition C.1 gives $\theta_N(g^\star) = \tau_N^E$.

Lemma C.6 (Fixed-cohort product tests for unequal means)**.** *Let $Z_1, \dots, Z_B$ be independent random variables in $[0,1]$, and set $\bar{\mu} = B^{-1}\sum_k \mathbb{E}Z_k$. For deterministic fractions $0 < \lambda_j < 1$ and weights $v_j > 0$ with $\sum_j v_j = 1$, define, for $0 < m < 1$,*

$$E_B^+(m) = \sum_j v_j \prod_{k=1}^{B}\left(1 + \lambda_j \frac{Z_k - m}{m}\right),$$
$$E_B^-(m) = \sum_j v_j \prod_{k=1}^{B}\left(1 - \lambda_j \frac{Z_k - m}{1-m}\right). \tag{27}$$

Then $\mathbb{E}E_B^+(m) \le 1$ if $\bar{\mu} \le m$, and $\mathbb{E}E_B^-(m) \le 1$ if $\bar{\mu} \ge m$. Inverting both tests at threshold $2/\alpha$ gives an interval containing $\bar{\mu}$ with probability at least $1-\alpha$.

Proof. Every factor is positive, since it is bounded below by $1 - \lambda_j$. For each positive product, independence and the arithmetic–geometric mean inequality give

$$\mathbb{E}\prod_k \left(1 + \lambda_j \frac{Z_k - m}{m}\right) = \prod_k \left(1 + \lambda_j \frac{\mathbb{E}Z_k - m}{m}\right)$$
$$\le \left(1 + \lambda_j \frac{\bar{\mu} - m}{m}\right)^B \le 1 \qquad (\bar{\mu} \le m).$$

For the negative product the same calculation gives the upper bound $\{1 - \lambda_j(\bar{\mu} - m)/(1-m)\}^B \le 1$ when $\bar{\mu} \ge m$. Summing with the fixed weights proves both expectation bounds. At an interior true mean $m = \bar{\mu}$, Markov's inequality bounds each rejection probability by $\alpha/2$.

For fixed observations, each positive factor is nonincreasing in m and each negative factor is nondecreasing. Their joint nonrejection set is therefore an interval. At the sample mean, the average factor in each product equals one, so the arithmetic–geometric mean inequality also shows that both observed capitals are at most one; the interval is nonempty. We take the closed hull of the accepted set, which can only increase coverage. When $\bar{\mu} = 0$, boundedness implies $Z_k = 0$ almost surely for every k, and the interval includes zero; the case $\bar{\mu} = 1$ is symmetric. For arbitrary data the endpoint zero is retained exactly when all observations are zero, and endpoint one exactly when all are one. Combining the two tail bounds proves coverage. □

This is a fixed-cohort use of bounded-mean betting (Waudby-Smith & Ramdas, 2024). The arithmetic–geometric mean step uses a fixed fraction within each product and permits unequal component means at fixed B. Optional stopping applies to the subsequent audit index t through Proposition C.5.

Lemma C.7 (Design-certified clipping bias)**.** *For component k and candidate g, let s index the finite assignment states with known probabilities $p_k(s)$. Suppose the design and known outcome support give deterministic bounds $\ell_{kg}(s) \le X_{kg}(s) \le u_{kg}(s)$. Fix $a_g < b_g$, set $C_g(x) = \max\{a_g, \min(b_g, x)\}$ and $d_g(x) = x - C_g(x)$, and define*

$$\bar{\beta}_g^- = \frac{1}{B}\sum_{k,s} p_k(s) d_g\{\ell_{kg}(s)\}, \qquad \bar{\beta}_g^+ = \frac{1}{B}\sum_{k,s} p_k(s) d_g\{u_{kg}(s)\}. \tag{28}$$

Then

$$\bar{\beta}_g^- \le \theta_N(g) - \frac{1}{B}\sum_k \mathbb{E}C_g(X_{kg}) \le \bar{\beta}_g^+.$$

Consequently, if $[L_g, U_g]$ covers the average clipped mean at level $1-\alpha$, then

$$I_g = [L_g + \bar{\beta}_g^-,\ U_g + \bar{\beta}_g^+] \cap \mathcal{T} \tag{29}$$

covers $\theta_N(g)$ at that level for any deterministic interval $\mathcal{T}$ known to contain this target.

Proof. The function d_g is nondecreasing: its slopes are one below a_g, zero inside the clip, and one above b_g. Hence

$$d_g\{\ell_{kg}(s)\} \le X_{kg}(s) - C_g\{X_{kg}(s)\} \le d_g\{u_{kg}(s)\}.$$

Taking the known-design expectation and averaging over components proves the bias bounds. On the event that the clipped mean is in $[L_g, U_g]$, adding these bounds encloses $\theta_N(g)$. Intersecting with a known target range cannot remove the true target on this event. □

For known outcome support $[y_-, y_+]$, design-only score bounds follow directly from the augmented estimator. At assignment s, write

$$h_{ig}(s) = \frac{\mathbf{1}\{g_i(s) = 1\}}{\pi_{i1}(g)} - \frac{\mathbf{1}\{g_i(s) = 0\}}{\pi_{i0}(g)},$$
$$d_{ig}(s) = m_{i1} - m_{i0} - \frac{\mathbf{1}\{g_i(s) = 1\}m_{i1}}{\pi_{i1}(g)} + \frac{\mathbf{1}\{g_i(s) = 0\}m_{i0}}{\pi_{i0}(g)}.$$

Then $X_{kg} = 70^{-1}\sum_i\{d_{ig} + h_{ig}Y_i^{\mathrm{obs}}\}$, and valid bounds are

$$\ell_{kg}(s) = \frac{1}{70}\sum_i \left[d_{ig}(s) + \min\{h_{ig}(s)y_-, h_{ig}(s)y_+\}\right],$$
$$u_{kg}(s) = \frac{1}{70}\sum_i \left[d_{ig}(s) + \max\{h_{ig}(s)y_-, h_{ig}(s)y_+\}\right].$$

These extrema use the assignment law, augmentation, and declared outcome range, not the true candidate or unrealized outcomes. In the experiments, $[y_-, y_+] = [-.1, .8]$, so every candidate conditional-mean contrast lies in $\mathcal{T} = [-.9, .9]$.

Apply Lemma C.6 to $Z_{kg} = \{C_g(X_{kg}) - a_g\}/(b_g - a_g)$, rescale, and then use Lemma C.7. This gives the marginal bands in Proposition C.5. Replacing α by α/M and taking a union bound gives the simultaneous-band ablation without changing the clipping points.

Numerical specification. The product mixture uses equal weights on fractions

$$\{1/64, 1/32, 1/16, 1/8, 1/4, 1/2, .70, .85, .95, .99\}.$$

Monotonic roots are bracketed about the observed mean and computed by 43 outward-rounded bisection steps. Clip centers are the design-only augmentation contrast. Half-widths are selected from $\{.1, .15, .25, .35, .5, .75, 1, 1.5, 2, 3, 5, 10, 100\}$ by minimizing the certified bias width plus the betting interval width of a hypothetical constant-score vector at the center. This is a prespecified design proxy. Selection uses only the fixed cohort size and design. The executable experiment compares simultaneous and marginal betting with controlled clipping rules; the aggregate comparison series also includes an empirical-Bernstein reference with its specified design-only proxy.

D Design Guarantees, Composite Channels, and Failure Semantics

The input bands may use simultaneous calibration or the modelwise calibration of Proposition C.5; the stated coverage guarantee determines which is required. Here B_{aud} is an audit budget, not the independent-component cohort size B in Appendix F.3.1.

D.1 Version-set semantics

Proposition D.1 (Nesting and visible model failure)**.** *The sequence $\mathcal{V}_t(\delta)$ is nested and may be empty, but under the exact candidate/channel model*

$$\mathbb{P}\{\mathcal{V}_t(\delta) \neq \varnothing \text{ for every } t\} \geq 1 - \delta.$$

Proof. Every running maximum is nondecreasing. On the retention event in (10), $g_n^\star$ belongs to every version set, so no set is empty. Its complement has probability at most δ. □

Proposition D.2 (Robust-union and oracle limits)**.** *At $t = 0$, Algorithm 1 reports the hull over all candidates. For a singleton retained set it reports the corresponding candidate interval.*

Proof. At $t = 0$, every likelihood and e-process equals one, so $\mathcal{V}_0 = \mathcal{G}_n$. Both claims follow by substitution in (5). □

Propositions D.1 and D.2 formalize the two visible boundary cases used by Algorithm 1: persistent nonemptiness on the retention event and reduction to the unaudited or singleton hull.

Algorithm 1 Nested Target-Aware Bridge Auditing (TAB)

Require: $\mathcal{G}_n, \mathcal{Q}_n$, calibrated bands $I_n(g)$, BSC channels K_q, costs c_q, weights w_g, $\alpha, \delta, \varepsilon$, sequence $\{\beta_s\}_{s\geq 1} \subset (0,1)$, budget B_{aud}, optional decision threshold λ.

Initialize $t = 0$, $L(g) = 1$, $\mathsf{E}^{\max}(g) = 1$, $\mathcal{V} = \mathcal{G}_n$, and spent cost $b = 0$; set $\mathcal{C} = \mathcal{C}_{n,0}$ from (5) and report it.
while true **do**
  if $\mathcal{V} = \varnothing$ **then**
    return MODEL FAILURE; suspend causal reporting.
  end if
  if $W_n(\mathcal{V}) \leq 2\varepsilon$ **then**
    return PRECISION.
  end if
  if λ is supplied and ($\inf \mathcal{C} > \lambda$ or $\sup \mathcal{C} < \lambda$) **then**
    return DECISION.
  end if
  Let $\mathcal{Q}_b = \{q \in \mathcal{Q}_n : c_q \leq B_{\mathrm{aud}} - b\}$.
  if $\mathcal{Q}_b = \varnothing$ **then**
    return BUDGET.
  end if
  Choose $q \in \arg\max_{q \in \mathcal{Q}_b} \mathrm{Score}_t(q)$ using prespecified ties.
  if $\mathrm{Score}_t(q) \leq 0$ **then**
    return ZERO ONE-STEP GAIN.
  end if
  Set $r = \min\{r_q(\beta_{t+1}), \lfloor (B_{\mathrm{aud}} - b)/c_q \rfloor\} \geq 1$.
  for $j = 1, \ldots, r$ **do**
    Set $t \leftarrow t + 1$; execute q; observe A_t; update every $L(g)$, the mixture m_t, $\mathsf{E}^{\max}(g)$, and $\mathcal{V}$; set $b \leftarrow b + c_q$.
    Report the audit record.
    if $\mathcal{V} = \varnothing$ **then**
      return MODEL FAILURE; suspend causal reporting.
    end if
    Set $\mathcal{C} = \mathcal{C}_{n,t}$ and report it; recheck all nonfailure stopping rules.
    if a nonfailure stopping rule now holds **then**
      return its corresponding output.
    end if
  end for
end while

D.2 SAFE INFERENCE UNDER BOUNDED CHANNEL ERROR

Suppose an exact likelihood is unavailable, but candidate g entails the predictable bound

$$\mathbb{P}\{A_t \neq a_{Q_t}(g) \mid \mathcal{F}_{t-1}\} \leq \nu_{Q_t}, \qquad 0 \leq \nu_{Q_t} < 1/2.$$

For predictable finite $\lambda_t(g) \geq 0$ satisfying $\lambda_t(g)\nu_{Q_t} \leq 1$, define

$$M_t(g) = \prod_{s=1}^{t} \left[1 + \lambda_s(g)\{\mathbf{1}(A_s \neq a_{Q_s}(g)) - \nu_{Q_s}\}\right]. \tag{30}$$

The conditional bound determines a valid betting factor without an exact likelihood.

Proposition D.3 (Safe structural confidence set)**.** *If g° satisfies the conditional error bound at every time, then $M_t(g^\circ)$ is a nonnegative supermartingale and*

$$\mathcal{V}_t^{\mathrm{safe}}(\delta) = \{g : \max_{s \leq t} M_s(g) < 1/\delta\}$$

retains g° for every t with probability at least $1 - \delta$.

Proof. Each factor is nonnegative because its minimum is $1 - \lambda_t \nu_{Q_t} \geq 0$. Its conditional expectation under g° is at most one. The product is therefore a nonnegative supermartingale starting at one, and Ville's inequality proves the claim. □

Average mismatch under the main randomization is not enough for this result: the bound must hold conditionally along the adaptively selected bridge path.

D.3 Sensitivity to bridge-mechanism changes

Bridge protocols can change physical receipt even when the sensor remains calibrated. We hold the main experiment, candidate class, and causal target fixed and quantify deviations between main-experiment exposure and bridge exposure.

For candidate h, write $p_{t,h}(a) = K_{Q_t}\{a \mid a_{Q_t}(h)\}$ for the nominal conditional label law and $\widetilde{p}_{t,h}$ for the actual law under the bridge protocol. Both are evaluated at the actual past history. On a finite alphabet, define $\mathrm{TV}(p,q) = \frac{1}{2}\sum_a |p(a) - q(a)|$.

Theorem D.4 (Bridge-transport sensitivity)**.** *Suppose the nominal channels have full support. For each h, let $\xi_t(h) \geq 0$ be predictable, and assume the true candidate $g^\star$ satisfies*

$$\mathrm{TV}(\widetilde{p}_{t,g^\star}, p_{t,g^\star}) \leq \xi_t(g^\star) \quad \textit{conditionally at every audit time.}$$

Put $p_{t,\min}(h) = \min_a p_{t,h}(a) > 0$ and

$$D_t(h) = \prod_{s=1}^{t}\left(1 + \frac{\xi_s(h)}{p_{s,\min}(h)}\right), \qquad \mathsf{E}_t^{\mathrm{tr}}(h) = \frac{\mathsf{E}_t(h)}{D_t(h)}. \tag{31}$$

Then $\mathsf{E}_t^{\mathrm{tr}}(g^\star)$ is a nonnegative supermartingale starting at one. The nested set $\mathcal{V}_t^{\mathrm{tr}} = \{h : \max_{s\leq t}\mathsf{E}_s^{\mathrm{tr}}(h) < 1/\delta\}$ retains $g^\star$ for all t with probability at least $1-\delta$. Combining this set with true-candidate marginal coverage $1-\alpha$ gives all-time physical-target coverage at least $1-\alpha-\delta$, under main-model exactness and outcome sufficiency.

For the uncorrected rule, suppose instead that auditing stops by a fixed horizon H and the deviations are bounded by deterministic $\bar{\xi}_1, \ldots, \bar{\xi}_H$. Its all-time coverage through H is at least

$$\left[1 - \alpha - \delta - \min\left\{1, \sum_{t=1}^{H}\bar{\xi}_t\right\}\right]_+. \tag{32}$$

The finite-horizon statement compares the same predictable policy and initial information law under actual and nominal bridge kernels.

Proof. Let $\overline{p}_t(a) = m_{t-1}^{-1}\sum_g w_g L_{t-1}(g)p_{t,g}(a)$. Full support makes all denominators positive, and $\overline{p}_t$ is a predictable probability distribution. The one-step factor of $\mathsf{E}_t(h)$ is $R_t(h,a) = \overline{p}_t(a)/p_{t,h}(a)$, whose nominal expectation is one. Total variation bounds each coordinate difference, so under the true candidate

$$\widetilde{p}_{t,h}(a) \leq p_{t,h}(a) + \xi_t(h) \leq \left(1 + \frac{\xi_t(h)}{p_{t,\min}(h)}\right)p_{t,h}(a).$$

Multiplying by the nonnegative $R_t(h,a)$ and summing gives

$$\mathbb{E}_h\{R_t(h,A_t) \mid \mathcal{F}_{t-1}\} \leq 1 + \xi_t(h)/p_{t,\min}(h).$$

Dividing each factor by this predictable upper bound proves the supermartingale claim. Ville's inequality gives structural retention; a union bound with failure of the true-candidate interval gives causal coverage. No simultaneous event over the other candidate targets is needed for this last step.

For the uncorrected rule, couple the actual and nominal initial information identically, including the policy's random seed. Until the first differing label, their selected queries coincide. A maximal coupling of their finite conditional label laws differs at step t with probability at most $\bar{\xi}_t$, conditional on the preceding histories agreeing. Thus the probability of any divergence by H is at most $\min\{1, \sum_{t\leq H}\bar{\xi}_t\}$. Every reported set and stopping decision coincides on the complementary event. Apply the nominal coverage bound and subtract the coupling failure probability to obtain (32). After stopping, an identical absorbing record may be used in both processes, so the argument also covers stopping before H. □

Physical transport followed by a calibrated sensor. Let E_t^{br} be the bridge's physical event and suppose

$$\mathbb{P}\{E_t^{\mathrm{br}} \neq a_{Q_t}(g^\star) \mid \mathcal{F}_{t-1}, Q_t\} \leq \rho_{Q_t}^{\mathrm{br}}.$$

Let the recorded label be $A_t = E_t^{\mathrm{br}} \oplus N_t$, where $\oplus$ is addition modulo two and N_t is independent BSC sensor noise conditionally on the past and the physical event, with error $0 < \eta_{Q_t} < 1/2$. Writing r_t for the actual conditional transport-disagreement probability, the total error relative to the main mechanism is

$$\mathbb{P}\{A_t \neq a_{Q_t}(g^\star) \mid \mathcal{F}_{t-1}, Q_t\} = \eta_{Q_t} + (1 - 2\eta_{Q_t})r_t.$$

Hence $\xi_t = (1 - 2\eta_{Q_t})\rho_{Q_t}^{\mathrm{br}}$ suffices in (31), with $p_{t,\min} = \eta_{Q_t}$. Alternatively, when $\nu_t = \eta_{Q_t} + (1 - 2\eta_{Q_t})\rho_{Q_t}^{\mathrm{br}} < 1/2$, replace ν_{Q_t} in (30) by this predictable ν_t; the same conditional-expectation proof gives safe retention without a complete transport law. At $\rho^{\mathrm{br}} = 0$, the likelihood correction reduces exactly to the nominal rule.

The transport bounds are sensitivity inputs requiring external calibration or substantive justification. Audit labels alone confound changed physical receipt with sensor error. A disagreement rate averaged over unselected queries does not imply the conditional bound, and a persistent positive bound can make (32) uninformative as H grows.

D.4 Sequential pair resolution

We use the sequential probability ratio test of Wald (1945), with explicit error and overshoot bounds below.

For BSC(η) with $0 < \eta < 1/2$, put $\ell_\eta = \log\{(1 - \eta)/\eta\}$ and $S_t = \log\{L_t(1)/L_t(0)\}$ for two opposite predictions. Stop at $T = \inf\{t : |S_t| \geq b\}$ and select its sign.

Proposition D.5 (Sequential pair resolution)**.** *For $\beta \in (0, 1)$ and $b = \log(1/\beta)$, the wrong-boundary probability is at most β, and*

$$\max_{e \in \{0,1\}} \mathbb{E}_e T \leq \frac{b + \ell_\eta}{\kappa_\eta}.$$

Proof. Under truth $e \in \{0, 1\}$, orient the log-likelihood ratio as $S_t^{(e)} = \log\{L_t(e)/L_t(1 - e)\} = (2e - 1)S_t$. The likelihood ratio for the other prediction is a nonnegative martingale. Crossing the wrong boundary requires it to reach e^b, so Ville's inequality bounds the probability by $e^{-b} = \beta$. The process $S_t^{(e)}$ has i.i.d. increments $\pm\ell_\eta$ with mean $\kappa_\eta > 0$. For $T_m = T \wedge m$, optional stopping for the bounded time T_m gives $\kappa_\eta \mathbb{E}_e T_m = \mathbb{E}_e S_{T_m}^{(e)} \leq b + \ell_\eta$, because before stopping $S_t^{(e)} < b$ and the upward overshoot is at most one increment. Monotone convergence yields $\mathbb{E}_e T \leq (b + \ell_\eta)/\kappa_\eta < \infty$ for both e, proving the claim. □

Equation (7) uses the leading expected-length term of Proposition D.5 as a batch-price proxy. Fixed-sample error control instead uses the majority count in Corollary B.3.

Proposition D.6 (Expected cost of ordered sequential resolution)**.** *Use the ordered blocks of Corollary B.3, with $K \geq 2$ and $d = \lceil \log_2 K \rceil$. At each visited cut v, choose one available BSC channel with error $0 < \eta_v < 1/2$ and cost c_v before collecting that cut's labels. Keep that channel fixed until its log-likelihood ratio exits $(-b, b)$, with $b = \log(d/\delta)$. Suppose fresh labels obey the calibrated channel condition. If*

$$A_\star = \sup_{v,h} \frac{c_{v,h}\{b + \ell_{\eta_{v,h}}\}}{\kappa_{\eta_{v,h}}} < \infty,$$

where the supremum ranges over all cuts and channels the policy can select, then the procedure returns the true block with probability at least $1 - \delta$, stops almost surely, and satisfies $\sup_g \mathbb{E}_g \mathsf{Cost}_\mathsf{T} \leq dA_\star$. When every cut offers the same finite channel menu $\mathcal{H}$, choosing a minimizer at every cut gives

$$A_\star = \min_{h \in \mathcal{H}} \frac{c_h\{\log(d/\delta) + \ell_{\eta_h}\}}{\kappa_{\eta_h}}.$$

Proof. At any visited cut, conditional on its starting history, the true noiseless response is fixed. Proposition D.5, applied conditionally, bounds its error probability by δ/d and its expected cost by $A_\star$. This remains true on histories following an earlier wrong cut: the current physical response is still one of the two possible responses. Balanced search visits at most d cuts. A conditional union bound gives probability at most δ of any wrong response. On the complementary event all branch choices are correct and the returned block contains the true candidate. Let C_s be the cost of the sth cut, or zero if fewer cuts are visited. The conditional expected-cost bound and the tower property give $\mathbb{E}_g C_s \le A_\star$. Therefore $\mathbb{E}_g \mathsf{Cost}_\mathsf{T} = \sum_{s=1}^{d} \mathbb{E}_g C_s \le dA_\star < \infty$, and each of the finitely many conditional tests terminates almost surely by Proposition D.5. Choosing a minimizing channel in the common finite menu gives the last display. Each conditional resolver is almost surely finite; at most d such resolvers are visited. □

D.5 Search at capacity scale and sequential boundary confirmation

We prove Theorem 5.6 by combining a constant-error search with repeated boundary confirmation. The search is a version of the backtracking comparison tree of Feige et al. (1994). Expected-query bounds for noisy search also attain the noise and error scales (Zhu et al., 2023); here supplied target cells and feasible cuts connect those scales to causal auditing. Its target cells are supplied population objects, as in Corollary B.3; every query below is a cut between consecutive cells. Fix $K \ge 2$, let $d = \lceil \log_2 K \rceil$, and write

$$s = 1 - 2\eta, \qquad r_0 = \left\lceil \frac{2 \log 12}{s^2} \right\rceil_{\text{odd}}, \qquad L = 16d.$$

By Lemma A.6, the majority of r_0 fresh labels has error at most $1/12$ for any cut, conditionally on all preceding observations.

The search walk. Construct a balanced binary tree whose leaves are the K cells in order, with depth at most d. Attach an infinite one-child chain to every leaf. Only L transitions will be used, so these chains need not be stored beyond depth L. Each node inherits a consecutive cell interval $[l, r]$; chain nodes inherit the singleton interval of their leaf. At a node, check membership in $[l, r]$ using the two exterior cuts: the lower cut, when $l > 1$, must answer zero; the upper cut, when $r < K$, must answer one. Each check uses a fresh r_0-label majority. If a check fails, move to the parent. Otherwise, at an internal tree node use another majority to choose the child at its split; at a leaf or chain node move to its sole child. Missing exterior cuts are accepted without measurement. The root therefore always moves to a child. Each transition uses at most $3r_0$ labels.

After L transitions, propose the cell of the current leaf or chain node. At an internal node, propose any predetermined cell. This completes one search call; it uses at most $48dr_0$ labels.

Lemma D.7 (Finite-time backtracking search)**.** *Under fresh BSC(η) labels, one search call proposes the true cell with probability at least $1 - \exp(-49d/32) > 3/4$, conditionally on the history before the call.*

Proof. Fix the true cell and its infinite root-to-chain ray. At a node v on this ray, set $\phi(v) = \mathrm{depth}(v)$. Off the ray, let u be the last common ancestor of v and the ray and set

$$\phi(v) = 2\,\mathrm{depth}(u) - \mathrm{depth}(v).$$

Moving one edge changes ϕ by either $+1$ or -1. If all majorities in a transition are correct, the walk moves toward the true ray when off it, and forward when on it. Thus ϕ increases by one. This also covers visits to an incorrect leaf or its chain: at least one exterior cut rejects membership, so the correct move is to the parent.

Let G_t indicate that all majorities in transition t are correct. There are at most three such tests, each conditionally wrong with probability at most $1/12$. Hence $\mathbb{E}(G_t \mid \text{past}) \ge 3/4$, and

$$\phi(v_L) \ge 2 \sum_{t=1}^{L} G_t - L.$$

Put $D_t = G_t - \mathbb{E}(G_t \mid \text{past})$. Conditional Hoeffding's lemma gives $\mathbb{E}\{\exp(-\lambda D_t) \mid \text{past}\} \le \exp(\lambda^2/8)$. Iteration and Markov's inequality, optimized at $\lambda = 4a/L$, give

$$\mathbb{P}\left\{\sum_{t=1}^{L} G_t < 3L/4 - a\right\} \le \exp(-2a^2/L), \qquad a > 0.$$

Take $a = L/4 - d/2 = 7d/2$ and $L = 16d$. Then

$$\mathbb{P}\{\phi(v_L) < d\} \le \exp(-49d/32) < 1/4.$$

Every off-ray node has potential at most $d-1$, and every internal node on the ray has depth at most $d-1$. Therefore $\phi(v_L) \ge d$ places the walk at the true leaf or on its chain. The proposed cell is then correct. The argument is conditional on an arbitrary initial history and uses only fresh channel labels. □

Lemma D.8 (Asymmetric boundary confirmation)**.** *For a proposed noiseless answer e, let $S_t = \log\{L_t(e)/L_t(1-e)\}$ and stop at $T_{a,b} = \inf\{t : S_t \ge a \text{ or } S_t \le -b\}$, with $a, b > 0$. Accept at the upper boundary and reject at the lower. Under BSC(η),*

$$\mathbb{P}_{1-e}(\textit{accept}) \le e^{-a}, \qquad \mathbb{P}_e(\textit{reject}) \le e^{-b},$$

$$\mathbb{E}_e T_{a,b} \le \frac{a + \ell_\eta}{\kappa_\eta}, \qquad \mathbb{E}_{1-e} T_{a,b} \le \frac{b + \ell_\eta}{\kappa_\eta},$$

where $\ell_\eta = \log\{(1-\eta)/\eta\}$. In particular, the test stops almost surely under either answer.

Proof. Under answer $1-e$, $\exp S_t$ is a mean-one nonnegative martingale; under answer e, so is $\exp(-S_t)$. Ville's inequality gives the two error bounds. Under answer e, S_t has independent increments $\pm\ell_\eta$ with mean κ_η. At the bounded time $T_{a,b} \wedge m$, the martingale identity gives

$$\kappa_\eta \mathbb{E}_e(T_{a,b} \wedge m) = \mathbb{E}_e S_{T_{a,b} \wedge m} \le a + \ell_\eta.$$

The inequality uses the one-increment overshoot at the upper boundary; it also holds before either boundary is reached. Monotone convergence proves integrability and the first expected-length bound. Under answer $1-e$, apply the same argument to $-S_t$, whose upper boundary is b, to prove the second bound. Each finite expectation implies almost sure termination. All statements remain valid conditional on the history before a fresh test. □

Restart and confirmation rule. At attempt $j = 1, 2, \ldots$, run a fresh search call. For its proposed cell, check each available exterior boundary using Lemma D.8, oriented toward the answer asserting membership in that cell. Use

$$a_j = \log \frac{2j(j+1)}{\delta}, \qquad b = \log 8.$$

Reject the proposal and start the next attempt as soon as a boundary rejects. If all available boundaries accept, return the proposed cell's target hull. Each attempt starts from the root and uses fresh labels for both search and confirmation. This reset is part of the procedure.

Proof of Theorem 5.6. A false proposal has at least one exterior boundary whose asserted answer is false. Conditional on the proposed cell and all preceding observations, a false acceptance at any tested boundary has probability at most $e^{-a_j} = \delta/\{2j(j+1)\}$. A union bound over at most two boundaries and all attempts yields

$$\mathbb{P}(\text{a false cell is returned}) \le \sum_{j=1}^{\infty} \frac{\delta}{j(j+1)} = \delta.$$

This argument allows early rejection and random numbers of attempts.

By Lemma D.7, the search proposes the true cell with conditional probability at least $3/4$. Given a true proposal, each available boundary rejects with probability at most $e^{-b} = 1/8$. Their joint acceptance probability is at least $3/4$. Thus the conditional probability of stopping with the correct cell in an attempt is at least

$$p_\star = (3/4)^2 = 9/16.$$

If $J_\star$ is the number of attempts, induction gives $\mathbb{P}(J_\star \geq j) \leq (1-p_\star)^{j-1}$. Every individual confirmation test terminates almost surely by Lemma D.8. Consequently the entire procedure terminates almost surely, and its error bound gives the stated probability of returning the true hull.

To bound the expected labels, each attempted search uses at most $48dr_0$. Since $\delta < 1/2$, $a_j \geq a_1 = \log(4/\delta) > b$. Under either possible true answer, the conditional expected length of a confirmation test is at most $(a_j + \ell_\eta)/\kappa_\eta$. There are at most two tests per attempt. Conditioning at the start of each test and summing the nonnegative costs therefore gives

$$\begin{aligned}\sup_g \mathbb{E}_g \mathsf{T} &\leq \frac{48dr_0}{p_\star} + \frac{2}{\kappa_\eta}\sum_{j=1}^{\infty}(1-p_\star)^{j-1}(a_j+\ell_\eta) \\ &\leq \frac{48dr_0}{p_\star} + \frac{2}{\kappa_\eta}\left\{\frac{\log(4/\delta)+\ell_\eta}{p_\star} + \frac{2(1-p_\star)}{p_\star^2}\right\}. \end{aligned} \tag{33}$$

The last step uses $\log\{j(j+1)\} \leq \log 2 + 2(j-1)$ and the two geometric-series identities. In particular, the bound includes failed searches, rejected confirmations, and all retries.

It remains to express the bound in information units, uniformly in η. With $s = 1 - 2\eta \in (0,1)$,

$$C_\eta = \sum_{m=1}^{\infty}\frac{s^{2m}}{2m(2m-1)} \leq s^2\log 2, \qquad \kappa_\eta = s\log\frac{1+s}{1-s} \geq 2s^2, \qquad \frac{\ell_\eta}{\kappa_\eta} = \frac{1}{s}.$$

The capacity series follows by integrating the power series for $\operatorname{arctanh}(s)$; its coefficient sum at $s=1$ is $\log 2$. Also

$$d \leq \frac{2\log K}{\log 2}, \qquad r_0 \leq \frac{2\log 12 + 2}{s^2}, \qquad \frac{1}{\kappa_\eta} \leq \frac{\log 2}{2C_\eta}, \qquad \frac{\ell_\eta}{\kappa_\eta} \leq \frac{\log 2}{C_\eta}.$$

Substitution into (33), with $p_\star = 9/16$ and $\log K \geq \log 2$, proves (15). For example, $A_0 = 1200$ and $A_1 = 4$ suffice: the coefficient of $\log K/C_\eta$ is at most

$$\frac{96(2\log 12+2)}{p_\star} + \frac{\log 4 + 2}{p_\star} + \frac{2(1-p_\star)}{p_\star^2} < 1200,$$

and the coefficient of $\log(1/\delta)/\kappa_\eta$ is $2/p_\star < 4$. These constants are explicit bounds for this construction. For $K = 1$ the hull is already resolved; for $\eta = 0$, ordinary binary search takes at most d labels. □

D.6 Hard-decoding contraction benchmark

Define an auxiliary absorbing process by $\widetilde{V}_0 = V_0$ and

$$\widetilde{V}_t = \begin{cases} (\widetilde{V}_{t-1})_{q_t,\widehat{a}_t}, & \widetilde{V}_{t-1} \neq \varnothing,\ W_n(\widetilde{V}_{t-1}) > 2\varepsilon, \\ \widetilde{V}_{t-1}, & \text{otherwise}, \end{cases}$$

where $\widehat{a}_t$ is a majority-decoded response. This set is used only to analyze geometric efficiency.

Theorem D.9 (Pathwise contraction for a hard-decoding benchmark)**.** *Fix $0 < \rho < 1$ and $0 < \delta < 1$, and suppose $g_n^\star \in V_0$ almost surely. Suppose every selected query obeys*

$$\max_a W_n\{(\widetilde{V}_{t-1})_{q_t,a}\} \leq (1-\rho)W_n(\widetilde{V}_{t-1})$$

until width 2ε is reached. If, conditional on the past, the fresh replicate errors are independent with probabilities at most $0 \leq \bar{\eta} < 1/2$, and odd replication counts satisfy, for $0 < \beta_t < 1$,

$$r_t \geq \frac{2\log(1/\beta_t)}{(1-2\bar{\eta})^2}, \qquad \sum_{t\geq 1}\beta_t \leq \delta, \tag{34}$$

then, with probability at least $1-\delta$, $g_n^\star \in \widetilde{V}_t$ for every t and

$$W_n(\widetilde{V}_t) \leq \max\{(1-\rho)^t W_n(\widetilde{V}_0), 2\varepsilon\} \quad \textit{for all } t. \tag{35}$$

Proof. Under the replication rule (34), conditional Hoeffding inequality for the independent, not necessarily identically distributed replicate errors gives

$$\mathbb{P}(\widehat{a}_t \neq a_t^\star \mid \mathcal{F}_{t-1}) \leq \exp\{-r_t(1-2\bar{\eta})^2/2\} \leq \beta_t.$$

A union bound over the summable sequence gives simultaneous correct decoding with probability at least $1-\delta$. Since $g_n^\star \in V_0$, correct decoding retains it in every child. All sets on this event are nonempty, and each selected child satisfies the assumed contraction until absorption. Induction proves (35). □

Theorem D.9 concerns $\widetilde{V}_t$, not the inferential set $\mathcal{V}_t$. A correctly decoded batch need not immediately eliminate the opposite cell under the e-process threshold.

D.7 Aligned geometry, computation, and audit records

Proposition D.10 (When count and interval-width balancing align)**.** *At a retained set $V \neq \varnothing$, suppose $0 < a \leq b < \infty$ and $a|U| \leq W_n(U) \leq b|U|$ for every nonempty feasible child $U = V_{q,e}$. If a query's worst-cell count is within a factor $c \geq 1$ of the minimum possible count, its worst-response interval width is within a factor bc/a of the target-optimal width.*

Proof. Writing $m(q) = \max_a |V_{q,a}|$ and $\Gamma(q) = \max_a W_n(V_{q,a})$, a count query q_c and width-optimal q_w satisfy

$$\Gamma(q_c) \leq bm(q_c) \leq bc \min_q m(q) \leq (bc/a)\Gamma(q_w).$$

□

The comparability condition in Proposition D.10 bounds the loss from using counts in place of width. With a precomputed $Q \times M_n$ signature matrix, direct score evaluation costs $O(QM_n)$ per decision and a likelihood update costs $O(M_n)$ per label. The immutable record contains the main-data hash, timestamped query, cost, channel calibration, raw label, likelihood update, retained set, interval, and stopping reason. These records allow the query sequence, likelihood updates, and stopping decision to be reconstructed.

E Proof of the Dependency-Graph Central Limit Theorem

Let

$$S_n = \sum_{i=1}^{n} \big[\psi_i(g_n^\star) - \mathbb{E}\{\psi_i(g_n^\star) \mid \mathcal{H}_n^{\mathrm{aug}}\}\big], \qquad s_n^2 = \mathrm{Var}(S_n \mid \mathcal{H}_n^{\mathrm{aug}}),$$

and assume the triangular array and augmentation sigma-fields are defined on a common probability space.

Condition E.1 (Nondegenerate local-dependence sequence)**.** With probability one over the augmentation sequence, $\sup_n R_{g_n^\star}/\underline{\pi}_{g_n^\star} < \infty$, $s_n^2 \geq cn$ for some (possibly realization-dependent) $c > 0$, and $D_n^{\cup} + 1 = o(n^{1/4})$.

Theorem E.2 (Conditional dependency-graph central limit theorem)**.** *Under Assumptions 3.2 to 3.4 and condition E.1,*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{n\{\widehat{\tau}_n(g_n^\star) - \tau_n^E\}}{s_n} \leq x \middle| \mathcal{H}_n^{\mathrm{aug}} \right\} - \Phi(x) \right| \longrightarrow 0 \quad \textit{almost surely}, \tag{36}$$

where Φ is the standard normal distribution function.

We use the dependency-graph cumulant route of Janson (1988). For a random variable X with moments of all orders, let $\kappa_r(X)$ denote its rth cumulant. For a collection $(X_1, \ldots, X_r)$, let $\kappa(X_1, \ldots, X_r)$ denote the joint cumulant.

Lemma E.3 (Cumulant bound under a dependency graph)**.** *Let $X_1, \dots, X_n$ be centered random variables with a dependency graph of maximum degree D and $|X_i| \le K$ almost surely. For every integer $r \ge 2$, there exists a finite constant C_r, depending only on r, such that*

$$\left| \kappa_r \left(\sum_{i=1}^{n} X_i \right) \right| \le C_r n (D+1)^{r-1} K^r.$$

Proof. Multilinearity of joint cumulants gives

$$\kappa_r \left(\sum_i X_i \right) = \sum_{i_1, \dots, i_r \in [n]} \kappa(X_{i_1}, \dots, X_{i_r}).$$

A standard factorization property of cumulants implies that a joint cumulant is zero whenever its arguments can be partitioned into two nonempty independent subcollections. Form an augmented graph on the r argument positions: connect two positions when they carry the same index or when their indices are adjacent in the dependency graph. If this augmented graph is disconnected, its two components are independent; therefore only tuples with connected augmented position graph contribute.

For fixed r, the moment–cumulant formula writes a joint cumulant as a signed sum over set partitions of $[r]$. Since every mixed moment is bounded by K^r, there exists $c_r < \infty$ such that

$$|\kappa(X_{i_1}, \dots, X_{i_r})| \le c_r K^r$$

uniformly over the indices.

It remains to count tuples with connected augmented graph. Choose the first index in n ways. Every connected graph on the r labeled positions contains a spanning tree. For each labeled tree, expose vertices from a root. A child position must receive either its parent's index or an index adjacent to it, giving at most $D+1$ choices. Summing over the finitely many labeled trees yields at most $c_r' n (D+1)^{r-1}$ tuples, where c_r' depends only on r. Multiplication by the uniform joint-cumulant bound proves the lemma with $C_r = c_r c_r'$. □

E.1 Proof of Theorem E.2

Proof. Fix an augmentation realization in the probability-one event of Condition E.1; throughout the proof, expectations, variances, and cumulants are taken under the resulting conditional laws. Let

$$X_{ni} = \psi_i(g_n^\star) - \mathbb{E}\{\psi_i(g_n^\star) \mid \mathcal{H}_n^{\mathrm{aug}}\}, \qquad S_n = \sum_{i=1}^{n} X_{ni}.$$

By the bounded residual and positivity conditions, there exists a constant $K < \infty$ independent of n such that $|X_{ni}| \le K$. The variables admit the common dependency graph of degree $D_n^{\cup}$. By definition, $\mathbb{E}S_n = 0$ and $\mathrm{Var}(S_n) = s_n^2$.

For every fixed $r \ge 3$, Lemma E.3 and $s_n^2 \ge cn$ imply

$$|\kappa_r(S_n/s_n)| \le C_r' n^{1-r/2} (D_n^{\cup} + 1)^{r-1}. \tag{37}$$

For $r = 3$, the right-hand side is $O\{(D_n^{\cup}+1)^2 n^{-1/2}\} = o(1)$ because $D_n^{\cup} = o(n^{1/4})$. For $r \ge 4$, use $D_n^{\cup} + 1 = O(n^{1/4})$ and

$$1 - \frac{r}{2} + \frac{r-1}{4} = \frac{3-r}{4} < 0.$$

Thus the right-hand side of (37) tends to zero for every fixed $r \ge 4$. The first cumulant is zero, the second cumulant equals one, and all cumulants of fixed order $r \ge 3$ converge to zero.

Every fixed moment is a polynomial in the cumulants of orders no larger than that moment. Hence the moments of S_n/s_n converge to the corresponding standard normal moments. The sequence is tight because its variance is one. The standard normal distribution is moment determinate, so the method of moments yields, under the fixed conditional-law sequence,

$$S_n/s_n \xrightarrow{d} N(0,1).$$

Finally,

$$S_n = n\{\widehat{\tau}_n(g_n^\star) - \tau_n^E\}$$

by exact conditional unbiasedness. Because the Gaussian limit distribution is continuous, Pólya's theorem upgrades conditional weak convergence to uniform convergence of the conditional distribution functions. The argument applies to every augmentation realization in the probability-one event, proving the almost-sure assertion in (36).

□

F Experiments and Reproducibility

The experiments study structural reporting and target geometry, network acquisition and calibration, and real-outcome assumption sensitivity. The first two use fully specified synthetic designs; the third uses weighted response cells from a public randomized study. Configurations, raw records, output hashes, and independent verifiers are provided in the artifact.

F.1 Structural validity, target geometry, and ordered search

The structural suite uses PCG64 seed 20260915 with a fixed serial stream order. The artifact contains its configuration, per-path records, aggregates, figure sources, and repeat-execution checks.

The two-candidate experiment has target values -1 and 1, interval half-width $.2$, uniform mixture weights, BSC$(.35)$ labels, and $\delta = .05$. Each of $20{,}000$ paths contains 50 labels. Reporting begins after the first label, including ties resolved toward candidate zero. The running maximum of each mixture likelihood ratio makes rejection permanent. Full-path union coverage is 0.97655, single-map full-path coverage is 0.3899, and single-map terminal coverage is 0.98485.

The misaligned geometry experiment assigns 16 candidates to each of the targets -1 and 1 and 96 to zero. Four group-bit queries balance candidate count; two membership queries isolate the extreme groups. Both deterministic policies are evaluated on every possible truth and stop at target diameter at most $.1$ (equivalently zero for these three target values). Target-directed queries cost 1.875 on average (maximum 2), compared with 5.875 (maximum 6) for count balancing. The aligned control uses ordered integer targets, all threshold cuts, and stopping diameter at most one. Both policies cost exactly 6. These controls concern population diameter Δ_n, not estimated confidence width W_n.

The ordered-search grid crosses $K \in \{8, 32, 128\}$, $\eta \in \{.10, .30, .45\}$, and $\delta \in \{.10, .05, .01\}$, with $5{,}000$ paths per cell. Each cut uses fresh BSC labels and the least odd count in Corollary B.3. The exact dr budgets range from 33 to $9{,}177$ labels. The largest empirical error-to-δ ratio is 0.2; the largest Wilson-upper-limit ratio is 0.3678. Configurations and all $135{,}000$ path outcomes are recorded separately by cell. The artifact's error/budget panel evaluates this repeated-majority benchmark; the backtracking cost experiment is specified below. Assertions check nesting, retention/coverage equivalence, minimal odd counts, and label budgets.

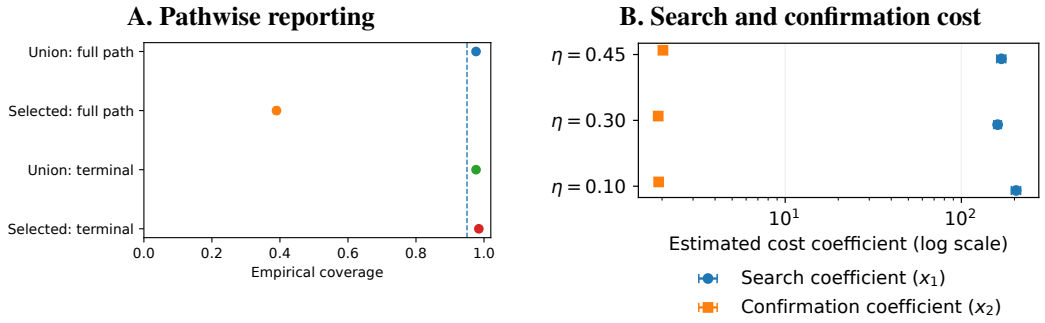


Figure 4: **Reporting validity and noisy-search costs.** A: $20{,}000$ two-candidate paths with ordinary Wilson intervals. B: cost coefficients of $x_1 = \log K/C_\eta$ and $x_2 = \log(1/\delta)/\kappa_\eta$ from 48 network-embedded configurations, 400 paths each; paired-cluster 95% Monte Carlo intervals on a logarithmic coefficient axis. Appendix F.2 gives all phase and cross-effect estimates.

F.2 Separating search and confirmation cost in a network construction

A common-law network with feasible ordered cuts. Take a star with an external source and $n = K - 1$ receiver units. The source switch Z_0 is Bernoulli$(1/2)$. Under candidate $k \in \{1, \dots, K\}$, receiver i has physical receipt

$$g_i^k(z) = z_0 \oplus \mathbf{1}\{i < k\}, \qquad Y_i(z) = 1 - z_0.$$

Set $Y_i^{E,k}(e) = e$ for $i < k$ and $Y_i^{E,k}(e) = 1 - e$ otherwise. Substitution gives $Y_i^{E,k}\{g_i^k(z)\} = 1 - z_0$ for both assignments. Thus every candidate satisfies exposure sufficiency and positivity $\pi_{ie} = 1/2$, with the same complete main-data law, but

$$\theta_n(g^k; P_0) = \tau_n^{E,k} = \frac{2(k-1) - n}{n}.$$

A bridge turns the source switch on and measures receiver $i = q$, giving $a_q(g^k) = \mathbf{1}\{k \le q\}$. Every ordered cut is therefore a feasible receiver measurement. Taking $\varepsilon = 1/(2n)$ gives K separated singleton cells. Complementary delivery gates on the branches yield the supplied ordered population geometry assumed by Theorem 5.6. Each reported cell is the hull of its population targets.

Procedure and factorial design. We execute the proof's backtracking procedure exactly: $16\lceil \log_2 K \rceil$ transitions per search call, the prescribed r_0-label majorities, exterior boundary tests at $a_j = \log\{2j(j+1)/\delta\}$ and $b = \log 8$, and a fresh search after rejection. The grid crosses $K \in \{8, 32, 128, 512\}$, $\eta \in \{.10, .30, .45\}$, and $\delta \in \{.1, .01, .001, .0001\}$, with 400 independent uniform-truth trials per cell. Across the four δ values, each trial shares its truth, main assignment, and phase/attempt-indexed latent random stream. Each policy observes only the labels it acquires. All failed proposals, rejected tests, and restarts enter realized unit cost. The 19,200 paths used up to five attempts; none was truncated.

A fixed majority batch is sampled through its exact binomial error count, which is sufficient for its branch decision; its full label count is charged. Every sequential confirmation bit is saved. An independent interval-stack implementation replays all 23,347 search calls, 4,444,648 majority batches, and 4,984,641 confirmation bits, and checks the network equations for every candidate and both main assignments. All returns in this finite grid were correct. The theorem certifies the requested error levels; 400 trials per cell estimate realized costs.

Separate information scales. For each η and each phase, regress trial-level cost on an intercept, $x_1 = \log K/C_\eta$, and $x_2 = \log(1/\delta)/\kappa_\eta$. Table 4 reports every slope, including cross-effects and total cost. Monte Carlo standard errors cluster the four paired δ observations for each (K, trial). The regressions describe phase-cost scaling within this factorial grid.

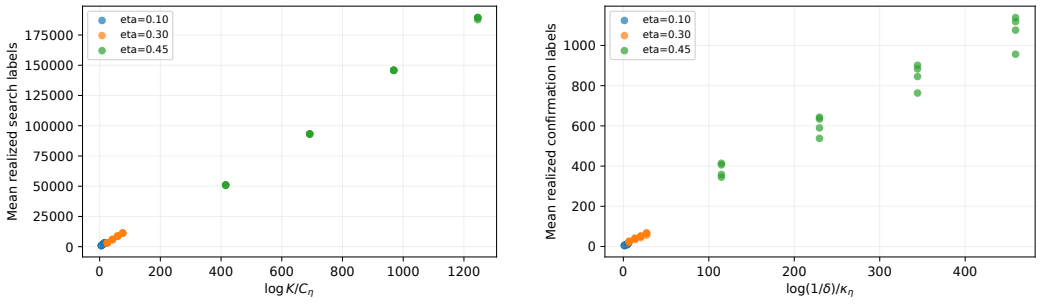


Figure 5: Realized phase costs versus the two information scales. Each point is a 400-trial cell mean; color identifies the channel error rate. Points retain the separate factorial settings. Search cost includes every restarted search; confirmation includes rejected tests and both exterior boundaries when present.

Search slopes on x_1 range from 160.54 to 204.42; confirmation slopes on x_2 range from 1.90 to 2.02. Search also increases with the confidence requirement through restarts, with a positive cross-slope at $\eta = .45$. Confirmation depends weakly on K through boundary availability and restarts. The phases have distinct dominant scales, with cross-effects from restarts. The proof constants

Table 4: Within-channel cost slopes with paired-cluster 95% Monte Carlo intervals. Both predictors are included in every regression.

η	Cost component	Coefficient of x_1	Coefficient of x_2
0.10	Search	204.42 [192.07, 216.78]	0.76 [-0.73, 2.24]
0.10	Confirmation	0.09 [0.05, 0.13]	1.91 [1.88, 1.95]
0.10	Total	204.51 [192.13, 216.90]	2.67 [1.18, 4.16]
0.30	Search	160.54 [152.55, 168.53]	0.77 [-0.33, 1.87]
0.30	Confirmation	0.14 [0.09, 0.19]	1.90 [1.86, 1.94]
0.30	Total	160.68 [152.66, 168.70]	2.67 [1.57, 3.77]
0.45	Search	168.63 [158.67, 178.60]	1.44 [0.15, 2.73]
0.45	Confirmation	0.14 [0.09, 0.19]	2.02 [1.98, 2.07]
0.45	Total	168.77 [158.78, 178.76]	3.47 [2.17, 4.76]

$A_0 = 1200, A_1 = 4$ give a uniform upper bound; the fitted slopes summarize realized costs under the common-channel design.

F.3 End-to-end synthetic-network audit and ablations

Finite design and causal loop. The executable experiment fixes a 70-node stochastic-block network with 14 blocks of size five and 178 edges, then independently treats blocks with probability .35. The candidate class comprises 10 one- and two-hop fractional-threshold exposure maps. Exhaustive enumeration of all 16,384 block assignments gives exact candidate exposure probabilities and pseudo-targets. Within each truth-specific synthetic finite population, the exact-design $1 - \alpha$ quantile of the maximum absolute candidate error defines one common simulation-oracle radius; the minimum simultaneous design coverage over truths is 0.950005. Outcomes combine a bounded baseline, heterogeneous physical exposure effects, and no post-assignment residual noise.

Audits and comparators. The 34 nonconstant prediction signatures are each available through rapid $(\eta, c) = (.30, 1)$, verified $(.18, 1.4)$, and gold $(.08, 2.5)$ binary-symmetric channels. We set $\alpha = \delta = .05$, $\beta = .10$, and budget 30; every raw label updates the mixture e-process. Truth candidates are balanced over 600 fixed-seed replications. All methods share the truth and main assignment, but their audit streams differ unless stated otherwise. The target-score and reliability-price ablations and the fixed-query baseline retain the same e-process and visible empty-set semantics. The selected-map reporter alone reuses TAB's query and raw-label path exactly; it begins after the first label, thereby excluding the arbitrary no-label likelihood tie. The generator uses the same physical mechanism for main outcomes and bridge predictions, followed by the stated sensor channel. The paired comparisons measure the effects of target scoring and pricing within this query menu.

Table 5 summarizes the width, coverage, and realized-cost endpoints.

Table 5: Oracle-band acquisition benchmark over 600 fixed-seed replications. Mean widths use nonempty reports: 600 runs for each method except fixed querying, which has 598 reports and two model failures. Coverage uses all 600 runs, counting model failure as noncoverage. "Path" includes every reporting time. Monte Carlo intervals are provided in the result CSV.

Method	Terminal width	Contraction	Path cov.	Terminal cov.	Mean cost
TAB	0.554	32.5%	0.997	0.997	22.45
− target-aware score	0.595	28.4%	0.998	0.998	26.68
− reliability pricing	0.618	24.6%	0.997	0.997	26.96
Fixed query	0.783	5.5%	0.997	0.997	29.39
ML report on TAB labels[†]	0.543	—	0.930	0.997	22.45

[†] The ML reporter uses TAB's complete query and label path, with the same mean cost 22.45. It reports after the first label (cost 1.40) and selects the current maximum-likelihood candidate at each subsequent time.

Figure 7 displays the complete cost path, replication-level heterogeneity, reporting-time coverage contrast, and realized channel choices.

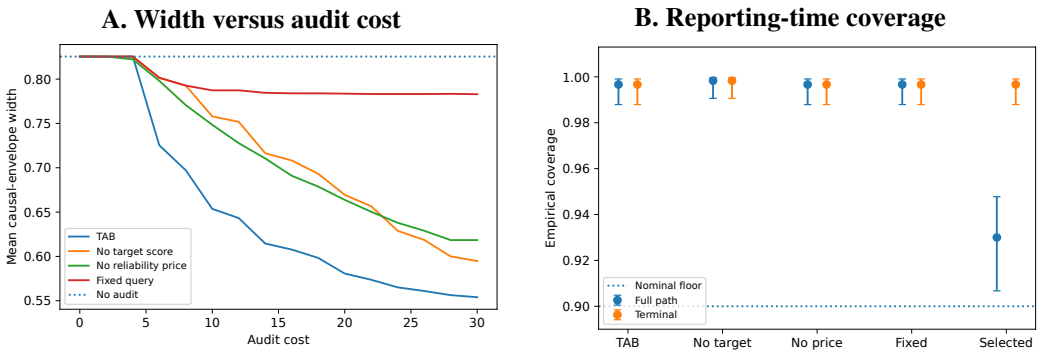


Figure 6: **Network outcome–audit loop.** Oracle-calibrated bands isolate acquisition in 600 paired populations. Wilson intervals describe coverage. The selected-map report shares TAB queries: its path coverage is 0.930, terminal coverage 0.997, and nominal floor 0.90.

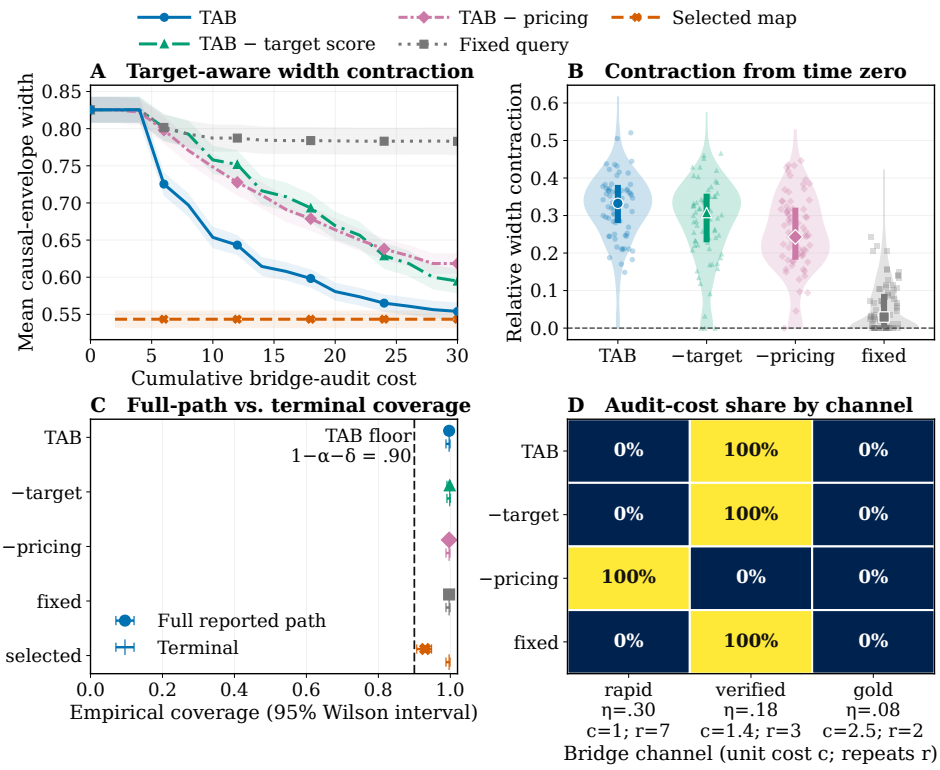


Figure 7: **Network acquisition, reporting, and channel use.** **A**, mean widths among reporting paths with 95% Monte Carlo intervals. **B**, replication-level contractions from time zero; selected-map reporting starts after one label and has constant width thereafter. **C**, full-path and terminal coverage with ordinary 95% Wilson intervals; the vertical line is $1 - \alpha - \delta = .90$. **D**, realized cost by channel. TAB selected the verified channel, while the no-pricing ablation selected the rapid channel. The common candidate bands are simulation-oracle exact-design bands.

Acquisition and reporting. TAB contracted mean width from 0.825 to 0.554 (32.5%), compared with terminal widths 0.595, 0.618, and 0.783 under the three ablations. Its full-path and terminal coverage were both 0.997. The current maximum-likelihood single-map reporter had path coverage 0.930 but terminal coverage 0.997: among 42 path failures, 40 recovered by the terminal report. Path coverage is 0.930 against the nominal 0.90 floor; terminal coverage masks the transient misses.

Analytic-band replay and information prices. A separate audit constructs the union dependency graph from overlapping two-hop randomization-block supports and verifies a proper coloring with 69 colors. The data-generating process gives a common residual bound $R = 0.536937$ across both outcome values, augmentation arms, and all candidates. Applying (24) gives candidate half-widths from 5.132 to 41.037. Exhaustive assignment enumeration confirms simultaneous coverage one for every truth. Reconstructing the e-process from each recorded raw label reproduces every retained-set size. The analytic envelope's mean width decreases from 82.074 to 35.972 on these queries. The replay holds the original queries fixed and changes only the candidate bands.

Paired Monte Carlo differences use the same main population and assignment but policy-specific audit noise. Both score ablations have all 600 pairs. Fixed querying reports in 598 of 600 runs; its paired width comparison uses those 598 nonfailure pairs, with the failure rate reported separately. TAB and selected-map reporting share every query and label.

The rapid, verified, and gold channels have capacity per cost 0.0823, 0.1584, and 0.1658 nats, and pairwise KL per cost 0.3389, 0.6932, and 0.8206. Gold has the largest continuous information rates, whereas the integer resolution-batch costs are 7, 4.2, and 5. The score therefore chooses verified. Integer batching changes the acquisition choice relative to continuous information pricing. The implementation starts only complete affordable batches; the unused budget is retained in the cost record.

Reproduction and path checks. The network reproduction script regenerates the CSVs, numerical macros, and vector figures twice and compares their hashes. The execution manifests enumerate the checked outputs. A separate invariant audit checks 61,851 rows across 3,000 method paths, including 600 exactly paired TAB/selected-map paths, for version-set nesting, realized budgets, failure semantics, band/envelope arithmetic, coverage reductions, and post-label pairing.

F.3.1 INDEPENDENT-COMPONENT CALIBRATION

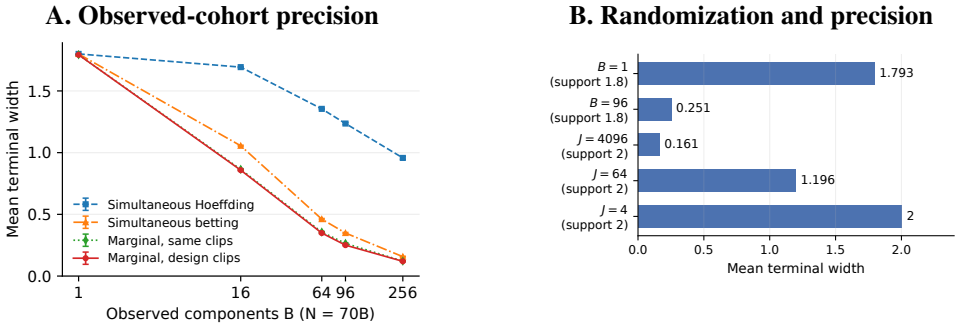


Figure 8: **Non-oracle precision.** Mean terminal widths, 600 trials per setting. A: independent components, original schedule, budget 30; 95% Monte Carlo intervals. B: original component schedule and connected cycles (effect 0.30), with the effect-range support recorded beside each setting; the final bar equals its full support.

Design and observable inputs. Each trial observes $B \in \{1, 16, 64, 96, 256\}$ disconnected copies of a 70-unit network, with independent block assignments and one shared exposure mechanism. Hence $N = 70B$ is the observed population size. We fix B, the candidate class, augmentation, outcome support $[-.1, .8]$, and query menu before drawing assignments. The original outcome schedule and graph construction are specified in Appendix F.3. All ten mechanisms are equally represented among the 600 original-schedule trials, 200 trials for each additional outcome schedule, and 100 for each additional topology. These are Monte Carlo repetitions, separate from the B observed components.

The simulator generates observed assignments and outcomes, from which it computes the component scores X_{kg}. Inference uses the observed scores and design-certified bounds. A query rule receives candidate intervals, feasible signatures, accumulated labels, and remaining cost. The audit clock uses the same calibrated channels and budget 30 as the connected-network design; its tolerance is $\varepsilon = .02$. Fresh random streams are shared across methods to form paired comparisons. Every cohort size is analyzed separately, without selecting B from the observed outcomes.

Calibration and acquisition. Four bands separate concentration, error allocation, and clipping. Simultaneous Hoeffding intervals use the independent-component radius

$$(b_g - a_g)\sqrt{\frac{\log(2M/\alpha)}{2B}},$$

centered at the average clipped score, with the two bias corrections in Equation (28). Simultaneous betting inverts Equation (27) at error α/M; its clips minimize the design-only constant-score proxy at that allocation. Marginal betting first keeps those same clips and changes only the allocation to α. The design-clipped marginal variant reoptimizes the same proxy at α. All bands cover the unchanged target after bias correction, and use the known contrast support $[-.9, .9]$. The primary paired contrast compares simultaneous betting with design-clipped marginal betting under the same acquisition rule.

The plateau-aware rule applies the TAB score while a feasible query has positive one-step width gain; if all such gains are zero within 10^{-14}, it applies the count score. Target-only TAB and count-based querying provide acquisition controls under the same marginal bands. All three rules use the same information-adjusted batch prices. The e-process updates after every raw label, including labels within a batch, and reports the resulting hull at each update. An empty hull counts as noncoverage and receives the support diameter 1.8 in the efficiency comparison. All trials enter the means, with structural failure and reporting status recorded separately. Because the candidate bands are fixed and the retained sets are nested, the reported union intervals are nested. Consequently, terminal coverage at a stopping time equals full-path coverage on each trial. Selected-map reports need not be nested, so their two coverage measures can differ.

Table 6: Independent-component experiment, original schedule, 600 paired trials per cohort size. Width and cost are unconditional means. The primary marginal method uses design-selected clips and plateau-aware acquisition. The final column is a descriptive width-0.25 attainment rate.

B	N	Sim. width	Marginal width	Cost	Path cov.	$W \le .25$
1	70	1.7993	1.7927	24.997	1.0000	0.000
16	1,120	1.0557	0.8582	25.046	1.0000	0.000
64	4,480	0.4612	0.3487	23.947	0.9917	0.167
96	6,720	0.3487	0.2506	23.317	0.9917	0.657
256	17,920	0.1572	0.1195	23.107	0.9867	0.993

What produces the precision gain? At $B = 96$, marginal design-clipped bands give mean terminal width 0.2506, compared with 0.3487 for simultaneous betting: a 28.1% reduction. Mean audit cost is 23.317, and full-path coverage is 0.9917. The true-candidate coverage argument in Proposition C.5 justifies the allocation change; simultaneous statements about every candidate target still require simultaneous bands. Width contraction along the audit path in Figure 9 then quantifies the information gained from new exposure labels.

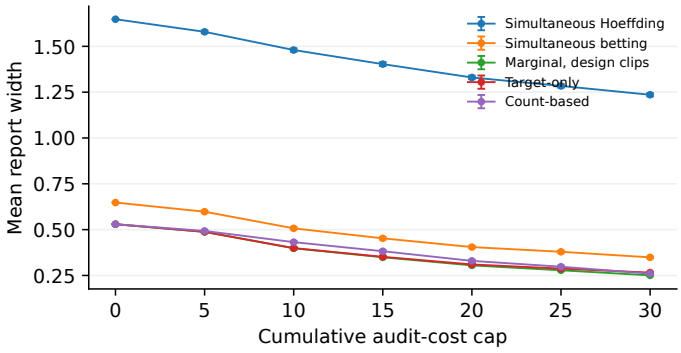


Figure 9: Mean width along budget-30 policies at $B = 96$, original schedule, 600 trials. A stopped report is carried forward to subsequent cost caps. Error bars are 95% pointwise Monte Carlo intervals with truth-stratified standard errors.

Table 7: Paired differences: design-clipped marginal minus simultaneous betting, with the same plateau-aware rule. Normal Monte Carlo intervals use within-truth-stratum variances and a Bonferroni adjustment over the ten displayed width/cost comparisons. At $B = 256$, improved width costs more on average.

B	Metric	Mean difference	Adjusted MC interval
1	Width	-0.00663	[-0.00677, -0.00650]
1	Cost	0.25900	[-0.11448, 0.63248]
16	Width	-0.19748	[-0.20091, -0.19406]
16	Cost	-0.96600	[-1.37091, -0.56109]
64	Width	-0.11254	[-0.11600, -0.10907]
64	Cost	-0.68600	[-0.93839, -0.43361]
96	Width	-0.09810	[-0.10253, -0.09367]
96	Cost	-0.51800	[-0.83359, -0.20241]
256	Width	-0.03769	[-0.03998, -0.03539]
256	Cost	0.30100	[0.05971, 0.54229]

Let n_g denote the number of paired trial differences in truth stratum g, $R = \sum_g n_g$, and s_g^2 their within-stratum sample variance. For the overall paired mean $\bar{D}$, the reported Monte Carlo variance is

$$\widehat{\mathrm{Var}}(\bar{D}) = \sum_{g=1}^{10} (n_g/R)^2 s_g^2/n_g.$$

The adjusted intervals are $\bar{D} \pm z_{1-.05/(2\cdot 10)}\widehat{\mathrm{se}}(\bar{D})$, where z_p is the standard-normal p-quantile. They quantify simulation error in these fixed comparisons. For a rate $\hat{p}$ based on R independent, possibly nonidentical trials, Figure 10 uses the distribution-free bounds $[\hat{p} \pm \sqrt{\log(40)/(2R)}] \cap [0, 1]$. No pooled-binomial assumption about the ten truth strata is required.

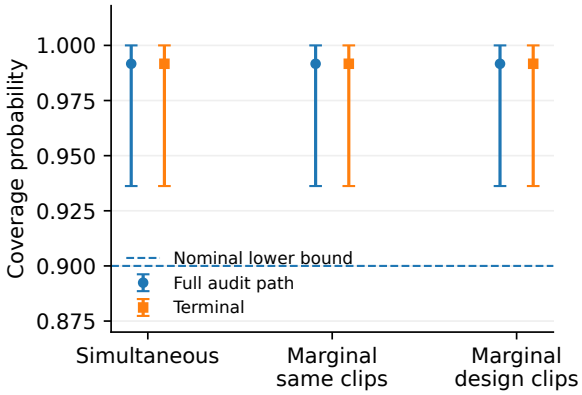


Figure 10: Audit-time coverage at $B = 96$ on the original independent-component schedule, 600 paired trials. Intervals use independent-nonidentical Hoeffding bounds.

Acquisition controls and low-information regimes. Table 8 holds the primary bands fixed. Count-based querying is narrower at $B = 16$, while target-only querying spends less but leaves wider reports. At $B = 256$, the primary method improves on simultaneous betting in width but uses an additional 0.301 mean cost (adjusted Monte Carlo interval $[0.060, 0.542]$). With $B = 1$, its mean width is 1.7927 at cost 24.997, close to the support diameter 1.8. Within-component dependence limits precision in this single-component setting.

Outcome and topology controls. Let y_{0i}, τ_i be the original fixed schedule, and index components by $k = 0, \ldots, B-1$. The heterogeneous schedule uses baseline $y_{0i} + .035\sin(2\pi k/7)$ and effect $\tau_i\{.6 + .4(k \bmod 5)/4\}$. The null schedule sets all effects to zero; the weak schedule uses $.04\tau_i/\bar{\tau}$, where $\bar{\tau} = 70^{-1}\sum_{i=1}^{70}\tau_i$. The support-stress schedule sets $y_{0i} = -.09 + .005\sin(.7i)$ and $y_{1i} = .79 -$

Table 8: Acquisition controls under the same marginal, design-clipped bands. Original schedule, 600 trials per row. "Marginal, design clips" denotes plateau-aware acquisition.

B	Policy	Width	Cost	Path coverage
16	Marginal, design clips	0.8582	25.046	1.0000
16	Target-only	0.9958	9.772	1.0000
16	Count-based	0.8462	26.670	1.0000
96	Marginal, design clips	0.2506	23.317	0.9917
96	Target-only	0.2656	20.790	0.9933
96	Count-based	0.2599	26.670	0.9900
256	Marginal, design clips	0.1195	23.107	0.9867
256	Target-only	0.1202	22.960	0.9867
256	Count-based	0.1369	26.670	0.9833

$.005\cos(.9i)$. All schedules respect the declared support, and the heterogeneous target is averaged over the actually observed components. Three further fixed topologies use seeds 2026091801, 2026091803, and 2026091807. The inference is conditional on each network, not on a population distribution of networks.

At $B = 256$, weak-effect positive decisions occur in only 31.5% of trials, despite full-path coverage one. The null schedule gives $0/200$ positive decisions. Table 9 reports every configured schedule; the artifact contains all $51{,}000$ trial–method–cohort records.

Table 9: All outcome and topology controls for the primary marginal method. W_B is mean terminal width at cohort size B. Coverage and positive-decision frequency use $B = 256$. The complete five-size, six-method grid is included as CSV.

Schedule	Trials	W_{96}	W_{256}	Path cov.$_{256}$	Positive$_{256}$
original	600	0.2506	0.1195	0.987	1.000
heterogeneous	200	0.2586	0.1168	0.980	1.000
null	200	0.2747	0.1422	1.000	0.000
weak_effect	200	0.2635	0.1316	1.000	0.315
support_stress	200	0.2559	0.1462	0.980	1.000
topology_1	100	0.2231	0.0951	1.000	1.000
topology_2	100	0.2382	0.0977	0.970	1.000
topology_3	100	0.2460	0.1098	0.980	1.000

Exact ties in the acquisition controls. Two case–cohort comparisons have identical complete paths in every trial. For topology 1 at $B = 96$, the simultaneous- and marginal-allocation clipping objectives select the same grid points. Thus marginal betting with fixed clips and with reoptimized clips has identical bands and follows identical paths in all 100 trials. For topology 1 at $B = 256$, plateau-aware and target-only acquisition use the same bands and choose the same queries on all 100 paths; the fallback does not change any query in this control. These equalities result from the clipping choices and visited queries in the specified controls. The artifact records pathwise equality counts for all 600 method-pair/case/cohort comparisons.

Executable records. The artifact includes the generator, observable-only inference and acquisition functions, observed component scores, assignment identifiers, all candidate intervals, and every selected raw audit label and report. An independent implementation replays the likelihood process and reconstructs the hulls and paired comparisons from those records. Two complete executions agree on the numerical arrays and CSVs. The configuration was fixed before these draws. The independent series and the separately retained aggregate comparison series use different data and calibration controls.

F.3.2 CONDITIONAL CHANNEL CALIBRATION UNDER DRIFT

Design. We retain all 600 original-schedule main-data realizations at $B = 96$ and their design-clipped marginal intervals, and draw fresh paired audit streams. This holds outcome information and band calibration fixed while changing the label channel. Each method reselects queries from its own retained set using the plateau-aware rule. The budget is 60 unit-cost raw labels, with $\delta = .05$ and $\varepsilon = .02$; this normalized single-channel stress test is separate from the heterogeneous cost menu above. Every report prefix is stored, including those preceding an early stop or an empty set.

Nominal likelihood updates assume BSC(.18). The bounded-error method assumes only the conditional upper bound $\nu = .35$ and averages seven instances of (30), with $\lambda_j = f_j/\nu$ and

$$(f_1, \ldots, f_7) = (.02, .05, .10, .20, .40, .70, .95).$$

Each product is a nonnegative supermartingale under the conditional bound; their fixed uniform mixture is also a nonnegative supermartingale. The retained set uses its running maximum and threshold $1/\delta$, as in Proposition D.3. A known-drift likelihood diagnostic receives the actual error probability at each selected query. It measures the cost of specifying only an error bound rather than the complete channel law. In the calibrated control both likelihood rules use error .18; coupled uniforms and identical tie-breaking make their selected queries, labels, retained sets, and intervals identical on all 600 paths. Their equal rows are therefore an implementation control, not separate evidence of superiority.

Channel laws. The calibrated control has error .18. Step drift changes .18 to .35 after the first ten labels. History/query drift has error .35 when the selected query index plus the previous observed label is even, and .08 otherwise; the initial previous label is zero. The bound-edge control has error .35 throughout. These four regimes satisfy the conditional upper bound. Two negative controls do not: one has error .45 throughout; the other assigns error .49 to the query selected before the first label and .005 to each of the other 33 queries. The latter's uniform-menu average error is only .0193, but its selected-query error need not be below .35.

Table 10: Channel-calibration stress test, 600 paired main-data trials per row. Retention and coverage refer to the entire reported path. Width and label count are unconditional terminal means; an empty report receives width 1.8. The last two regimes violate the bounded-error method's assumption. All methods use the same main-data bands and predictable acquisition rule.

Channel regime	Retention rule	Retention	Coverage	Width	Labels
Calibrated	Nominal likelihood	0.980	0.987	0.218	16.64
	Bounded-error mixture	1.000	1.000	0.300	46.75
	Known-drift diagnostic	0.980	0.987	0.218	16.64
Step drift	Nominal likelihood	0.888	0.942	0.222	22.25
	Bounded-error mixture	0.998	0.998	0.337	49.36
	Known-drift diagnostic	0.982	0.988	0.244	35.05
History/query drift	Nominal likelihood	0.960	0.978	0.217	18.44
	Bounded-error mixture	0.997	0.997	0.311	47.94
	Known-drift diagnostic	0.987	0.990	0.217	14.43
Bound edge	Nominal likelihood	0.707	0.832	0.234	28.20
	Bounded-error mixture	0.978	0.980	0.355	51.95
	Known-drift diagnostic	0.977	0.978	0.308	51.63
Bound violated	Nominal likelihood	0.340	0.532	0.255	33.28
	Bounded-error mixture	0.843	0.867	0.412	54.60
	Known-drift diagnostic	0.997	0.997	0.521	60.00
Menu-average only	Nominal likelihood	0.535	0.617	0.230	18.01
	Bounded-error mixture	0.670	0.727	0.367	52.90
	Known-drift diagnostic	1.000	1.000	0.529	60.00

Retention and its cost. At the bound edge, nominal likelihood retains the true mechanism on .707 of paths, compared with .978 for bounded-error retention. Their full-path causal coverages are .832 and

.980. With 600 independent, possibly nonidentical trials, the Hoeffding upper limit for the former is .887, below the nominal coverage lower bound .90. Bounded-error retention is more expensive: its mean label count is 51.95, versus 28.20 for nominal likelihood, and its terminal width is .355 rather than .234. The narrower nominal reports trade away coverage. Even under a correctly specified .18 channel, bounded-error retention uses 46.75 labels on average, versus 16.64 for calibrated likelihood.

The menu-average negative control reduces bounded-error retention to .670 and coverage to .727. An average calibration guarantee over a query menu cannot replace the selected-query conditional bound. The known-drift diagnostic retains every true mechanism in these 600 trials but leaves mean width .529 at the full budget; a nearly uninformative selected query may preserve coverage without resolving the target.

Records and uncertainty. The generator saves 10,800 method–trial paths and 417,644 report prefixes. A separate implementation reconstructs query choices, channel probabilities, labels, evidence, running maxima, hulls, stopping, all 18 summary rows, and 48 paired contrasts. Two executions agree on every numerical array and CSV. Rate intervals use $\sqrt{\log(40)/(2 \cdot 600)}$ as their half-width; paired descriptive intervals use the truth-stratified Monte Carlo variance given above. The artifact includes all six configured regimes. The artifact records the input-array provenance and frozen channel specification.

F.3.3 Physical bridge changes with a calibrated sensor

The common-law star construction above separates a change in physical receipt from recording error. We use $K = 8$, supplied population targets, 1,000 independent uniform-truth trials, $\delta = .05$, and at most 80 unit-cost labels. The main mechanism is unchanged. At each query its physical bridge receipt is flipped with conditional probability r_t; a separate independent BSC(.10) sensor then records that receipt. All methods query the lower median retained cell and stop at a singleton, an empty set, or the budget. Shared latent physical and sensor uniforms pair methods, while each chooses queries from its own observed history.

Nominal likelihood ignores physical transport. Corrected likelihood uses (31) with $D_t = \{1+8\rho^{\mathrm{br}}\}^t$. Bounded-error evidence uses the fixed mixture of seven bets from the channel experiment, with $\nu = .1 + .8\rho^{\mathrm{br}}$. A known-transport diagnostic receives the actual conditional rate. The first four regimes have constant r_t; history drift uses .30 when the selected query plus the preceding label is even and zero otherwise. The negative control has $r_t = .40$ but supplies bound .15.

Table 11: Bridge-transport sensitivity on the common-law star. Each entry is full-path structural retention / mean labels, over 1,000 paired trials. The last row violates the supplied transport bound.

Regime (bound)	Nominal	Corrected	Bounded-error	Known rate
Invariant (0)	0.973 / 8.69	0.973 / 8.69	0.981 / 12.17	0.973 / 8.69
0.05 (0.05)	0.948 / 9.48	0.977 / 15.07	0.985 / 17.66	0.985 / 12.54
0.15 (0.15)	0.834 / 11.34	0.978 / 62.75	0.969 / 30.20	0.951 / 19.66
0.30 (0.30)	0.545 / 14.23	0.984 / 80.00	0.971 / 79.43	0.961 / 60.81
History (.30)	0.825 / 11.48	0.990 / 79.26	0.996 / 69.24	0.976 / 16.61
.40 (.15)	0.306 / 16.17	0.877 / 79.94	0.409 / 36.53	0.988 / 80.00

At transport rate .15, corrected retention is .978 versus nominal .834, but mean cost rises from 11.34 to 62.75. At rate .30, correction uses the full budget and leaves mean target width 1.603 out of the initial 2; bounded-error evidence leaves width .294 at cost 79.43. A valid transport bound can therefore preserve coverage while leaving the target unresolved. At zero transport, nominal, corrected, and known-rate paths are identical by construction and by record comparison. The negative control quantifies retention loss when the supplied bound understates physical drift.

Every physical flip, sensor bit, query, retained set, and stopping decision is recorded for all 24,000 paths. A separate direct-product calculation replays the simulator's log-space updates and recomputes all 24 summary rows. An ordinary 95% Hoeffding rate interval has half-width .043; full paired comparisons and all widths are retained in the artifact. These sensitivity experiments condition on the supplied transport bounds in Theorem D.4.

F.4 NON-ORACLE INFERENCE ON ONE CONNECTED NETWORK

Design and observed data. A cycle of $J = n/b$ blocks has b units per block and all edges between adjacent blocks. This is one connected network; each block receives an independent Bernoulli$(1/2)$ assignment U_j. At a unit in block j, the six candidate exposures are the left signal U_{j-1}, the right signal U_{j+1}, their OR, AND, XOR, and XNOR. Indices are cyclic. The corresponding exposure probabilities are $(1/2, 1/2, 3/4, 1/4, 1/2, 1/2)$, obtained from the four local assignments. For unit $i = 0, \dots, n-1$, the finite population is

$$Y_i^E(0) = .25 + .10\cos(2\pi i/n) + .02\sin(6\pi i/n), \qquad Y_i^E(1) - Y_i^E(0) = \bar{\tau} + .04\sin(4\pi i/n).$$

The inference inputs are the randomized assignments, observed block-mean outcomes, the known range $[0, 1]$, and constant augmentation $m_{ie} = 1/2$. Band construction and query selection use only these observed and design inputs.

Dependence-aware calibration. All unit scores in block j depend only on (U_{j-1}, U_{j+1}). Blocks at cyclic distance two share an assignment. For J divisible by four, the color $b(\lfloor j/2 \rfloor \bmod 2) + r$ for within-block index $r = 0, \dots, b-1$ gives $2b$ colors with disjoint score supports inside each color. Thus Theorem C.2, with $R = 1/2$, gives half-width

$$r_g(a) = \frac{1}{2\underline{\pi}_g}\sqrt{\frac{4b\log(2/a)}{n}} = \frac{1}{\underline{\pi}_g}\sqrt{\frac{\log(2/a)}{J}}.$$

Intervals are intersected with the known target range $[-1, 1]$. The modelwise rule uses $a = \alpha = .05$ and Proposition C.5; the simultaneous comparison uses $a = \alpha/6$. This applies the dependence bound to one main network using design and outcome-support information. The independence control deliberately substitutes one color for $2b$.

Adaptive acquisition and comparisons. The four queries set the two neighboring block assignments to $00, 01, 10, 11$; all other block assignments are zero. Each query returns a fresh unit-cost BSC$(.18)$ label. Uniform mixture weights, $\delta = .05$, and the running maximum of the likelihood evidence define the retained set. Target acquisition minimizes the worst-child envelope width, with count balance on zero-gain plateaus. Count acquisition and fixed cycling through the four queries use the same modelwise intervals. Each policy recomputes its queries after every label. Auditing stops at width .20, a singleton retained set, model failure, or 30 labels. Policies share the main data and query-indexed label noise.

The grid crosses $n \in \{256, 1024, 4096\}$ and $b \in \{1, 4, 64\}$ at $\bar{\tau} = .30$; two weak-effect configurations use $n = 4096$, $b \in \{1, 64\}$, and $\bar{\tau} = .05$. Each configuration balances 600 trials across the six mechanisms. All 55 method/configuration summaries and paired contrasts are provided with the raw observations and paths.

Table 12: Connected-network non-oracle results, 600 paired trials per configuration. Width averages nonempty reports; coverage counts every trial. The supplied summaries also report model failures and Monte Carlo intervals. "Simultaneous" changes only the candidate error allocation. "Indep. control" uses an incorrect independence assumption; the other rows use the certified $2b$-color calibration.

n	b	$\bar{\tau}$	Rule	Width	Path coverage	Labels
4,096	1	0.30	Target	0.161	0.983	16.10
4,096	1	0.30	Count	0.160	0.987	14.05
4,096	1	0.30	Fixed cycle	0.212	0.987	24.20
4,096	1	0.30	Simultaneous	0.195	0.982	15.85
4,096	4	0.30	Target	0.324	0.988	16.02
4,096	64	0.30	Target	1.196	1.000	13.78
256	64	0.30	Target	2.000	1.000	13.83
256	64	0.30	Indep. control	0.452	0.737	14.84
4,096	64	0.05	Target	1.271	1.000	14.19
4,096	64	0.05	Indep. control	0.166	0.693	6.87

Precision, dependence, and query design. At $n = 4096$, $b = 1$, target acquisition gives mean width .161, compared with .195 under simultaneous calibration and .212 under fixed queries (Table 12).

Count acquisition has width .160 at 14.05 labels versus 16.10 for target acquisition. Increasing b from 1 to 64 at the same population size widens the valid report from .161 to 1.196. At $n = 256, b = 64$, it equals the full support diameter 2. In this family, with fixed exposure probabilities, $J = n/b$ controls the certified radius. Other designs also change overlap and score dependence; J alone is not a universal precision formula. Ignoring that dependence gives path coverage .737 at $n = 256$, $b = 64$ and .693 in the weak-effect $n = 4096$, $b = 64$ control. With $b = 1$ and the weak effect, valid target acquisition has width .161 but makes a positive decision in only 5.5% of trials.

Record checks. An independent implementation checks the graph coloring, all observed scores, every chosen query and label, likelihood rejection, stopping, and reported hulls, then recomputes the summaries and paired differences. Equal endpoints and identical query histories are checked separately. On configurations with zero width gain throughout, target acquisition uses its count fallback, so coincident policy paths are expected. The 95% Monte Carlo coverage intervals use a familywise Hoeffding bound over the 55 rows and allow unequal truth-specific success probabilities. These experiments specify a structured connected-graph family; the radius explicitly retains its dependence factor.

F.5 Matched-budget passive and fixed-horizon comparisons

Information and resource matching. We use every main-data realization and query-indexed label table from Appendix F.4: eleven configurations and 600 trials per configuration. The candidate intervals, $\alpha = \delta = .05$, BSC$(.18)$, and four-query menu are unchanged. Active methods take exactly $H = 30$ unit-cost labels, including after a singleton or precision threshold is reached. Passive inference takes no new labels. Target queries use the nested likelihood rule for acquisition, with count balance on plateaus and fixed cycling after no informative split remains. Three reporting rules share that entire target-query path; a fourth uses fixed cycling from the start. The experiment separates added measurements, query choice, and reporting calibration under the same causal target and main assumptions.

Conventional controls. The passive union retains every supplied candidate. Fixed-horizon inference reports only at H, using $\mathcal{V}_H^{\mathrm{term}} = \{g : \mathsf{E}_H(g) < 1/\delta\}$, either on fixed or target queries. The Bonferroni union predeclares H looks and uses $\mathcal{V}_t^{\mathrm{Bonf}} = \{g : \max_{s \le t} \mathsf{E}_s(g) < H/\delta\}$. The anytime rule uses $\mathcal{V}_t$ from (4). The multiplicity correction is over inspection times, not over mutually exclusive true candidates. All three active report constructions have causal coverage at least $1 - \alpha - \delta$ at their designated reporting times; the passive union has coverage at least $1 - \alpha$.

Proposition F.1 (Fixed-horizon comparison rules)**.** *Under the main assumptions and true-candidate band coverage $1 - \alpha$, the terminal rule is valid at prespecified H, and the Bonferroni rule is valid simultaneously for $0 \le t \le H$. On a common label path,*

$$\mathcal{V}_H \subseteq \mathcal{V}_H^{\mathrm{term}}, \qquad \mathcal{V}_H \subseteq \mathcal{V}_H^{\mathrm{Bonf}}.$$

Thus a nonempty anytime envelope is no wider than either comparator on that path. This inclusion concerns these particular likelihood rules; empty reports remain coverage failures.

Proof. For the true candidate h, $\mathbb{E}_h \mathsf{E}_s(h) \le 1$ under any predictable query rule. Markov's inequality at H bounds terminal rejection by δ. At each of H looks it bounds rejection at threshold H/δ by δ/H; a union bound gives the Bonferroni claim. Combining with the true-candidate band gives causal coverage. A running maximum exceeds its terminal value, and $H/\delta \ge 1/\delta$, which proves both inclusions. □

At $n = 4096, b = 1$, width changes from 0.591 for the passive union to 0.228 for fixed-query terminal inference. On the shared target-query path, terminal, Bonferroni, and anytime widths are 0.194, 0.244, and 0.161. The anytime rule's empirical coverage is 0.983, while the more conservative target-terminal and Bonferroni rows cover all 600 trials. Set inclusion predicts this ordering; the experiment quantifies its magnitude. The $b = 64$ rows remain much wider under every reporting rule: additional labels resolve mechanism uncertainty, not the main-design information deficit. An independent probability-domain replay checks all queries, answers, retained sets, 33,000 method–trial records, and 627,000 reported intervals, with terminal-only entries left undefined before H. Paired normal Monte Carlo intervals use Bonferroni correction across all 176 configured contrasts; simultaneous Hoeffding intervals for the empirical coverage rows are also supplied.

Table 13: Matched-budget controls on the connected network, 600 paired trials per configuration. Every active row uses all 30 labels. Width averages nonempty terminal reports; all trials enter coverage. A dash means the method reports only at H, so path coverage is not assigned. All eleven configurations, failures, and paired Monte Carlo intervals are provided in the artifact.

Method	Labels	Width	Terminal coverage	Path coverage
$n = 4096$, $b = 1$, $\bar{\tau} = .30$				
Passive union	0	0.591	1.000	1.000
Fixed queries, terminal	30	0.228	0.998	—
Target queries, terminal	30	0.194	1.000	—
Target queries, Bonferroni	30	0.244	1.000	1.000
Target queries, anytime	30	0.161	0.983	0.983
$n = 4096$, $b = 64$, $\bar{\tau} = .30$				
Passive union	0	1.877	1.000	1.000
Fixed queries, terminal	30	1.391	1.000	—
Target queries, terminal	30	1.256	1.000	—
Target queries, Bonferroni	30	1.386	1.000	1.000
Target queries, anytime	30	1.197	0.997	0.997

These controls propagate the same finite candidate uncertainty and physical-effect estimand. The propensity-sensitivity bounds of Schröder et al. (2026) concern direct and spillover contrasts under a different uncertainty model. A width ranking against those bounds would therefore change the estimand or assumptions as well as the information collected. The passive union is the matching fixed-information reference for the experiment here.

Enumeration and computational scaling. Let $Q = |\mathcal{Q}_n|$. A precomputed binary prediction matrix costs QM_n bits in principle, or QM_n bytes in the supplied uint8 implementation. Direct scoring costs $O(QM_n)$ per decision; a label update costs $O(M_n)$. Mechanisms with identical complete query signatures can share likelihood updates, summing their original mixture weights and retaining the extrema of their candidate intervals. This compression preserves the envelope but does not establish that the omitted mechanisms are physically adequate.

Table 14: Single-thread precomputed-signature kernel benchmark at $Q = 256$. Times are medians of 15 repetitions after compilation and warm-up; full timing distributions and the software environment are recorded. Signature memory excludes interval and likelihood arrays.

M	Q	Signature MiB	Scoring ms	Update ms
16	256	0.004	0.014	0.001
64	256	0.016	0.025	0.002
256	256	0.062	0.087	0.005
1024	256	0.250	0.312	0.015
4096	256	1.000	1.313	0.069

The benchmark crosses $M_n \in \{16, 64, 256, 1024, 4096\}$ and $Q \in \{16, 64, 256\}$. Independently computed scores and likelihood updates agree at every configuration. This isolates enumeration after prediction precomputation. Exact exposure probabilities, design integration, and a valid dependence bound are separate costs: the 70-node study enumerates 2^{14} assignments, whereas the connected-cycle study uses four local assignment patterns. Extending the former enumeration to unrestricted large networks would remain exponential. Tractable local designs or certified probability bounds are needed before the reported kernel costs become relevant.

F.6 Real-outcome boundary audit: Twitter

The original study distinguishes potential tweet visibility from measured receipt (Coppock et al., 2016, p. 125). The calculations below quantify ambiguity using its recorded opportunity and reshare variables.

Source and sample. Study 2 of Coppock et al. (2016) has a public CC0 replication archive on Harvard Dataverse (Coppock et al., 2015), version 1.0. Its source sample has 8,507 accounts; the strict-overlap subset contains 1,975 with $0 < p_i < 1$ and complete fields. Our aggregate calculations use the supplied weighted (Z, D, Y) cells for this subset. The artifact's source manifest records the six source files and their hashes, including the account-level input.

The archive's `exposure` indicator means that an account followed at least one signer randomly encouraged to tweet: an assigned opportunity. `actual_exposure` means following at least one signer who eventually tweeted: a realized-reshare proxy. Physical receipt is latent: the archive contains neither delivery timestamps nor impression or reading telemetry. It supplies follower-based indicators rather than raw follower edges.

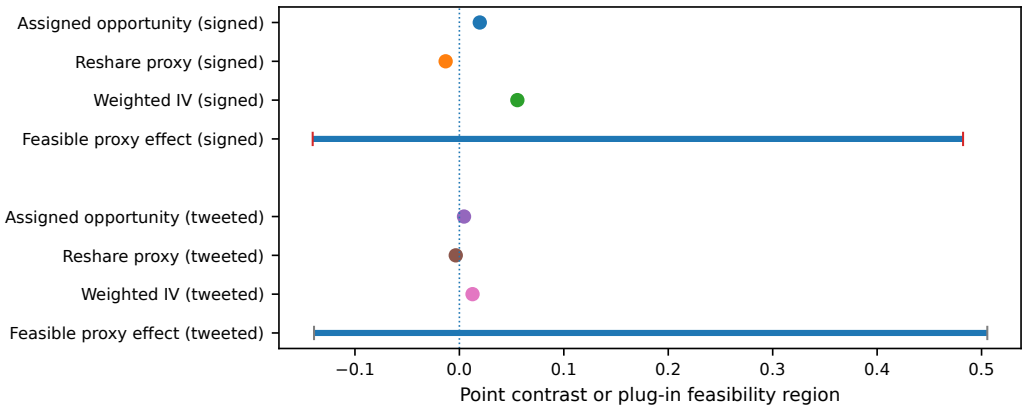


Figure 11: **Real-outcome boundary from weighted response cells.** Points are contrasts computed from recorded opportunity and resharing. Segments are binary-IV plug-in feasibility regions under exclusion and no monotonicity. The segments quantify identification ambiguity conditional on the estimated cells; they are not sampling confidence intervals.

Table 15: Response-cell contrasts and plug-in identification regions. Each value is recomputed from the included weighted cells.

Representation	Petition signed	Petition reshared
Assigned-opportunity Hájek contrast	0.0195	0.0044
Realized-reshare proxy; descriptive	−0.0132	−0.0035
Weighted IV coefficient; assumption-sensitive	0.0555	0.0126
Binary-IV proxy effect; plug-in region	$[-0.1405, 0.4823]$	$[-0.1392, 0.5055]$

Estimands and assumptions. The supplied inverse-probability weights give a first-stage Hájek contrast of 0.3517. The opportunity contrast and the descriptive reshare contrast are weighted differences; dividing the opportunity contrast by the first stage gives the just-identified IV coefficient. A causal LATE reading of the IV coefficient further requires instrument independence, exclusion, relevance, and monotonicity (Imbens & Angrist, 1994). The last row uses the binary response-type linear-programming formulation of Balke & Pearl (1997), enumerating all sixteen types under exclusion and a common response-type distribution without monotonicity. Its extrema condition on the estimated cells; sampling uncertainty is not included. Identity, number, timing, and other-neighbor assignment paths are suppressed by that binary model. Receipt-based analysis would additionally require an exposure measurement protocol and channel calibration.

Exclusion sensitivity from response cells. Let x_u be the population mass of binary response type

$$u = (D(0), D(1), Y(0,0), Y(0,1), Y(1,0), Y(1,1)) \in \{0,1\}^6.$$

Instrument independence is represented by one type distribution across Z. The observed conditional cells impose

$$\sum_u x_u \mathbf{1}\{D_u(z) = d,\ Y_u(z,d) = y\} = \widehat{p}(d, y \mid z), \qquad x_u \geq 0, \quad \sum_u x_u = 1.$$

Let $v_u = \mathbf{1}\{Y_u(0,0) \neq Y_u(1,0) \text{ or } Y_u(0,1) \neq Y_u(1,1)\}$. For a sensitivity level ν, add $\sum_u v_u x_u \leq \nu$ and minimize/maximize $\sum_u x_u\{Y_u(0,1) - Y_u(0,0)\}$. This defines a proxy effect at reference assignment $z = 0$. At $\nu = 0$, the 64-type formulation reduces to the 16-type exclusion model; no monotonicity constraint is imposed. Increasing ν nests the feasible sets, so the lower bound cannot increase and the upper bound cannot decrease.

We solve both extrema for both outcomes at $\nu \in \{0, .01, .025, .05, .10, .20\}$. All 24 optimizations reproduce their observed cells, satisfy primal–dual gaps below 1.2×10^{-16}, and exhibit maximum cell residual below 2.8×10^{-15}. At $\nu = .05$, the signing region is $[-.1473, .5450]$ and the resharing region is $[-.1418, .5573]$. The feasible regions condition on the supplied cell estimates.

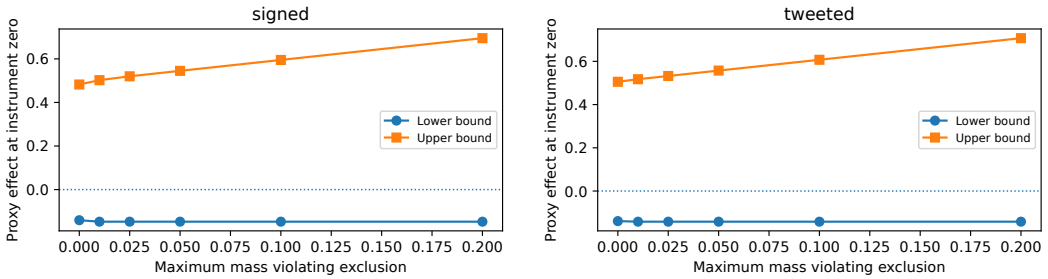


Figure 12: Response-type exclusion sensitivity. The shaded regions are feasible proxy effects at reference assignment $z = 0$, obtained from the supplied conditional cells. The horizontal axis bounds the fraction of types violating exclusion.

Reproduction. The aggregate-audit script recomputes contrasts, solves the response-type programs, saves primal/dual diagnostics, and compares two executions. Script locations, input hashes, and environment requirements are listed in the artifact's reproduction guide.

F.7 Published-design overlap and data requirements

Assignment, measurement, and analysis population. Jones et al. (2017) report a randomized Facebook intervention with administratively matched turnout for 15,060,897 individuals. Their Table 1 assigns 96% to both banner and feed, 2% to banner only, 1% to feed only, and 1% to neither, giving marginal banner probability .98. The all-matched population is the design-based starting point; restricting to post-assignment platform use requires additional justification. The calculation below uses those published design probabilities and outcome support, with hypothetical degree configurations.

Overlap before scale. Under independent subject-level banner draws, an exposure requiring at least k of d friends to be assigned has

$$\pi_{d,k}(1) = \sum_{j=k}^{d} \binom{d}{j} p^j (1-p)^{d-j}.$$

For $p = .98, d = 5, k = 1$, $\pi_{d,k}(0) = .02^5 = 3.2 \times 10^{-9}$. A hypothetical homogeneous population of the reported size would then have only .0482 unexposed units in expectation. With $Y \in [0,1]$, augmentation $1/2$, and candidate level $a = .05$, the optimistic one-color version of Theorem C.2 gives

$$r^{\text{opt}}_{d,k} = \frac{1}{2\underline{\pi}_{d,k}} \sqrt{\frac{2\log(2/a)}{15{,}060{,}897}}.$$

Table 16 varies only the exposure definition at fixed sample size and assignment probability. These one-color values are optimistic design diagnostics for this particular band; network-specific calibration requires a valid dependence factor, which can only increase them.

Table 16: Published-design overlap diagnostic at $p = .98$. Degrees and exposure thresholds are hypothetical; probabilities are calculated exactly.

d	k	$\underline{\pi}_{d,k}$	Expected rarer exposure	$r^{\mathrm{opt}}_{d,k}$
1	1	0.02	3.012×10^5	0.0175
2	1	0.0004	6024	0.8749
5	1	3.200×10^{-9}	0.04819	1.094×10^5
5	3	7.762×10^{-5}	1169	4.509
5	5	0.09608	1.447×10^6	0.003642

Why assignment summaries do not replace topology. Two connected three-regular graphs on eight vertices—the cube and the Möbius ladder—share degrees and treated-neighbor counts $(1, 1, 2, 2, 1, 1, 2, 2)$ at assignment $(0, 0, 1, 1, 0, 0, 1, 1)$. Under independent Bernoulli$(1/2)$ randomization, their any-treated-neighbor exposures both have marginal probability $7/8$. For constant outcomes one and augmentation $1/2$, exact enumeration of all 256 assignments nevertheless gives variances $32/49$ and $30/49$ for the average candidate score. Thus even identical marginal exposure probabilities and observed neighbor counts need not identify the joint design law needed for variance calculation. The artifact supplies both edge sets, assignments, scores, and an independent verifier.

Measurement needed for an empirical audit loop. Recomputing candidate bands requires the randomization law and score-dependence information, as well as outcomes; an edge list or a certified alternative can supply the latter. Physical-exposure audits additionally require labels and conditional sensor/transport calibration. In a separate experiment, Bakshy et al. (2012) randomized subject–URL feed display and logged feed exposures, censored exposures, and clicks. Display is an observed delivery event; cognitive attention is a different exposure. Such logs can anchor a display-based candidate class when their assignment support and calibration match the proposed audit protocol. Retrospective policy evaluation further requires known logging probabilities for the queried actions.

Execution. Using published design probabilities and explicit finite constructions, the artifact checks 18 overlap configurations, 512 graph–assignment rows, and $4{,}096$ omitted-truth label paths (Example A.1). These records assess design feasibility and candidate scope.