LEARNING AFTER STRUCTURAL POISONING: LABEL COVERAGE AND LABEL TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Structural poisoning can erase target information without destroying realizability; labels can recover this information, yet training may fail to transfer it. On a sparse-tree family, we derive an exact coverage-based minimax law and a depth-dependent capacity law. A separate connected, bounded-degree construction preserves the realizing teacher under legal edge additions and has 20% cross-block neighbor-message mass at receivers. On this family, we prove a nonvanishing expected excess-risk lower bound over the information reference for finite-width, one-hop nonlinear graph neural networks with shared representations and trainable biases. The bound holds for GD and Adam at every fixed training horizon under the specified independent initialization and equally weighted hit-role logistic loss. A permutation symmetry obstructs label transfer even when cross-role logit-gradient inner products remain strictly positive. In a linear specialization, a positive-definite gradient preconditioner attains the information reference in one step without changing the function class, loss, or initial predictor. In a matched nonlinear study on the connected family, coupled Adam updates reduce overall test error by 10.90 percentage points across 96 independent problems, with all observed labels fitted. These results establish a separation between label coverage and label transfer, and show how optimizer design can improve transfer within a fixed model class.

1 INTRODUCTION

An edge edit can change many message-passing representations. That observation alone does not say how many labels become statistically less useful. When a known feature rule determines targets, graph edits may leave prediction unaffected. When clean structure encodes targets, an attack can instead make distinct clean worlds observationally identical. Learning then depends on which of those worlds sampled labels can distinguish.

Our main experiment removes one observable sufficient statistic: bridge presence encodes a block's common label, and adding the absent bridges erases that signal. Each block remains representable after the attack, so labels are needed to resolve information loss rather than a failure of the target class. A sampled label recovers the entire block. The exact risk is half the expected uncovered mass. Optimizing the legal block size gives $\mathfrak{R}_L^\star \asymp \min\{1, bQ_L/N, b/m\}$, where $b = \min\{B, d\}$ and Q_L is the depth-dependent construction capacity. This separates capacity, coverage and packing limits and yields label complexity $\Theta(b/\varepsilon)$ in a specified coverage-limited regime.

Our central question is what makes an identified label transfer. A second, connected construction keeps degree bounded and makes 20% of each receiver's neighbor message come from other blocks. Its clean and poisoned targets have the same realizing teacher. We prove that a finite-width one-hop nonlinear GNN, including a shared head and trainable biases, has a nonvanishing expected excess over the information reference. A permutation symmetry prevents it from assigning some observed partner labels to their corresponding inputs, even with strictly positive cross-role gradient interactions. The bound holds at every fixed training horizon under the stated label-independent initialization distribution. In the linear specialization, a coupled gradient update removes the gap in one step with the same function class, loss and initial predictor.

In a fresh shared-head nonlinear study, the matched update intervention reduces overall error by 10.90 percentage points across 96 independent problems (95% interval [10.09, 11.75]), with every observed

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.

label fitted. Separate experiments compare stacked GraphSAGE and test a sufficient architectural transfer condition on the sparse-tree family. Our main contributions are as follows.

- **Information and capacity.** A legal hiding construction yields an exact minimax reference, a depth-dependent capacity law, and label-complexity bounds in the coverage-limited regime (Sections 4–5).
- **Connected trained failure and repair.** A legal connected family supports a nonlinear training lower bound and an exact same-class linear optimizer remedy (Theorem 6.1).
- **Matched nonlinear intervention.** A frozen-protocol study tests the optimizer intervention in the shared-head model; separate experiments test a sufficient architectural transfer condition (Section 7).
- **Refined information laws.** Exact deterministic and noisy-label risks and a separate budget-constrained relational bound distinguish prediction, supervision and latent dependence (Section 8).

The paper connects an exact information reference under legal, teacher-preserving structural edits to the behavior of specified training procedures. Its connected construction supports a finite-width nonlinear training lower bound despite shared parameters and positive logit-gradient overlap, together with an exact same-class linear optimizer intervention.

2 Related work

Transductive and graph learning theory. Transductive complexity (El-Yaniv & Pechyony, 2009), realizable minimax bounds (Tolstikhin & Lopez-Paz, 2016), and robust transduction under shared corrupted observations (Montasser et al., 2022) are established. Unlabeled structure can help when linked to labels (Ben-David et al., 2008; Göpfert et al., 2019), while GNN generalization and representation are studied by Garg et al. (2020); Xu et al. (2019). Whereas Ghasemi & Pishro-Nik (2026) study a squared-error spectral setting, our binary experiment characterizes observational equivalence induced by a legal structural attack. Smoothing and communication constraints are studied by Oono & Suzuki (2020); Alon & Yahav (2021).

Label use and trained-model generalization. Algorithmic alignment connects architecture to sample complexity (Xu et al., 2020). Feature and label propagation have distinct influence properties (Wang & Leskovec, 2020); Correct and Smooth uses label-based post-processing (Huang et al., 2021). Jia & Benson (2022), Sections 2.4–2.5, distinguish conditioning on features from conditioning also on observed labels in a Gaussian MRF and derive residual correction. Our binary legal-poisoning experiment gives an exact finite information reference and quantifies the excess risk of specified training procedures. Yang et al. (2024), Sections 3–5, study function-space gradient dynamics and kernel–graph alignment, including an overparameterized regime. Bechler-Speicher et al. (2024) show that gradient descent can prefer graph-dependent solutions even when ignoring the graph is better. Theorem 6.1 establishes an information–training separation on a connected graph with shared finite-width nonlinear representations and gives a same-class linear optimizer intervention. The analysis builds on implicit bias and symmetry-based training lower bounds (Soudry et al., 2018; Abbe & Boix-Adsera, 2022; Lahoti et al., 2024); Appendix D develops complementary linear examples. Cai et al. (2021) obtain transfer guarantees under expansion and consistency assumptions. For homogeneous rank-one blocks, our invariant gives a sufficient transfer condition that we test separately from the linear dynamics.

Structural attacks, certification, and semantics. Graph attacks and defenses include training-time perturbations, scalable attacks, certified training and refinement (Dai et al., 2018; Zügner et al., 2018; Zügner & Günnemann, 2019b; Bojchevski & Günnemann, 2019a; Geisler et al., 2021; Zügner & Günnemann, 2019a; Jin et al., 2020). Certification treats local/global budgets (Bojchevski & Günnemann, 2019b) and node-feature poisoning (Gosch et al., 2025), with objectives distinct from our latent-world minimax experiment. Reference-based semantics distinguish label preservation from observable semantic content (Gosch et al., 2023). Our attack model preserves each world's targets while making distinct clean worlds observationally identical after release.

Readout agreement and supporting relational analysis. Version-space agreement (El-Yaniv & Wiener, 2010) and conic separation (Boyd & Vandenberghe, 2004) underlie the frozen-representation

audit (Proposition B.1); Lipschitz perturbation bounds underlie its fixed-weight stability counterpart. The relational analysis uses signed-path inference, cardinality counting and cut arguments (Cesa-Bianchi et al., 2012; Tarlow et al., 2012; Abbe et al., 2014; Chen et al., 2016; Foster et al., 2018); Appendix K.1 compares their precise protocols.

3 TRANSDUCTIVE STRUCTURAL-COVERAGE MODEL

Protocol and risk. Let $N \geq 2$, $B, d \geq 0$, $\Delta \geq 2$, $L \geq 1$, and $m \in \{0, \ldots, N-1\}$ be integers. Here B is the total edge-addition budget, d is the feature dimension, and Δ bounds the maximum degree of both clean and released graphs. In the constructions below, each vertex is incident to at most one added edge. Let $V = [N]$ and $X \in \mathbb{R}^{N \times d}$. A labeled set S is sampled uniformly from all m-subsets of V; $U = V \setminus S$ and $u = N - m$. The learner receives features X, the attacked graph $\widetilde{G}$, and $(Y_i)_{i \in S}$, then predicts the clean targets on U. For randomized decision rules $\mathcal{A}$, risk is expected test error $R(\mathcal{A}) = \mathbb{E}\left[\frac{1}{u}\sum_{i \in U} \mathbf{1}\{\widehat{Y}_i \neq Y_i\}\right]$. For a fixed clean world, the expectation is over the sampled label set and the learner's private randomness. When a hidden-bit prior is introduced below, we explicitly average over that prior as well. The minimax risk takes the infimum over $\mathcal{A}$ and the supremum over the model class described below. The minimax reference allows unrestricted learners, while depth L constrains the realizable target family. Section 6 separately analyzes specified GNN training procedures. Randomized rules define the main minimax value; Proposition E.3 treats deterministic prediction.

Depth-L graph-mediated teacher. Let A be the binary adjacency matrix of a simple undirected graph and $D_\lambda = \mathrm{diag}(\deg_A(v) + \lambda : v \in V)$. Let $P_\lambda(A) = D_\lambda^{-1}(A + \lambda I)$ be row-normalized propagation with $0 < \lambda \leq 1$ fixed and known, and $Z_\lambda^{(L)} = P_\lambda(A)^L X$. Binary clean targets are $Y_v = \mathbf{1}\{(Z_\lambda^{(L)} w)_v > 0\}$. The construction uses K disjoint recognizable blocks $C_1, \ldots, C_K$, each of size Q, plus known-zero filler nodes. Block k carries a bit $\theta_k \in \{0, 1\}$ and $Y_v = \theta_k$ for $v \in C_k$.

To realize the bit, C_k contains a payload vertex p_k with feature e_{j_k} and a pruned $(\Delta - 1)$-ary tree rooted at r_k. The clean bridge $\{p_k, r_k\}$ is present iff $\theta_k = 1$. For $K > 0$, set

$$w^\theta = K^{-1/2} \sum_{k=1}^{K} \theta_k e_{j_k}, \qquad \|w^\theta\|_2 \leq 1.$$

For $K = 0$, use the all-zero target and readout. For $K > 0$, nonnegative propagation makes every node in the block positive exactly when $\theta_k = 1$. Negative active nodes have score zero and are classified by the zero-to-negative threshold convention; this sparse realization therefore has no positive uniform signed margin. Every integer block size up to

$$Q_L = 1 + \sum_{t=0}^{L-1} (\Delta - 1)^t \tag{1}$$

is realized by pruning the tree while keeping depth at most $L - 1$.

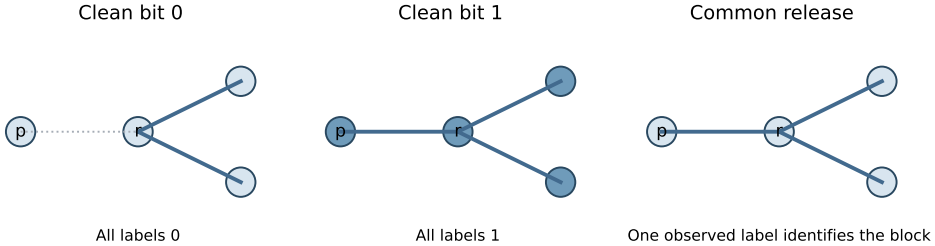

Figure 1: Bridge completion erases the structural bit. The same released block is compatible with either common label; observing one of its labels identifies all its targets, but a trained predictor need not transfer that information.

Canonicalizing attack. The attacker adds each missing payload–root bridge. All clean bit vectors now map to one common released graph in which every bridge is present. At most one edge is added per block; bridge endpoints are disjoint, so local edit incidence is one. The rule is fixed and applied before the random label set S is drawn; it observes neither S nor any labels. The same w^θ still realizes the clean targets on the released graph: a zero bit uses a zero readout coordinate even though its bridge was added. The readout is unknown to the learner and may differ between clean worlds; within each world, it is unchanged by the attack. Thus the ambiguity is statistical rather than a failure of realizability on the released graph. Without attack, bridge presence reveals every θ_k, so a learner recovers all targets from the clean graph with zero risk even when $m = 0$. The positive label requirement is therefore created by canonicalization rather than assumed in the clean problem.

Formal family and observable parser. For each feasible K, Q, choose disjoint sets $C_k \subseteq [N]$ of size Q, a payload $p_k \in C_k$, and a rooted tree on $C_k \setminus \{p_k\}$ obtained by pruning the depth-$(L-1)$, $(\Delta-1)$-ary tree. Assign distinct coordinates $j_k \in [d]$, put $X_{p_k} = e_{j_k}$, and set all other rows to zero. Vertices outside the blocks are isolated fillers with target zero. The clean edge set consists exactly of the tree edges and bridges with $\theta_k = 1$; the released edge set contains every bridge. Allow every such labeled layout, coordinate assignment, pruned shape, and $\theta \in \{0,1\}^K$ satisfying $K \le \min\{B, d, \lfloor N/Q \rfloor\}$ and $2 \le Q \le \min\{Q_L, N\}$. Together with the empty, all-zero instance, these tuples define $\mathcal{P}_L$; the attack is the specified completion map. Its nontrivial released connected components are exactly the C_k, each containing one nonzero feature row. Connected-component traversal therefore recovers the blocks and their sizes without an oracle or knowledge of θ.

Scope of the exact law. Any ambient legal-poisoning class containing $\mathcal{P}_L$ inherits the lower bound. The matching decoder uses the observable homogeneous blocks of $\mathcal{P}_L$. Appendix V gives the exact heterogeneous-layout optimization within the same construction.

4 Exact minimax risk

For fixed Q, define $p_0(Q) = \Pr(S \cap C_k = \varnothing) = \frac{\binom{N-Q}{m}}{\binom{N}{m}}$, the exact finite-population probability that a block receives no label. The model class $\mathcal{P}_L$ is the explicitly parsed union in Section 3.

Theorem 4.1 (Exact optimized structural-coverage law)**.** *On $\mathcal{P}_L$, the randomized minimax test risk is*

$$\boxed{\mathfrak{R}_L^\star = \max_{2 \le Q \le \min\{Q_L, N\}} \frac{Q \min\{B, d, \lfloor N/Q \rfloor\}}{2(N-m)} \frac{\binom{N-Q}{m}}{\binom{N}{m}}.} \tag{2}$$

If $Q_L < 2$, $B = 0$, or $d = 0$, the attack-induced minimax risk is zero.

Proof sketch. Fix Q and take $K_Q = \min\{B, d, \lfloor N/Q \rfloor\}$. Put independent fair priors on the K_Q block bits. Once the attack canonicalizes the bridges, the released graph and features are independent of every θ_k. If S misses C_k, the full observation is also independent of θ_k, so its posterior remains fair and the Bayes expected mistake count on that block is $Q/2$. Linearity of expectation gives the lower bound $K_Q Q p_0(Q)/(2u)$; maximizing over Q lower bounds the union.

For the upper bound, a single decoder works on the entire union because the blocks are recognizable. It copies any observed label across a hit block and makes a fair blockwise guess when the block is missed; filler nodes receive their known label. For every fixed bit vector, its expected number of errors is $Q/2$ on each missed block, where the expectation is over the decoder's fair guess. Hence its worst-case value equals the lower bound for the realized Q, and taking the worst union member proves equality. Appendix E gives the complete proof.

The same argument immediately handles heterogeneous recognizable blocks:

$$\mathfrak{R}^\star_{\{Q_k\}} = \frac{1}{2u} \sum_{k=1}^{K} Q_k \frac{\binom{N-Q_k}{m}}{\binom{N}{m}} = \frac{1}{2} \mathcal{C}_m(\{Q_k\}),$$

where $\mathcal{C}_m$ is the probability-weighted uncovered block mass. Equation (2) is the optimized homogeneous case. For this family, the sufficient statistic for expected risk is the probability-weighted mass of blocks missed by supervision.

5 DEPTH-DEPENDENT CONSTRUCTION CAPACITY

Write $b = \min\{B, d\}$. The exact equality exposes three competing scales: the largest block reachable at depth L, the coverage scale N/m, and the packing scale N/b.

Corollary 5.1 (Phase law)**.** *In the principal transductive regime $1 \le m \le N/4$, $1 \le b \le N/4$, $Q_L \ge 2$, and under the integer feasibility of the bridge construction,*

$$\boxed{\mathfrak{R}_L^\star \asymp \min\left\{1, \frac{bQ_L}{N}, \frac{b}{m}\right\}.} \tag{3}$$

The comparison constants may be chosen as $1/16$ and $2/3$ (lower and upper, respectively).

The upper bound uses $p_0(Q) \le e^{-mQ/N}$; the lower bound chooses Q within a constant factor of $\min\{Q_L, N/m, N/b\}$ so that $K_Q = b$ and $p_0(Q) \ge 1/2$ (Appendix E). For $\Delta > 2$, capacity Q_L grows exponentially with L. Risk grows as bQ_L/N while depth limits block size, saturates at b/m when coverage limits it, and can remain of constant order when $b \gtrsim m$ permits packing a constant fraction of the population.

Corollary 5.2 (Coverage-limited label complexity)**.** *Define the principal-regime threshold $m_L^\star(\varepsilon) = \min\{m \in \{1, \dots, \lfloor N/4 \rfloor\} : \mathfrak{R}_L^\star(m) \le \varepsilon\}$. If $0 < \varepsilon \le 1/64$, $1 \le b \le \varepsilon N/16$, $Q_L \ge 32\varepsilon N/b$, and the integer bridge construction is feasible, then*

$$\frac{b}{32\varepsilon} \le m_L^\star(\varepsilon) \le \frac{2b}{\varepsilon}.$$

These laws compare admissible target families as depth varies. Corollary 5.2 concerns the positive-label, coverage-limited regime; when $bQ_L/(2N) \le \varepsilon$, zero labels already suffice in the unrestricted experiment. Appendix W illustrates the capacity and packing regimes, and Proposition V.2 analyzes dense feature support.

6 WHEN IDENTIFIED LABELS TRANSFER

Under independent fair block bits, any learner using the permitted transcript has error probability $1/2$ on each missed active test vertex. With $\mathcal{M}$ the missed active test set,

$$\mathbb{E}_\theta R(\mathcal{A}) - \frac{\mathbb{E}|\mathcal{M}|}{2u} = \mathbb{E}[\text{covered and filler contributions to overall test error}].$$

Proposition V.1 is an identity averaged over independent fair block bits; it also holds for bit-independent stratified sampling of fixed size with the corresponding coverage reference. We first bound this gap for specified learners, then give a sufficient transfer condition.

6.1 CONNECTED GRAPHS: INFORMATION AVAILABLE TO TRAINING

We now use a separate connected family in the same ambient addition model, at teacher depth one, with edit and feature budgets at least K. A simple loop-free directed routing graph on K indices has in- and out-degree $r \ge 1$ and connected underlying graph. Block i has $a \ge 4r$ sources P_i with feature e_i and $b \ge 2$ receivers R_i with feature zero; $N = K(a + b)$. Include the undirected edges $P_i \times R_i$ and, for each routing arc $j \to i$, edges from one designated source in P_j to all of R_i. One designated internal edge is absent iff the block bit is zero. Adding the missing edges erases every bit using at most K additions, with local incidence one. Both clean and released graphs are connected and have degree at most $\max\{a + r, b(r + 1)\}$. The same unit-norm signed readout realizes all block labels before and after attack. Membership is observable from features and the dominant coordinate of each receiver's neighbor mean.

Let $p_m(s) = \binom{N-s}{m}/\binom{N}{m}$ and $u = N - m$. For released mean-neighbor propagation M, consider the one-hop nonlinear model $f_v = F_\phi(X_v W_s + (MX)_v W_n)$, with a shared nodewise network F_ϕ, including biases. Graph aggregation occurs once; subsequent layers are nodewise.

Theorem 6.1 (Connected training gap and an optimizer remedy)**.** ***(i) Information reference.*** *The connected family's exact randomized minimax risk is*

$$R_{\text{info}} = \frac{K(a + b)}{2u} p_m(a + b).$$

(ii) Nonlinear training. *Put $R_{\mathrm{role}} = K\{ap_m(a) + bp_m(b)\}/(2u)$. Initialize all parameters independently of the hidden bits and sampled label set. The rows of W_s are iid, and W_s is independent of (W_n, ϕ). Under independent fair bits, full-batch GD or Adam on equally weighted hit-role logistic loss, without regularization, satisfies at every fixed horizon*

$$\mathbb{E}R_T \geq R_{\mathrm{role}} - \varepsilon_{K,m}, \qquad \varepsilon_{K,m} = \frac{a\sqrt{Kp_m(a)} + b\sqrt{K[p_m(b) - p_m(a+b)]}}{2u}. \tag{4}$$

All weights and biases are trained, with zero optimizer moments and fixed common hyperparameters. The lower bound holds at any finite width; two ReLU units suffice to realize all block-label assignments in the one-hidden-layer model below. For disjoint branches $g_\alpha(X_v) + h_\beta((MX)_v)$ satisfying $g_\alpha(0) = h_\beta(0) = 0$, label-independent initialization, and the branch-local update conditions in Appendix C.2, the stronger bound $\mathbb{E}R_T \geq R_{\mathrm{role}}$ holds, with equality if all observations are fitted in every run.

(iii) Linear intervention. *For $f = X\xi + MX\zeta$, initialize $\xi = \zeta = 0$ and use the sum of the within-role mean logistic losses defined in Appendix C.4. One GD step of size two fits every observation and has prior-average risk R_{role}. Precondition this step by $H = \begin{pmatrix} I_K & \frac{2}{3}I_K \\ \frac{2}{3}I_K & I_K \end{pmatrix}$. The resulting predictor is correct on every covered block and has prior-average risk R_{info}. Replacing its predictions on wholly missed blocks by fair blockwise guesses attains this risk in every world and hence attains the randomized minimax value. The linear function class, loss, and initial predictor are unchanged; matching the two nonzero update norms preserves the classifications.*

Positive logit-gradient overlap does not guarantee transfer. Unused self-matrix rows remain iid, and permuting source-only labels together with their initial rows leaves the receiver training distribution unchanged. Shared parameters can transmit class counts without resolving those assignments. For $F(z) = v^T \mathrm{ReLU}(z+b) + c$ with trainable biases, every cross-role logit-gradient inner product is at least one, throughout training. Thus the bound does not require zero training interaction. Receiver messages also have 20% foreign mass when $a = 4r$. Fix r and $a = b = q \geq 4r$. As $K \to \infty$ with $m/[K(2q)] \to \rho \in (0,1)$, $\varepsilon_{K,m} = O(K^{-1/2})$, and the limiting lower bound on prior-average excess risk is $\frac{1}{2}[(1-\rho)^{q-1} - (1-\rho)^{2q-1}] > 0$. At $r = 1, q = 4, \rho = 1/4$ it is 14.42 percentage points of all test nodes. Appendix C proves the theorem and tests the nonlinear intervention. Appendix D gives complementary linear results, including the dependence of transfer time on initialization.

6.2 A SUFFICIENT CONDITION FOR TRANSFER

A useful distinction is whether propagation preserves a single score direction within each block. Separate self/neighbor matrices and biases can destroy that property even when sampled labels are fitted perfectly.

Proposition 6.2 (A sufficient label-transfer invariant)**.** *On the original disjoint block family, consider an ℓ-layer bias-free network with hidden updates $H^{j+1} = \mathrm{ReLU}(PH^jW_j)$ for $j < \ell - 1$ and final logits $f = PH^{\ell-1}W_{\ell-1}$, where P is nonnegative and block diagonal, and $P^\ell \mathbf{1}_{p_k}$ is strictly positive throughout each active block. Then, for every $v \in C_k$, the logits have the form $f_v = a_v t_k$ with $a_v > 0$. Consequently, if every sampled label is fitted, all covered blocks are predicted correctly. The same conclusion holds for any fixed rank-one block inputs with nonnegative amplitudes and strictly positive final propagation coefficients. The result applies to the bias-free GCN and SGC implementations used here (Kipf & Welling, 2017; Wu et al., 2019), as well as to bias-free tied-weight SAGE with $P = I + M$, whenever the final propagation coefficients are strictly positive.*

A proof based on positive homogeneity appears in Appendix B, which also characterizes agreement among fitted readouts of a frozen representation and certifies transfer stability under fixed weights.

7 EMPIRICAL TESTS OF TRANSFER

We report overall test error with denominator $u = N - m$. A covered-error contribution uses the same denominator; conditional covered error divides instead by the number of covered test vertices. Percentages are labeled explicitly. Pooled node rates and means of problem-level rates are distinct summaries, as specified with each result.

Table 1: Shared-head connected study: overall test error (%). R_{info} is the exact information risk; the training lower bound is $R_{\text{role}} - \varepsilon_{K,m}$ from Theorem 6.1. Each empirical cell averages 24 independent problems and two initializations. All 48 runs per method and cell fit every observed label.

K	r	R_{info}	Nonlinear bound	Adam	Coupled Adam
16	1	6.18	12.21	21.53	9.46
16	3	6.31	11.61	19.39	10.01
64	1	6.55	16.73	21.27	9.07
64	3	6.65	15.48	18.89	8.91

Shared-head connected study. A separately frozen study uses $K \in \{16, 64\}$, $r \in \{1, 3\}$, $a = b = 4r$, and $m = 2K$, with 24 independent world/split problems per cell and two initializations. The one-hop ReLU GNN has width 32 and a trainable shared head and biases. Both arms use 1,000 Adam steps (Kingma & Ba, 2015). The intervention couples the self/neighbor matrix updates with coefficient $2/3$ and rescales them to the norm of the uncoupled update computed at the intervention arm's current state. All 384 runs fit every observed label. The paired reduction in overall test error is 10.90 percentage points (stratified 95% bootstrap interval $[10.09, 11.75]$); every cell's paired interval is positive (Table 17, Appendix C.6). The nonlinear intervention reduces, but does not eliminate, covered error. The exact one-step guarantee concerns the linear specialization. Appendix C reports the separate 576-run branch study, the stacked-GraphSAGE comparison, and the prespecified sparse-tree forecast study.

Original sparse-feature result. At $N = 256, K = Q = 8$, depth 3, $m \in \{16, 32, 64\}$ and 20 samples per label count, mean SAGE contributes 0.03010 covered error to total test loss (seed-cluster SE .00392; 95% t interval $[.02190, .03830]$). The conditional covered rate is $403/2012 = 20.03\%$, with zero training classification error in all 60 fits; bias-free GCN and SGC have zero covered test errors. Perfect fit means thresholded classification, not zero BCE or positive signed margin. In the original factorial, covered contributions to total error are .0301 (separate/biased), .0185 (separate/bias-free), .0090 (tied/biased) and zero (tied/bias-free). The separate/bias-free model fits 59/60 samples; the others fit 60/60. Thus the combined restriction enforces the invariant, while either restriction alone leaves errors. The gap persists across four tested learning rates and when additional known-zero filler nodes increase N (Appendix J).

Stratified matched controls. Twelve independent problems with two initializations each hold eight active labels fixed while varying filler loss, geometry, and features. Removing fillers reduces SAGE's pooled conditional covered error from 27.7% to 13.9%, while its contribution to overall test error changes from 0.0445 to 0.0893 because the test denominator also changes. Masking filler loss makes the active logits identical to those of the no-filler run. Biased GCN also has covered errors while fitting all labels. Appendix W reports the 2,136-run study and its separate maintained-operator replication. A block feature floor nearly removes SAGE's gap but also adds block-specific local information, tested below.

Simple information baselines. Component copying and harmonic propagation eliminate covered error when the components are observably homogeneous. Component correction retains hard SAGE decisions on missed active nodes; its mean test error is $.05524554$. Paired clean runs also show transfer errors despite exact structural identifiability: poisoning creates an information requirement, while the specified training procedure can leave identifiable labels unused in both settings.

Feature-only information controls. At floor .05, a feature-only MLP reduces its conditional covered error from 53.82% to 1.62%, while a shared floor leaves 41.55%. An adaptively added bias-free feature-only linear diagnostic has zero covered error and 24/24 perfect fits with the block floor. For every positive block floor, normalized feature directions identify membership without edges, so even a feature-direction copy rule suffices. The block floor supplies local block identity as well as changing feature geometry. These comparisons therefore assess the combined intervention; they do not isolate a margin effect. At zero floor, feature-only models fail to fit all observations, so their test errors include fitting failures. Appendix A reports all six floor values and all fits.

Fresh tests of the transfer invariant. We freeze the sufficient-invariant prediction, then draw 12 fresh samples with two initializations. Across base, increased depth, increased block size at the same

depth, path geometry, and reduced active-label coverage, SAGE retains covered errors while tied bias-free propagation has zero (Figure 2). Both methods fit every observed label in all 120 fits each. Base uses 1,000 steps and other settings 200; the paired depth comparison uses the base's 200-step checkpoint. The fresh configurations test the invariant's conditional zero-covered-error prediction.

Equal tuning budget and paired uncertainty. On 12 separate fresh problems, each architecture receives the same four learning-rate/L2 candidates, using 8 held-out labels from the original 32 for validation, then refits on all 32. No extra labels or test-based choices are used. SAGE retains 16.44% pooled conditional covered error; bias-free SGC, GCN and tied propagation have zero, all with 24/24 perfect fits. The paired reduction from SAGE to tied propagation is 15.23 percentage points (95% seed-cluster t interval $[7.82, 22.65]$). We average the two initializations within each sampled problem before calculating paired intervals. SAGE errors occur in 9 of the 12 independent selected problems. The paired intervals are exploratory and unadjusted, and tuning is restricted to the stated four-candidate grid. Training trajectories show decreasing BCE alongside persistent covered classification errors.

Table 2: Equal-budget selection on 12 independent problems with two initializations each. Both error columns are conditional on coverage. Pooled rates weight covered nodes; problem means average initializations within each problem and then weight problems equally. Intervals are unadjusted 95% t intervals over problem means. All 24 refits per method fit every observed label. The paired SAGE-to-tied reduction is 15.23 percentage points, with interval $[7.82, 22.65]$.

Configuration	Pooled (%)	Perfect fits	Seed mean [95% CI]
SAGE with bias	16.44	24/24	15.23 [7.82,22.65]
SGC, bias-free	0.00	24/24	0.00 [0.00,0.00]
GCN, bias-free	0.00	24/24	0.00 [0.00,0.00]
Tied, bias-free	0.00	24/24	0.00 [0.00,0.00]

An adaptive sensitivity analysis validates only on active nodes with retained observed block partners. Pooled conditional covered error is 16.09% for biased SAGE, 8.91% for bias-free SAGE, and 9.84% for biased GCN. Both SAGE configurations fit all 24 observed label sets; biased GCN fits 23 of 24. Bias-free GCN, SGC and tied propagation retain zero. These matched settings support a configuration-level conclusion; Appendix B retains failed fits and gives fit-conditioned rates and paired intervals.

Readout and connected-operator audits. Across 864 covered prediction instances from the selected SAGE fits, the floating-point conic audit identifies 248 as forced by the observed decisions. Each of the other 616 admits an individually verified final-head reversal preserving those decisions; 474 are currently correct. For fixed tied-network weights, the stability bound certifies all covered nodes at mixing parameter $\alpha = .01$ and 82.23% at $\alpha = .05$; no covered errors occur in either test. At $\alpha = .2$, no covered nodes are certified, and pooled conditional covered error is $18/816 = 2.21\%$ ($6/640 = 0.94\%$ among the 20 of 24 models that still fit every observation). The cross-block row-mass fraction is $\alpha/(2-\alpha)$. A prospective bound covers the declared transport family and bounded input changes, with median certified cross-block mass 0.111%. These are weighted, directed operator-sensitivity tests with fixed weights and a known homogeneous reference, distinct from legal simple-edge poisoning and retraining. Appendices A–B report the certificate and error results. Under known block homogeneity, direct agreement with an observed label determines retrospective correctness; the prospective envelope instead controls future changes within its declared class.

8 Deterministic, noisy, and relational information

The coverage calculation also distinguishes the roles of prediction randomness, label reliability and shared latent structure. These results characterize information and decision rules independently of a training algorithm.

Exact deterministic prediction. For fixed recognizable block sizes Q_k, set $u = N - m$, $D = \binom{N}{m}$ and $A = \sum_k Q_k \binom{N-Q_k}{m}$. Proposition E.3 gives

$$R^*_{\mathrm{det}} = \begin{cases} N^{-1}\sum_k \lceil Q_k/2 \rceil, & m = 0, \\ \lceil A/2 \rceil/(uD), & 1 \le m < N, \end{cases} \qquad R^*_{\mathrm{rand}} = \frac{A}{2uD}.$$

For positive m, the gap is at most $1/(2uD)$. The construction uses random sampled positions and observed labels on paired splits to cancel odd-block residual errors for every fixed world. Rounding each block separately can be suboptimal; Appendix F gives an explicit two-block example.

Noisy supervision. Flip each observed label independently with probability $\eta \in [0, 1/2]$, and evaluate against clean targets. Let $H \sim \mathrm{Hypergeom}(N, Q, m)$ count labels in a block, and let $e_h(\eta)$ be the majority-vote error with fair ties ($e_0 = 1/2$). Then

$$R^*_\eta(K, Q, m) = \frac{K}{N - m}\mathbb{E}[(Q - H)e_H(\eta)].$$

For $1 \le m, b \le N/4$ and $Q_L \ge 2$, the optimized value satisfies (Theorem G.1)

$$\frac{\Psi}{128} \le R^*_{\eta,L} \le \Psi, \qquad \Psi = \min\left\{1, \frac{bQ_L}{N}, \frac{b}{m(1-2\eta)^2}\right\}.$$

The last term is $+\infty$ at $\eta = 1/2$. The noisy-label law incorporates the standard binary-testing penalty into the legal graph construction; Corollary G.2 gives the resulting label complexity in the coverage-limited regime.

Relational labels under an edit budget. A separate one-layer family has r anchors, r pendants and one query node for each anchor pair, so $N = 2r + \binom{r}{2}$. The presence of each anchor–pendant edge encodes a bit; query labels indicate bit equality. Adding at most B missing pendant edges leaves a hidden edit mask of weight at most B. Observed query labels constrain relative bits, while the budget constrains component orientations (Appendix O). For integers $r \ge 16$, $B \ge 0$, $0 \le m < N$, put $\bar{B} = \min\{B, \lfloor r/2 \rfloor\}$. When $\bar{B} \ge 1$, Theorem O.1 supplies an explicit decoder with

$$\ell_m \le R^*(B, m) \le R_{\mathrm{dec}}(B, m) \le 113\ell_m, \qquad \ell_m = \frac{\bar{B}(r - \bar{B})}{2(N - m)} \frac{\binom{N-r+1}{m}}{\binom{N}{m}}.$$

The factor improves to $2e^3$ for $\bar{B} \le r/16$; $B = 0$ has zero risk. For $m = o(r^{3/2})$, $R^* = \Theta((\bar{B}/r)e^{-2m/r})$. Writing $h_r = \log(\bar{B}/(r\varepsilon_r))$, the conditions $h_r \to \infty$ and $h_r = o(\sqrt{r})$ give $m^*(\varepsilon_r) = (r/2)h_r + O(r)$. This gives a fine-accuracy label threshold when zero-label risk vanishes. The relational family extends the information analysis to dependent labels under a cardinality-constrained edit mask; its degree and feature dimension grow with r. Appendices P and R characterize how dependency geometry and observable membership affect prediction.

9 Discussion and limitations

The coverage law quantifies information erased by legal structural edits. The connected theorem shows that graph connectivity and substantial cross-block messages do not ensure that training uses every informative observed label. Under the stated initialization and training conditions, shared nonlinear representations retain a permutation-induced risk gap. A same-class linear optimizer intervention attains the information reference. The matched nonlinear study supports the specified update intervention; the separate stacked-GraphSAGE comparison examines an architecture with repeated graph aggregation.

The constructions use controlled feature encodings whose dimension grows with the number of hidden bits. The connected training theorem concerns one graph aggregation, the stated independent initialization, fixed schedules, and equally weighted hit-role loss; its positive teacher margins decrease with problem size. The linear intervention attains the prior information reference, and fair guessing on wholly missed blocks gives its worst-world minimax guarantee. Nonlinear improvement is empirical and leaves residual error.

The tree depth law compares admissible target families. The relational extension allows growing degree and dimension, and the sufficient transfer invariant assumes homogeneous rank-one reference blocks. Experiments test constructed graph families; prevalence on natural graphs remains untested.

Exact information laws and a nonlinear training lower bound separate label coverage from label transfer. A one-step linear remedy and matched nonlinear intervention show that optimizer design can improve transfer within a fixed model class.

REPRODUCIBILITY STATEMENT

Sections 3–8 present the model classes, graph constructions, attack and sampling protocols, risk definitions, and principal results. Detailed proofs, supporting derivations, and additional theoretical results are provided in the appendices. The experimental appendices specify architectures, optimization settings, randomization, fitting outcomes, and uncertainty calculations.

The anonymous supplementary code archive includes implementations, study protocols, pre-run manifests where applicable, sampled instances, raw outcomes, stored model outputs, analysis scripts, and verification utilities. Its README provides installation and reproduction commands, and a manuscript-to-code map links the reported results to the corresponding scripts and outputs. The archive also includes file checksums and a report of the executed reproduction checks.

The two separately frozen connected studies contain 960 training runs in total, with 96 independent problems per study. Their uncertainty calculations average initializations within each problem before resampling problems. Adaptive diagnostics and the earlier sparse-tree forecast study are documented separately.

ETHICS STATEMENT

This work studies attacks to clarify robustness limits. The current experiments use synthetic graphs and no human-subject data collection. Releasing attacks has dual-use risk; we mitigate it by focusing on research benchmarks, bounded perturbations, transparent threat models, and reproducible evaluation. The results should not be used to attack deployed graph systems without authorization.

STATEMENT ON THE USE OF AI TOOLS

Generative AI tools were used to aid and polish the writing of this manuscript. The authors reviewed and verified the resulting text and retain full responsibility for the content of the submission.

REFERENCES

Emmanuel Abbe and Enric Boix-Adsera. On the non-universality of deep learning: quantifying the cost of symmetry. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/file/6d9aac9407bcb1a5957401fa0b8de693-Paper-Conference.pdf.

Emmanuel Abbe, Afonso S. Bandeira, Annina Bracher, and Amit Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22, 2014. URL https://arxiv.org/abs/1404.4749.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021.

Maya Bechler-Speicher, Ido Amos, Ran Gilad-Bachrach, and Amir Globerson. Graph neural networks use graphs when they shouldn't. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3284–3304, 2024. URL https://proceedings.mlr.press/v235/bechler-speicher24a.html.

Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, pp. 33–44, 2008. URL https://www.learningtheory.org/colt2008/papers/92-Ben-David.pdf.

Aleksandar Bojchevski and Stephan Günnemann. Adversarial attacks on node embeddings via graph poisoning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 695–704, 2019a.

Aleksandar Bojchevski and Stephan Günnemann. Certifiable robustness to graph perturbations. In *Advances in Neural Information Processing Systems*, volume 32, 2019b. URL https://arxiv.org/abs/1910.14356.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 1170–1182. PMLR, 2021. URL https://proceedings.mlr.press/v139/cai21b.html.

Nicolò Cesa-Bianchi, Claudio Gentile, Fabio Vitale, and Giovanni Zappella. A correlation clustering approach to link classification in signed networks. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pp. 34.1–34.20. PMLR, 2012. URL https://proceedings.mlr.press/v23/cesa-bianchi12.html.

Yuxin Chen, Changho Suh, and Andrea J. Goldsmith. Information recovery from pairwise measurements. *IEEE Transactions on Information Theory*, 62(10):5881–5905, 2016. URL https://arxiv.org/abs/1504.01369.

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1115–1124, 2018.

Ran El-Yaniv and Dmitry Pechyony. Transductive Rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Dylan Foster, Karthik Sridharan, and Daniel Reichman. Inference in sparse graphs with pairwise measurements and side information. In *Proceedings of AISTATS*, volume 84, pp. 1810–1818, 2018. URL https://proceedings.mlr.press/v84/foster18a.html. Detailed comparison uses the expanded arXiv version 3, Theorem 4 and Example 2.

Vikas K. Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3419–3430, 2020.

Simon Geisler, Tobias Schmidt, Hakan Şirin, Daniel Zügner, Aleksandar Bojchevski, and Stephan Günnemann. Robustness of graph neural networks at scale. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7637–7649, 2021.

Ahmad Ghasemi and Hossein Pishro-Nik. Minimax sample complexity of graph neural networks: Lower bounds and structural effects. In *International Conference on Learning Representations*, 2026.

Christina Göpfert, Shai Ben-David, Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, and Ruth Urner. When can unlabeled data improve the learning rate? In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1500–1518, 2019. URL https://proceedings.mlr.press/v99/gopfert19a.html.

Lukas Gosch, Daniel Sturm, Simon Geisler, and Stephan Günnemann. Revisiting robustness in graph machine learning. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2305.00851.

Lukas Gosch, Mahalakshmi Sabanayagam, Debarghya Ghoshdastidar, and Stephan Günnemann. Provable robustness of (graph) neural networks against data poisoning and backdoor attacks, 2025. URL https://arxiv.org/abs/2407.10867. Published in Transactions on Machine Learning Research; arXiv version 3.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. Combining label propagation and simple models out-performs graph neural networks. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.13993.

Junteng Jia and Austin R. Benson. A unifying generative model for graph learning algorithms: Label propagation, graph convolutions, and combinations. *SIAM Journal on Mathematics of Data Science*, 4(1):100–125, 2022. doi: 10.1137/21M1395351. URL https://doi.org/10.1137/21M1395351.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 66–74, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Aakash Sunil Lahoti, Stefani Karp, Ezra Winston, Aarti Singh, and Yuanzhi Li. Role of locality and weight sharing in image-based tasks: A sample complexity separation between CNNs, LCNs, and FCNs. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/71b17f00017da0d73823ccf7fbce2d4f-Abstract-Conference.html.

Colin McDiarmid, Alex Scott, and Paul Withers. The component structure of dense random subgraphs of the hypercube, 2021. URL https://arxiv.org/abs/1806.06433. arXiv:1806.06433, version 3.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Transductive robust learning guarantees. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 11461–11471, 2022. URL https://proceedings.mlr.press/v151/montasser22a.html.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

Nikita Polyanskii and Yihan Zhang. Codes for the Z-channel. *IEEE Transactions on Information Theory*, 69(10):6340–6357, 2023. doi: 10.1109/TIT.2023.3292219. URL https://doi.org/10.1109/TIT.2023.3292219.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Daniel Tarlow, Kevin Swersky, Richard S. Zemel, Ryan Prescott Adams, and Brendan J. Frey. Fast exact inference for recursive cardinality models. In *Proceedings of UAI*, pp. 825–834, 2012. URL https://arxiv.org/abs/1210.4899.

Ilya Tolstikhin and David Lopez-Paz. Minimax lower bounds for realizable transductive classification, 2016. URL https://arxiv.org/abs/1602.03027.

Hongwei Wang and Jure Leskovec. Unifying graph convolutional neural networks and label propagation. *arXiv preprint arXiv:2002.06755*, 2020. URL https://arxiv.org/abs/2002.06755.

Felix Wu, Tianyi Zhang, Amauri Holanda de Souza, Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6861–6871, 2019. URL https://proceedings.mlr.press/v97/wu19e.html.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S. Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. What can neural networks reason about? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJxbJeHFPS.

Chenxiao Yang, Qitian Wu, David Wipf, Ruoyu Sun, and Junchi Yan. How graph neural networks learn: Lessons from training dynamics. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 56594–56623. PMLR, 2024. URL https://proceedings.mlr.press/v235/yang24ae.html.

Jialong Zhou, Xing Ai, Yuni Lai, Tomasz Michalak, Gaolei Li, Jianhua Li, Di Tang, Xingxing Zhang, Mengpei Yang, and Kai Zhou. Adversarial robustness of link sign prediction in signed graphs, 2025. URL https://arxiv.org/abs/2401.10590v3. arXiv:2401.10590, version 3, September 2025.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 912–919, 2003. URL https://mlg.eng.cam.ac.uk/zoubin/papers/zgl.pdf.

Daniel Zügner and Stephan Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 246–256, 2019a.

Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *International Conference on Learning Representations*, 2019b.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2847–2856, 2018.

TABLE OF CONTENTS OF APPENDICES

A Feature information, predictive tests, and transfer stability **16**

A.1 Feature-only controls and block-identity information 16

A.2 Fresh configurations and equally budgeted selection 17

A.3 Statistics and training trajectories . . . 18

A.4 Retrospective agreement and prospective stability 19

A.5 Connected propagation sensitivity and a uniform bound 19

A.6 Hard-label component correction . . . 21

B Readout geometry and prospective neighborhoods **21**

B.1 An exact boundary for a fixed representation 21

B.2 The scope of the exact readout boundary 21

B.3 Transfer stability beyond disconnected propagation 22

B.4 A set-valued prospective envelope . . . 23

B.5 Validation focused on already covered active targets 24

C Connected legal attacks, nonlinear training, and an optimizer intervention **25**

C.1 Construction, legality and observable information 25

C.2 Information reference and separated nonlinear branches 26

C.3 Shared nonlinear representations with positive gradient interaction 26

C.4 One-step intervention in the same linear function class 28

C.5 Exact verification of the linear intervention 29

C.6 Fresh shared-head nonlinear experiment 29

C.7 Prospective nonlinear intervention and an out-of-scope comparator 30

D Exact trained risks and finite-horizon transfer **32**

D.1 An exact risk gap for specified trained learners 32

D.2 Legal paired construction and training convention 32

D.3 Proof of the exact untied and tied risks 33

D.4 Positive overlap: delay and population risk 33

D.5 A general fixed-representation overlap law 34

D.6 A separate affine-head failure selected by training 35

D.7 Scope and finite computational checks 35

E Complete proofs **36**

E.1 Bridge-tree realization 36

E.2 Fixed homogeneous family 36

E.3 Proof of Theorem 4.1 37

E.4 Heterogeneous blocks and decision-rule scope 37

E.5 Proof of Corollary 5.1 38

E.6 Proof of Corollary 5.2 38

F Worked deterministic cancellation **39**

G Exact noisy risk and optimized phase **39**

G.1 Exact value and random-test-index identity 40

G.2 Proof of Theorem G.1 40

H Trained block experiment: protocol and full results **41**

I Architecture and graph-structure stress tests: protocol and results **43**

J Propagation diagnostics and learning-rate sensitivity **45**

J.1 Additional population-size baseline . . 47

J.2 Original factorial transfer intervention . 47

K Budget-constrained relational ambiguity **48**

K.1 Detailed comparison with relational inference 49

K.2 What the poisoning budget contributes 50

L Partial attacks: realization and finite minimax game **50**

L.1 Realization under every legal release . 51

L.2 Reduction and exact finite game 51

L.3 Exact endpoints 52

M Polynomial posterior rule **52**
M.1 Exact posterior computation 52
M.2 One-edit diagnostic 53

N Complete proof of the sparse-budget theorem **53**
N.1 Lower bound 53
N.2 Low-label branch 53
N.3 High-label branch and component errors 53
N.4 Fixed-size small-component bound . . 54
N.5 Asymptotics and first-crossing threshold 54

O Complete dependencies at every budget scale **55**
O.1 Lower bound and low-label branch . . 56
O.2 A component bound up to half the latent population 56
O.3 High-label risk and consequences . . . 57

P Regular dependency graphs: a supporting extension **57**
P.1 A sufficient condition using connected-set boundaries 58
P.2 A separation at half labeling 58

Q Value and limits of the mask constraint **60**
Q.1 Exact unrestricted risk and sparse-class comparison 60
Q.2 Exact identification from a feasible transcript 61

R Coverage without identifiable membership **62**

S Hidden membership: compact subclass and full family **62**
S.1 Proof of Proposition R.1 62
S.2 The distinct full-family minimax game 63
S.3 Feature dimension of the full product family 64

T Relational decoder evaluation **64**

U Relational verification and evaluation details **65**

V Identification, stable features, and decision-class boundaries **66**
V.1 An exact prior-averaged excess decomposition 66
V.2 A positive-margin feature variant . . . 66
V.3 Approximate transfer and an explicit failure 67
V.4 Heterogeneous layouts and the ambient problem 67
V.5 Active labels and partial canonicalization 68

W Matched controls and executable evidence **68**
W.1 Sampling, estimands, and matched objectives 68
W.2 Architecture and feature controls . . . 69
W.3 Information baselines and paired clean tasks 71
W.4 Original off-family boundary tests . . . 71
W.5 Exact capacity optimization 72
W.6 Maintained-library and worst-world checks 72
W.7 Closest-work comparison and precision of bounds 72

A FEATURE INFORMATION, PREDICTIVE TESTS, AND TRANSFER STABILITY

This appendix reports feature-information controls, fresh tests of the transfer invariant, and fixed-weight stability audits. These use the sampling protocols specified below and are separate from the uniform-node study in Appendix H and the stratified controls in Appendix W. The study protocol specifies seeds, conditions, fitting budgets, and analysis choices, which are detailed below. No test labels enter model selection or certificates.

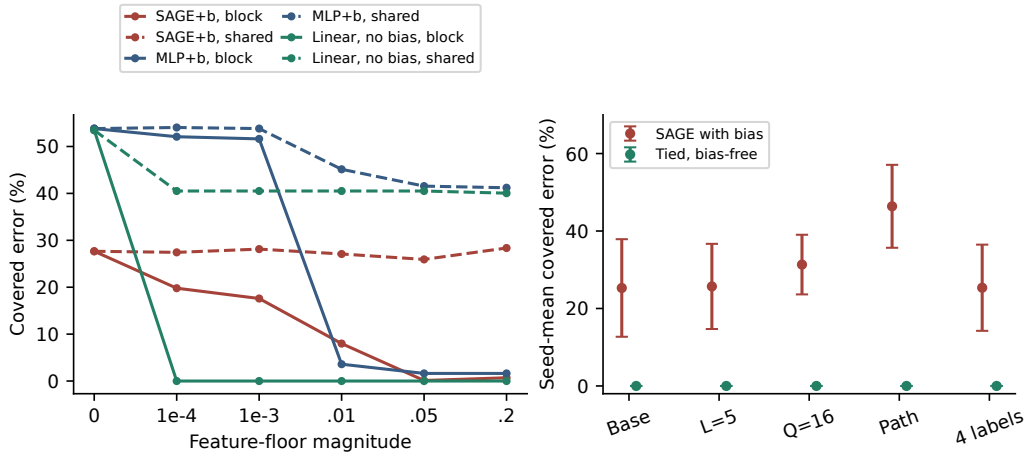

Figure 2: Left: block-specific feature floors allow feature-only prediction; shared floors do not convey the same local identity. Bias-free linear is an explicitly adaptive diagnostic. Right: fresh configurations test the transfer invariant; all plotted models fit the observed labels. Base uses 1,000 steps and other configurations 200; common-horizon comparisons are reported separately. Left rates pool covered nodes descriptively. Right points are means over 12 independent problems after averaging initializations; bars are exploratory 95% t intervals. Configurations are discrete; intervals summarize the 12 sampled problems.

A.1 FEATURE-ONLY CONTROLS AND BLOCK-IDENTITY INFORMATION

The original 12 stratified sampled problems and two initializations are reused for paired comparisons. With $K = Q = 8$, $F = 192$, and 32 observed labels (8 active, 24 filler), we sweep $\delta \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, .05, .2\}$. A block-specific floor sets $X_v = \delta e_k$ at nonpayload nodes in block k, retaining $X_{p_k} = e_k$. A shared floor instead adds $\delta \mathbf{1}/\sqrt{K}$ at every active node to the original sparse features. The shared nonpayload vector contains no block identity. These are controlled feature changes, not equivalent hypotheses under a fixed feature norm constraint. Fillers remain zero. Four initial methods are biased SAGE, bias-free SGC, biased feature-only logistic regression, and a biased feature-only three-layer ReLU MLP. All use width 32 where applicable, Adam .02 and 200 steps. Bias-free feature-only logistic regression was added as an adaptive diagnostic after the initial biased model fit poorly. All methods and their fitting counts are reported below.

An exact feature-only observation. For every $\delta > 0$, normalized nonzero feature rows in the block-floor family are exactly e_k. A learner can match their directions to observed feature rows, copy the observed label, and use a fair guess on unmatched directions, without reading a single edge. On every covered block this yields zero error. Equivalently, a bias-free linear readout with a coordinate of the appropriate sign per covered block fits all its nonzero amplitudes. At $\delta = 0$, nonpayload rows are all zero; they cannot be distinguished by a shared feature-only function. Thus the floor changes local information availability discontinuously at zero, even though its magnitude can be arbitrarily small. Proposition V.2 establishes realizability for the dense-feature variant. The experiments show that the block floor also supplies local identity, so its empirical effect combines an information change with changes in feature scale and geometry.

Table 3: Pooled conditional covered error (percent); parentheses give perfect fits out of 24. Linear and MLP use no edges. Bias-free linear was added after inspecting the initial runs; the other four comparisons were fixed beforehand.

Floor	SAGE	SGC	Linear+b	MLP+b	Linear, no bias
zero	27.66 (24)	0.00 (24)	53.47 (0)	53.82 (0)	53.47 (0)
block 0.0001	19.79 (24)	0.00 (24)	46.06 (0)	52.08 (0)	0.00 (24)
block 0.001	17.59 (24)	0.00 (24)	46.06 (0)	51.62 (0)	0.00 (24)
block 0.01	7.99 (24)	0.00 (24)	46.06 (0)	3.59 (24)	0.00 (24)
block 0.05	0.12 (24)	0.00 (24)	46.06 (0)	1.62 (24)	0.00 (24)
block 0.2	0.69 (24)	0.00 (24)	46.06 (0)	1.62 (24)	0.00 (24)
shared 0.0001	27.43 (24)	0.00 (24)	51.39 (0)	54.05 (0)	40.51 (0)
shared 0.001	28.12 (24)	0.00 (24)	51.39 (0)	53.82 (0)	40.51 (0)
shared 0.01	27.08 (24)	0.00 (24)	51.39 (0)	45.14 (0)	40.51 (0)
shared 0.05	25.93 (24)	15.28 (8)	51.39 (0)	41.55 (0)	40.51 (0)
shared 0.2	28.36 (24)	27.78 (4)	43.75 (0)	41.20 (0)	40.05 (0)

Table 4: Mean across fits of the minimum signed logit in each indicated set. The linear model is the adaptively added bias-free feature-only diagnostic. These scale-dependent trained-model scores are descriptive diagnostics; normalized teacher margins and causal effects are separate quantities. Zero-feature fillers are excluded.

Features	Method	Observed active	Covered test
zero	SAGE	9.0303	-13.2913
zero	MLP	-1.8408	-3.1710
zero	Linear, bias-free	0.0000	0.0000
block 0.001	SAGE	9.5061	-8.3642
block 0.001	MLP	-1.8030	-2.1028
block 0.001	Linear, bias-free	0.0034	0.0031
block 0.01	SAGE	9.8301	-2.8440
block 0.01	MLP	3.2942	0.5892
block 0.01	Linear, bias-free	0.0338	0.0313
block 0.05	SAGE	12.3601	5.9037
block 0.05	MLP	6.5291	5.4706
block 0.05	Linear, bias-free	0.1691	0.1560
shared 0.05	SAGE	9.0540	-12.1440
shared 0.05	MLP	-0.8189	-26.7273
shared 0.05	Linear, bias-free	-0.3508	-3.0908

A.2 Fresh configurations and equally budgeted selection

Fresh seeds 100–111 provide 12 independent problems per condition and two initializations. Base uses $K = Q = 8, L = 3$ and 1,000 steps. Other conditions use 200 steps: $Q = 8, L = 5$; $Q = 16, L = 5$; a path with $Q = 8, L = 7$; and only 4 active labels at $Q = 8, L = 3$. All have 192 fillers and 24 observed filler labels. The depth comparison uses the base 200-step checkpoint. Increasing Q at fixed $L = 5$ separates a size comparison from a depth comparison. Tied bias-free propagation uses $P = I + M$ and one optimized matrix per layer. The prespecified prediction is zero covered error conditional on fitting and positive final profiles.

Selection uses separate seeds 200–211, the base sparse geometry, two initializations, and a common four-candidate grid $\mathrm{lr} \in \{.005, .02\}$, $\mathrm{L2} \in \{0, .001\}$. For each SGC, GCN, SAGE and tied predictor, 2 of 8 observed active labels and 6 of 24 observed filler labels are held out internally. Fit each candidate on the remaining 24 for 200 steps, choose minimum validation BCE, and refit on all 32 for 200 steps. The test set is never consulted. Budgets count training candidates and steps, not equal FLOPs; parameter counts and wall times differ. SGC and GCN are bias-free, SAGE biased, and tied propagation bias-free. This does not optimize all possible bias/normalization settings within each architecture name.

Table 5: Fresh invariant tests: pooled conditional covered error (%). Each setting has 12 independent problems and two initializations. All fits are retained; every run fits all observed labels. Base uses 1,000 steps and other settings 200; the paired depth comparison uses the base 200-step checkpoint.

Fresh condition	SAGE error	Fits	Tied error	Fits
base	26.96	24/24	0.00	24/24
depth5	26.96	24/24	0.00	24/24
Q16 L5	31.14	24/24	0.00	24/24
path L7	46.08	24/24	0.00	24/24
labels4	26.04	24/24	0.00	24/24

Table 6: Equal four-candidate validation budgets followed by one refit per model/sample/initialization. Errors are percentages on covered test nodes. SAGE uses biases; the other listed models do not. Seeds 200–211 are separate from the original and feature-sweep samples.

Model	Pooled error	Fits	BCE	Seed-mean error [95% CI]
SGC, bias-free	0.00	24/24	0.6492	0.00 [0.00,0.00]
GCN, bias-free	0.00	24/24	0.5241	0.00 [0.00,0.00]
SAGE with bias	16.44	24/24	0.0027	15.23 [7.82,22.65]
Tied, bias-free	0.00	24/24	0.5217	0.00 [0.00,0.00]

A.3 Statistics and training trajectories

The independent unit is the sampled problem. For each comparison, we average the two initialization-specific conditional covered rates within each seed, subtract paired conditions, and form a Student-t interval over 12 seed differences. Intervals are exploratory and unadjusted for multiplicity. Pooled rates weight samples by covered-node counts, whereas problem means weight independent problems equally. For SAGE, nonzero seed-mean covered error occurs in 10/12 base problems, 10/12 depth-five problems, 12/12 larger-block problems, 12/12 path problems, 9/12 reduced-label problems, and 9/12 selected problems. All fresh and selected runs fit the observed labels. Zero-error instances and failed feature-only/shared-floor fits are retained, and fitting status is reported separately from covered error. The summaries also report fit-conditioned rates.

Table 7: Paired changes in seed-mean conditional error, in percentage points: right minus left. Two initializations are averaged within each of 12 sampled problems. Intervals are unadjusted exploratory summaries. Linear+b denotes the biased feature-only logistic model.

Matched comparison	Difference	95% interval
SAGE zero to block 0.05	-27.62	[-36.31,-18.94]
SGC zero to block 0.05	0.00	[0.00,0.00]
Linear+b zero to block 0.05	-7.52	[-9.49,-5.55]
MLP zero to block 0.05	-52.61	[-62.65,-42.58]
SAGE shared 0.05 to block 0.05	-25.41	[-32.96,-17.85]
Linear+b shared 0.05 to block 0.05	-5.38	[-6.85,-3.91]
MLP shared 0.05 to block 0.05	-39.06	[-45.88,-32.25]
SAGE base to depth5	0.98	[-5.53,7.49]
TIED base to depth5	0.00	[0.00,0.00]
SAGE to TIED base	-25.30	[-37.89,-12.70]
SAGE to TIED depth5	-25.69	[-36.69,-14.70]
SAGE to TIED Q16 L5	-31.34	[-39.03,-23.64]
SAGE to TIED path L7	-46.38	[-57.09,-35.67]
SAGE to TIED labels4	-25.36	[-36.48,-14.23]
SAGE to TIED selected	-15.23	[-22.65,-7.82]

Checkpoints at steps 1, 20, 50, 100, and 200, and additionally at steps 500 and 1,000 in the base setting, store all logits, training BCE, training error and signed normalized scores $(2Y_v - 1)f_v/a_v$. The latter use test labels only for evaluation after fitting. Distance, class and whether the payload was

observed are also tabulated; these are retrospective diagnostics, not inputs to selection or certification.

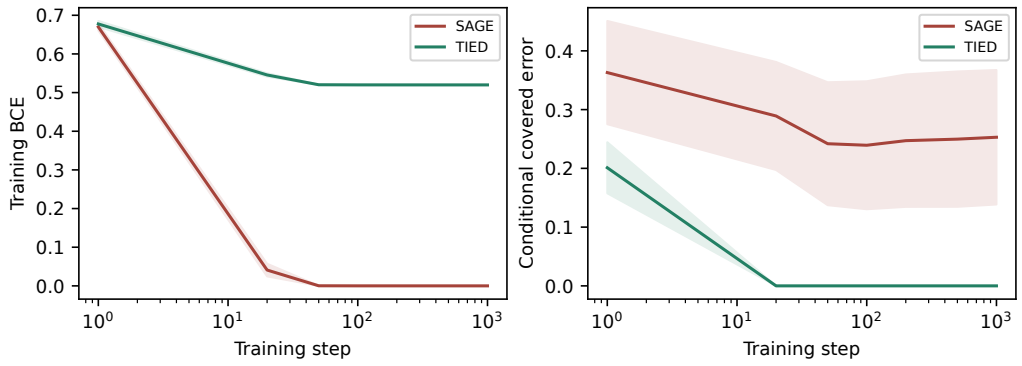

Figure 3: Fresh base trajectories, with the two initializations averaged within sample before aggregation. Continued BCE reduction can coexist with covered classification errors. Bands show two seed-cluster standard errors. Bias-free zero-feature fillers fix a BCE contribution $(24/32)\log 2 \approx .520$; this plateau is not classification failure.

Table 8: Descriptive covered-node counts across 24 fresh base fits. Groups overlap across the three analyses and nodes are not independent inferential units. Payload distance does not give a monotone error pattern; no unique range explanation follows.

Fresh base SAGE subgroup	Nodes	Errors	Rate (%)
label 0	384	116	30.21
label 1	432	104	24.07
payload labeled False	718	193	26.88
payload labeled True	98	27	27.55
distance 0	110	42	38.18
distance 1	92	20	21.74
distance 2	214	74	34.58
distance 3	400	84	21.00

A.4 Retrospective agreement and prospective stability

Fix a profile before examining outcomes: $a = (I + M)^L \mathbf{1}_{\text{payload}}$ for SAGE and tied propagation, and $a = P^L \mathbf{1}_{\text{payload}}$ for other methods. Within each known block, set $t_k = \sum_v a_v f_v / \sum_v a_v^2$ and $\epsilon_v = |f_v - a_v t_k|$. Compute the certificate using actual observed signed scores and Proposition V.3, accepting only slack above 10^{-10}. Profiles and residuals use no labels, but the final test uses observed labels and assumes common-label blocks. The certificate uses observed labels and known block homogeneity. Under that assumption, direct agreement with an observed block label determines retrospective correctness and certifies at least as many correct predictions as the residual test. The layerwise extension in Proposition B.2 serves a different purpose: it bounds changes under specified operator perturbations before evaluating the changed predictor.

A.5 Connected propagation sensitivity and a uniform bound

On the 24 fresh base tied fits, hold weights fixed and let $P^0 = I + M$. On active rows use $P_\alpha = (1 - \alpha)P^0 + \alpha C$, where C shifts a vertex to the same relative position in the next block cyclically; filler rows remain unchanged. Test $\alpha \in \{0, 10^{-4}, 10^{-3}, .01, .05, .2\}$. For $\alpha > 0$ the active operator is connected across blocks, weighted and directed. Its row sum is $2 - \alpha$, not one; the theorem requires only nonnegativity. This experiment perturbs weighted directed operators while holding learned weights fixed and supplying the homogeneous reference partition. Its guarantee concerns this

Table 9: Percent of covered test nodes certified by the fixed-profile residual test, versus directly agreeing with an observed block label. All 1656 final-model audits had no false certificates. Both checks use the same known block-homogeneity assumption.

Setting	Certified	Directly correct	False certificates
floor zero SAGE	22.92	72.34	0
floor block 0.05 SAGE	33.80	99.88	0
floor shared 0.05 SAGE	25.69	74.07	0
floor zero SGC	100.00	100.00	0
fresh base SAGE	26.47	73.04	0
fresh base TIED	100.00	100.00	0
tune selected SAGE	23.38	83.56	0

sensitivity class; legal simple-undirected edge-budget attacks and retraining require separate analysis.

Table 10: Fixed-weight connected-operator sensitivity. Certified fractions and conditional covered-error rates are percentages of covered test nodes. The cross-block row-mass fraction is $\alpha/(2-\alpha)$. Certification uses the layerwise bound and observed labels. At $\alpha = .2$, no covered nodes are certified, and four models cease to fit every observed label. The perturbed operators are weighted and directed.

α	Certified	Covered error	Fits after perturbation
0	100.00	0.00	24/24
0.0001	100.00	0.00	24/24
0.001	100.00	0.00	24/24
0.01	100.00	0.00	24/24
0.05	82.23	0.00	24/24
0.2	0.00	2.21	20/24

Table 11: Covered-node error in the original cyclic connected stress test, separated by observed-label fitting. Four perfectly fitting models across three seeds have covered errors. Models share seeds; pooled rates are descriptive.

At mixing .2	Models	Errors/nodes	Error (%)
All	24	18/816	2.2059
Still fitting all labels	20	6/640	0.9375
No longer fitting	4	12/176	6.8182

For completeness, a coarser bound can certify a whole operator neighborhood. Suppose inputs agree, $\|X\|_{\mathrm{row},2} \le x$, $\|P_j^0\|_\infty \le p$, $\|P_j - P_j^0\|_\infty \le \zeta$, and $\|W_j\|_2 \le w_j$, with the infinity matrix norm the maximum absolute row sum. Then

$$\|f - f^0\|_\infty \le x\Big(\prod_{j=1}^{L} w_j\Big)((p+\zeta)^L - p^L).$$

Indeed $\|H_j^0\|_{\mathrm{row},2} \le xp^j \prod_{i\le j} w_i$. Substituting this into the recursion in Proposition B.2 gives $E_j \le (p+\zeta)w_j E_{j-1} + \zeta x p^{j-1} \prod_{i\le j} w_i$; induction yields the displayed difference of powers, also when $p = 0$ or $\zeta = 0$. Reference logits, their positive profiles and observed labels yield reference signed block scores without any unseen label. A reference signed margin greater than this uniform bound preserves that block's predictions for every operator in the neighborhood. The table uses the tighter layerwise recursion for each specified operator. Appendix B.4 now evaluates prospective neighborhood radii and compares them with this generic bound and a stronger scalar row-sum bound. At $\alpha = .05$, the fraction of active row mass crossing blocks is $.05/(2-.05) = 2.56\%$. At .2 the fit-conditioned split in Table 11 separates transfer failures after fitting from failures to retain the observed labels.

A.6 HARD-LABEL COMPONENT CORRECTION

The component-corrected predictor first computes $\widehat{Y}_v = \mathbf{1}\{f_v > 0\}$, copies observed common labels onto hit components, and assigns zero to known-zero fillers. It retains SAGE's hard predictions on missed active components. Its mean test error is 0.05524554. For comparison, sampling the untouched predictions from their sigmoid probabilities gives 0.05344421 for a different rule. All 2,136 control summaries use the stated hard-label correction. Copy/fair and harmonic baselines use fair randomized predictions on missed components and are evaluated by expected loss. These operations leave the trained weights unchanged.

B READOUT GEOMETRY AND PROSPECTIVE NEIGHBORHOODS

B.1 AN EXACT BOUNDARY FOR A FIXED REPRESENTATION

Rank-one inputs are sufficient, not necessary. Freeze any learned representation $h_v \in \mathbb{R}^p$ and allow every linear final readout $f_v = h_v^\top w$; an affine readout is included by appending a coordinate of one. This leaves the graph, hidden weights and observed labels fixed. The following is a direct application of classical conic separation, closely related to version-space agreement (El-Yaniv & Wiener, 2010; Boyd & Vandenberghe, 2004); its use here is to distinguish transfer forced by a representation from the behavior of a selected readout.

Proposition B.1 (Fixed-representation transfer boundary)**.** *Let $z_s = y_s h_s$ for observed $y_s \in \{-1, 1\}$ and suppose some readout has $z_s^\top w > 0$ for every observed constraint. For a specified target label y_v and $z_v = y_v h_v \neq 0$, every such readout has $z_v^\top w > 0$ if and only if*

$$z_v \in \operatorname{cone}\{z_s : s \in S\}.$$

If membership fails, a readout fits all observations with strict margin and misclassifies v. If $z_v = \sum_s \lambda_s z_s$ with $\lambda_s \geq 0$, then $z_v^\top w \geq \sum_s \lambda_s \gamma_s$ whenever $z_s^\top w \geq \gamma_s > 0$.

Proof. A nonzero conic combination has a positive coefficient, proving sufficiency and the margin bound. Outside the finitely generated closed cone, separation gives a vector d with $z_s^\top d \geq 0$ for all s and $z_v^\top d < 0$. For any strictly fitting w_0, $w_0 + td$ still fits every observation and has negative target score for sufficiently large t. This proves necessity. □

With one same-label observation and a free homogeneous readout, the boundary is exactly a positive ray. For a free affine readout and only same-label constraints it is the convex hull of observed representations, so one observation forces transfer only to an identical representation. Multiple observations can force transfer on higher-rank representations. The criterion applies to arbitrary graphs and hidden maps; it does not assert that training finds a favorable map. On covered blocks y_v is supplied by an observed block label, not a test label. Identically zero negative filler rows impose no restriction under the paper's zero-to-negative convention; the strict-margin audit omits only those rows.

Proof of Proposition 6.2. Within block k, $X_v = \mathbf{1}\{v = p_k\} e_k$. By induction and positive homogeneity of ReLU, $H_v^j = (P^j \mathbf{1}_{p_k})_v h_{j,k}$, where $h_{j+1,k} = \mathrm{ReLU}(h_{j,k} W_j)$. The final linear layer gives the stated factorization. All positive coefficients have the same thresholded sign; hence one correctly fitted block label fixes every prediction in that block. Zero-feature isolated fillers retain zero logits. This is a sufficient structural property, not a characterization of all networks that generalize. □

B.2 THE SCOPE OF THE EXACT READOUT BOUNDARY

Proposition B.1 is an algebraic statement about a frozen feature map and unrestricted final linear weights. The conic formulation applies version-space agreement (El-Yaniv & Wiener, 2010) and Farkas separation (Boyd & Vandenberghe, 2004) to the learned representation. It shows why identifying a block label statistically need not place every block member inside the agreement region of fitted readouts.

For completeness, imposing normalized margins gives the linear-programming dual identity, when z_v belongs to the cone,

$$\min_{w:\ z_s^\top w \geq 1\ (s \in S)} z_v^\top w = \max_{\lambda \geq 0:\ \sum_s \lambda_s z_s = z_v} \sum_s \lambda_s.$$

The primal is feasible by scaling a strictly fitting readout. A conic representation bounds the primal below; finite-dimensional LP duality gives the attained common value. Outside the cone the primal is unbounded below by the separating-direction proof. Zero target vectors are excluded from the strict-positive equivalence; classification at zero continues to follow the paper's zero-to-negative convention.

If all constraints have label y, applying the cone condition to augmented vectors $(h_s, 1)$ and $(h_v, 1)$ forces $\sum_s \lambda_s = 1$. This is precisely convex-hull membership. Without a bias coordinate and with one observation, it is positive proportionality. For example, observing positive labels at $(1, 0)$ and $(0, 1)$ forces a homogeneous readout to be positive at $(1, 1)$; these three vectors are not rank one. With a free bias the corresponding safe point is $(1/2, 1/2)$.

Saved-model audit. For all 96 validation-selected fits we reconstruct the exact final design matrix. For biased SAGE it is $[H, MH, \mathbf{1}]$ and the two final branch weights are free; for tied propagation it is $(I + M)H$; for GCN it is PH; for SGC it is $P^L X$. Hidden states and all earlier weights stay fixed. Every informative observed row has strictly positive signed margin in these fits. Bias-free zero filler rows with negative labels are automatically correct for every readout. All other observed constraints, including biased filler rows, are retained.

We use positive row normalization and a nonnegative feasibility LP for each covered target. A separate LP constructs a readout with observed normalized signed margins at least one and target margin at most minus one whenever possible. Every returned counterexample is checked against all original observed decisions with a fresh matrix multiplication. The saved coefficients, heads and residuals permit independent checking. Cone membership is checked numerically at residual tolerances of 10^{-8}, or 10^{-10} for explicit positive rays. These are floating-point checks of an exact mathematical criterion. The mathematical positive-ray guarantee applies independently to the invariant-preserving models. All 248 SAGE agreement cases have an exactly identical same-label observed design row

Table 12: Fixed-hidden-state audit of selected models. Agreement is checked with explicit conic coefficients; each reversal has a separate final-head witness preserving every observed decision. Floating-point certificates are distinguished from the mathematical equivalence.

Frozen model	Covered nodes	Conic agreement	Reversible	Correct but reversible
GCN, bias-free	864	864	0	0
SAGE with bias	864	248	616	474
SGC, bias-free	864	864	0	0
Tied, bias-free	864	864	0	0

in the computed representation. The 248 agreement cases therefore use equal-row transfer. The remaining 616 SAGE predictions include 142 existing errors and 474 currently correct predictions, each with an individually verified final-head reversal. Each witness keeps hidden states fixed and preserves every observed decision while reversing one target. The witnesses need not reverse all targets jointly and impose no norm bound on the readout; they characterize the frozen representation rather than the probability of an optimizer selecting a particular head.

B.3 Transfer stability beyond disconnected propagation

The exact invariant can be relaxed without assuming that the perturbed propagation remains block diagonal. This addresses robustness of a fitted predictor, not identifiability of an unknown partition.

Proposition B.2 (Computable stability of label transfer)**.** *Let P_j^0 be nonnegative block-diagonal reference operators and P_j any nonnegative operators on the same vertices. Use the same fixed matrices W_j in two bias-free networks with hidden ReLU and a scalar linear last layer. Assume the reference inputs have nonnegative-amplitude rank-one blocks and its final profiles a_v are strictly positive. Write H_j^0 for reference hidden states, $f_v^0 = a_v t_k$ for reference logits, and f for actual logits.*

If the input row differences are bounded by $e_{0,v}$, define, for $j = 1, \dots, L$,

$$e_j = \|W_j\|_2 P_j e_{j-1} + \mathrm{rownorm}_2((P_j - P_j^0)H_{j-1}^0 W_j). \tag{5}$$

Then $|f_v - f_v^0| \le e_{L,v}$. On a known common-label block with an observed label $y_k \in \{-1, 1\}$, every vertex satisfying

$$a_v \max_{s \in S \cap C_k} \frac{y_k f_s - e_{L,s}}{a_s} > e_{L,v}$$

is classified correctly. No unobserved labels enter the bound.

Proof. Subtract the two preactivations as $P_j(H_{j-1} - H_{j-1}^0)W_j + (P_j - P_j^0)H_{j-1}^0 W_j$. Nonnegativity of P_j, the rowwise triangle inequality, and the operator norm bound yield the recursion. ReLU is 1-Lipschitz, and the last layer is linear. Induction bounds each row difference. Positive homogeneity supplies the reference factorization; the observed score lower bounds $y_k t_k$, giving the stated test. □

The inequalities use standard Lipschitz perturbation reasoning. Their role here is to turn the exact block invariant into an executable transfer-stability test. The actual network may mix blocks, but the reference partition and its label homogeneity must remain known assumptions. We do not derive a minimax law or simple-edge attack guarantee for the connected operators from this statement.

B.4 A SET-VALUED PROSPECTIVE ENVELOPE

Fix the reference tied network and let $P^0 = I + M$. For each active vertex v in reference block k, each layer independently allows

$$(P_j)_{v,:} = (1 - \alpha_{j,v})P^0_{v,:} + \alpha_{j,v} C_{j,v,:}, \quad 0 \le \alpha_{j,v} \le \bar{\alpha},$$

where $C_{j,v,:}$ is any probability distribution over active vertices outside block k. Filler rows remain unchanged. Actual inputs may differ from reference inputs by at most η per active row in Euclidean norm; filler inputs remain zero. Perturbed inputs may have signed coordinates, and their rows and hidden states may have rank greater than one. The guarantee holds over the declared operator/input sensitivity class relative to fixed known reference labels. This class is distinct from simple-undirected poisoning and does not assert realizability for a newly attacked teacher.

Write $Z_j^0 = H_{j-1}^0 W_j$, $w_j = \|W_j\|_2$, and define, on active rows,

$$g_{j,v} = \max_{u \notin C_k,\ u \text{ active}} \|Z^0_{j,u} - (P^0 Z_j^0)_v\|_2, \qquad q_v(e) = \max_{u \notin C_k,\ u \text{ active}} e_u.$$

Set $g_{j,v} = q_v(e) = 0$ on fillers and $e_{0,v} = \eta$ on active rows (zero on fillers). A bound simultaneously valid for *every* change in this class is

$$e_j = w_j \left[P^0 e_{j-1} + \bar{\alpha}(q(e_{j-1}) - P^0 e_{j-1})_+ \right] + \bar{\alpha} g_j. \tag{6}$$

Indeed $P_j e \le P^0 e + \bar{\alpha}(q(e) - P^0 e)_+$ for nonnegative e. Convexity of the Euclidean norm bounds the forcing term at row v by $\bar{\alpha} g_{j,v}$. Substitute both inequalities into Proposition B.2 and induct. All quantities in (6) are computed from the reference network and the declared perturbation class, before evaluating any changed predictor. A reference signed score $y_k f_v^0 > e_{L,v}$ therefore guarantees correctness for all future changes in that class. No target label is needed beyond an observed common block label.

Two scalar comparisons under the same metric. The disjoint supports give $\|P_j - P^0\|_\infty \le 3\bar{\alpha}$, active row sums $2 - \alpha_{j,v}$, and cross-block row-mass fraction $\alpha_{j,v}/(2 - \alpha_{j,v})$. In particular .05 corresponds to 2.56% cross-block mass, not 5%. With $\|X^0\|_{\mathrm{row},2} = 1$, the earlier global bound, including input perturbations, is

$$E_{\mathrm{generic}} = \Big(\prod_j w_j \Big) [(2 + 3\bar{\alpha})^L (1 + \eta) - 2^L].$$

Exploiting the additional row-sum bound $\|P_j\|_\infty \le 2$ improves this to

$$E_{\mathrm{scalar}} = 2^L \Big(\prod_j w_j \Big) (\eta + \tfrac{3}{2} L \bar{\alpha}).$$

For the latter, the recurrence is $E_j \le 2 w_j E_{j-1} + 3\bar{\alpha}\, 2^{j-1} \prod_{i \le j} w_i$ with $E_0 = \eta$. Induction proves the formula. The following table compares the rowwise envelope with both the generic scalar bound and the stronger row-sum scalar bound.

Table 13: Prospective radii over 24 fixed tied networks. First three rows give mixing radii with unchanged inputs; last gives input radius with unchanged operators. Bounds hold over all declared transports, not just the sampled paths. Radii use reference scores only. The median rowwise cross-block mass is 0.111%, so this is a small neighborhood.

All-covered guarantee	Minimum	Median	Maximum
Generic scalar	0.000045	0.000306	0.000921
Row-sum scalar	0.000045	0.000306	0.000922
Rowwise envelope	0.000352	0.002212	0.007210
Input-only η radius	0.000201	0.001379	0.004149

Radius and sampled-path protocol. For each of 24 saved fresh-base tied models, bisection finds the largest $\bar{\alpha} \in [0, 1]$ for which every covered reference signed score exceeds the bound by 10^{-10} at $\eta = 0$; monotonicity of the nonnegative recurrence justifies bisection. A separate η radius fixes $\bar{\alpha} = 0$. Values are numerical evaluations of proved bounds, not formally rounded interval certificates. The same calculation gives the certified covered-node fraction when an all-covered guarantee is unavailable. No changed logits enter the radius calculation.

For an empirical comparison, eight fixed random transport sequences per model are tested at $\alpha = 0, .01, \ldots, .50$, each layer using its own transport. The first failing grid point is descriptive and not a worst-case threshold; sign errors need not be monotone along a path. Of 192 paths, 191 have a first failure on this grid, with minimum $.01$, median $.05$ among failures and maximum $.35$; one path is censored at $.50$. The bounds are conservative, not tightness claims. We also evaluate four random full-rank input perturbations per model jointly with transport: α is half its operator radius and η one quarter its input radius. All 96 cases have rank eight within every input block, all 3,264 covered predictions are certified, and none err. These are fixed-weight stress cases, not 96 new training runs. The neighborhood theorem is stronger than this finite check, but still relies on a correct known homogeneous reference.

B.5 Validation focused on already covered active targets

The original internal validation contains two active and six filler nodes. Only 15 of 24 active validation nodes across the 12 problems have an inner-training observation in the same block. We retain that legal procedure and add an explicitly adaptive sensitivity analysis on the same problems and candidate budget. Using only membership and observed-node indices, choose up to two active validation nodes while retaining an observed partner in each block; retain all 24 filler labels in inner training. Select by active-validation BCE, with the same four candidates, 200-step fits and 200-step full refit. Nine problems supply two eligible validation nodes; three supply only one, as allowed by the fixed protocol. All 21 active validation nodes are covered internally. No problem needs the no-eligible-node fallback. Thus the total observed budget remains 32, but the inner training and validation objectives differ deliberately. This adaptive sensitivity analysis reuses the same problems and keeps the total label budget and four-candidate grid fixed while changing the inner validation objective. It adds 576 candidate fits and 144 refits, including a failed biased-GCN refit. All refits are retained, with pooled and fit-conditioned errors reported separately.

Table 14: Adaptive covered-active validation sensitivity. Same total labels and four candidate settings; validation uses two active labels with retained observed partners in nine problems and one in three. All refits are retained. Rates pool covered test nodes descriptively.

Configuration	Error (%)	Perfect fits	Fit-only errors/nodes	Fit-only (%)
SGC, bias-free	0.00	24/24	0/864	0.00
GCN, bias-free	0.00	24/24	0/864	0.00
GCN with bias	9.84	23/24	80/824	9.71
SAGE, bias-free	8.91	24/24	77/864	8.91
SAGE with bias	16.09	24/24	139/864	16.09
Tied, bias-free	0.00	24/24	0/864	0.00

Conclusions concern the specified configurations. The independent unit is the sampled problem: paired intervals average initializations within each problem and are exploratory, without multiplicity adjustment.

Table 15: Exploratory paired seed-mean conditional-error differences over 12 problems, averaging two initializations first. Bias-setting comparisons include retraining and model selection under the stated protocol.

Paired comparison	Change (pp)	95% interval
Active validation minus original, biased SAGE	-0.76	[-4.85,3.32]
Tied minus biased SAGE, active validation	-14.47	[-21.88,-7.06]
Bias-free minus biased SAGE, active validation	-6.47	[-11.76,-1.17]
Bias-free minus biased GCN, active validation	-8.68	[-16.95,-0.41]

Empirical zero describes these samples; the invariant guarantees transfer conditional on its structural and fitting assumptions.

C Connected legal attacks, nonlinear training, and an optimizer intervention

This appendix proves Theorem 6.1. The family is distinct from $\mathcal{P}_L$: its teacher has depth one, and neither its capacity nor its training risk is asserted to satisfy the original depth-dependent formula. Here R_T denotes the realized test-error fraction before averaging over the hidden-bit prior, sampled indices and initialization.

C.1 Construction, legality and observable information

Fix integers $K \geq 2$ and $1 \leq r < K$, and a simple loop-free routing digraph on $[K]$ with in- and out-degree r and connected underlying graph. Its adjacency convention is $A_{ij} = 1$ for $j \to i$. The physical graph constructed below is undirected. Choose integers $a \geq 4r$, $b \geq 2$. Each block consists of sets P_i, R_i of sizes a, b, and designated vertices $p_i \in P_i, t_i \in R_i$. Fix the vertex layout and designated vertices as part of the known experiment. Features are e_i on P_i and zero on R_i. A world is specified by $\theta \in \{0,1\}^K$, with block-constant signed target $y_i = 2\theta_i - 1$.

The released edges are all $P_i \times R_i$, together with $\{\{p_j, v\} : v \in R_i, A_{ij} = 1\}$. The clean graph removes $\{p_i, t_i\}$ iff $\theta_i = 0$. The fixed attack adds all such missing edges before S is sampled. It sees designated-edge presence, not sampled labels. At most K edges are added; each vertex is incident to at most one added edge. The graph is simple because the routing graph has no loops or multiple arcs. Degree is bounded by $\max\{a + r, b(r + 1)\}$, independently of K. Each complete bipartite internal graph remains connected upon deletion of one edge, and permanent routing edges join the blocks. Hence every clean world and the common release are connected. No filler vertices are used.

Let $w = K^{-1/2} \sum_i y_i e_i$. Then $\|w\|_2 = 1$, and every feature row has norm at most one. Under P_λ at depth one, a source's signed score is $\lambda/[(\deg(v) + \lambda)\sqrt{K}] > 0$: all neighbors have feature zero. A receiver's signed numerator is at least $(a - 1 - r)/\sqrt{K} > 0$, because at least $a - 1$ own-block source features remain and at most r foreign features have opposite sign. Thus the same w realizes the clean targets on both graphs for every fixed $\lambda \in (0, 1]$. The teacher margins are strictly positive for each finite K, with lower bounds that decrease with K.

Clean designated-edge presence reveals every bit without supervision. Released features identify source groups. Define

$$B = \frac{aI + A}{a + r}, \qquad \gamma = \frac{a}{a + r} \geq \frac{4}{5}.$$

Each receiver in R_i has mean-neighbor feature B_i, while every source has mean-neighbor feature zero. The largest receiver coordinate is uniquely i, so group membership is observable without labels. B is nonnegative and doubly stochastic. Foreign coordinates contribute total mass $1 - \gamma$, exactly $1/5$ when $a = 4r$. Foreign messages carry features from independently labeled blocks and account for 20% of receiver neighbor mass when $a = 4r$.

C.2 Information reference and separated nonlinear branches

Write $u = N - m$ and $p_m(s) = \binom{N-s}{m}/\binom{N}{m}$, with an impossible binomial coefficient interpreted as zero. A block hit by S reveals its bit. A missed block has an independent fair bit under the product prior, since released inputs are the same in all worlds. Summing its Bayes error and dividing by u gives R_{info}. Copying an observed label within covered blocks and making fair guesses on missed blocks attains this risk in every world. This proves the exact randomized minimax statement.

For the training result, require $g_\alpha(0) = h_\beta(0) = 0$ for all parameter values and disjoint parameter sets. A pointwise loss may have nonnegative weights determined by released inputs and S, and separate regularization is allowed. Each branch update must use only its own gradient history, state and bit-independent auxiliary randomness. The joint initialization is independent of the hidden bits and sampled label set; the branches may nevertheless have correlated or identical initial parameter values. Step counts, step sizes and sampling schedules may depend on released inputs and S, but not on hidden bits or cross-branch label information. Label-independent minibatch sampling is allowed. The branch-local theorem requires each branch update to depend only on its own history and permitted bit-independent randomness. Global gradient clipping, loss-dependent shared schedules, shared readouts, and validation-based stopping require separate analysis.

On a source observation the receiver branch is identically zero, so its parameter derivative is zero; the reverse holds on a receiver observation. Induction on optimizer steps shows that the self branch state depends only on labels observed in source roles, and the neighbor branch state only on labels observed in receiver roles. Conditional on $S \cap P_i = \varnothing$, the self branch output at e_i is independent of y_i, even if S hits R_i. Averaging the independent bit gives error exactly $1/2$. The same argument holds on R_i when S misses that role. Neighbor inputs may mix coordinates, and the neighbor branch may share all its parameters across blocks; neither supplies the missing bit.

Every missed source role contributes a test vertices and every missed receiver role contributes b. Their probabilities are $p_m(a)$ and $p_m(b)$, respectively. Adding the nonnegative test error on hit roles proves the sharper identity

$$\mathbb{E}R_T = R_{\mathrm{role}} + \mathbb{E}E_{\mathrm{hit}}(T) \geq R_{\mathrm{role}}. \tag{7}$$

All vertices in a role have the same predictor input; therefore perfect fitting on every run makes the hit-role error zero. The identity averages over the full prior and all runs. Equality follows when every run fits its observations; conditioning afterward on successful fits can change the bit distribution. The prior lower bound also implies a worst-world lower bound for each specified algorithm.

When $a = b = q$ is fixed and $m/N \to \rho$, the finite-population miss probability tends to $(1-\rho)^s$ for fixed s. Since $N/(N-m) \to 1/(1-\rho)$,

$$R_{\mathrm{role}} - R_{\mathrm{info}} \longrightarrow \frac{(1-\rho)^{q-1} - (1-\rho)^{2q-1}}{2} > 0.$$

For $q = 4, \rho = 1/4$ the limit is 0.144195556640625. The family remains bounded-degree and connected with constant foreign message mass. The construction uses K feature coordinates and encodes each clean block label in a designated edge, making the information erased by the attack explicit.

C.3 Shared nonlinear representations with positive gradient interaction

We prove the finite-width nonlinear statement in Theorem 6.1. Use $f_v = F_\phi(X_v W_s + (MX)_v W_n)$, with $W_s, W_n \in \mathbb{R}^{K \times h}$ and an arbitrary shared nodewise network F_ϕ. Biases are allowed and all parameters are trained. The single graph aggregation occurs before F_ϕ; there is no subsequent message passing. Initialize all parameters independently of the hidden bits and sampled label set. The rows of W_s are iid, and W_s is independent of the remaining parameters. Gaussian initialization with independent self and neighbor matrices is one example. Use full-batch GD or coordinatewise Adam, zero optimizer moments, fixed common hyperparameters and no regularization. Each hit source or receiver role has equal total logistic-loss weight. A global normalization by the number of hit roles is allowed. The fixed-horizon theorem does not require successful fitting.

Condition on S. Let J_P index missed source roles, J_S index blocks whose source role is hit but receiver role is missed, and J_0 index wholly missed blocks; write their cardinalities as n_P, n_S, n_0.

Put

$$e(n) = \mathbb{E}\min\{Z, n - Z\} = \frac{n}{2} - \frac{1}{2}\mathbb{E}|2Z - n|, \qquad Z \sim \mathrm{Bin}(n, 1/2).$$

We first prove the sharper conditional bound

$$\mathbb{E}[R_T \mid S] \geq \frac{ae(n_P) + be(n_S) + bn_0/2}{u}. \tag{8}$$

Missed source roles. For $i \in J_P$, row i of W_s enters no observed score: receivers have zero self input, and another source uses only its own row. Its loss gradient is identically zero, so it stays at initialization. The trained state of all other parameters is independent of these unused iid rows. Conditional on that state and all labels, source predictions in J_P have a common marginal distribution, though realized predictions can differ. If their target vector has j positives, averaging the unused rows gives at least $\min\{j, n_P - j\}$ erroneous roles. This remains true when observed receiver partners have affected shared weights. Averaging the independent bits gives $e(n_P)$, each role contributing a test vertices.

Receivers with an observed source partner. Fix all observed receiver labels and the other observed source labels. Permute the target assignments within J_S and apply the same permutation to the initial self-matrix rows indexed by J_S. Each observed source score and its label move together. Equal role weighting leaves the total objective and all neighbor/shared-parameter gradients identical, while self-matrix gradients undergo the same row permutation. GD and coordinatewise Adam preserve this relation inductively. All receiver predictions therefore remain identical under the joint transformation. Iid initialization is invariant under the row permutation, so the distribution of the complete receiver prediction vector depends on the J_S labels only through their count of positives. Conditional on that count, their assignments remain equiprobable. Every receiver decision has Bayes error at least the minority fraction; averaging gives $e(n_S)$ erroneous roles, each with b test vertices. The symmetry jointly permutes labels and initial self-matrix rows and holds in distribution.

Wholly missed blocks and averaging. Every receiver in J_0 has an independent unobserved bit, giving error $1/2$. The three evaluated sets are disjoint; remaining errors are nonnegative. This proves (8). Cauchy–Schwarz gives $\mathbb{E}|2Z - n| \leq \sqrt{n}$. Moreover,

$$\mathbb{E}n_P = Kp_m(a), \quad \mathbb{E}n_S = K[p_m(b) - p_m(a + b)], \quad \mathbb{E}(n_S + n_0) = Kp_m(b).$$

Apply Jensen's inequality to $\mathbb{E}\sqrt{n_P}$ and $\mathbb{E}\sqrt{n_S}$ in (8) to obtain (4). For fixed r and $a = b = q \geq 4r$, as $K \to \infty$ with $m/N \to \rho \in (0, 1)$, the correction is $O(K^{-1/2})$. Combining with the information formula gives the positive asymptotic lower bound on excess risk in the main text. The bound averages over the fair-bit prior and implies a worst-world lower bound for each algorithm.

Strictly positive cross-role logit-gradient overlap. Take the finite-width special case $F_\phi(z) = v^T \mathrm{ReLU}(z + \beta) + c$, with trainable v, β, c. Let h_s, h_t be the nonnegative hidden activations at a source and receiver, and q_s, q_t the ReLU gates (derivative zero at a kink). At every parameter value the Euclidean logit-gradient inner product is

$$\langle \nabla f_s, \nabla f_t \rangle = 1 + \langle h_s, h_t \rangle + \sum_j v_j^2 q_{s,j} q_{t,j} \geq 1. \tag{9}$$

The self/neighbor matrix contributions vanish across roles, but the shared output bias, head and hidden bias give the displayed terms. Thus the permutation restriction persists despite uniformly positive cross-role logit-gradient inner products. The shared output bias supplies the constant term one. For logistic losses, the corresponding loss-gradient inner product is multiplied by

$$\frac{y_s y_t}{(1 + e^{y_s f_s})(1 + e^{y_t f_t})},$$

which changes sign for opposite labels. Equation (9) establishes finite-width logit-gradient positivity without a kernel-limit assumption.

Representation and scope. The model can represent every block-label assignment with two ReLU units: use first-layer columns $\pm y$ for W_s and $\pm B^{-1}y$ for W_n, head weights $(1, -1)$ and zero biases. Since B is invertible, the logits are y_i on both roles. Failure is therefore not misspecification. Coupling corresponding self/neighbor updates breaks the independent row permutation symmetry without changing this function class. The nonlinear intervention is evaluated empirically; Section C.4 proves exact one-step transfer in the linear specialization. The permutation argument relies on equal hit-role weights, independent self/neighbor initialization, and a single graph aggregation. Additional aggregation can expose unused self rows to observed receiver losses.

C.4 One-step intervention in the same linear function class

Take $f = X\xi + MX\zeta$ and initialize both parameter vectors to zero. Use the sum of the observable within-role mean logistic losses

$$\mathcal{L} = \sum_{C \in \{P_i, R_i\}_{i=1}^{K} :\, S \cap C \neq \varnothing} \frac{1}{|S \cap C|} \sum_{s \in S \cap C} \log(1 + e^{-y_s f_s}).$$

The objective gives each observed role equal total weight, computed from membership and sampled indices. Set $U_i = y_i \mathbf{1}\{S \cap P_i \neq \varnothing\}$ and $V_i = y_i \mathbf{1}\{S \cap R_i \neq \varnothing\}$. At zero the gradients are $-U/2$ and $-B^T V/2$. Ordinary GD with step size two gives $\xi = U, \zeta = B^T V$. Source scores are U, receiver scores CV, where $C = BB^T$. The matrix C is nonnegative and doubly stochastic, and

$$c := C_{ii} = \frac{a^2 + r}{(a + r)^2} \geq \gamma^2 \geq \frac{16}{25}, \qquad \sum_{j \neq i} C_{ij} = 1 - c.$$

Every observed source role has signed score one. Every observed receiver role has signed score at least $2c - 1 > 0$. Thus every observation is fitted for every world and sample. The preceding identity gives prior risk R_{role}, including the empty sample under a fixed zero-score tie rule.

Instead apply the positive-definite preconditioner $H = \left(\begin{smallmatrix} I & \kappa I \\ \kappa I & I \end{smallmatrix}\right)$, $\kappa = 2/3$, to this same gradient step. Then

$$\xi = U + \kappa B^T V, \qquad \zeta = B^T V + \kappa U, \qquad f_P = U + \kappa B^T V, \quad f_R = CV + \kappa BU.$$

Table 16 gives signed-score lower bounds obtained from nonnegativity and the row and column sums of B and C.

Table 16: Signed-score lower bounds after one preconditioned gradient step in the linear model. A role is hit when it contains at least one sampled vertex. With $\kappa = 2/3$, $\gamma \geq 4/5$, and $c \geq \gamma^2$, every bound is strictly positive, establishing correct predictions on all vertices of every covered block.

Target role	Observed information	Signed-score lower bound
P_i	P_i hit	$1 - \kappa(1 - \gamma) > 0$
P_i	only R_i hit	$\kappa(2\gamma - 1) > 0$
R_i	R_i hit	$2c - 1 - \kappa(1 - \gamma) > 0$
R_i	only P_i hit	$\kappa(2\gamma - 1) - (1 - c) \geq 1/25$

Own-coordinate contributions from a second hit role have the correct sign and can be dropped. Foreign contributions are bounded by the off-diagonal row mass. Thus all vertices of all covered blocks are correct. On a wholly missed block, its bit is independent of the trained prediction, giving error $1/2$ under the prior. The raw predictor attains R_{info} under the fair-bit prior. Fair blockwise guesses on wholly missed blocks make the risk equal to R_{info} in every world, yielding the randomized minimax guarantee.

Positive scalar multiplication of the whole parameter update preserves all post-step signs. For $m > 0$ the ordinary update is nonzero: either $U \neq 0$ or $V \neq 0$ and the strictly diagonally dominant B is invertible. The preconditioned update is also nonzero. Rescaling it to the ordinary update's Euclidean norm therefore preserves the result. For $m = 0$, both are zero and the fair fallback suffices. The intervention changes update direction while holding the function class, initial predictions and, optionally, step norm fixed.

C.5 EXACT VERIFICATION OF THE LINEAR INTERVENTION

The executable checks verify graph legality, connectivity and the same-readout teacher for all worlds at $(K, r, a, b) = (3, 1, 4, 4), (4, 2, 8, 3), (5, 3, 12, 12)$. Integer arithmetic verifies observed-role fitting and covered-block transfer over 299,520 world/role-hit cases across four configurations. These checks evaluate the closed-form linear optimizer outputs. On the $N = 24, K = 3, a = b = 4$ instance, all $5^6 = 15,625$ physical role-count vectors are weighted by their exact binomial multiplicities. Summing over all eight worlds verifies both risk formulas as rational numbers for every $m = 0, \ldots, 23$, covering all $2^{24} - 1$ subsets with a nonempty test set. The remaining subset, with $m = 24$, has no test risk to evaluate.

C.6 FRESH SHARED-HEAD NONLINEAR EXPERIMENT

The shared-head protocol was fixed after the separated-branch study and the derivation in Section C.3, before generating its own outcomes. It was not externally preregistered. A pre-run manifest records the code and protocol versions used for the study.

Design and matched intervention. Use the same four graph configurations, $K \in \{16, 64\}$, $r \in \{1, 3\}$, $a = b = 4r$, $N = 2aK$ and $m = 2K$, with the directed circulant routing described below. Draw 24 new independent fair-bit worlds and uniform physical-vertex samples in each cell, with two initializations per problem. The model is

$$f = \mathrm{ReLU}(XW_s + MXW_n + b)v + c,$$

with hidden width 32, independent Gaussian W_s, W_n of standard deviation $1/\sqrt{K}$, Gaussian v of standard deviation $1/\sqrt{32}$, and zero initial biases. Every parameter, including both biases, is trained. Both arms use 1,000 fixed full-batch Adam steps, learning rate .01, $(\beta_1, \beta_2) = (.9, .999)$, numerical constant 10^{-8} and no regularization. The objective averages equally weighted hit-role logistic losses. There is no validation, tuning, early stopping or fit-based exclusion.

At the intervention arm's current parameters and optimizer state, compute the ordinary first-matrix Adam updates d_s, d_n. Put $\widetilde{d} = (d_s + 2d_n/3, d_n + 2d_s/3)$ and apply

$$d^{\mathrm{coupled}} = \frac{\|(d_s, d_n)\|_2}{\|\widetilde{d}\|_2}\widetilde{d}$$

when $(d_s, d_n) \neq 0$, using zero otherwise. The norm match is within this arm at its current state, not between the two evolving trajectories. Shared-head and bias updates remain ordinary Adam. The arms share their function class, observed labels and initial predictor. This intervention breaks the independent source-row permutation symmetry; its nonlinear success is not assumed by the lower-bound theorem.

Predetermined outcome and all results. The primary endpoint is paired overall-error reduction. Initializations are averaged within each independent problem. Per-cell intervals use 10,000 percentile bootstrap resamples of problems; the pooled interval resamples within each cell and averages the four cells equally. All intervals are unadjusted and refer to this fixed constructed-family protocol.

Table 17: Shared-head connected study: conditional covered error (%) and paired reduction in overall test error (percentage points). Each cell contains 24 independent problems with two initializations per method. Intervals use problem-level percentile bootstrapping, with equal cell weights for the pooled reduction.

		Covered error (%)		Overall reduction	95% interval
K	r	Adam	Coupled	(percentage points)	
16	1	15.98	2.34	12.07	[9.64, 14.58]
16	3	14.89	2.95	9.38	[7.81, 10.94]
64	1	16.74	2.27	12.20	[11.13, 13.26]
64	3	13.79	2.01	9.98	[8.81, 11.17]
Equal-cell pooled overall reduction				10.90	[10.09, 11.75]

Every one of the 384 runs fits all observed labels. Table 1 reports overall errors, the exact information reference and the analytic nonlinear lower bound. The bound is not a prediction of the numerical error; the baseline is above it in all four empirical cell means. The coupled model still has covered errors and does not attain the exact information reference.

Implementation and mechanism checks. The maximum absolute finite-difference gradient discrepancy is 5.10×10^{-11} on the validation fixture. A source-only label swap paired with the same permutation of initial self-matrix rows leaves ordinary Adam's receiver outputs, neighbor weights and shared parameters bitwise identical after 100 steps. Unobserved self rows remain unchanged. Under coupled updates, the same transformation changes receiver logits by up to 3.949, as expected when the permutation symmetry is broken. The analytic cross-role gradient formula (9) agrees with a direct parameter-Jacobian dot product to 3.56×10^{-15}; its minimum over all source/receiver pairs on the fixture is 5.258, consistent with the universal lower bound one. These are mechanism checks, not independent proof validation or evidence about multi-hop stacked GraphSAGE.

Relation to the earlier connected experiment. The separated-branch experiment below uses different fresh seeds and a different architecture. Its stronger exact role bound and empirical 12.24-point reduction remain useful supporting results. They are not pooled with this shared-head study. Its stacked-GraphSAGE comparator shows explicitly why repeated message passing must remain outside the present theorem.

C.7 Prospective nonlinear intervention and an out-of-scope comparator

The separated-branch study protocol was fixed before generating its outcomes. The supplementary materials include the protocol, a pre-run SHA-256 checksum manifest, all raw outcomes, random seeds, and analysis code. These connected-family experiments are separate from the original tree-block experiments.

Fixed design. The four cells have $K \in \{16, 64\}$ and $r \in \{1, 3\}$, with directed circulant routing arcs from $i - j$ to i for $j = 1, \ldots, r$ modulo K. Take $a = b = 4r$, $N = 2aK$, and $m = 2K$ uniformly sampled physical vertices. The foreign message fraction is 20% in every cell. Each cell uses 24 independent fair-bit worlds and splits, with two initializations per problem; all three arms are paired on the same observations. No validation, hyperparameter selection, early stopping or fit-based exclusion is used.

The branch model uses separate bias-free three-layer ReLU MLPs, with width 32 and a scalar linear output. At initialization their matrices are identical Gaussian draws with standard deviation $1/\sqrt{\text{fan-in}}$, but their trainable parameters are disjoint. This aligns coordinates for the intervention; it is not a claim about all relative initializations. Full-batch Adam uses learning rate .01, $(\beta_1, \beta_2) = (.9, .999)$, numerical constant 10^{-8}, no weight decay and 1,000 fixed steps. The loss averages over hit roles, each with equal total weight, and uses only observed labels.

For the coupled arm, calculate the usual Adam updates d_s, d_n for all corresponding branch parameters at that arm's current parameters and optimizer state. Replace them by $(d_s + 2d_n/3, d_n + 2d_s/3)$ and rescale the joint vector to the norm of (d_s, d_n), using zero if both are zero. This matches the ordinary update available at the intervention arm's current state, not the separate baseline arm's update norm. Everything else, including the initial predictor, is unchanged. This nonlinear intervention is empirical; the one-step linear proof does not establish its success.

The third arm is ordinary three-layer mean GraphSAGE with self/neighbor matrices and biases at each layer, width 32, ReLU after the first two layers, and the same Adam settings and role-weighted loss. It is a boundary comparator, not a matched causal comparison with the branch model. Its repeated graph aggregation falls outside both the separated-branch assumptions and the one-aggregation shared-head theorem. Later aggregation can expose self-matrix rows to observed receiver losses. Duplicate inputs are compressed exactly. For stacked SAGE the three cells per block are the designated source, other sources and receivers; their equitable quotient has the same mean-aggregation computations as the full graph. Training and test weights retain actual sampled and unsampled vertex counts.

Predetermined endpoint and uncertainty. The primary endpoint is the paired overall-error reduction from independent to coupled branches. Average initializations within each problem before

forming paired differences. Per-cell intervals use 10,000 percentile bootstrap resamples of the 24 independent problems; the pooled interval resamples within each cell and weights the four cells equally. Intervals are unadjusted and concern the stated constructed-family protocol. The predeclared pooled-positive criterion is satisfied, and every cell's interval is also positive:

Table 18: Separated-branch connected study: paired reduction in overall test error from ordinary to coupled Adam, in percentage points. Initializations are averaged within each problem before bootstrap resampling; the pooled reduction weights cells equally.

K	r	Reduction (percentage points)	95% interval
16	1	14.15	$[11.63, 16.75]$
16	3	11.72	$[9.80, 13.64]$
64	1	12.20	$[10.98, 13.37]$
64	3	10.90	$[9.85, 12.02]$
Equal-cell pooled		12.24	$[11.37, 13.14]$

All outcomes, including failed fits, are retained. The branch model has 190/192 perfect fits, the coupled model 191/192, and stacked SAGE 192/192. The following rates are means of per-run conditional covered errors, not pooled node counts:

Table 19: Separated-branch connected study: mean per-run conditional covered error (%) and counts of perfectly fitted runs. All 48 runs per method and cell are retained, including failed fits.

		Covered error (%)			Perfect fits / 48		
K	r	Branch	Coupled	Stacked	Branch	Coupled	Stacked
16	1	17.31	2.21	3.76	48	48	48
16	3	15.76	2.57	1.38	48	48	48
64	1	16.87	2.46	0.61	46	47	48
64	3	14.19	2.14	0.54	48	48	48

Table 20: Separated-branch connected study: overall test error (%). References are exact prior expectations; empirical means average 24 problems and two initializations. Stacked SAGE provides a separate architecture comparison.

K	r	R_{info}	R_{role}	Branch	Coupled	Stacked SAGE
16	1	6.18	20.76	21.35	7.20	9.16
16	3	6.31	18.90	19.89	8.17	8.62
64	1	6.55	21.01	21.13	8.94	7.25
64	3	6.65	19.12	19.53	8.63	7.24

Table 20 reports overall test errors and exact prior-risk references for the separated-branch study. Coupling reduces overall error in all four cells, while residual covered errors remain. Stacked SAGE improves on independent branches in every cell. Coupled branches have lower mean overall error at $K = 16$, whereas stacked SAGE has lower means at $K = 64$.

Mechanism and implementation checks. Analytic gradients agree with centered finite differences to maximum absolute error below 10^{-10} on deterministic fixtures. The equitable quotient's propagation agrees with the full graph. A fixed $K = 8, r = 1$ counterfactual observes a source in block zero but no receiver there, and holds all other observed labels and the initialization fixed. After 100 steps, flipping only that block's bit leaves the independent receiver branch's weights and outputs bitwise identical. The receiver-zero logit changes by -7.633 under coupled Adam and $+1.186$ under stacked SAGE. The nonzero changes establish sensitivity to the flipped observed label. The invariant independent-branch output verifies the branch-local mechanism on this fixture; correctness of transfer is assessed by the study's error rates.

Sparse-tree forecast and intervention study. A separate study used the original sparse tree blocks, three-layer stacked SAGE, and 312 fits, with four development and eight held-out problems and two initializations. A matched, norm-controlled update intervention reduced mean conditional covered error from 15.3125% to 8.28125%, with a paired interval for the reduction of $[-4.65, 18.71]$ percentage points. Overall error fell from 5.80% to 5.02%, with an interval for the reduction of $[-.58, 2.15]$ points. Both intervals include zero. Adding early training-influence features worsened held-out margin-forecast MSE from 48.8625 to 53.8512 and error Brier score from .103373 to .106794. Three of six joint success criteria passed. The study did not meet its prespecified joint success criterion and did not establish the proposed nonlinear optimization explanation. It is analyzed separately from the connected-family studies.

D Exact trained risks and finite-horizon transfer

D.1 An exact risk gap for specified trained learners

Specialize the legal family to $K \geq 1$ pairs ($Q = 2$) and $N - 2K$ fillers, with $K \leq \min\{B, d, \lfloor N/2 \rfloor\}$. Let M be released mean-neighbor propagation, with zero filler rows. Compare the linear scores

$$f_{\mathrm{un}} = Xa + MXb, \qquad f_{\mathrm{tie}} = (X + MX)t, \qquad f_\delta = f_{\mathrm{un}} + \delta(X + MX)c.$$

These are one-layer linear learners, not the nonlinear SAGE of Section 7. Use no bias or regularization, summed logistic loss, and full-batch gradient descent with step size one. Initialize coordinates of a, b, t independently and uniformly in $\{-1/4, 1/4\}$, independently of the world and sample; set $c = 0$ and take $\delta \geq 0$. An empty sample gives zero updates.

Proposition D.1 (Trained risk and a finite-horizon plateau)**.** *For every fixed hidden world and $0 \leq m < N$, all three learners classify every observed label correctly after any $T \geq 1$ steps. Averaging over the sample and initialization, for all such T,*

$$R_{\mathrm{un}} = \frac{K}{N}, \qquad R_{\mathrm{tie}} = R_2^\star := \frac{K(N - m - 1)}{N(N - 1)}. \tag{10}$$

Here $R_2^\star$ is the fixed-layout paired-family minimax value. For $\delta > 0$, put $q = 1 + \delta^2$, $r = \delta^2$ and

$$\tau_-(\delta) = \frac{e^{-1/4}(e^{q/(4r)} - 1)}{qe^q}. \tag{11}$$

For every integer $1 \leq T < \tau_-(\delta)$, $R_\delta(T, m) = K/N$, so both R_{un} and $R_\delta(T, m)$ have excess $Km/[N(N - 1)]$ over $R_2^\star$. For each fixed $\delta > 0$, $R_\delta(T, m) \to R_2^\star$ as $T \to \infty$.

The untied learner leaves an unobserved endpoint's coordinate unchanged; tying lets either label update both endpoints. Positive overlap eventually corrects a wrong initialized target, but its correction is proportional to a logarithmically growing training margin. At $\delta = .1$, the sufficient plateau horizon exceeds 2.59×10^{10} steps; this is an analytic bound, not a simulated runtime or a forecast for SAGE. Appendix D gives the proof, a general overlap law and an affine-head example. Zero initialization instead gives immediate covered transfer at positive overlap. Thus the result specifies a training-dependent excess risk, not an optimizer-independent barrier or an explanation of the nonlinear experiments. Kernel-mediated dynamics and logistic implicit bias are established (Yang et al., 2024; Soudry et al., 2018).

D.2 Legal paired construction and training convention

Use the fixed layout of Section 3 with $Q = 2$: block k has payload p_k with feature e_k, root r_k with feature zero, and common label θ_k. Unused feature coordinates may be omitted. There are $F = N - 2K$ isolated zero-feature fillers with label zero. The clean edge $\{p_k, r_k\}$ is present exactly when $\theta_k = 1$; the attacker adds each missing designated edge before sampling S. The released graph is independent of the bits. This uses at most $K \leq B$ additions and at most one incident addition per vertex. Maximum degree one satisfies the declared degree upper bound $\Delta \geq 2$.

For any fixed $0 < \lambda \leq 1$, the original teacher with $w^\theta = K^{-1/2} \sum_k \theta_k e_k$ realizes the targets on both clean and released graphs. At depth one a positive block has positive scores at both endpoints; positive

self-loops preserve positivity for any greater depth. Negative blocks use zero readout coefficient and fillers have zero feature. The readout norm is at most one. Clean bridge presence identifies every bit; released connected components identify the pairs but reveal no bit. The teacher's zero scores on negative blocks remain a limitation of this sparse witness: this is not a positive-margin extension.

Let $y_s = 2Y_s - 1$. For each learner in Section D.1, optimize

$$\mathcal{L}(w) = \sum_{s \in S} \log(1 + e^{-y_s f_w(s)}), \qquad w_{t+1} = w_t - \nabla \mathcal{L}(w_t).$$

All parameters are finite at each finite step. With $S = \varnothing$ the sum and gradient are zero. For $m > 0$, mean loss with step size m is equivalent to this summed-loss convention; using a fixed step size with mean loss gives a different iteration scale. Predictions are $\mathbf{1}\{f > 0\}$. Perfect fit means zero classification error: a sampled negative filler has logit zero, loss $\log 2$, and zero gradient. It is correctly classified by the declared tie rule, without a positive signed margin.

D.3 Proof of the exact untied and tied risks

On a released pair, $f_{\mathrm{un}}(p_k) = a_k$ and $f_{\mathrm{un}}(r_k) = b_k$. Observing one endpoint updates only its own coordinate. An observed signed coordinate z evolves as

$$z_{t+1} = z_t + \frac{1}{1 + e^{z_t}}.$$

Initially $z_0 = \pm 1/4$. Both initial values give $z_1 > 0$, and later updates are positive. An unobserved endpoint's coordinate never changes and has an independent fair sign. For each fixed world and sample, each active test vertex therefore has error probability $1/2$ over initialization. Uniform sampling leaves $2K(N-m)/N$ active test vertices in expectation. Dividing their expected error by $N - m$ proves $R_{\mathrm{un}} = K/N$.

For f_{tie} both endpoints have score t_k. A covered pair supplies one or two identical labels, multiplying the scalar update by one or two. The first signed update is strictly positive and stays so. A missed pair retains one fair random sign and contributes one expected mistake, for every fixed value of its bit. Its miss probability is $\binom{N-2}{m}/\binom{N}{m}$. Thus

$$R_{\mathrm{tie}} = \frac{K}{N-m} \frac{\binom{N-2}{m}}{\binom{N}{m}} = \frac{K(N-m-1)}{N(N-1)}.$$

The independent fair-bit lower bound and the copy/fair decoder give the same fixed-layout minimax value. This equality does not optimize over Q or over arbitrary graph classes. Subtraction proves the stated excess risk.

D.4 Positive overlap: delay and population risk

First consider one observed point s and an unobserved same-label target v with fixed design rows h_s, h_v. Other observations have design rows orthogonal to both. Put $q = \|h_s\|^2 > 0$, $r = \langle h_v, h_s \rangle > 0$, $z_t = y f_t(s)$, and $y f_0(v) = -a_0 < 0$. For summed logistic GD with step size $\eta > 0$ and $\alpha = \eta q$, the exact recurrences are

$$z_{t+1} = z_t + \frac{\alpha}{1 + e^{z_t}}, \qquad y f_t(v) = -a_0 + \frac{r}{q}(z_t - z_0). \tag{12}$$

Writing $d_t = \alpha/(1 + e^{z_t}) \le \alpha$ and using $e^d - 1 \le d e^d$ for $d \ge 0$ gives

$$e^{z_{t+1}} - e^{z_t} \le e^{z_t} d_t e^{d_t} \le \alpha e^{\alpha}, \qquad z_t \le \log(e^{z_0} + t\alpha e^{\alpha}).$$

Hence a strictly positive target score requires

$$t > \frac{e^{z_0 + a_0 q/r} - e^{z_0}}{\alpha e^{\alpha}}. \tag{13}$$

If $z_0 > 0$ the observed prediction stays strictly correct during this delay. Also $z_t \to \infty$: if this increasing sequence had a finite upper bound, its increments would be bounded below by a positive constant. Thus transfer eventually occurs whenever $r > 0$ in this single-observation setting.

For the graph learner f_δ, the two design rows within a block, restricted to its three coordinates, are $(1, 0, \delta)$ and $(0, 1, \delta)$. Thus $q = 1 + \delta^2$, $r = \delta^2$, and different blocks are orthogonal. If an endpoint is observed, its first signed update is at least its diagonal contribution $q/(1 + e^{z_0})$; a second observed endpoint only adds a nonnegative same-label term. Since $q \geq 1$ and $z_0 = \pm 1/4$, every observed active score is strictly correct after one step and stays so.

To prove the population plateau, partition pairs by whether they have zero, one or two observations. A missed pair keeps its initial endpoint scores, each with error probability $1/2$. A pair with two observations has no test endpoint. In a singleton-covered pair, an initially positive signed target score stays positive. For an initially negative target score $-1/4$, the observed signed initial score is independently $z_0 \in \{-1/4, 1/4\}$. Its target-score correction is at most

$$\frac{r}{q} \log\left(1 + Tqe^q e^{-z_0}\right).$$

The largest bound occurs at $z_0 = -1/4$. For $T < \tau_-(\delta)$ it is strictly less than $1/4$, so every initially wrong target remains strictly wrong. Each singleton-covered test endpoint therefore also has error probability exactly $1/2$. Every active test vertex contributes $1/2$ in expectation, and the same count as in Section D.3 proves $R_\delta(T, m) = K/N$. The strict inequality on T avoids the zero-logit boundary. This holds for every fixed world, not only a prior average.

For each fixed positive δ, (12) eventually corrects every singleton-covered target, while missed-pair predictions stay unchanged. Bounded convergence over the finite sample and initialization spaces yields $R_\delta(T, m) \to R_2^\star$. This completes the proof of Proposition D.1.

Quantitative and initialization boundaries. At $\delta = .1$, (11) is approximately 2.5966×10^{10}. Only finite verification runs are reported below; this large horizon is derived, not simulated. With zero initialization, one observation instead gives $yf_1(v) = \eta r/2 > 0$. Weak overlap alone does not force a classification delay. Iteration counts depend on initialization, loss normalization, parameterization and optimizer. For example, a unit-norm gradient-direction update changes the target signed score by $\eta r/\sqrt{q}$ per step in the single-observation case; its crossing time is proportional to $1/r$. The exponential delay above is therefore not a computational lower bound for all learning rules.

D.5 A GENERAL FIXED-REPRESENTATION OVERLAP LAW

Let the observable homogeneous blocks be C_k with independent fair bits. Fix representations h_v, independently of the bits and the label sample, with nonzero active rows, zero filler rows, and Gram matrix $G_{vs} = \langle h_v, h_s \rangle$. Assume zero Gram entries across blocks and nonnegative entries within blocks. Define $O_v = \{s \in C_k : G_{vs} > 0\}$ for $v \in C_k$; then $v \in O_v$.

Proposition D.2 (Overlap support and prior-average risk)**.** *Train a bias-free linear readout from zero by summed logistic GD with any strictly positive finite step sizes. At every finite $T \geq 1$, every observed label is correctly classified; an active target has strictly correct signed score exactly when $S \cap O_v \neq \varnothing$, and otherwise has score zero. Its fair-bit prior-average test risk is*

$$\mathbb{E}R_T = \frac{\sum_{v \in \cup_k C_k} \binom{N-|O_v|}{m}}{2(N-m)\binom{N}{m}}. \tag{14}$$

The excess over the fixed-layout coverage reference is

$$\frac{\sum_k \sum_{v \in C_k} [\binom{N-|O_v|}{m} - \binom{N-|C_k|}{m}]}{2(N-m)\binom{N}{m}}.$$

Proof. For every observed s, accumulated update coefficients $c_s(T)$ are strictly positive. Zero initialization gives $w_T = \sum_{s \in S} c_s(T) y_s h_s$. The cross-block assumption and the common label within C_k imply

$$y_k f_T(v) = \sum_{s \in S \cap C_k} c_s(T) G_{vs}.$$

Nonnegativity makes this positive exactly when the sample hits O_v; otherwise it is zero. An observed active row has a positive diagonal, and fillers are correct by the zero-to-negative convention. Avoiding

O_v also makes v unobserved, since $v \in O_v$. Its probability is $\binom{N-|O_v|}{m}/\binom{N}{m}$. On that event the target score is zero for either value of its fair bit and errs only for the positive bit. Summing and dividing by $N-m$ proves the formula. Subtracting the coverage reference proves the excess expression. □

This is a prior-average statement, not the fixed-world statement of Proposition D.1. Negative kernel entries, cross-block label interactions, and label-dependent learned representations need not satisfy its sign characterization. The underlying kernel update is standard (Yang et al., 2024); the formula combines it with the finite sampling law.

D.6 A SEPARATE AFFINE-HEAD FAILURE SELECTED BY TRAINING

Suppose fixed rows are $h_v = a_v e_k$ on block C_k, with $a_v > 0$, and zero on known-negative fillers. Assume at least one filler is observed. Consider the hard-margin problem with the bias included in the norm,

$$\min_{w,b} \tfrac{1}{2}(\|w\|^2 + b^2) \quad \text{subject to} \quad y_s(w^\top h_s + b) \geq 1 \quad (s \in S).$$

The unique solution has $b^* = -1$. For an observed positive block let $a_k^{\min} = \min_{s \in S \cap C_k} a_s$; then $w_k^* = 2/a_k^{\min}$. All other coordinates are zero. Indeed the negative filler forces $b \leq -1$. Writing $b = -t$, $t \geq 1$, the minimum-norm feasible positive-block coordinate is $(1+t)/a_k^{\min}$; zero is feasible and minimum norm elsewhere. The resulting objective increases strictly with t, proving the claim. Thus a positive covered target is misclassified exactly when $2a_v \leq a_k^{\min}$ under the specified threshold.

The original family contains a strict-error example: $Q = 2, K = 1, F = 1, L = 1$ and $\lambda = 1/4$ give propagated amplitudes $(.2, .8, 0)$ at payload, root and filler. Observe the positive root and the negative filler. The solution $(w, b) = (2.5, -1)$ has scores $(-.5, 1, -1)$, so it fits both labels but misclassifies the identified payload. Logistic GD on augmented rows $(h_v, 1)$ converges in direction to the hard-margin solution under a suitable step size (Soudry et al., 2018, Theorem 3). Hence it eventually agrees on targets with nonzero maximum-margin score. Directional convergence alone does not settle a zero limiting score or provide a finite crossing time. The supplied step-size-one trajectory from zero initially predicts the payload correctly and makes it incorrect by step five. This is an affine-head example, not an explanation of bias-free SAGE or a theorem for Adam.

For completeness, the hard-margin head also has a conditional exact risk. Let $A_{\text{act}} = \sum_k |C_k|$, let E mean that at least one filler is observed, and assume $D_E = \binom{N}{m} - \binom{A_{\text{act}}}{m} > 0$. Define $T_v = \{s \in C_k : a_s < 2a_v\}$. For independent fair block bits,

$$\mathbb{E}[R_{\text{MM}} \mid E] = \frac{\sum_{v \text{ active}} [\binom{N-|T_v|}{m} - \binom{A_{\text{act}}-|T_v|}{m}]}{2(N-m)D_E}.$$

Avoiding T_v, which contains v, makes a positive test target wrong. Subtracting samples containing only active nodes imposes E. Negative targets are correct. This counts the conditional risk of the SVM itself; its asymptotic GD interpretation additionally requires nonzero limiting test scores.

D.7 SCOPE AND FINITE COMPUTATIONAL CHECKS

For the paired learner the excess is bounded by $\min\{B, d, N/2\}m/[N(N-1)]$. Constant overall risk thus requires K, and the permitted dimension and edit-budget bounds, to scale with N. The theorem does not establish a trained-risk analogue of the larger-block depth-capacity law. It also does not show failure exclusive to poisoning: if K_+ blocks are positive, the same untied learner on clean pairs has $R_{\text{un,clean}} = (K + K_+)/(2N)$. Each positive pair has two independent endpoint coordinates, whereas a negative clean pair has only one nonzero design row. Counting their unobserved random signs proves the formula. When all bits are positive, no edge is added and $R_{\text{un,clean}} = K/N$. The unrestricted clean-graph structural decoder still has zero risk.

The accompanying `theory_checks` directory verifies the new formulas. Exhaustive paired enumeration uses $N = 7, K = 3, F = 1$, all worlds, all samples with $m < N$, all sign initializations, and horizons one and seven: 146,304 optimizer trajectories give maximum risk discrepancy 1.12×10^{-16}. A separate checker verifies the positive-overlap population plateau across 6,160 trajectories with (K, N, δ, T) equal to $(1, 2, .1, 50)$, $(1, 4, .2, 50)$, $(2, 5, .2, 40)$ and $(2, 6, .1, 17)$, enumerating all worlds, samples and initializations in each case. The overlap-law checker uses 6,120

trajectories; 99 independently solved affine quadratic programs agree with the closed form to below 6.21×10^{-12} in parameter coordinates. Five weak-overlap trajectories are checked through 20,000 steps; unobserved crossings are censored. These are finite implementation checks, not independent statistical replications or replacements for proofs. No new nonlinear or natural-graph training is claimed by these calculations.

E COMPLETE PROOFS

This appendix makes explicit the construction and all steps suppressed in the main text. Throughout, S is uniform over the m-subsets of $[N]$, $U = [N] \setminus S$, and $u = N - m$.

E.1 BRIDGE-TREE REALIZATION

For maximum degree $\Delta \geq 2$ and depth $L \geq 1$, let

$$q_L = \sum_{t=0}^{L-1} (\Delta - 1)^t, \qquad Q_L = 1 + q_L.$$

Start with a rooted $(\Delta - 1)$-ary tree of depth $L - 1$ and add one payload vertex joined to the root by a bridge. The tree has q_L vertices and the complete block has Q_L vertices. For any integer $2 \leq Q \leq Q_L$, repeatedly remove deepest leaves while preserving the root. The resulting tree has $Q - 1$ vertices, depth at most $L - 1$, root degree at most $\Delta - 1$, and all other degrees at most Δ. After the payload bridge is added, the released block has maximum degree at most Δ and every block node lies within distance L of the payload.

Relabel the chosen distinct coordinates as e_k for this proof. Use one feature coordinate e_k for block k: $X_{p_k} = e_k$ and all other block features are zero. Because $P_\lambda(A)$ is entrywise nonnegative and has positive self-loop and edge entries, if the bridge is present then every node in the block receives a strictly positive multiple of e_k after exactly L propagation steps. If the bridge is absent, no tree node receives this coordinate. For $w^\theta = K^{-1/2} \sum_k \theta_k e_k$, thresholding at zero gives the common block target $Y_v = \theta_k$. The payload node itself remains positive by self-loops when $\theta_k = 1$. The readout norm is at most one and payload feature norms equal one.

The attacker adds every bridge missing under θ before the random labeled set S is drawn. It reads neither S nor labels. The number of edits is $K - \|\theta\|_0 \leq K \leq B$. Since blocks and their bridges are disjoint, every vertex is incident to at most one edit. On the common released graph, w^θ still outputs one precisely in blocks with $\theta_k = 1$, so every clean target remains realizable.

Proposition E.1 (The attack creates the label requirement)**.** *In the corresponding unpoisoned experiment, the minimax test risk is zero for every split, including* $m = 0$.

Proof. In the clean graph, a connected component containing a nonzero payload feature and at least two vertices has its bridge present and hence target one throughout. An isolated payload has an absent bridge and target zero. Every component without a payload has target zero. This decoder reads only clean adjacency and features; it does not need to assign a disconnected zero tree to its isolated payload. It works for every pruned block shape and for the known-zero fillers. Canonicalization makes all these bridges present and removes exactly this structural sufficient statistic. □

E.2 FIXED HOMOGENEOUS FAMILY

Proposition E.2 (Fixed-family minimax equality)**.** *For K recognizable disjoint blocks of size Q satisfying $KQ \leq N$, the randomized minimax test risk is*

$$\mathfrak{R}^\star_{K,Q} = \frac{KQ}{2u} \frac{\binom{N-Q}{m}}{\binom{N}{m}}.$$

Proof. For the lower bound, place independent $\mathrm{Bernoulli}(1/2)$ priors on $\theta_1, \ldots, \theta_K$. After canonicalization, $(\widetilde{G}, X)$ does not depend on θ. If $S \cap C_k = \varnothing$, no observed label depends on θ_k, so its posterior conditional on the complete observation remains fair. The entire block then lies in U, and

the minimum posterior expected number of zero-one mistakes on its Q common labels is $Q/2$. If $H_Q = \sum_k \mathbf{1}\{S \cap C_k = \varnothing\}$, conditional Bayes risk is at least $QH_Q/(2u)$. For every fixed block,

$$\Pr(S \cap C_k = \varnothing) = \frac{\binom{N-Q}{m}}{\binom{N}{m}},$$

because all m labels must be chosen from the other $N - Q$ vertices. Linearity of expectation (independence of miss indicators is unnecessary) gives the displayed lower bound. Minimax risk is at least Bayes risk.

For the upper bound, parse the recognizable blocks. If block C_k contains a revealed label, copy it to every node in the block; realizability and common labels make this exact. If it contains no revealed label, draw one fair bit and predict it throughout the block. Predict the known zero on fillers. Conditional on S, this rule has expected test error $QH_Q/(2u)$ for every fixed θ, hence its worst-case expected risk equals the lower bound after averaging over S. □

E.3 Proof of Theorem 4.1

For feasible Q, put $K_Q = \min\{B, d, \lfloor N/Q \rfloor\}$. Proposition E.2 supplies the exact value

$$R_Q = \frac{K_Q Q}{2u} \frac{\binom{N-Q}{m}}{\binom{N}{m}}$$

on that subclass, so the union minimax risk is at least $\max_Q R_Q$.

For the reverse inequality, use the same block decoder from the proposition. The released graph reveals which union member occurred, including Q and K, so no oracle knowledge of the clean bit vector is used. On every member, its risk is exactly that member's fixed-family value. The worst risk of this one rule over the union is $\max_Q R_Q$. This matches the lower bound. When the feasible set is empty or $K_Q = 0$ everywhere, no active ambiguity block exists and the attack-induced risk is zero.

E.4 Heterogeneous blocks and decision-rule scope

Let disjoint recognizable blocks have sizes $Q_1, \ldots, Q_K$. The preceding prior argument applies independently to each hidden bit, and the block decoder applies to each realized block. Therefore

$$\mathfrak{R}^{\star}_{\{Q_k\}} = \frac{1}{2u} \sum_{k=1}^{K} Q_k \frac{\binom{N-Q_k}{m}}{\binom{N}{m}}.$$

Proposition E.3 (Exact deterministic value)**.** *For a fixed recognizable layout, put $u = N - m$, $D = \binom{N}{m}$, and $A = \sum_k Q_k \binom{N-Q_k}{m}$. The randomized value is $A/(2uD)$, and the exact deterministic value is*

$$R^{*}_{\mathrm{det}} = \begin{cases} N^{-1} \sum_k \lceil Q_k/2 \rceil, & m = 0, \\ \lceil A/2 \rceil/(uD), & 1 \le m < N. \end{cases} \tag{15}$$

For the homogeneous union and $m \ge 1$, its deterministic value is

$$\max_Q \frac{\left\lceil K_Q Q \binom{N-Q}{m}/2 \right\rceil}{(N-m)\binom{N}{m}}, \qquad K_Q = \min\{B, d, \lfloor N/Q \rfloor\},$$

where $2 \le Q \le \min\{Q_L, N\}$. At $m = 0$ it is $\max_Q K_Q \lceil Q/2 \rceil / N$. Empty families have zero value.

Proof. For $m = 0$, a deterministic rule predicting one on a_k vertices of block k incurs $\max\{a_k, Q_k - a_k\} \ge \lceil Q_k/2 \rceil$ mistakes for the worse bit, chosen independently across blocks. Balanced predictions attain the sum. For $m \ge 1$, put $t_k = \binom{N-Q_k}{m}$, $D = \binom{N}{m}$, and $A = \sum_k Q_k t_k$. The fair-bit prior gives the lower bound $A/(2uD)$. For every fixed world, a deterministic rule's total mistakes summed over the D equally likely splits is an integer. Its worst numerator is therefore at least $\lceil A/2 \rceil$.

We construct a matching rule. Copy hit-block labels; predict zero on fillers. On a missed even block predict each label on half its vertices. On an odd block, choose a majority bit h and predict it on

$(Q_k + 1)/2$ vertices, using fixed node order. Writing $s_b = 2b - 1$, its excess over $Q_k/2$ mistakes is $-s_{\theta_k} s_h/2$.

If t_k is even, alternate majority bits zero and one over the ordered collection of splits missing block k; its total excess is zero for either target bit. Pair the odd blocks with odd t_k. For a pair (i, j) choose a distinguished split S_i missing i but containing a fixed vertex of j, and a split S_j missing j but containing a fixed vertex of i. Such splits exist because $m \geq 1, t_i, t_j > 0$, and the blocks are disjoint. Complete each split with $m - 1$ other vertices outside its missed block. On S_i choose majority $h_i = \theta_j$; on S_j choose $h_j = 1 - \theta_i$. These labels are observed on the respective splits. The two excesses are $-s_{\theta_i} s_{\theta_j}/2$ and $+s_{\theta_i} s_{\theta_j}/2$, canceling for every fixed world. Alternate majorities on the remaining, evenly many, missing splits in each block. If one residual block is unpaired, choose majority zero on one distinguished split and alternate on all others. Its worst excess is $1/2$.

The number of unpaired residuals is A modulo two. Thus the maximum total numerator is $\lceil A/2 \rceil$, as required. The cancellations occur across splits for a fixed world; the adversary never observes the split. Lexicographic subset ranking by binomial counts implements the alternation without enumerating all splits. Deleting a distinguished split subtracts one from subsequent ranks. These ranks and binomial coefficients have polynomial bit complexity in N. The decoder uses only parsed blocks, sampled positions, and observed labels.

For the union, the parser identifies the layout; apply this decoder on each member. Restricting to any fixed layout gives the reverse bound. For each Q the value is nondecreasing in K, hence K_Q is worst. This proves both union formulas. In particular, two size-three blocks at $N = 6, m = 1$ have deterministic value $3/10$, whereas separately balancing each block across splits gives $1/3$ (and rounding independently on every split gives the weaker $2/5$). The positive label count is essential: at $m = 0$ the same layout has value $2/3$. □

E.5 Proof of Corollary 5.1

Put $Q_c = \min\{Q_L, N\}$ and $b = \min\{B, d\}$. If $m > N - Q$, $p_0(Q) = 0$. Otherwise every factor below is nonnegative, and

$$p_0(Q) = \prod_{j=0}^{m-1} \left(1 - \frac{Q}{N - j}\right) \leq e^{-mQ/N}.$$

Thus this exponential upper bound holds also when the block cannot be missed. Since $u \geq 3N/4$, $K_Q Q \leq N$, $K_Q \leq b$, and $xe^{-x} \leq 1$ for $x \geq 0$, the exact risk is at most

$$\frac{2}{3} \min\{1, bQ_L/N, b/m\}.$$

For the lower bound define

$$q_0 = \min\{Q_c, N/m, N/b\}, \qquad Q = \max\{2, \lfloor q_0/2 \rfloor\}.$$

The principal regime gives $N/m, N/b \geq 4$, and hence $q_0 \geq 2$. If $q_0 < 4$ then $Q = 2 \leq Q_c, N/(2m), N/(2b)$; if $q_0 \geq 4$ these inequalities follow by rounding down $q_0/2$. In both cases $Q \geq q_0/4$, and Q is an admissible integer. Consequently $K_Q = b$, and a union bound for a fixed block gives $p_0(Q) \geq 1 - mQ/N \geq 1/2$. Since $u \leq N$,

$$\mathfrak{R}_L^\star \geq \frac{bQ}{4N} \geq \frac{bq_0}{16N} = \frac{1}{16} \min\{1, bQ_L/N, b/m\}.$$

In the last equality, truncating Q_L at N does not affect the minimum, because $b \geq 1$. This proves the claimed constants, including rounding cases.

E.6 Proof of Corollary 5.2

Take $\varepsilon \leq 1/64$, $1 \leq b \leq \varepsilon N/16$, and $Q_L \geq 32\varepsilon N/b$. These imply $N \geq 1024$ and $b \leq N/4$. For $m_+ = \lceil b/\varepsilon \rceil$,

$$1 \leq m_+ \leq N/16 + 1 \leq N/4, \qquad \mathfrak{R}_L^\star(m_+) \leq \tfrac{2}{3} b/m_+ \leq \tfrac{2}{3}\varepsilon.$$

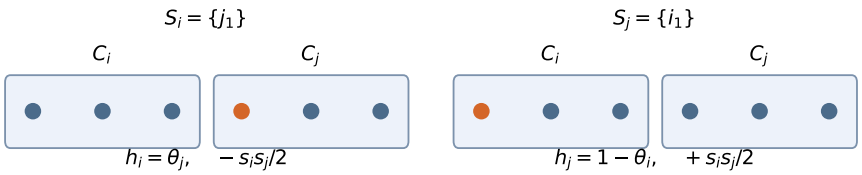

Figure 4: Deterministic residual pairing. Orange denotes the single observed vertex. The two panels are different possible splits under the same fixed world, not simultaneous observations. The displayed signed excess errors cancel across the split distribution.

Thus the threshold exists and is at most $2b/\varepsilon$. For every principal-regime integer $m < b/(32\varepsilon)$, all three terms in the phase minimum are at least 32ε: $1 \geq 64\varepsilon$, $bQ_L/N \geq 32\varepsilon$, and $b/m > 32\varepsilon$. The lower bound is therefore at least $2\varepsilon > \varepsilon$. It follows that $m_L^\star \geq b/(32\varepsilon)$. These sufficient constants are conservative; finite numerical comparisons outside their hypotheses are descriptive, not validation of this sufficient regime.

Endpoint and monotonicity check. Define $\binom{a}{j} = 0$ for integer $j < 0$ or $j > a$, with $a \geq 0$. At $m = N - 1$ every feasible block of size at least two is hit, so risk is zero. For $m \leq N - 2$ and $p_0(Q; m) > 0$, the fixed-Q risk ratio is

$$\frac{R_Q(m+1)}{R_Q(m)} = \frac{N - Q - m}{N - m - 1} \leq 1.$$

If $p_0(Q; m) = 0$, both risks vanish. The maximum over the same feasible Q set is consequently nonincreasing in m. The case $m = N$ is excluded because the test set is empty; no division by zero or positive endpoint risk is claimed.

F Worked deterministic cancellation

Worked cancellation: two odd blocks. Take $N = 6$, $Q_i = Q_j = 3$, $m = 1$. Each block is missed on three splits, so $A = 18$, $uD = 30$. Rounding each block's total across splits separately costs $\lceil 9/2 \rceil + \lceil 9/2 \rceil = 10$, giving $1/3$; cross-block pairing attains $9/30 = 3/10$. For a missed three-node block, predict its majority h on two nodes and $1 - h$ on one. Relative to $3/2$ mistakes its excess is $-s_\theta s_h/2$, where $s_b = 2b - 1$. On the distinguished splits below, use the observed partner bit to set h. The two excesses cancel for every fixed (θ_i, θ_j). On each block's two remaining missed splits, alternate $h = 0, 1$ to cancel again. This explains why balancing across blocks improves even the stronger per-block, across-split baseline; per-split rounding alone would give $2/5$.

G Exact noisy risk and optimized phase

Suppose each sampled label is independently flipped with probability $\eta \in [0, 1/2]$, independently of the split and clean world. Test error is still measured against clean targets. Let $\delta = 1 - 2\eta$ and let $e_h(\eta) = \Pr\{\mathrm{Bin}(h, \eta) > h/2\} + \frac{1}{2}\Pr\{\mathrm{Bin}(h, \eta) = h/2\}$, with $e_0 = 1/2$. Majority voting with fair ties gives the exact fixed-family randomized value

$$R_\eta^*(K, Q, m) = \frac{K}{N - m} \sum_{h=0}^{\min(Q,m)} \frac{\binom{Q}{h}\binom{N-Q}{m-h}}{\binom{N}{m}} (Q - h) e_h(\eta). \tag{16}$$

The independent fair-bit prior matches this decoder for every world. Maximizing Equation (16) over feasible Q with $K = K_Q$ gives the exact optimized union value. Noisy observations change the learning experiment, while the clean targets and their graph realizations remain the same.

Theorem G.1 (Optimized noisy coverage phase)**.** *For $1 \leq m \leq N/4$, $1 \leq b = \min\{B, d\} \leq N/4$, and $Q_L \geq 2$, put $\Psi = \min\{1, bQ_L/N, b/(m\delta^2)\}$, interpreting the last entry as $+\infty$ if $\delta = 0$. Then*

$$\frac{\Psi}{128} \leq R_{\eta,L}^* \leq \Psi.$$

The proof uses the affinity $\sqrt{1-\delta^2}$ of two Bernoulli observation laws and sampling without replacement; it is proved below. The coverage scale becomes $N/(m\delta^2)$. In a qualified nondegenerate regime, Corollary G.2 below also gives label complexity $\Theta(b/(\varepsilon\delta^2))$. At $\eta = 1/2$, labels carry no block-bit information. The factor δ^{-2} is the standard binary-testing information scale, expressed here through the legal graph construction and its optimized block capacity. These randomized noisy results do not transfer Equation (15) to nonuniform noisy transcripts.

G.1 Exact value and random-test-index identity

Conditional on a block receiving h labels, the two possible bits pass through h independent binary symmetric channels. Their likelihood ratio is the majority statistic. The fair-bit Bayes error is $e_h(\eta)$; majority with fair ties has that error for either true bit. Multiplying by the $Q-h$ unobserved nodes and averaging the hypergeometric count proves Equation (16). Independent fair priors on all bits supply one common lower bound, and the block majority decoder attains it on every world. Known-zero fillers require no inference. The observable parser supplies the union upper bound, as in Theorem 4.1. For heterogeneous layouts sum the one-block expressions.

Choosing a uniform test node first gives the equivalent identity

$$R^*_\eta(K,Q,m) = \frac{KQ}{N}\mathbb{E}[e_H(\eta)], \qquad H \sim \mathrm{Hypergeom}(N-1, Q-1, m).$$

Indeed the test node is in an active block with probability KQ/N, and the m training positions are a uniform subset of the other $N-1$ nodes. This identity handles the changing denominator exactly.

G.2 Proof of Theorem G.1

Put $\alpha = 2\sqrt{\eta(1-\eta)} = \sqrt{1-\delta^2}$. For two equally likely product observation laws, $\min\{a,b\} \le \sqrt{ab}$ bounds Bayes error by $\alpha^h/2$. Use $\alpha^0 = 1$ when $\eta = 0$. The quantity $\mathbb{E}[\alpha^H]$ is the degree-m elementary symmetric mean of $Q-1$ weights α and $N-Q$ weights one. This is at most the mth power of the arithmetic mean. To see this, hold the sum of any two nonnegative weights fixed and average them. The only changing part of the symmetric polynomial is their product times a nonnegative coefficient; averaging increases it. Repeating pair averaging and taking a limit gives equal weights. Therefore

$$\mathbb{E}[e_H] \le \frac{1}{2}\left(1 - \frac{Q-1}{N-1}(1-\alpha)\right)^m \le \frac{1}{2}\exp\left(-\frac{m\delta^2 Q}{4N}\right),$$

since $1-\alpha \ge \delta^2/2$ and $(Q-1)/(N-1) \ge Q/(2N)$ for $Q \ge 2$. Using $KQ \le N$, $K \le b$, and $\max_{x\ge 0} xe^{-ax} = 1/(ae)$ gives simultaneous upper bounds $1/2$, $bQ_L/(2N)$, and $(2/e)b/(m\delta^2)$. At $\delta = 0$ use the first two. These imply the claimed upper bound.

For the lower bound set $q_0 = \min\{Q_L, N/b, N/(m\delta^2)\}$, omitting the last entry if $\delta = 0$, and choose $Q = \max\{2, \lfloor q_0/16 \rfloor\}$. The principal regime gives $q_0 \ge 2$, $Q \le q_0$, and $Q \ge q_0/32$, so $K = b$ is feasible. If $Q = 2$, the test node's partner is unobserved with probability $1 - m/(N-1) \ge 2/3$; thus its Bayes error is at least $1/3$. If $Q > 2$, then $\delta^2\mathbb{E}H \le 2m\delta^2 Q/N \le 1/8$. The affinity of the two joint observation laws, including the positions, is $\mathbb{E}[\alpha^H]$. Since $1-\alpha^H \le H(1-\alpha) \le H\delta^2$, it is at least $7/8$. For probability masses p, q, Cauchy–Schwarz applied to $|p-q| = |\sqrt{p}-\sqrt{q}|(\sqrt{p}+\sqrt{q})$ gives total variation at most $\sqrt{2(1-\sum\sqrt{pq})}$. The variation is therefore at most $1/2$, and the equal-prior Bayes error is at least $1/4$. Consequently

$$R^*_{\eta,L} \ge \frac{bQ}{4N} \ge \frac{bq_0}{128N} = \frac{\Psi}{128}.$$

This proves all parameter cases, including $\eta = 0$ and $\eta = 1/2$.

Corollary G.2 (Qualified noisy label complexity)**.** *Let $0 \le \eta < 1/2$, $0 < \varepsilon \le 1/512$, $1 \le b \le \varepsilon\delta^2 N/16$, and $Q_L \ge 256\varepsilon N/b$. The first label count in $\{1, \dots, \lfloor N/4 \rfloor\}$ achieving risk at most ε obeys*

$$\frac{b}{256\varepsilon\delta^2} \le m^*_\eta(\varepsilon) \le \frac{2b}{\varepsilon\delta^2}.$$

Proof. Below the lower threshold, all three scales in Ψ are at least 256ε, so risk exceeds ε. At $m = \lceil b/(\varepsilon\delta^2) \rceil$ the upper bound is at most ε. The population condition makes this count feasible and bounds it by twice the unrounded quantity. The conditions imply $N \geq 8192$, so the rounding cannot leave the principal interval. At $\eta = 1/2$ no such label-complexity statement is made. □

The supplement enumerates noisy transcripts and independently checks the random-test-index identity and optimized phase. These are implementation checks, not substitutes for the preceding proofs. The model does not cover adversarial label corruption or noisy relational parity propagation.

H TRAINED BLOCK EXPERIMENT: PROTOCOL AND FULL RESULTS

Each graph contains eight full bridge trees, with zero-feature isolated fillers. A breadth-first enumeration assigns tree parents; the released graph completes all bridges. Independent fair bits specify targets and the corresponding clean bridges. We sample m physical nodes uniformly without replacement, independently of bits. Each cell uses 20 independent seeds. Features contain only one payload coordinate per block; training does not receive the clean bridge mask.

Write $\widehat{P} = \widehat{D}^{-1/2}(A + I)\widehat{D}^{-1/2}$. SGC trains $\widehat{P}^L X w$ from $w = 0$. The GCN computes $H^{j+1} = \mathrm{ReLU}(\widehat{P}H^j W_j)$ for $j < L - 1$ and uses a linear final graph-convolution output. Hidden width is 32; there are no biases or dropout. Each weight is initialized independently with variance $2/\text{fan-in}$. Zero logits predict zero. Both methods minimize binary cross entropy only on S, using 200 full-batch Adam updates with learning rate 0.02, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and zero weight decay. No validation split or early stopping is used. These are specified adaptations of the cited architectures, rather than a claim to reproduce their benchmark hyperparameters. A NumPy/SciPy implementation checks every weight gradient on a small graph by central differences; the maximum absolute discrepancy is 7.56×10^{-11}. The supplied check also verifies that changing test labels cannot affect the training loss or gradients.

The population-test loss decomposes into errors on hit blocks, missed blocks, and fillers, each divided by $N - m$. The reference decoder copies observed labels and integrates a fair guess on missed blocks. Its split-conditional risk is $QH_Q/[2(N - m)]$. Our empirical comparison samples the proven least-favorable prior; it is not an empirical maximization over clean worlds. The entire protocol was written before this run, but this is a development experiment, not independently preregistered confirmation. No hyperparameters were selected after observing it.

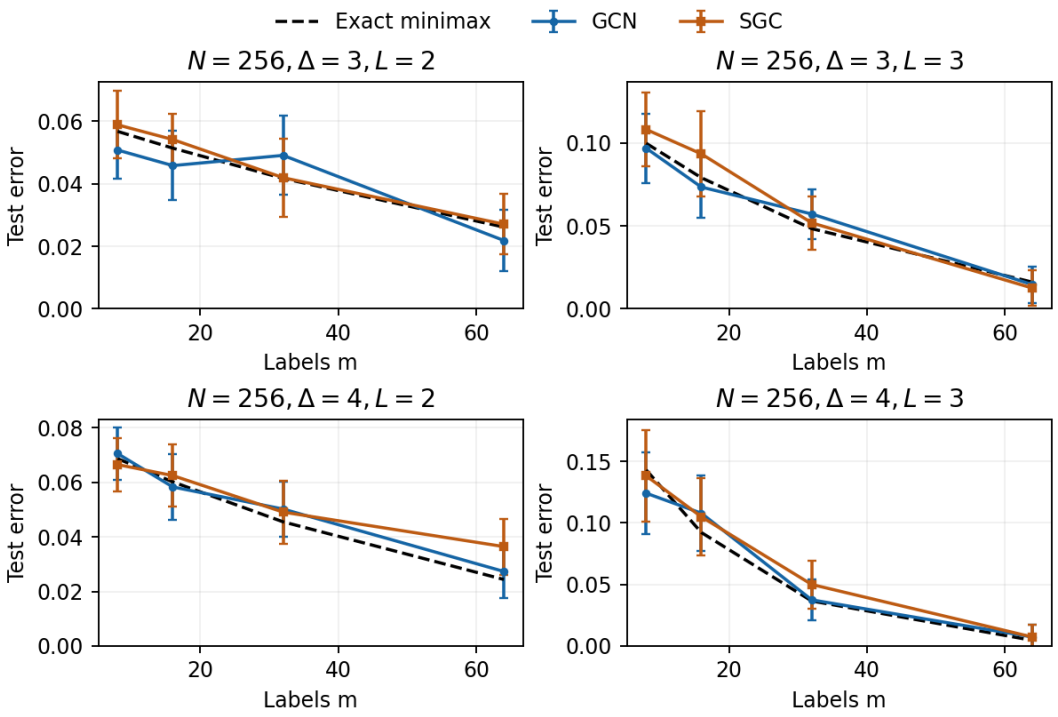

Figure 5: Original experiment: the four declared $N = 256$ slices, with 20-sample means and two-standard-error bars against the fixed-family reference. The complete original grid is reported in this appendix.

Table 21: Original block calibration: fixed-family information risk and mean test-error fractions. Each cell has $K = 8$ blocks of size $Q = Q_L$ and 20 sampled problems. Parentheses give standard errors. Covered-block and filler errors are zero in all listed cells.

N	Δ	L	m	Exact	GCN (SE)	SGC (SE)
256	3	2	8	0.0568	0.0508 (0.0046)	0.0589 (0.0054)
256	3	2	16	0.0514	0.0458 (0.0055)	0.0542 (0.0042)
256	3	2	32	0.0417	0.0491 (0.0063)	0.0420 (0.0063)
256	3	2	64	0.0262	0.0219 (0.0049)	0.0271 (0.0048)
256	3	3	8	0.0997	0.0968 (0.0105)	0.1081 (0.0110)
256	3	3	16	0.0790	0.0733 (0.0092)	0.0933 (0.0129)
256	3	3	32	0.0483	0.0571 (0.0075)	0.0518 (0.0080)
256	3	3	64	0.0161	0.0146 (0.0055)	0.0125 (0.0053)
256	4	2	8	0.0687	0.0706 (0.0047)	0.0665 (0.0049)
256	4	2	16	0.0602	0.0583 (0.0060)	0.0625 (0.0057)
256	4	2	32	0.0455	0.0502 (0.0051)	0.0491 (0.0057)
256	4	2	64	0.0244	0.0273 (0.0048)	0.0365 (0.0051)
256	4	3	8	0.1431	0.1242 (0.0167)	0.1383 (0.0185)
256	4	3	16	0.0922	0.1079 (0.0154)	0.1050 (0.0156)
256	4	3	32	0.0366	0.0375 (0.0084)	0.0500 (0.0097)
256	4	3	64	0.0046	0.0073 (0.0050)	0.0073 (0.0050)
512	3	2	8	0.0298	0.0310 (0.0025)	0.0270 (0.0017)
512	3	2	16	0.0284	0.0274 (0.0018)	0.0310 (0.0020)
512	3	2	32	0.0257	0.0225 (0.0029)	0.0263 (0.0027)
512	3	2	64	0.0209	0.0250 (0.0031)	0.0223 (0.0032)
512	3	3	8	0.0559	0.0516 (0.0049)	0.0651 (0.0053)
512	3	3	16	0.0500	0.0468 (0.0040)	0.0484 (0.0029)
512	3	3	32	0.0396	0.0475 (0.0047)	0.0408 (0.0041)

Continued on next page

Table 21 (continued)

N	Δ	L	m	Exact	GCN (SE)	SGC (SE)
512	3	3	64	0.0244	0.0232 (0.0041)	0.0277 (0.0053)
512	4	2	8	0.0367	0.0352 (0.0029)	0.0372 (0.0037)
512	4	2	16	0.0344	0.0358 (0.0031)	0.0363 (0.0031)
512	4	2	32	0.0301	0.0260 (0.0031)	0.0286 (0.0043)
512	4	2	64	0.0228	0.0223 (0.0030)	0.0273 (0.0019)
512	4	3	8	0.0889	0.1028 (0.0083)	0.0972 (0.0084)
512	4	3	16	0.0720	0.0790 (0.0093)	0.0706 (0.0095)
512	4	3	32	0.0467	0.0452 (0.0068)	0.0379 (0.0060)
512	4	3	64	0.0188	0.0109 (0.0047)	0.0312 (0.0060)

All covered-block and filler errors are zero in this table.

Average and worst-world performance. Zero-initialized SGC leaves an unobserved block's coordinate at zero: its feature has no support on the training set, so that coordinate has zero gradient. On a hit block, its scalar logistic gradient has the sign that learns the common bit. Consequently the rule is a Bayes-optimal default under the fair-bit prior, but on the all-one world every missed block is wrong. Its worst-world risk is $KQp_0(Q)/(N-m)$, twice the randomized minimax value. Thus zero-initialized SGC attains the fair-prior reference while its worst-world risk is twice the randomized minimax value.

I Architecture and graph-structure stress tests: protocol and results

The protocol was fixed before executing this development run. It is not independent preregistration. Every condition has $N = 256$, $K = 8$, $Q = 8$, and the same depth-three binary bridge trees. Baseline uses eight payload coordinates and three prediction layers. Shared maps block k to coordinate $k \bmod 2$. Deep uses eight prediction layers on the unchanged graph, keeping construction depth three. The *attached-path* condition joins one path of 24 known-zero filler vertices to each payload. Every filler lies in a nontrivial component, but the graph still has eight connected components. The *combined* condition uses shared coordinates, eight prediction layers, and these attached paths. No graph, feature assignment, or hyperparameter was selected using test performance.

For each $m \in \{16, 32, 64\}$, draw 20 independent fair bit vectors and uniform physical-node label sets. Pair the same bits, indices, and initialization seeds across all five conditions and three architectures. Each method receives only released features, adjacency and sampled labels. The partition reference receives the eight original blocks and the known-zero filler identities. Its conditional expected risk is half the missed active mass divided by $N - m$; it averages 0.04511 over these samples. In baseline, shared, and deep, connected-component parsing recovers this partition. The attached-path reference receives the original block partition and filler identities as oracle information. Shared-feature and attached-path conditions test modified representation settings; the original one-coordinate-per-block teacher guarantee applies to the baseline and deep conditions. The attack completes the same designated bridges before sampling, while the feature and target-realizability assumptions differ in these modified settings.

GCN and SGC use the implementation in Appendix H. The added full-neighbor mean GraphSAGE adaptation (Hamilton et al., 2017) uses

$$H^{j+1} = \mathrm{ReLU}\left(H^j W_{s,j} + M H^j W_{n,j} + b_j\right), \qquad M_{vu} = \frac{\mathbf{1}\{u \sim v\}}{\max\{1, \deg(v)\}},$$

with a linear final layer. Hidden width is 32, all biases start at zero, and self and neighbor weights are independently Gaussian with variance $1/\text{fan-in}$. There is no neighbor subsampling or embedding normalization; this is an explicitly specified adaptation, not the original benchmark recipe. All methods use 200 Adam updates and the previous learning rate and optimizer parameters. There is no dropout, validation, early stopping, or test-based selection. A predeclared sensitivity repeats deep GCN and SAGE for 1,000 steps. The 60 bit/split instances are evaluated under five graph conditions and three architectures, giving 900 primary fits. The two 1,000-step depth sensitivities add 120 fits on the same underlying instances.

Implementation checks. Central differences check every weight and bias in a small two-layer SAGE network, with maximum discrepancy 5.20×10^{-10}. The inherited GCN check has discrepancy 7.56×10^{-11}. Changing every held-out target while keeping sampled labels fixed leaves both predictions and training loss exactly unchanged for each of the three training APIs. Every reported test error equals the sum of covered-block, missed-block, and filler contributions. Standard errors in the main table cluster by seed after averaging its three label-count cells, retaining pairing across conditions and methods. The next table reports all primary cells; each entry averages 20 samples.

Table 22: Complete five-condition block results. Entries are mean fractions over 20 problems per cell. Covered, missed, and filler contributions all use denominator $N - m$ and sum to test error; training error uses denominator m. The same 60 bit/split instances are paired across conditions and methods.

Condition	Method	m	Test	Covered	Missed	Filler	Train
baseline	SGC	16	0.0733	0.0000	0.0733	0.0000	0.0000
baseline	GCN	16	0.0783	0.0000	0.0783	0.0000	0.0000
baseline	SAGE	16	0.1033	0.0285	0.0748	0.0000	0.0000
baseline	SGC	32	0.0357	0.0000	0.0357	0.0000	0.0000
baseline	GCN	32	0.0518	0.0000	0.0518	0.0000	0.0000
baseline	SAGE	32	0.0926	0.0451	0.0475	0.0000	0.0000
baseline	SGC	64	0.0104	0.0000	0.0104	0.0000	0.0000
baseline	GCN	64	0.0125	0.0000	0.0125	0.0000	0.0000
baseline	SAGE	64	0.0289	0.0167	0.0122	0.0000	0.0000
shared	SGC	16	0.0996	0.0146	0.0850	0.0000	0.0312
shared	GCN	16	0.0963	0.0146	0.0817	0.0000	0.0312
shared	SAGE	16	0.0975	0.0242	0.0733	0.0000	0.0125
shared	SGC	32	0.1011	0.0547	0.0464	0.0000	0.0797
shared	GCN	32	0.1011	0.0547	0.0464	0.0000	0.0797
shared	SAGE	32	0.1100	0.0627	0.0473	0.0000	0.0484
shared	SGC	64	0.0755	0.0630	0.0125	0.0000	0.0734
shared	GCN	64	0.0755	0.0630	0.0125	0.0000	0.0734
shared	SAGE	64	0.0977	0.0828	0.0148	0.0000	0.0570
deep	SGC	16	0.0733	0.0000	0.0733	0.0000	0.0000
deep	GCN	16	0.0783	0.0000	0.0783	0.0000	0.0000
deep	SAGE	16	0.1102	0.0287	0.0815	0.0000	0.0000
deep	SGC	32	0.0357	0.0000	0.0357	0.0000	0.0000
deep	GCN	32	0.0464	0.0000	0.0464	0.0000	0.0000
deep	SAGE	32	0.0938	0.0420	0.0518	0.0000	0.0000
deep	SGC	64	0.0104	0.0000	0.0104	0.0000	0.0000
deep	GCN	64	0.0125	0.0000	0.0125	0.0000	0.0000
deep	SAGE	64	0.0357	0.0253	0.0104	0.0000	0.0000
attached paths	SGC	16	0.0952	0.0029	0.0733	0.0190	0.0125
attached paths	GCN	16	0.1223	0.0029	0.0767	0.0427	0.0125
attached paths	SAGE	16	0.1121	0.0248	0.0733	0.0140	0.0000
attached paths	SGC	32	0.0804	0.0165	0.0357	0.0281	0.0312
attached paths	GCN	32	0.0964	0.0181	0.0429	0.0355	0.0328
attached paths	SAGE	32	0.1033	0.0404	0.0355	0.0275	0.0000
attached paths	SGC	64	0.0708	0.0279	0.0104	0.0326	0.0375
attached paths	GCN	64	0.0708	0.0279	0.0104	0.0326	0.0375
attached paths	SAGE	64	0.0755	0.0323	0.0099	0.0333	0.0000
combined	SGC	16	0.1863	0.0190	0.0833	0.0840	0.0813
combined	GCN	16	0.1863	0.0190	0.0833	0.0840	0.0813
combined	SAGE	16	0.1333	0.0267	0.0783	0.0283	0.0125
combined	SGC	32	0.1661	0.0547	0.0464	0.0650	0.1250
combined	GCN	32	0.1661	0.0547	0.0464	0.0650	0.1250
combined	SAGE	32	0.1400	0.0652	0.0480	0.0268	0.0484
combined	SGC	64	0.1190	0.0896	0.0104	0.0190	0.1180
combined	GCN	64	0.1190	0.0896	0.0104	0.0190	0.1180
combined	SAGE	64	0.1099	0.0867	0.0138	0.0094	0.0641

Representation audits. For each sampled world, we test full-label homogeneous linear realizability on $Z = \widehat{P}^{\ell} X$ using the feasibility problem $Z_v w \geq 1$ for $Y_v = 1$ and $Z_v w \leq 0$ for $Y_v = 0$. For

a finite node set this is equivalent to strict-positive threshold realizability: any positive minimum margin can be rescaled to one. This uses all labels only as a diagnostic; the feasible readout is never passed to training. The solver finds feasibility in all 60 baseline and deep samples, one of 60 shared samples, and none of the attached-path or combined samples. These numerical checks agree with structural obstructions. In the shared condition, isomorphic blocks with the same payload coordinate have identical corresponding representations, so opposite bits cannot be realized. In an attached-path condition, a positive block and a nearby zero filler both receive a positive multiple of the same payload coordinate, precluding their separation by a homogeneous linear score. The shared and attached-path tests therefore probe representation constraints as well as training behavior.

We also initialize colors with exact input features and vertex degree, then refine by the node's previous color and the multiset of neighboring colors for the specified number of layers. Nodes in the same resulting class have equal outputs for all three tested architectures. For GCN, this follows inductively because degrees, self-loop normalization and neighbor contributions agree. On each test set, summing the minority-label count within each class gives a label-aware lower bound on errors of any such predictor. This floor averages 0.07254 for shared and combined, and zero otherwise. Every fit respects it. The refinement floor supplies a sufficient obstruction when positive; a zero value leaves realizability unresolved. The original/deep family does have an explicit realization for all three architectures: preserve the eight payload coordinates through positive self/neighbor propagation, then select coordinates by the world bits.

Training fit and depth sensitivity. SAGE has zero training error in all baseline, deep, and attached-path samples but positive errors on covered test nodes. This is a generalization/inductive-bias gap despite available realizations in baseline and deep, not a failure to fit sampled labels. Baseline covered-test contribution is 0.03010 for SAGE, versus zero for GCN and SGC. Attached-path contributions (covered, filler) are $(0.01577, 0.02655)$, $(0.01629, 0.03692)$ and $(0.03250, 0.02492)$ for SGC, GCN and SAGE. Combined contributions are $(0.05441, 0.05597)$, $(0.05441, 0.05597)$ and $(0.05952, 0.02150)$, respectively. Eight-layer GCN has zero covered-block errors at both 200 and 1,000 steps; its test predictions are unchanged. Deep SAGE error changes from 0.07988 to 0.08034, with zero training error at both budgets. We therefore do not infer oversmoothing from depth or claim that more training removes the gap. These are synthetic stress tests with fixed architecture choices, not worst-world estimates or evidence about prevalence in natural graphs.

J Propagation diagnostics and learning-rate sensitivity

This development analysis reuses the original 60 baseline bit/split samples. Its protocol fixed the learning rates $\{0.005, 0.01, 0.02, 0.05\}$, three methods, 200 updates and a separate/tied by bias/no-bias factorial before execution. There are 720 learning-rate fits and 180 additional factorial fits; the original SAGE setting is shared between the two analyses, giving 900 distinct executed fits. No test-based selection or validation is used. The baseline 0.02 results for all three methods reproduce the previously reported loss components and training errors exactly. All learning-rate and factorial fits reuse the same 60 bit/split instances. The supplied files include raw rows, predictions, learned SAGE matrices and biases for all four factorial settings, branch outputs' norm summaries, and per-distance error counts. Standard errors retain the prior seed-cluster rule.

Architectural intervention. Untied initialization and optimization match the earlier implementation. For a tied layer we retain the original self-weight draw and use it for both branches; the neighbor draw is consumed but unused, preserving subsequent random-number alignment. A single matrix receives the sum of self and neighbor gradients. No-bias variants keep biases at zero and omit them from optimization entirely. Tying changes both parameterization and the effective update. Together with bias removal, it enforces Proposition 6.2's algebraic invariant; the experiment tests that combined restriction. Central differences for the tied matrix gradient give maximum discrepancy 7.72×10^{-11}. Replacing all held-out labels leaves the newly implemented tied/no-bias training outputs exactly unchanged.

In this sparse coordinate-feature family, each covered block contributes to one readout coordinate. From zero initialization, every nonzero logistic update to that coordinate has the common block label's sign; an unobserved coordinate remains zero. Positive propagation amplitudes therefore give the same classifications at the tested positive learning rates and propagation depths. This explains the

Table 23: Original learning-rate sensitivity on shared samples. Excess and SE are paired with the sample-conditional reference.

LR	Method	Test error	Covered contribution	Excess (SE)	Perfect fits
0.005	SGC	0.0398	0.0000	-0.0053 (0.0045)	60/60
0.005	GCN	0.0464	0.0000	0.0013 (0.0028)	60/60
0.005	SAGE	0.0667	0.0217	0.0216 (0.0033)	60/60
0.010	SGC	0.0398	0.0000	-0.0053 (0.0045)	60/60
0.010	GCN	0.0464	0.0000	0.0013 (0.0028)	60/60
0.010	SAGE	0.0710	0.0255	0.0259 (0.0037)	60/60
0.020	SGC	0.0398	0.0000	-0.0053 (0.0045)	60/60
0.020	GCN	0.0475	0.0000	0.0024 (0.0032)	60/60
0.020	SAGE	0.0750	0.0301	0.0298 (0.0039)	60/60
0.050	SGC	0.0398	0.0000	-0.0053 (0.0045)	60/60
0.050	GCN	0.0475	0.0000	0.0024 (0.0031)	60/60
0.050	SAGE	0.0886	0.0403	0.0435 (0.0050)	60/60

identical baseline and deep SGC entries in Appendix I; it is specific to the stated feature design and optimization rule.

Denominators and sampling. The baseline covered contribution 0.03010 is averaged with denominator $N - m$, as in the theorem's loss. It is not a 3.010% error rate conditional on being covered. Across the baseline SAGE samples, 403 of 2,012 covered test predictions are wrong, a pooled conditional error rate of 20.03%. The pooled count weights samples by their covered population, whereas the reported total-normalized loss averages the specified per-sample risk. Both conventions are identified explicitly. All 60 samples are included; no perfect-fit filtering is applied to the factorial.

Weight norms and functional contributions. Frobenius norms are measured separately at each layer. Branch-output norms are $\|HW_s\|_F$ and $\|MHW_n\|_F$ restricted to the 64 active vertices; bias norms are reported separately. We average each fitted model's ratio, not a ratio of pooled norms. Mean weight ratios (self/neighbor) are $(0.7944, 0.7901, 0.8567)$, while activation ratios are $(2.3832, 1.2558, 0.9120)$. These quantities depend on parameterization and feature magnitude, and are not causal attribution scores. Neither set of ratios supports replacing branch interventions with a norm-only story.

Table 24: Supporting numerical results under the protocol stated in this subsection. Rates use the denominators specified in the text.

Post-fit intervention	Test error	Covered contribution	Training error
None	0.07496	0.03010	0.00000
Zero self weights	0.09315	0.04686	0.06016
Zero neighbor weights	0.11492	0.07509	0.11615
Zero biases	0.07673	0.03131	0.00000

Zeroing biases after fitting is different from retraining without biases: the former leaves the original gap and training fit largely intact, while the latter reduces the gap but fails to fit one sample perfectly. Removing either branch changes intermediate distributions; these interventions establish sensitivity, not a unique mechanistic decomposition.

Table 25: Supporting numerical results under the protocol stated in this subsection. Rates use the denominators specified in the text.

Distance from payload	Covered test nodes	Errors	Conditional error
0	262	75	0.2863
1	239	38	0.1590
2	509	138	0.2711
3	1002	152	0.1517

Errors are not confined to the greatest distance, so these measurements also rule out the narrow explanation that only the final propagation radius fails. The factorial ablation and sufficient invariant offer a more precise account: coverage specifies what the observations identify, whereas architecture and fitting determine whether predictions share a sign across the identified block. This remains a controlled synthetic result for the stated implementations, not a general failure theorem for all GraphSAGE variants or natural graphs.

J.1 Additional population-size baseline

We retain the original 64 active nodes, eight feature coordinates, depth three, width 32, 200 Adam steps and learning rate 0.02, and increase isolated known-zero fillers from 192 to 448. Label counts $32, 64, 128$ match the original fractions at $N = 256$. Twenty newly drawn samples per label count give 60 samples shared across the three methods. No method or sample is selected after fitting. All methods fit all 60 training sets perfectly. These cross-size samples are not paired; SEs below cluster by seed, averaging label counts within each cluster. The reference is the split-specific fair-bit prediction error on unseen blocks. The pooled conditional covered error is $407/1935 = 21.03\%$ for

Table 26: Supporting numerical results under the protocol stated in this subsection. Rates use the denominators specified in the text.

Method	Test error	Covered contribution	Excess	SE
SGC	0.02298	0	-0.00133	0.00199
GCN	0.02439	0	0.00009	0.00150
SAGE	0.03951	0.01559	0.01521	0.00262

SAGE and zero for GCN and SGC. Mean covered contributions use $N - m$ per sample; conditional rates pool only covered test nodes, so they have different denominators. Increasing the population through fillers reduces active-node mass; this does not test growth in active component size, feature dimension or graph complexity. The sufficient transfer proposition has no population-size restriction but retains its input and propagation hypotheses. Earlier calibration already included $N = 512$ for SGC and GCN; the additional comparison supplies SAGE's covered-error measurement. All fitted weights, logits, splits and per-sample metrics are supplied, with independent forward-pass and aggregation checks.

Table 27: Supporting numerical results under the protocol stated in this subsection. Rates use the denominators specified in the text.

Labels	Method	Test error	Covered contribution	Perfect fits
32	SGC	0.04333	0.00000	20/20
32	GCN	0.04417	0.00000	20/20
32	SAGE	0.05448	0.01156	20/20
64	SGC	0.02143	0.00000	20/20
64	GCN	0.02589	0.00000	20/20
64	SAGE	0.04531	0.02009	20/20
128	SGC	0.00417	0.00000	20/20
128	GCN	0.00313	0.00000	20/20
128	SAGE	0.01875	0.01510	20/20

J.2 Original factorial transfer intervention

We test the two restrictions factorially, keeping the original splits, seeds, width, depth, optimizer and learning rate. A tied layer computes $(H + MH)W + b$ and optimizes a single matrix using the summed gradient. The original separate, biased architecture is the first row; all 60 samples are retained in each row. Removing biases or tying weights alone reduces error but does not eliminate it. Their combination enforces Proposition 6.2 and has zero covered errors with perfect training fit in all 60 samples. The separate bias-free model fits 59 of 60 samples. The combined restriction's zero covered error agrees with the sufficient invariant.

Table 28: Original SAGE factorial on 60 shared samples. Covered contribution uses the full test denominator; excess is paired with the split reference. Perfect fit means zero training classification error, not zero BCE.

Weights / biases	Test error	Covered contribution	Excess (SE)	Perfect fits
Separate / on	0.0750	0.0301	0.0298 (0.0039)	60/60
Separate / off	0.0632	0.0185	0.0181 (0.0033)	59/60
Tied / on	0.0494	0.0090	0.0043 (0.0034)	60/60
Tied / off	0.0467	0.0000	0.0016 (0.0036)	60/60

Approximate transfer. If $f_v = a_v t_k + e_v$, $a_v > 0$, $|e_v| \le \epsilon_v$, and an observed label has signed score g_s, then $y_k t_k \ge (g_s - \epsilon_s)/a_s$. Consequently $a_v \max_s(g_s - \epsilon_s)/a_s > \epsilon_v$ certifies that vertex's label (Proposition V.3). This gives a boundary in terms of margins and factorization error, not a guarantee that optimization achieves those bounds. An explicit two-node SAGE counterexample in Appendix V separates existence of bad fitted solutions from their frequency under training.

The original learning-rate and second-population studies remain in Appendix J. The latter changes only filler count: conditional covered error is 20.03% at $N = 256$ and 21.03% at $N = 512$; it is not active-geometry scaling. Our new matched controls vary active geometry and features explicitly (Appendix W).

K Budget-constrained relational ambiguity

We now study a separate family in which targets share latent variables across many overlapping relations. This is not an extension of the preceding depth law: the teacher is one propagation followed by two ReLUs, and degree and feature dimension grow. The actual edit mask is hidden.

Physical graph and permitted attacks. Let $H = K_r$ with integer $r \ge 2$. There are r anchors a_u, r pendants p_u, and one query q_{uv} per pair, so $N = 2r + \binom{r}{2}$. Each query is joined to its two anchors. In clean world $\theta \in \{0,1\}^r$, edge $a_u p_u$ is present exactly when $\theta_u = 1$. Query targets are $Y_{q_{uv}} = \mathbf{1}\{\theta_u = \theta_v\}$; auxiliary targets are zero. Features are $X_{a_u} = e_u$ and zero elsewhere. The attacker may add any subset of absent designated pendant edges, at most B, before the uniform sample of m physical-node labels. Writing the released pendant indicator as z, legality is $\theta \le z$ and $|z - \theta| \le B$. No other edits are permitted. The learner knows the construction and B, but neither the clean graph nor the actual addition mask.

Clean adjacency reveals θ, giving zero risk without labels. The attack sees no sampled indices or label observations, but clean adjacency contains target information. Local addition incidence is one; released maximum degree is r. For $s_u = (-1)^{\theta_u}$, the same world-specific score on every legal release is

$$f_\theta(v) = \left|(P_{1/2}(A_z)Xs/\sqrt{r})_v\right| - \frac{11}{15\sqrt{r}}.$$

It realizes all targets with margin at least $1/(15\sqrt{r})$ using two ReLUs. At the scale $\sqrt{r}$, query magnitudes are 0 or $4/5$, attached pendant magnitudes are $2/3$, and anchor magnitudes are at most $1/3$. Thus dimension is r and margin shrinks; these are not constant-margin asymptotics.

Proposition K.1 (Worst-release reduction)**.** *The minimax risk over all permitted attacks equals the minimax risk of predicting $Y'_{uv} = 1 \oplus T_u \oplus T_v$ for an unknown mask $|T| \le B$, given the same physical-node sampling protocol and known-zero auxiliary labels.*

Proof. For any release, $T = z \oplus \theta$ is supported on $\mathrm{supp}(z)$. XOR observed and predicted query labels with $z_u \oplus z_v$. This preserves loss and gives the displayed target. Every release yields a subset of the full mask family, so its minimax decoder upper-bounds every allowed release. Conversely, one legal attack fills every missing pendant edge when there are at most B missing edges and does nothing otherwise. Worlds $\theta = \mathbf{1} \oplus T$, $|T| \le B$, then share the all-present release. This fixed, pre-sample attack realizes the full subexperiment. The supremum over worlds may be restricted to these $\theta = \mathbf{1} \oplus T$ for a lower bound; worlds outside this set need not share the release. The rule

depends on observed clean adjacency, not on sampled labels, and is one fixed admissible attack on the entire clean family. This proves the reverse inequality. □

Why budget changes inference. Relative bits within each connected component of sampled queries follow from sign propagation, a classical property of balanced signed networks (Cesa-Bianchi et al., 2012). Between components, the budget constrains their possible orientations. For example, at $r = 4, B = 1$, two disjoint observed positive pairs force $T = \varnothing$: flipping either entire pair requires two edits. A path-only rule would still randomize on the four cross-pair queries. Its conditional error is $1/6$, whereas budget-aware prediction is exact. This is a feasible conditional example, not an unconditional error floor.

Theorem K.2 (Sparse-budget prediction)**.** *In this relational family, let $r \geq 16$, $1 \leq B \leq \lfloor r/16 \rfloor$ and $0 \leq m < N$. Define $n_0 = N - r + 1$, $\lambda_m = mr/n_0$, and*

$$A_m = \frac{\binom{n_0}{m}}{\binom{N}{m}}, \qquad \ell_m = \frac{B(r-B)}{2(N-m)} A_m.$$

There is a deterministic decoder with worst risk over all permitted attacks satisfying

$$\boxed{\ell_m \leq R^*_{\mathrm{rel}}(B, m) \leq R_{\mathrm{dec}}(B, m) \leq U_m \leq 2e^3 \ell_m,}$$

where

$$U_m = \begin{cases} B(r-B)/N, & \lambda_m \leq 3, \\ 16rBA_m/(N-m), & \lambda_m > 3. \end{cases}$$

The decoder uses $O(r + m + \binom{r}{2})$ graph operations. For $B = 0$, risk is zero.

Decoder and proof idea. Apply Proposition K.1. For $\lambda_m \leq 3$, predict $T = \varnothing$. Otherwise, propagate relative bits in each sampled component and choose its orientation with fewer mask members; ties are broken deterministically by the smallest-index root. Undo the known XOR transformation. The rule does not inspect true edits. Its estimate may ignore the extra support restriction supplied by z.

For the lower bound, choose a fixed support of size B and independent fair mask bits on it, before sampling. Each support vertex is isolated from sampled queries with probability A_m; then its $r - B$ relations to outside vertices remain fair. These are distinct unsampled targets, giving ℓ_m. For the upper bound, a wrongly oriented component has at most $2B$ vertices. If p_{small} is the probability a fixed vertex belongs to such a component, expected mistakes are at most $2rBp_{\mathrm{small}}$. In the high-label regime a fixed-size sampling and spanning-tree count gives $p_{\mathrm{small}} \leq 8A_m$. The low-label rule has risk $B(r-B)/N$. Appendix N proves every bound, including $A_m = 0$. The factor $2e^3 \approx 40.17$ bounds this explicit rule relative to the lower bound; it is not a claim of near-exact practical optimality.

Corollary K.3 (Fine-accuracy label threshold)**.** *Uniformly for $1 \leq B \leq r/16$ and $m = o(r^{3/2})$,*

$$R^*_{\mathrm{rel}}(B, m) = \Theta\big((B/r)e^{-2m/r}\big).$$

If $h_r = \log(B/(r\varepsilon_r)) \to \infty$ and $h_r = o(\sqrt{r})$, the first label count achieving risk at most ε_r satisfies

$$m^*_{\mathrm{rel}}(\varepsilon_r) = \frac{r}{2} \log \frac{B}{r\varepsilon_r} + O(r).$$

For fixed B, zero-label risk decreases with r; the threshold concerns accuracy ε_r small relative to B/r. It is a relational-family threshold, with the all-budget extension in Appendix O using $\bar{B} = \min\{B, \lfloor r/2 \rfloor\}$. Appendix L gives the exact finite minimax game and saturation result; Appendix M gives a polynomial posterior rule with factor $B + 1$ below saturation.

K.1 Detailed comparison with relational inference

Relational inference and closest comparisons. Cesa-Bianchi et al. (2012), Section 2, identify balanced two-cluster labels with products of signs along paths. Our switched labels are exactly such pairwise relations, restricted by a mask cardinality budget. Their Section 3.2, Theorem 6 bounds approximate correlation-clustering risk under uniform edge permutations with high probability; our

expected isolation rate uses a sparse mask and known-zero auxiliary nodes. Their active protocol chooses queries; ours samples physical nodes uniformly. Signed prediction is not new (Cesa-Bianchi et al., 2012).

Appendix M specializes established cardinality-counting methods to the posterior over component orientations (Tarlow et al., 2012). One-sided additions also resemble asymmetric binary corruption, as studied in Polyanskii & Zhang (2023). Here the goal is prediction of sampled/unobserved relations rather than decoding a designed codeword.

The threat model in Section IV-A of the study by Zhou et al. (2025) flips existing training-edge signs while preserving topology. Our legal physical edits add unsigned pendant edges and leave subsequently sampled target labels uncorrupted. Thus that protocol is different, although both concern adversarial relational learning.

Cuts, recovery, and the regular-graph extension. Small-cut ambiguity and expansion are established in Abbe et al. (2014), Theorem 4.3; Chen et al. (2016), Section 4.1 and Theorems 4–7, relate exact recovery to cut sizes and counts. Foster et al. (2018), Theorem 4 and Example 2, give boundary lower bounds and regular-graph counterexamples to degree-only rates. Their vertex Hamming loss and noisy measurements differ from our unobserved-query loss, hard mask budget, and noiseless fixed-size sampling. Graph cuts and hypercube component behavior are not new (McDiarmid et al., 2021); Appendix P compares geometries under our poisoning and observation protocol.

K.2 What the poisoning budget contributes

The reduction is a risk-preserving equivalence to sparse parity inference with the released support and mask constraint. This permits exact analysis of the information erased by the legal graph attack.

Proposition K.4 (Cost of enlarging the mask class)**.** *For a known dependency graph H, let R_B^* denote the switched experiment with $|T| \le B$ and let R_{all}^* allow every binary mask. If F is the random graph of sampled query edges, then*

$$R_{\text{all}}^* = \frac{1}{2(N-m)} \sum_{uv \in E(H)} \Pr\{u \not\leftrightarrow_F v\}.$$

where $u \not\leftrightarrow_F v$ means that u and v lie in different connected components of F, and the probability is over the sampled query edges. A path rule that predicts known within-component relations and guesses fairly between components has this risk for every fixed mask, including masks of size at most B. For the hypercube parameters of Proposition P.2,

$$\frac{R_{\text{all}}^*}{R_B^*} \ge \frac{r}{32B}.$$

The proposition compares nested admissible mask classes and the path rule's risk on the sparse class. The unrestricted formula follows from classical component parity propagation, while the rate comparison uses the sparse bound. The minimum-cardinality decoder does not require the numerical value of B; the result quantifies the constraint on possible worlds rather than the value of supplying B as an additional input.

The budget also gives an exact observable identification criterion. In the all-present release, let component i have two relative-bit orientations of sizes $a_i \le b_i$. Set $c = \sum_i a_i$ and $\tau = B - c$. A feasible transcript has $\tau \ge 0$. An unobserved query between distinct components i, j has a unique label over all feasible masks if and only if

$$b_i - a_i > \tau \quad \text{and} \quad b_j - a_j > \tau.$$

Within-component relations are identified. Appendix Q extends the component-orientation criterion to partial releases. Using established cardinality-feasibility reasoning (Tarlow et al., 2012), the analysis characterizes which query labels remain ambiguous under the legal edit budget. The contribution of this extension is the teacher-preserving poisoning realization and its quantitative risk laws.

L Partial attacks: realization and finite minimax game

The construction extends to every connected simple dependency graph $H = ([r], E)$ with $r \ge 2$, with $N = 2r + |E|$. The learner observes z, features, and a uniform m-subset of physical labels. Its

minimax value is

$$\inf_{\delta} \sup_{\theta\in\{0,1\}^r} \sup_{z\geq\theta,\ |z-\theta|\leq B} \mathbb{E}\left[\frac{1}{N-m}\sum_{i\notin S}\mathbf{1}\{\delta_i \neq Y_i\}\right].$$

A feasible action at any world extends to an attack map by identity at all other worlds. Conversely every legal map selects a feasible action. Randomized attacks average deterministic-action risks and cannot increase this supremum. All choices precede S.

L.1 Realization under every legal release

Set $s_u = (-1)^{\theta_u}$ and $w = s/\sqrt{r}$, so $\|w\|_2 = 1$. Anchor feature norms are one; all other feature norms are zero. At a query, $\sqrt{r}(P_{1/2}Xw) = (s_u + s_v)/(5/2)$, whose magnitude is $4/5$ for equal bits and zero otherwise. At anchor u the value is $s_u/(2\deg(a_u)+1)$, with magnitude at most $1/3$ because connectedness and $r \geq 2$ give every anchor at least one incident query. Attached pendants have value $2s_u/3$; isolated pendants have zero. Subtracting $11/15$ after absolute value gives positive score at least $1/15$ for positive queries and negative score at most $-1/15$ for every auxiliary or negative query. Dividing back by $\sqrt{r}$ proves the margin. Absolute value is implemented as $\mathrm{ReLU}(x) + \mathrm{ReLU}(-x)$ with hidden weights $w, -w$ and output weights $(1, 1)$. Weights are preserved within each world on clean and released graphs, but differ between worlds and are not observed by the learner. Maximum degree is $\max\{\Delta(H) + 1, 2\}$ and local addition incidence is one. If the actual addition mask were supplied, $\theta = z \oplus T$ would be known and risk would vanish. Thus hidden edits are essential.

L.2 Reduction and exact finite game

Proposition K.1 applies to this general H with the same proof. Write $\mathcal{T}_B = \{T \subseteq [r] : |T| \leq B\}$ and $y_T(uv) = 1 \oplus T_u \oplus T_v$. For observed query subset $F \subseteq E$, label pattern t, and unobserved query i, define

$$W_a(F, t, i) = \{T \in \mathcal{T}_B : y_T|_F = t,\ y_T(i) = a\}, \quad a = 0, 1.$$

Set $w(F) = \binom{2r}{m-|F|}$ and $D = (N - m)\binom{N}{m}$, with invalid binomial coefficients interpreted as zero. The exact primal is

$$\begin{aligned}
&\min_{p,v} v, \qquad 0 \leq p(F, t, i) \leq 1,\\
&\text{subject to } \frac{1}{D}\sum_{F\subseteq E} w(F)\sum_{i\notin F}\big[y_T(i)(1 - p(F, y_T|_F, i))\\
&\qquad\qquad + (1 - y_T(i))p(F, y_T|_F, i)\big] \leq v \quad \text{for every } T \in \mathcal{T}_B.
\end{aligned}$$

Every randomized rule has these prediction marginals, and independent coins can implement any such marginals. Known auxiliary labels can be predicted without error. Their observed identities are ancillary and can be averaged out; $w(F)$ counts all compatible auxiliary samples. Under a prior π, minimizing each prediction loss gives the smaller of its two joint masses. Therefore finite minimax duality gives the identical value

$$\max_{\pi\in\Delta(\mathcal{T}_B)} \frac{1}{D}\sum_{F} w(F)\sum_{t}\sum_{i\notin F}\min\left\{\sum_{T\in W_0(F,t,i)}\pi_T, \sum_{T\in W_1(F,t,i)}\pi_T\right\}.$$

These are unnormalized joint masses; no posterior denominator is missing. One prior is chosen before the sample. Optimizing priors separately after seeing local sample counts would change the experiment. Cells with identical (W_0, W_1) may be aggregated by summing weights: weighted averaging of their prediction probabilities preserves every world risk. A Bayes rule for an optimal prior need not be minimax under arbitrary tie or off-support choices; the supplied primal certificates explicitly check all world risks.

For $H = K_r$, permutation averaging of a prior preserves its Bayes objective up to concavity, and hence cannot reduce it. An optimal prior may therefore be taken uniform within cardinalities, with arbitrary masses on $0, \ldots, B$.

L.3 EXACT ENDPOINTS

Put $e = \binom{r}{2}$, $t_B = \min\{B, \lfloor r/2 \rfloor\}$ and $d_B = t_B(r - t_B)$. Without labels,

$$R^*_{\mathrm{rel}}(B, 0) = \frac{\min\{d_B, e/2\}}{N}.$$

Predicting from z makes at most d_B query errors; fair predictions make $e/2$ expected errors. For a lower bound, a uniform t_B-mask crosses each query with probability $q = d_B/e$. If $q \le 1/2$, its Bayes loss is d_B. Otherwise mix this prior with the empty mask, assigning the former mass $1/(2q)$: every query is then fair and Bayes loss is $e/2$. Both priors are feasible under the fixed completion map. This includes zero budget.

For $B \ge \lfloor r/2 \rfloor$, every target vector is represented because T and its complement have the same pairwise labels. For the full family a uniform latent-bit prior leaves independent fair orientations on observed components; within-component targets are known. A decoder propagates signs within components and guesses fairly between them, attaining the same loss for every target. For K_r, writing the observed components as C_j, this yields

$$R^*_{\mathrm{rel}}(B, m) = \frac{r^2 - \mathbb{E}\sum_j |C_j|^2}{4(N - m)}.$$

The threshold is sufficient, not always necessary. The physical sample, including auxiliaries, determines the distribution of the observed query graph.

M POLYNOMIAL POSTERIOR RULE

For $B < \lfloor r/2 \rfloor$, take prior $\pi(T) = [(B+1)\binom{r}{|T|}]^{-1}$ on $|T| \le B$. Posterior-majority classification with fair ties is permutation equivariant, so its all-present-release risk L_t depends only on $t = |T|$. Its Bayes risk is $(B+1)^{-1}\sum_{t=0}^{B} L_t \le R^*_{\mathrm{rel}}$. Nonnegativity gives $\max_t L_t \le \sum_t L_t \le (B+1)R^*_{\mathrm{rel}}$. Switching extends the bound to every release. This is classification, not a posterior sample. At saturation, use the full-family path rule above, which is exact minimax.

M.1 EXACT POSTERIOR COMPUTATION

Propagate relative bits within observed components. The two orientations of component j have costs a_j, b_j in mask cardinality. Let

$$P(x) = \prod_j (x^{a_j} + x^{b_j}), \qquad Z = \sum_{t=0}^{B} \frac{[x^t]P(x)}{\binom{r}{t}}.$$

Coefficients count distinct masks, including multiplicity two when $a_j = b_j$. For a query across components c, d, remove their two factors and denote the remaining polynomial by $P_{-c,-d}$. Assign orientation costs $c_j(0) = a_j, c_j(1) = b_j$. The unnormalized mass of orientations (s, t) is

$$M_{s,t} = \sum_{h + c_c(s) + c_d(t) \le B} \frac{[x^h]P_{-c,-d}(x)}{\binom{r}{h + c_c(s) + c_d(t)}}.$$

The four masses sum to Z. Sum those yielding query equality, divide by Z, threshold at $1/2$, and undo switching. Within-component labels are forced.

Exact division by $x^a + x^b$ is polynomial: shift by $\min(a, b)$, then for $\delta = |a - b| > 0$ recursively subtract the quotient coefficient δ places earlier. If $\delta = 0$, divide coefficients by two. Building the product takes $O(r^2)$ arithmetic operations, and two divisions per component pair take $O(r)$, for total $O(r^3)$ including outputs. Coefficients are at most 2^r; a common denominator dividing $\prod_{t=0}^{B}\binom{r}{t}$ has $O(r^2)$ bits. Thus bit complexity is polynomial as well. Cardinality inference itself is established machinery (Tarlow et al., 2012).

M.2 One-edit diagnostic

For $B = 1, r \geq 4$, define

$$A = \frac{\binom{N-r+1}{m}}{\binom{N}{m}}, \quad D_2 = (r-1)\frac{\binom{N-2r+3}{m-1}}{\binom{N}{m}}, \quad U_1 = \frac{(r-1)A + (r-2)D_2}{N-m}.$$

The posterior rule has worst risk U_1 and $U_1/2 \leq R^*_{\text{rel}} \leq U_1$. To see this, the empty mask is always predicted correctly. With a singleton true mask at v, isolation leaves the posterior predicting all positive relations, giving $r-1$ errors. An observed negative component of size two leaves its two orientations tied, giving expected $r-2$ mistakes on exterior relations. A larger negative component identifies the singleton. The two events have probabilities A, D_2: the size-two event requires its one internal edge and excludes $2r-4$ boundary edges, leaving $N-2r+3$ available positions for the other $m-1$ labels. For an all-positive transcript with I isolated vertices, the empty mask's posterior weight is proportional to r, versus one for each singleton candidate; query crossing probability is at most $2/(r+I) < 1/2$ whenever both candidates can contribute. This verifies the stated decisions. The prior half on the empty mask and half on uniform singletons gives Bayes risk $U_1/2$, proving the lower bound.

N Complete proof of the sparse-budget theorem

Throughout $r \geq 16$, $1 \leq B \leq \lfloor r/16 \rfloor$ and $0 \leq m < N$. Use $n_0 = N - r + 1$, $A_m = \binom{n_0}{m}/\binom{N}{m}$, and $\lambda_m = mr/n_0$.

N.1 Lower bound

Fix $K \subseteq [r]$ of size B. Independently include each element of K in T with probability $1/2$ and exclude every other vertex. This single prior is feasible under the all-present-release attack. For $v \in K$, the event that no incident query is sampled has probability A_m. On this event every observation is independent of T_v, which remains fair. Its $r - B$ query targets to vertices outside K are unsampled and have Bayes error $1/2$. Those queries are distinct for different $v \in K$. Summing and dividing by $N - m$ gives ℓ_m. No independence of isolation events is assumed. Allowing the Bayes learner to know K only strengthens the lower-bound argument.

N.2 Low-label branch

For a mask of cardinality $t \leq B$, the empty-mask rule has expected test error $t(r-t)/N$, maximized at $t = B$. Retain the exact denominator cancellation:

$$\frac{A_m}{(N-m)/N} = \frac{\binom{n_0}{m}}{\binom{N-1}{m}} \geq (1 - m/n_0)^{r-2} \geq (1 - 3/r)^{r-2} \geq e^{-3},$$

when $\lambda_m \leq 3$. The final inequality follows from $\log(1-x) \geq -x - x^2/[2(1-x)]$:

$$(r-2)\log(1-3/r) \geq -3 + \frac{3r-18}{2r(r-3)} \geq -3.$$

Therefore $U_m/\ell_m \leq 2e^3$.

N.3 High-label branch and component errors

Within each sampled component the propagated relative bits are correct up to a common orientation. If minimum-cardinality orientation chooses incorrectly on a component of size k, the true mask contains at least $k/2$ members there. Thus $k \leq 2B$, including wrong ties. If $\widehat{T}$ is the estimate,

$$d := |T \triangle \widehat{T}| \leq 2 \sum_{v \in T} \mathbf{1}\{|C(v)| \leq 2B\}.$$

By permutation symmetry of the physical sample, the component-size probability $p_{\text{small}} = \Pr(|C(v)| \leq 2B)$ is common to all v. Hence $\mathbb{E}d \leq 2Bp_{\text{small}}$. Query errors form the cut of

the symmetric difference, with size $d(r-d) \le rd$. Consistency ensures all such errors are unsampled, so

$$R_{\mathrm{dec}} \le \frac{2rB}{N-m} p_{\mathrm{small}}.$$

N.4 Fixed-size small-component bound

If $A_m = 0$, then $m > n_0$. Any component of size $k \le 2B \le r/8$ needs $k(r-k) \ge r-1$ boundary queries absent, but fewer than $r-1$ physical positions are unsampled. Thus $p_{\mathrm{small}} = 0$.

Assume $A_m > 0$ and $\lambda_m > 3$. The singleton contribution is A_m. For $k \ge 2$, fix a k-set containing a specified root and a spanning tree on it. Let $l = k-1$ and $q = k(r-k)$. First exclude any $r-1$ boundary queries, an event of probability A_m. The remaining population has size n_0. Inclusion of the l tree queries has probability $(m)_l/(n_0)_l \le (m/n_0)^l$. Conditional on these inclusions, exclusion of the other $q-(r-1) = l(r-k-1)$ boundary queries has probability at most

$$\exp\left[-\frac{m-l}{n_0-l} l(r-k-1)\right].$$

This follows from the without-replacement avoidance product and $1-x \le e^{-x}$; if insufficient positions remain, the probability is zero and the inequality still holds. Required tree edges and forbidden boundary edges are disjoint. In this regime $m > 3n_0/r \ge 3r/2$, $l < m/12$, and $r-k-1 \ge 13r/16$. The exponent magnitude is at least $(143/192)\lambda_m l$. Counting sets and Cayley trees as above gives the size-k contribution divided by A_m at most

$$\left[e\lambda_m e^{-143\lambda_m/192}\right]^l.$$

Here $\binom{r-1}{l} \le r^l/l!$ and $(l+1)^{l-1}/l! \le e^l$; the latter follows from $l! \ge (l/e)^l$ and $(1+1/l)^{l-1}/l \le e/l \le 1$ for $l \ge 3$, with $l = 1, 2$ direct. The base decreases for $\lambda_m \ge 3$ and is at most $3e^{-79/64} < 7/8$. Indeed the degree-five Taylor polynomial of $e^{79/64}$ already exceeds $24/7$, an exact rational inequality. Summing gives $p_{\mathrm{small}} \le 8A_m$ and hence

$$R_{\mathrm{dec}} \le \frac{16rBA_m}{N-m} = U_m, \qquad \frac{U_m}{\ell_m} \le \frac{32r}{r-B} \le \frac{512}{15} < 2e^3.$$

The last inequality also follows from the degree-five Taylor polynomial of e^3. This proves the revised factor. Along $r \to \infty$ with $\lambda_m \to 3$ from below, the low-branch ratio to ℓ_m tends to $2e^3$; this identifies a limit of the selected branch relative to this lower bound, not an intrinsic minimax approximation factor. Building components needs $O(r+m)$ graph operations and writing all query predictions needs $O(\binom{r}{2})$. The implementation's arbitrary-length integer bit access is not being counted as unit bit complexity.

N.5 Asymptotics and first-crossing threshold

The finite result gives $R^*_{\mathrm{rel}} = \Theta(BrA_m/(N-m))$ with absolute constants. For $m = o(r^{3/2})$, expand the avoidance product:

$$\begin{aligned}\log A_m &= -m\sum_{j=0}^{r-2}\frac{1}{N-j} + O\left(m^2\sum_{j=0}^{r-2}\frac{1}{(N-j)^2}\right)\\ &= -\frac{2m}{r} + O(m/r^2 + m^2/r^3) = -2m/r + o(1).\end{aligned}$$

The expansion is uniform along these sequences because $m/(N-r+2) = o(1)$. Also $N-m \sim r^2/2$, proving Corollary K.3's rate.

Proposition N.1 (Monotonicity under uniform transductive sampling)**.** *Fix a finite family of worlds, a released observation fixed before label sampling, and binary node targets. For an unrestricted randomized nodewise learner, the normalized minimax test error is nonincreasing in $m = 0, \dots, N-1$. This holds for every fixed budget class in the relational experiment.*

Proof. Fix any prior π on the same world family for all m. Draw a test index J uniformly and, independently of the world conditional on J, a uniform permutation of $[N] \setminus \{J\}$. Let S_m contain its first m entries. For each $j \notin s$, $|s| = m$,

$$\Pr\{J = j, S_m = s\} = \frac{1}{N\binom{N-1}{m}} = \frac{1}{\binom{N}{m}(N-m)}.$$

This is the original split followed by uniform test-index sampling. Include the released observation in every information field and expose the permutation prefix in order. Its within-prefix ordering is independent of the world conditional on the unordered split, labels, release, and J, so revealing it does not alter the posterior. These ordered information fields $\mathcal{F}_m$ are nested. The tower property gives $p_m = \mathbb{E}[p_{m+1} \mid \mathcal{F}_m]$, where $p_m = \Pr_\pi(Y_J = 1 \mid \mathcal{F}_m)$. Since $g(p) = \min\{p, 1-p\}$ is concave,

$$r_\pi(m+1) = \mathbb{E}g(p_{m+1}) \le \mathbb{E}g(p_m) = r_\pi(m).$$

These scalar Bayes risks equal optimal normalized vector-prediction risks: additivity permits the learner to implement every node's marginal Bayes action simultaneously, without a constraint that predictions form a single model world. The finite game has finitely many observations and actions; randomized minimax duality therefore gives $R^*(m) = \max_\pi r_\pi(m)$ on the same simplex. Maximizing the pointwise inequality proves $R^*(m+1) \le R^*(m)$. For varying releases, include them in the finite world's observation; alternatively use the worst-release reduction to the fixed all-present subexperiment. □

Explicit first crossing. Let $b = \bar{B}$, $h_r = \log(b/(r\varepsilon_r)) \to \infty$ with $h_r = o(\sqrt{r})$. Uniformly over the all-budget range, the finite bounds give

$$(1/2 + o(1))\frac{b}{r}e^{-2m/r} \le R^*(B, m) \le (113 + o(1))\frac{b}{r}e^{-2m/r}$$

for $m = o(r^{3/2})$, since $b(r-b)/(2(N-m))$ lies between $(1/2 + o(1))b/r$ and $(1 + o(1))b/r$. Set $c = \log 452$ and $m_- = \lfloor (r/2)(h_r - c) \rfloor$, $m_+ = \lceil (r/2)(h_r + c) \rceil$. Eventually both are admissible, nonnegative and $o(r^{3/2})$. The lower bound at m_- is $(226 + o(1))\varepsilon_r > \varepsilon_r$; the upper bound at m_+ is $(1/4 + o(1))\varepsilon_r < \varepsilon_r$. Proposition N.1 excludes any earlier crossing and brackets the first successful label count between these two integers. Hence $m^*(\varepsilon_r) = (r/2)h_r + O(r)$, including rounding and normalization. The sparse case follows by taking $b = B$. This statement excludes fixed-error claims above the vanishing zero-label risk.

O Complete dependencies at every budget scale

Use the designated-pendant construction of Section K, with $H = K_r$ and $N = 2r + \binom{r}{2}$. The following theorem covers all edit-budget scales; its proof follows immediately.

Theorem O.1 (All-budget finite risk)**.** *In this designated-pendant experiment, let $r \ge 16$, $B \ge 0$ be integers, $0 \le m < N$, $\bar{B} = \min\{B, \lfloor r/2 \rfloor\}$, and $u = N - m$. For $\bar{B} \ge 1$, define*

$$n_0 = N - r + 1, \quad A_m = \frac{\binom{n_0}{m}}{\binom{N}{m}}, \quad \ell_m = \frac{\bar{B}(r-\bar{B})}{2u}A_m.$$

An explicit deterministic decoder satisfies

$$\ell_m \le R^*(B, m) \le R_{\mathrm{dec}}(B, m) \le 113\ell_m.$$

For $\bar{B} \le r/16$, the factor improves to $2e^3$. At $B = 0$ risk is zero; for $B \ge \lfloor r/2 \rfloor$ the randomized risk equals the unrestricted-mask value. Uniformly for $m = o(r^{3/2})$, the risk is $\Theta((\bar{B}/r)e^{-2m/r})$.

For the intermediate branch $\bar{B} > r/16$, use the empty switched mask if $\lambda_m = mr/n_0 \le 4$. Otherwise propagate observed parities, predict the identified within-component relations, and default to equality between components. Undo the known release-dependent XOR. The default predictions need not constitute one feasible mask; unrestricted nodewise decisions permit them.

O.1 Lower bound and low-label branch

Independent fair bits on a fixed support of size $\bar{B}$ give ℓ_m by the same isolation argument as Theorem K.2. A mask and its complement give identical pairwise targets, so every labeling has a representative of size at most $\lfloor r/2 \rfloor$, proving saturation. For $\lambda_m \le 4$, the empty-mask rule has worst risk $U = \bar{B}(r-\bar{B})/N$. Retaining the test-normalization cancellation gives

$$\frac{A_m}{u/N} \ge (1-4/r)^{r-2} \ge (3/4)^{14}, \qquad \frac{U}{\ell_m} \le 2(4/3)^{14} < 113.$$

The logarithm $(r-2)\log(1-4/r)$ is nondecreasing for $r \ge 16$: with $x = 4/r$, its derivative is $\log(1-x) + x(1-x/2)/(1-x) \ge 0$ by $\log(1-x) \ge -x - x^2/[2(1-x)]$.

O.2 A component bound up to half the latent population

We prove, for $\lambda_m \ge 4$,

$$p_{1/2} := \Pr\{|C(v)| \le r/2\} \le 3A_m. \tag{17}$$

If $A_m = 0$, fewer than $r-1$ physical positions are unobserved. Every nontrivial complete-graph cut has at least $r-1$ edges, so the sampled query graph is connected and $p_{1/2} = 0$. Otherwise, for a k-set containing the root, put $l = k-1$ and $d = l(r-k-1)$. Condition first on $r-1$ missing boundary edges. The remaining population has size n_0. A spanning-tree argument and, separately, a cut-only argument give the following two upper bounds on the size-k probability divided by A_m:

$$T_k(m) = \binom{r-1}{l} k^{k-2} (m/n_0)^l \exp\left[-\frac{m-l}{n_0-l} d\right],$$
$$C_k(m) = \binom{r-1}{l} \exp(-md/n_0).$$

Required tree edges and forbidden cut edges are disjoint. The usual without-replacement inclusion and exclusion products give the displayed bounds; if a required event is impossible its probability is zero. For $k \le r/2$ both expressions are nonincreasing in $m \ge 4n_0/r$. For T_k, the logarithmic derivative is $l/m - d/(n_0-l) \le 0$, using $4(r-k-1) \ge r$. Thus evaluate both at the real comparison point $m_0 = 4n_0/r$.

Split at $k = r/4$. For $2 \le k \le r/4$, $m \ge 2r$ and $l \le r/4 - 1$ give

$$(1-l/m)(1-k/r-1/r) \ge \left(\frac{7}{8} + \frac{1}{2r}\right)\left(\frac{3}{4} - \frac{1}{r}\right) \ge \frac{319}{512} =: c.$$

Using $n_0 - l \le n_0$ and $\binom{r-1}{l} \le r^l/l!$, we obtain

$$T_k(m) \le \frac{k^{k-2}}{(k-1)!} (\lambda_m e^{-c\lambda_m})^{k-1} \le \frac{k^{k-2}}{(k-1)!} x^{k-1}, \qquad x = 4e^{-319/128}.$$

The last inequality uses that $\lambda e^{-c\lambda}$ decreases for $\lambda \ge 4$. Let $T(x) = \sum_{k\ge1} k^{k-1}x^k/k!$ be the exponential generating function of labeled rooted trees. Cayley's count follows from Prüfer sequences, with k choices of root. Decomposing a rooted tree into its root and an unordered set of rooted subtrees gives $T = xe^T$. To justify convergence at this x, start $T_0 = 0$ and iterate $T_{h+1} = xe^{T_h}$. These height-truncated tree series increase to T and remain at most $2/3$, because $xe^{2/3} = 4e^{-701/384} < 2/3$. Monotone convergence gives $T \le 2/3$. Thus the singleton plus all smaller-size contributions, divided by A_m, are at most $T(x)/x = e^{T(x)} \le e^{2/3} < 2$.

For $r/4 < k \le r/2$, we have $k \ge 5$ and $2k \le r < 4k$. Retain the falling-factorial correction:

$$\binom{r-1}{l} = \frac{r^l}{l!} \prod_{j=1}^{l} (1-j/r) \le \frac{r^l}{l!} e^{-kl/(2r)}.$$

At m_0, this gives

$$C_k(m_0) \le \frac{r^l}{l!} \exp\left[-4l + \frac{(7k/2+4)l}{r}\right].$$

For real $r \in [2k, 4k]$, the logarithmic derivative of the right side is $l(r - 7k/2 - 4)/r^2$, so its maximum is at an endpoint. The endpoint $2k$ is larger, since the log ratio is $l(7/8 + 1/k - \log 2) > 0$. Consequently

$$C_k(m_0) \le b_k := \frac{(2k)^{k-1}}{(k-1)!} e^{-9(k-1)/4+2(k-1)/k}.$$

For $k \ge 5$,

$$\frac{b_{k+1}}{b_k} = 2(1 + 1/k)^k e^{-9/4+2/[k(k+1)]} \le 2e^{-71/60} < \frac{2}{3}, \qquad b_5 = \frac{1250}{3} e^{-37/5} < \frac{1}{3}.$$

Therefore the entire larger-size contribution is less than $\sum_{k\ge5} b_k < 1$. Adding both parts proves $p_{1/2} < 3A_m$ when $A_m > 0$, completing Equation (17) for every $r \ge 16$ without a finite population-size certificate.

For completeness, the numerical constants above follow from ordinary rational Taylor inequalities. With $P_n(t) = \sum_{j=0}^{n} t^j/j!$,

$$P_{12}(701/384) > 6, \quad P_{10}(71/60) > 3, \quad P_{24}(37/5) > 1250, \quad P_6(7/10) > 2.$$

For $z = 2/3$, $P_4(z) + (z^5/5!)/(1 - z/6) < 2$ bounds the remaining Taylor terms geometrically and proves $e^{2/3} < 2$. These fixed rational inequalities can be checked directly; no parameter-dependent computation is required by the proof.

O.3 High-label risk and consequences

There is at most one component larger than $r/2$. Every disconnected query therefore touches a vertex in a smaller component, so its expected count is at most $r^2 p_{1/2}$. The deterministic default makes at most one mistake per such query and is correct on all identified relations. Hence

$$R_{\text{all-budget}} \le \frac{3r^2 A_m}{u}, \qquad \frac{R_{\text{all-budget}}}{\ell_m} \le \frac{6r^2}{\bar{B}(r - \bar{B})} \le \frac{512}{5} < 113,$$

when $r/16 \le \bar{B} \le r/2$. Together with the low branch and the sparse result this proves Theorem O.1.

The avoidance expansion from Corollary K.3 is independent of $\bar{B}$, and $r\bar{B}/2 \le \bar{B}(r - \bar{B}) \le r\bar{B}$. Thus uniformly for $1 \le \bar{B} \le \lfloor r/2 \rfloor$ and $m = o(r^{3/2})$,

$$R^*(B, m) = \Theta\left((\bar{B}/r)e^{-2m/r}\right).$$

The same first-crossing proof gives $m^*(\varepsilon_r) = (r/2)\log(\bar{B}/(r\varepsilon_r)) + O(r)$ when the logarithm diverges and is $o(\sqrt{r})$. These are noiseless relational statements; the noisy-block theorem concerns a different observation channel.

P Regular dependency graphs: a supporting extension

This section extends the designated-pendant experiment of Section K to a known simple connected q-regular graph H on r latent vertices, with one query per edge of H. Thus $N = 2r + qr/2$. The allowed edits, targets, features, and sampling protocol are otherwise the same. Proposition K.1 applies verbatim: every release reduces to a subset of masks $|T| \le B$, and the same fixed pre-sample attack realizes all such masks at the all-present release. The learner need not know the attack or actual mask. We use parity $T_u \oplus T_v$ below; complementing query labels relative to Proposition K.1 preserves risk.

For completeness, the world-specific readout there still realizes all targets. At scale $\sqrt{r}$, propagated queries equal $2(s_u + s_v)/5$, anchors equal $s_u/[2(q + z_u) + 1]$, and pendants equal $2s_u/3$ when attached and zero otherwise. The threshold $11/15$ yields margin at least $1/(15\sqrt{r})$. Clean adjacency identifies the target with no labels. Feature dimension is r and maximum physical degree is $q + 1$. No common observable readout across worlds is claimed.

P.1 A SUFFICIENT CONDITION USING CONNECTED-SET BOUNDARIES

Put $s = N - m > 0$, $B \geq 1$, and

$$A_q = \frac{\binom{N-q}{m}}{\binom{N}{m}}, \qquad W_H(v, K) = \sum_{\substack{D \subseteq V(H),\ v \in D \\ H[D] \text{ connected},\ 1 \leq |D| \leq K}} \frac{\binom{N-|\delta_H(D)|}{m}}{\binom{N}{m}}.$$

Impossible binomial coefficients are zero. W_H is an upper-bound device, not an exact or universally matching risk invariant.

Proposition P.1. *A minimum-cardinality component decoder satisfies*

$$R_H^* \leq R_{\mathrm{mc}} \leq \frac{2qB}{s} \min\{1, \max_v W_H(v, 2B)\}.$$

If H has an independent set of size B, then

$$\frac{qB}{2s} A_q \leq R_H^* \leq \frac{qB}{N}.$$

In particular, $\max_v W_H(v, 2B) \leq C A_q$ supplies upper and lower bounds differing by factor at most $4C$ when $A_q > 0$.

Proof. Let F contain exactly the dependency edges whose query nodes were sampled. Within each connected component of F, propagate the observed relative bits and choose the orientation with fewer ones; break ties deterministically. The rule uses only permitted observations and costs $O(r + |E(H)|)$ operations. It may ignore the release's additional support constraint without invalidating an upper bound for an unrestricted learner.

If the estimate $\widehat{T}$ differs from T on a component C, it complements all its bits and minimality implies $|C| \leq 2|T \cap C| \leq 2B$. Therefore

$$|T \triangle \widehat{T}| \leq 2 \sum_{v \in T} \mathbf{1}\{|C_F(v)| \leq 2B\}.$$

At most $q|T \triangle \widehat{T}|$ query predictions differ, and sampled queries are always correct. If $C_F(v) = D$, then $H[D]$ is connected and no boundary query was sampled. Its probability is at most $\binom{N-|\delta_H(D)|}{m} / \binom{N}{m}$. Summing over D, bounding a probability also by one, taking expectation, and dividing by s proves the first bound.

For the lower bound, fix an independent set I of size B. Put independent fair mask bits on I and zero bits elsewhere, before sampling, with the common all-present release. If no incident query of $v \in I$ is sampled, its bit remains fair given every observation. Its q incident test targets each incur Bayes error $1/2$. These edge sets are disjoint because I is independent. Their isolation probability is A_q, proving the lower bound. This is one common prior, not a split-dependent adversary. A greedy selection guarantees such an I if $B \leq \lfloor r/(q+1) \rfloor$.

For the last upper bound, predicting $T = \varnothing$ makes errors on $\delta_H(T)$, whose size is at most qB. Uniform sampling leaves expected error fraction $|\delta_H(T)|/N$, giving qB/N. □

Fixed degree does not allow the same separation. For fixed $q \geq 2$, $m \leq (1-\eta)N$, and $N \geq 2(q-1)/\eta$, each factor $(s-i)/(N-i)$ for $1 \leq i < q$ is at least $\eta/2$. Hence the lower bound above is at least $\frac{1}{2}(\eta/2)^{q-1} qB/N$. Whenever the independent-set condition holds, all such graphs have risk $\Theta_{q,\eta}(qB/N)$. The degree-dependent constant is essential; the next result lets degree grow.

P.2 A SEPARATION AT HALF LABELING

Proposition P.2 (Supporting half-labeling comparison)**.** *For every integer $q \geq 14$, let*

$$r = 2^q, \quad B = \lceil (q+1)/2 \rceil, \quad N = r(q+4)/2, \quad m = N/2.$$

There exist two simple connected q-regular dependency graphs on r vertices, with the same designated-pendant poisoning protocol, such that

$$\frac{qB}{2Nr} \leq R_{\mathrm{good}}^* \leq \frac{8qB}{Nr}, \qquad R_{\mathrm{bad}}^* \geq \frac{N-2}{2N(N-1)}.$$

Consequently

$$\frac{R^*_{\text{bad}}}{R^*_{\text{good}}} \geq \frac{r(N-2)}{16qB(N-1)} \geq \frac{r}{32qB} = \Omega\left(\frac{r}{(\log r)^2}\right).$$

Proof. For the good graph use the hypercube Q_q. Its classical edge-isoperimetric bound is $|\delta(D)| \geq k(q - \log_2 k)$ for $|D| = k$. One proof bounds internal edges by $k \log_2 k/2$ inductively: split a coordinate into parts of sizes a, b; crossing edges form a matching, and

$$\frac{a \log_2 a + b \log_2 b}{2} + \min(a, b) \leq \frac{(a+b)\log_2(a+b)}{2}.$$

The inequality follows from binary entropy $h_2(a/(a+b)) \geq 2\min(a,b)/(a+b)$, by concavity and the chords through $0, 1/2, 1$; take $0 \log 0 = 0$.

In a maximum-degree-q graph, at most $(4q)^{k-1}$ connected k-sets contain a given root. Overcount by plane rooted spanning trees (at most 4^{k-1} shapes) and at most q choices for every tree edge. Repeated embeddings only increase this count. At half labeling the chance of missing b specified query nodes is $(N/2)_b/(N)_b \leq 2^{-b}$, or zero if $b > N/2$. With $K = 2B \leq q + 2$, these facts give

$$W_{Q_q}(v, K) \leq \sum_{k=1}^{K} \frac{(4q)^{k-1} k^k}{r^k} \leq \frac{2}{r}.$$

For the second inequality multiply by r. The first term is one; putting $a = 4qK/r$, the other terms are at most Ka^{k-1}. Since $r \geq 4qK(K+1)$, their geometric sum is at most $Ka/(1-a) \leq 1$. Indeed $2^{14} = 16384 \geq 4 \cdot 14 \cdot 16 \cdot 17$, and the ratio of successive polynomials $4q(q+2)(q+3)$ is $(q+1)(q+4)/(q(q+2)) < 2$ for $q \geq 3$.

Also

$$rA_q = \prod_{i=0}^{q-1} \left(1 - \frac{i}{N-i}\right) \geq 1 - \frac{q(q-1)}{2(N-q+1)} \geq \frac{1}{2}.$$

Here the product bound follows by induction for factors in $[0, 1]$, and $N - q + 1 \geq q(q-1)$ follows from the preceding exponential bound. The hypercube has an independent bipartition and $B \leq r/16$. Proposition P.1, with $s = N/2$, proves both good-graph bounds. Only small components are controlled; global connectivity is unnecessary.

For the bad graph, start with a clique on $l = q + 1$ vertices and a disjoint q-regular circulant on $M = r - q - 1$ vertices. For even q, use steps $\pm 1, \ldots, \pm q/2$ modulo M. For odd q, M is even; use steps $\pm 1, \ldots, \pm(q-1)/2$ and the antipodal matching. Since $M > q + 1$, the graph is simple. Delete edge $(0, 1)$ in each piece and cross-connect their endpoints. Degrees stay q. Both pieces remain connected: the clique minus an edge is connected and the circulant retains a spanning path along its nearest-neighbor cycle. The clique vertex set D now has exactly two boundary edges.

Partition D into masks T_0, T_1 of sizes $\lfloor (q+1)/2 \rfloor, \lceil (q+1)/2 \rceil$. Both meet the budget and their query targets differ exactly on $\delta(D)$. Under the fair prior on these two worlds and their common all-present release, the worlds are distinguished if either boundary query is sampled. Otherwise the expected number of test errors is one. The unconditional Bayes risk is

$$\frac{1}{s} \frac{\binom{N-2}{m}}{\binom{N}{m}} = \frac{s-1}{N(N-1)} = \frac{N-2}{2N(N-1)}.$$

This lower-bounds minimax risk. Dividing by the good-graph upper bound proves the ratio; $(N - 2)/(N-1) \geq 1/2$ gives its simpler bound. □

Interpretation. The construction holds dependency size and degree, physical population, budget bound, label count, features, and teacher architecture fixed while varying dependency geometry. Budget bounds are upper limits: the witness masks may have different cardinalities, so realized edit counts and clean edge counts need not coincide. The good risk is bounded within factor 16; the bad graph has the displayed lower bound and the generic upper bound qB/N. Both risks vanish as q grows, so the separation concerns relative risk and fine accuracy. Degree and feature dimension grow, and teacher margins decrease. The result specializes established cut-geometry and component principles (Abbe et al., 2014; Chen et al., 2016; Foster et al., 2018; McDiarmid et al., 2021) to this legal-poisoning experiment with known dependency graphs.

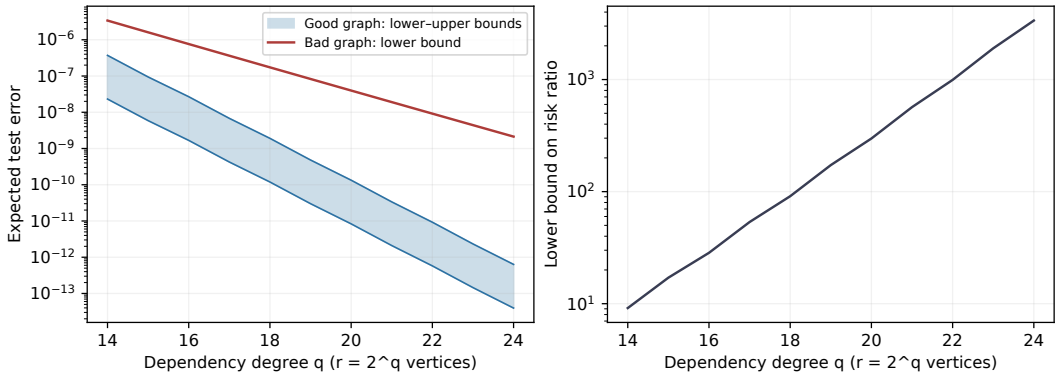

Figure 6: Deterministic bounds from Proposition P.2, with $r = 2^q$ and half the physical nodes labeled. Left: the blue interval bounds the good graph's minimax risk, and the red curve lower-bounds the bad graph's minimax risk. Right: a proved lower bound on their risk ratio. All curves are analytic bounds.

Reproducibility. The graph-feasibility script checks graph feasibility and decoder consistency and verifies the finite arithmetic underlying these bounds. It does not estimate large-graph minimax risk. The finite-graph audit additionally enumerates every sampled dependency subgraph and budget-feasible mask on the triangle, four-cycle, and K_4, checking the deterministic component argument and exact expected-risk bounds. The manuscript proof, not these finite checks, covers arbitrary q.

Proposition P.3 (A tunable label-fraction interval)**.** *Keep the two graphs, q, r, B, N of Proposition P.2, but allow any integer $N/2 \le m \le 3N/4$ and write $s = N - m$. Then*

$$R^*_{\text{good}}(m) \le \frac{4qB}{sr}, \qquad R^*_{\text{bad}}(m) \ge \frac{s-1}{N(N-1)}, \qquad \frac{R^*_{\text{bad}}(m)}{R^*_{\text{good}}(m)} \ge \frac{rs(s-1)}{4qBN(N-1)} \ge \frac{r}{128qB}.$$

Proof. For every cut, its avoidance probability is nonincreasing in m, by coupling uniform subsets as prefixes of a random permutation. The good graph's connected-set sum is therefore at most its value $2/r$ at $m = N/2$. The decoder bound $2qBW/s$ gives the first inequality. The same two bad-graph masks differ on two query targets. Until either is labeled their posterior remains fair, giving $\binom{N-2}{m}/[s\binom{N}{m}] = (s-1)/[N(N-1)]$. The good-graph isolation lower bound is positive throughout this interval, so the ratio is defined. Dividing proves the first ratio inequality. Since $s \ge N/4 \ge 2$, use $s - 1 \ge s/2$ and $N - 1 \le N$ for the last inequality. Thus the comparison holds on a label-fraction interval, although its absolute risks still vanish and no fixed-positive-accuracy separation is asserted. □

Q Value and Limits of the Mask Constraint

Q.1 Exact Unrestricted Risk and Sparse-Class Comparison

Proof of Proposition K.4. Put independent fair bits on all latent vertices before sampling. Conditional on the sampled edge set and its parity labels, the orientation of each connected component remains an independent fair bit. A query within a component is identified; a query between components is fair. Its endpoints cannot be connected by its own sampled edge, so every such query is a test node. Summing conditional Bayes errors and dividing by $N - m$ proves the lower bound. The path rule has exactly the same expected loss for every target mask, proving equality and its asserted fixed-world risk. Independent fair guesses per query suffice; no globally consistent prediction constraint is imposed.

For a q-regular graph, let A_q denote the probability that a fixed vertex is isolated in F. If either endpoint of an edge is isolated, that edge joins distinct components. Pointwise, the number of edges incident to at least one isolated vertex is at least half the sum of isolated-vertex degrees. Therefore

$$R^*_{\text{all}} \ge \frac{rqA_q}{4(N-m)}.$$

For the hypercube at half labeling, Appendix P proves $A_q \geq 1/(2r)$, so $R^*_{\text{all}} \geq q/(4N)$. Its sparse upper bound is $R^*_B \leq 8qB/(Nr)$. Division gives $r/(32B)$. The sparse lower bound is positive in these parameters, so the ratio is well defined. □

The comparison does not hold the allowed mask class fixed: R^*_{all} enlarges it. It does hold the dependency graph, features, query population, sampling protocol, and test loss fixed. The stated path rule has constant fixed-world risk and thus also supplies a concrete decoder comparison on the sparse family. Other procedures that do not explicitly receive B may perform better; no universal lower bound for such procedures follows. The sparse absolute risk vanishes; the ratio alone is not a fixed-accuracy sample-complexity theorem.

Q.2 Exact identification from a feasible transcript

Let z be any released support vector and consider the switched parity transcript. For a component C_i of sampled queries, the two candidate sets of ones are complementary subsets O_i^0, O_i^1 of C_i. Discard an orientation if it contains a vertex outside $\text{supp}(z)$. A feasible transcript leaves at least one orientation per component. Choose a minimum-size allowed orientation $O_i^{\min}$, let $c_i = |O_i^{\min}|$, and set

$$c = \sum_i c_i, \qquad \tau = B - c.$$

If a second orientation is allowed, let δ_i be its cardinality minus c_i; otherwise set $\delta_i = +\infty$.

Proposition Q.1 (Transcript identification)**.** *A feasible transcript has $\tau \geq 0$. A component's orientation is fixed over all feasible masks if and only if $\delta_i > \tau$. A query with endpoints in distinct components i, j is identified if and only if both orientations are fixed. A query within one component is always identified. These conclusions can be computed in $O(r + |E(H)|)$ operations.*

Proof. The minimum orientations together form an allowed mask of size c. Every other allowed mask is obtained by reversing a subset J of components with finite δ_i, and its cardinality is

$$c + \sum_{i \in J} \delta_i.$$

All increments are nonnegative. Thus component i can be reversed in a feasible mask precisely when $\delta_i \leq \tau$: sufficiency follows by reversing only i, and necessity follows since any subset containing i costs at least δ_i. If either endpoint component of a cross-component query can be reversed, reversing that component alone changes the query's parity while preserving feasibility. Otherwise both endpoint bits are fixed. Within one component, reversing all bits preserves every pairwise parity. A graph traversal constructs the two orientations; checking support membership, counting their sizes, and applying the criterion are linear operations. □

For $z = \mathbf{1}$, both orientations are allowed and $c_i = a_i$, $\delta_i = b_i - a_i$, recovering the main-text formula. Balanced components have $\delta_i = 0$ and are not fixed by a nonnegative slack budget. The criterion concerns agreement of all feasible targets on a particular query, not exact identification of the entire hidden mask.

For example, two observed positive pairs on four latent vertices have $a_1 = a_2 = 0$, $b_1 = b_2 = 2$. At $B = 1$, neither component can be reversed, so all cross-pair queries are identified despite the disconnected observation graph. At $B = 2$, either pair can be reversed, and those same queries become ambiguous. An example with nonzero minimum cost is a sampled three-vertex path with relative-bit classes of sizes 1 and 2 and one isolated vertex. Here $c = 1$. At $B = 1$ both components are fixed; at $B = 2$ each can be reversed individually. These are transcript statements, not unconditional error floors.

Identification and joint minimax prediction. Query ambiguity alone does not force conditional minimax error $1/2$ in the joint prediction problem, because cardinality constraints couple the possible labels. For $H = K_5$, $B = 1$, and no observed queries, every query is ambiguous, yet predicting the empty mask makes at most four of the ten query predictions wrong. The worst error normalized over queries is therefore at most $2/5$. The identification criterion characterizes certainty of individual query labels; the joint minimax decision problem additionally accounts for their coupled possibilities.

Statistical equivalence and the legal graph realization. For each fixed release, switching observed and predicted query labels is a loss-preserving bijection to sparse parity inference over $\{T \subseteq \mathrm{supp}(z) : |T| \le B\}$, with the same sampling protocol. At the all-present release, this gives the full sparse-mask experiment. The extension realizes that experiment through legal teacher-preserving graph edits and establishes its budget-dependent finite-risk and label-complexity bounds.

R COVERAGE WITHOUT IDENTIFIABLE MEMBERSHIP

The coverage equalities in Section 4 assume that the released observation identifies the true blocks. The following two-world construction removes that assumption using only two feature coordinates. It is a separate one-layer family and does not extend Corollary 5.1.

Proposition R.1 (Covered pairs can remain ambiguous)**.** *For every integer $q \ge 2$, there is a known two-world binary family on $2q$ vertices whose clean graphs are the two alternating perfect matchings of C_{2q}. Adding the absent matching uses q edges, has local incidence one, and produces the same released graph. A feature matrix in $\mathbb{R}^{2q\times 2}$ is fixed across worlds, with unit row norms. In each world a unit linear readout realizes its target on both graphs under $P_{1/2}$, with signed margin at least $1/(5\sqrt{2})$. Nevertheless, revealing all even-indexed labels covers every true pair in either world and leaves conditional randomized minimax test error $\lfloor q/2\rfloor/q$.*

The readout is preserved within a world but can differ between worlds; it is not given to the learner. Features expose both candidate target vectors, not the true selector. The error is conditional on the displayed label positions, not a constant unconditional floor. With $D = 2\lfloor q/2\rfloor$ disagreement vertices and known-zero fillers, its uniform-m risk is exactly $D\binom{N-D}{m}/[2(N-m)\binom{N}{m}]$. Appendix S gives the construction and matching proof.

The distinct full family allows arbitrary labels on pairs of either matching. It has an exact finite minimax game, but it cannot retain all those targets with constant-dimensional homogeneous linear features: for M active vertices in K cycles, $M/2 \le d_{\min} \le M - K$. Released cycles must still be recognizable. Thus compact features preserve the coverage counterexample, not the full arbitrary-pair-label family or a new unknown-membership depth law.

S HIDDEN MEMBERSHIP: COMPACT SUBCLASS AND FULL FAMILY

All results in this section use one propagation layer, $P_{1/2}$, binary threshold $\mathbf{1}\{z > 0\}$, and the same randomized test loss as the main theorem. They do not impose the main bridge family's relation between graph and target. The clean graphs here are matchings; the released graphs are recognizable cycles. The target family and the fixed feature design are known to the learner.

S.1 PROOF OF PROPOSITION R.1

Index vertices cyclically from 0 to $2q-1$. Define

$$M_0 = \{\{2i, 2i+1\} : 0 \le i < q\}, \qquad M_1 = \{\{2i+1, 2i+2 \pmod{2q}\} : 0 \le i < q\}.$$

Let $a_i = 2(i \bmod 2) - 1$, with subscripts interpreted modulo q when used below, and set

$$t^{(0)}_{2i} = t^{(1)}_{2i} = a_i, \quad t^{(0)}_{2i+1} = a_i, \quad t^{(1)}_{2i+1} = a_{i+1}.$$

World s has clean graph M_s and target $Y = (t^{(s)} + 1)/2$. Use fixed features $X_i = (t^{(0)}_i, t^{(1)}_i)/\sqrt{2}$ and readouts $w_0 = (1, 0)^\top$, $w_1 = (0, 1)^\top$. These have unit norms. Clean partners have equal signed targets, so their signed propagated scores are $1/\sqrt{2}$. Adding the absent matching yields C_{2q}. This rule reads only adjacency; it needs neither the targets nor sampled labels. It adds q distinct edges and exactly one incident edge per vertex. On the released graph, one neighbor is the true partner. If j is the other,

$$t^{(s)}_i (P_{1/2} X w_s)_i = \frac{3/2 + t^{(s)}_i t^{(s)}_j}{(5/2)\sqrt{2}} \ge \frac{1}{5\sqrt{2}}.$$

Thus the same readout within a world realizes both graphs' targets. It is not one shared readout across the two worlds.

For the displayed sample $S = \{0, 2, \ldots, 2q-2\}$, each true matching edge has a labeled endpoint. Both worlds give the same labels on S and the same released graph and features. The number of odd vertices on which their labels disagree is the number of transitions in the cyclic sequence a_i: $D = q$ for even q and $D = q-1$ for odd q, namely $2\lfloor q/2 \rfloor$. A fair prior on the worlds gives Bayes error $1/2$ on every disagreement and zero elsewhere. Fair guesses on those disagreements attain $D/(2q)$ for either world, proving the conditional minimax equality.

Proposition S.1 (Uniform-label two-world risk)**.** *With known-zero fillers, $N \geq 2q$ and $0 \leq m < N$, the same two-world family has*

$$R^*_{\text{two}}(N, m; q) = \frac{D}{2(N-m)} \frac{\binom{N-D}{m}}{\binom{N}{m}}, \qquad D = 2\lfloor q/2 \rfloor.$$

Its clean risk is zero without labels.

Proof. One sampled disagreement identifies the true target; otherwise the transcript is identical in both worlds and all D disagreements remain in the test set. The same fair prior and fair decoder match on that event, whose probability is the displayed binomial ratio. The clean adjacency identifies s directly. Fillers have zero features and target zero, so they introduce no extra uncertainty. □

S.2 The distinct full-family minimax game

Let $\mathcal{H}_q$ contain all binary vectors constant on M_0 or on M_1. Each matching supplies 2^q vectors. Their intersection consists of the two constant vectors, because the union of the matchings is connected. Thus $|\mathcal{H}_q| = 2^{q+1} - 2$. For K released cycles with vertex sets V_k of sizes $2q_k$, let the target family be $\prod_k \mathcal{H}_{q_k}$, with known-zero fillers. Set $M = \sum_k 2q_k \leq N$ and require $B \geq M/2$ so every product world admits the common attack. Labels on each true pair can vary arbitrarily; unlike the two-world subclass, clean adjacency alone does not identify those labels.

For $T \subseteq V_k$ put

$$a_k(T) = \frac{\binom{N-|V_k|}{m-|T|}}{\binom{N}{m}}.$$

This is the probability of the exact event $S \cap V_k = T$, not the probability of just its cardinality. For a prior π on $\mathcal{H}_{q_k}$ define

$$F_k(\pi) = \sum_{T \subseteq V_k} a_k(T) \sum_{z \in \{0,1\}^T} \sum_{i \in V_k \setminus T} \min_{b \in \{0,1\}} \sum_{\substack{y \in \mathcal{H}_{q_k}:\ y|_T = z \\ y_i = b}} \pi(y).$$

Impossible transcripts contribute zero. The inner minimum is the unnormalized Bayes mistake mass for the vertex; no posterior denominator is missing.

Proposition S.2 (Matching product-game characterization)**.** *For the full Cartesian-product family and $u = N - m > 0$,*

$$R^* = \frac{1}{u} \sum_{k=1}^{K} v_k, \qquad v_k = \max_{\pi \in \Delta(\mathcal{H}_{q_k})} F_k(\pi).$$

Proof. The local decision problem has finitely many worlds, sampled subsets, transcripts, and binary prediction vectors. Randomized rules form the convex hull of finitely many deterministic rules. Finite minimax duality therefore equates its minimax expected mistake count with the maximum Bayes count v_k. A componentwise rule attaining these local values has expected total mistakes at most $\sum_k v_k$ for every fixed product target, because the marginal law of its local sampled subset is a_k.

For the lower bound, select a maximizing prior for each component and take their product before sampling S. Such maxima exist by continuity on finite probability simplices. Given the sampled positions, other components' labels are independent of the local target. Therefore the posterior Bayes mistake counts add, and their expectation equals $\sum_k F_k(\pi_k) = \sum_k v_k$. Dividing by the fixed u gives a matching Bayes lower bound on minimax risk. Correlation among sampled subset sizes across components changes neither argument. Each component prior is fixed before the sampled subset and its size are observed. □

When the true matching is supplied, the $M/2$ recognizable pairs have exact risk $M\binom{N-2}{m}/[2u\binom{N}{m}]$ by the independent-bit coverage argument. For one four-node full family at $m = 2$, this value is $1/6$, whereas the hidden-matching full-family risk is $1/3$. The two-world subclass has risk $1/12$ at the same (N, m). These values correspond to the three distinct target classes.

Direct four-node hidden-risk proof. The six target vectors are $0000, 1111, 0011, 1100, 0110, 1001$. With two sampled vertices, an opposite pair (two of the six possible subsets) has worst-case one expected mistake among the two unsampled vertices: if its labels differ, either completion is possible; if they agree, the vector is constant and identified. For an adjacent sampled pair (four subsets), differing labels identify the completion. If its labels agree at b, choose the unsampled pair's common prediction by a fair coin. Then every constant target has average mistakes $(4/6)1 = 2/3$, and every nonconstant target has average mistakes $(2/6)1 + (2/6)1 = 2/3$. The uniform prior over the six distinct vectors gives Bayes count $2/3$: on an adjacent sampled pair only the two equal-label transcripts are ambiguous, each contributes $1/3$ mistake mass, whereas for an opposite sampled pair the two differing-label transcripts each contribute $1/3$. The resulting test risk is $(2/3)/2 = 1/3$.

S.3 Feature dimension of the full product family

The next bounds concern one fixed feature matrix across all worlds and homogeneous linear readouts on the common released graph. Fix one clean matching per component and select one representative from every pair. These $M/2$ vertices admit every binary labeling. Write their propagated rows as $z_1, \dots, z_{M/2} \in \mathbb{R}^d$. If they were linearly dependent, choose nonzero coefficients c_i with $\sum_i c_i z_i = 0$, reversing all signs if needed so at least one $c_i > 0$. Assign label one when $c_i > 0$, zero when $c_i < 0$, and arbitrary labels when $c_i = 0$. A realizing readout would satisfy

$$\sum_i c_i \langle z_i, w \rangle > 0,$$

since positive coefficients have strictly positive scores and negative coefficients have nonpositive scores. This contradicts the dependence. Thus $d \geq M/2$. Allowing an affine intercept instead yields $d + 1 \geq M/2$.

For the upper bound, in each component every signed target is orthogonal to the alternating vector $(1, -1, \dots, 1, -1)$. Use an orthonormal basis for that orthogonal complement, concatenate bases with disjoint coordinate supports, and set filler rows to zero. The resulting X has $d = M - K$ columns, orthonormal over the active rows, and row norms at most one. For a signed target t on active vertices, put $w = X^\top t/\sqrt{M}$. Then $\|w\| = 1$ and $Xw = t/\sqrt{M}$. The propagation calculation from the compact proof gives clean signed margin $1/\sqrt{M}$ and released margin at least $1/(5\sqrt{M})$. Hence $M/2 \leq d_{\min} \leq M - K$ for sign realization.

For exact target-score reconstruction, the two real spaces of vectors constant on the respective matchings have dimensions q and q, intersecting in the one-dimensional constant space. Their sum has dimension $2q - 1$; summing over components gives $M - K$. Signed pair-constant vectors span those spaces. Thus the upper design has minimum rank for exact target-score reconstruction. This rank fact does not prove that $M - K$ is the minimum sign-only dimension.

Scope of the comparison. The compact construction preserves a conditional membership ambiguity with constant dimension and margin. Preserving the full arbitrary-pair-label family requires growing feature dimension for this linear teacher. Both results still recognize released supercomponents and allow only two clean matchings in each. Neither is a universal nonlinear-GNN impossibility or an unknown-membership depth-phase law. Within each world a readout preserves the target, but the learner is not given that readout. A common observable reference classifier cannot simultaneously realize opposite targets on the same transcript.

T Relational decoder evaluation

The historical grid uses $r \in \{16, 32, 64\}$, $B = r/16$, $m/r \in \{1/2, 1, 2, 3, 4\}$, all cardinalities $t = 0, \dots, B$, and 60 paired physical-node samples per cell. The original 9,000 losses cover the threshold-four proof decoder, the cardinality-prior posterior rule, and a path-only fair rule. We rerun

the revised threshold-three decoder on exactly the same 3,000 label sets, with the same representative masks. Randomized baseline tie losses are integrated. All raw losses and paired summaries are supplied. These are conditional sample means, not empirical minimax values; representative masks and fixed tie rules do not establish a worst-world estimate.

Table 29: Retained relational evaluation under the sampling and mask protocol stated above.

$r = 64, B = 4, t = 4$	Old proof rule	Revised rule	Posterior	Path-only
$m = 64$	0.11196	0.11196	0.02400	0.19876
$m = 128$	0.11209	0.00259	0.00259	0.01648
$m = 192$	0.00000	0.00000	0.00000	0.00108

At $r = 64, B = 4, t = 4, m = 128$, the threshold change reduces mean error from 0.11209 to 0.00259, matching the displayed posterior mean; path-only error is 0.01648. At $m = 64$, the revised rule still uses its low-label default and has error 0.11196, compared with 0.02400 for the posterior. The threshold-three rule improves this finite-instance comparison, but the posterior rule still performs better in some other cells. At $m = 192$ the two budget-aware rules make no errors in the 60 sampled cases, not necessarily in expectation. Every loss on an unchanged branch matches the historical value exactly. The new comparison uses an existing diagnostic grid after algorithm revision and is not independent confirmation. Complete per-cardinality results and adverse cases are retained.

U Relational verification and evaluation details

The exact finite solver is a discovery tool; its verifier uses rational arithmetic without importing an optimization solver. It checks probability feasibility, prior normalization, every world risk, and equality of primal maximum loss and dual Bayes loss. Reverification passed 435 certificate sets containing 32,295 aggregated cells, 1,089 physical world/release pairs, 4,062 decoder/sample checks, and 306 complete world-risk checks. These cover small finite instances and do not replace the arbitrary-parameter proofs. Table 30 gives exact examples for $r = 5$ and $N = 20$. The risk saturates at $B = 2$, so the final column also applies to larger budgets. Budgets $B \geq 2$ saturate here. These

Table 30: Retained small-instance verification results; finite checks do not replace general proofs.

m	$B = 0$	$B = 1$	$B = 2$
0	0	$1/5$	$1/4$
2	0	$46/285$	$14/57$
5	0	$4207/46512$	$3013/15504$
10	0	$20051/1108536$	$4855/92378$
19	0	0	0

cells are outside the sparse theorem's $r \geq 16$ regime and are not a verification of that theorem's hypotheses.

The sparse-bound checker counts connected labeled graphs recursively and computes component probabilities under fixed-size sampling. It verifies 1,099 graph cases through five vertices and 100 exact bound cells at $r = 16, 32, 48, 64$. Four constructed component events check wrong ties, majority-corrupted components, and components too large to flip; eight further decoder examples test valid inputs. These are bound and implementation checks, not exact large-instance minimax calculations.

The historical evaluation comprises 3,000 sampled label sets and 9,000 algorithm executions: $r = 16, 32, 64$, all $t = 0, \ldots, r/16$, five label counts, and 60 samples. At each cardinality the first t latent vertices are the hidden mask; the release is all-present. A seed indexed by $(r, m, \text{replicate})$ pairs all methods and cardinalities. Each loss averages over the $N - m$ test nodes, integrating the prediction coin when needed. The CSV reports means and standard errors separately for every $(r, B, m, t, \text{method})$. No selection of an empirically worst cardinality is presented as an exact minimax value. No GNN optimization or runtime benchmark claim is made. Evaluation is descriptive and

was not independently preregistered; the earlier block audit is retained as a historical implementation diagnostic.

The revised threshold-three decoder adds 3,000 executions on the same label sets. Its new raw losses and paired summaries are supplied separately; the old decoder results remain unchanged.

V IDENTIFICATION, STABLE FEATURES, AND DECISION-CLASS BOUNDARIES

V.1 AN EXACT PRIOR-AVERAGED EXCESS DECOMPOSITION

Proposition V.1 (Identification and excess prediction error)**.** *Fix a layout and independent fair block bits, uniform label sampling, and a learner whose prediction uses only the permitted transcript and independent randomness. Let H, M, and F denote respectively the covered active, missed active, and filler test vertices. Then*

$$\mathbb{E}R(\mathcal{A}) - \frac{1}{2u}\mathbb{E}|M| = \frac{1}{u}\mathbb{E}\left[\sum_{v\in H}\mathbf{1}\{\widehat{Y}_v \neq Y_v\} + \sum_{v\in F}\mathbf{1}\{\widehat{Y}_v \neq 0\}\right].$$

The same identity holds with any bit-independent distribution over label sets of the same fixed size m, using its corresponding coverage reference. This is a prior-average identity, not a fixed-world or minimax excess identity.

Proof. Condition on the sampled indices and all revealed bits. A bit on a missed block remains an independent fair coin, also independent of the learner's private randomness. Each prediction on that block therefore has conditional error $1/2$, whatever the learner does. Sum the errors on the disjoint sets H, M, F and average. Individual realized errors on M need not equal $|M|/2$. □

More generally, for transcript T and a uniformly chosen test index v, put $p_v = \Pr(Y_v = 1 \mid T)$, $q_v = \Pr(\widehat{Y}_v = 1 \mid T)$, and $b_v = \mathbf{1}\{p_v > 1/2\}$. The conditional excess over the Bayes rule is

$$p_v(1-q_v) + (1-p_v)q_v - \min(p_v, 1-p_v) = |2p_v - 1|\Pr(\widehat{Y}_v \neq b_v \mid T).$$

At $p_v = 1/2$ the expression is zero regardless of the tie convention. At $p_v \in \{0,1\}$ it is exactly prediction error on an identified target. This elementary conditional-risk identity is not claimed as new statistical machinery. It connects the block case ($p_v \in \{0, 1/2, 1\}$) with cardinality posteriors in the relational family, which can also give intermediate values. Posterior-identifiability must always be defined relative to a specified prior; it is not automatically identifiability under every compatible world.

V.2 A POSITIVE-MARGIN FEATURE VARIANT

Proposition V.2 (Active-node margin with bounded feature perturbations)**.** *Fix the original block geometry and $K \geq 1$, $d \geq K$, $0 < \delta \leq 1$. Replace the sparse features on C_k by $X_v^0 = [\delta + (1-\delta)\mathbf{1}\{v = p_k\}]e_k$, leaving fillers zero. Let O be any fixed orthogonal $d \times d$ matrix and let E be fixed, independent of the hidden bits, with $E_v = 0$ on fillers and $\max_v \|E_v\|_2 \leq \eta < \delta/\sqrt{K}$. Set $X = (X^0 O + E)/(1+\eta)$ and $w^\theta = O^\top \sum_k (2\theta_k - 1)e_k/\sqrt{K}$. The feature row norms and readout norm are at most one. For every clean and completed graph, every $L \geq 1$, and every fixed $\lambda > 0$, the signed scores on all active vertices obey*

$$(2Y_v - 1)(P_\lambda^L X w^\theta)_v \geq \gamma := \frac{\delta/\sqrt{K} - \eta}{1+\eta} > 0.$$

Thus the exact fixed-layout coverage law is unchanged, including for features with within-block rank greater than one. An additional score perturbation bounded by ρ, and threshold τ, preserve active labels whenever $\rho + |\tau| < \gamma$.

Proof. The row norms of X^0 are at most one, orthogonal rotation preserves them, and the triangle inequality proves the feature bound. The readout has norm one. In each clean or released connected

piece of block k, every row of X^0 has only coordinate k and that coordinate is at least δ. Row-stochastic propagation takes convex combinations within that same block, even if the clean bridge is absent. Hence the unperturbed signed score is at least $\delta/\sqrt{K}$. Convexity and Cauchy–Schwarz give $|(P_\lambda^L E w^\theta)_v| \le \eta$. Division by $1+\eta$ yields the bound. The threshold claim is another triangle inequality. The released graph and fixed features still do not depend on θ, and the component parser still recovers the blocks. The fair-prior lower bound and copy/fair upper bound are therefore unchanged. Clean adjacency still reveals the bits in the specified layout. Fillers have known label zero and score zero; the asserted strict margin is only on active nodes. □

Dense feature support and depth capacity. The dense-feature family preserves the coverage law and gives positive active-node margins under bounded perturbations. Because every active node receives block-specific feature support, realization no longer requires propagation from a single payload. The original sparse-support capacity Q_L is therefore a geometric cap for this variant rather than an intrinsic depth restriction. Its margin bound decreases as $K^{-1/2}$.

V.3 Approximate transfer and an explicit failure

Proposition V.3 (A margin certificate for approximate factorization)**.** *On a common-label block C_k, suppose logits satisfy $f_v = a_v t_k + e_v$, where $a_v > 0$ and $|e_v| \le \epsilon_v$. For observed $s \in C_k$, write $y_k = 2Y_s - 1$ and assume the actual signed training score satisfies $y_k f_s \ge g_s$. Set*

$$\ell_k = \max_{s \in S \cap C_k} \frac{g_s - \epsilon_s}{a_s}.$$

Every test vertex satisfying $a_v \ell_k > \epsilon_v$ is predicted correctly. In particular, the covered-node contribution to overall test error is bounded by the number of covered test vertices for which this inequality fails, divided by u.

Proof. For each observed s, $a_s y_k t_k = y_k f_s - y_k e_s \ge g_s - \epsilon_s$. Take the maximum lower bound and substitute into $y_k f_v \ge a_v \ell_k - \epsilon_v$. Strict positivity suffices for both classes under the convention $\mathbf{1}\{f_v > 0\}$. □

This is a deterministic, potentially a posteriori certificate. Without an independently useful bound on ϵ_v it is not a learning-rate theorem. For example, least-squares fitting can compute residuals but does not guarantee they are small, and zero classification error need not give positive g_s.

A bad solution exists without invoking any optimizer. On a two-vertex block with feature column $(1,0)^\top$, both labels one, and only the payload labeled, a one-layer mean-SAGE score with self weight 1, neighbor weight -1, and zero bias is $(1,-1)^\top$. Training classification error is zero and covered test error is one. The target itself is realizable by positive tied propagation. This head alone proves only possibility of failed transfer, not its frequency under training or necessity of a restriction. Proposition D.1 adds an exact trained-risk statement for specified linear learners and a finite-horizon positive-overlap extension; it does not analyze nonlinear SAGE.

V.4 Heterogeneous layouts and the ambient problem

For an arbitrary ambient class of legal graph-world/attack pairs, define the minimax risk using the same pre-split attack protocol and observations as Section 3. If this class contains a fixed block subfamily, its risk is at least the subfamily risk by inclusion. No ambient upper bound follows: the copy/fair rule requires observable label-homogeneous components.

The exact heterogeneous optimization over at most b blocks of sizes $2 \le q \le Q_{\max}$, with total size at most N, is a finite resource allocation. Let $w(q) = q\binom{N-q}{m}$ and initialize $D_0(n) = 0$ for $0 \le n \le N$. For the bridge-tree family, $Q_{\max} = \min\{Q_L, N\}$. Interpret an inner maximum over an empty set as $-\infty$, so the option $D_{j-1}(n)$ remains available when $n < 2$. For $j = 1, \dots, b$, use

$$D_j(n) = \max\left\{ D_{j-1}(n), \max_{2 \le q \le \min(Q_{\max}, n)} [w(q) + D_{j-1}(n-q)] \right\}, \qquad R^*_{\text{het}} = \frac{D_b(N)}{2u\binom{N}{m}}.$$

The recurrence enumerates whether an additional block is used and its size; induction on j proves optimality. The original homogeneous maximum need not be this exact value. In the principal regime

the same phase order nevertheless holds: homogeneous layouts give its lower bound, while total mass at most N, $K \le b$, and $qe^{-mq/N} \le N/(em)$ give the three upper bounds separately. This still optimizes a constructed family, not arbitrary poisoned graphs.

V.5 Active labels and partial canonicalization

For a fixed recognizable noiseless layout, sort $Q_1 \ge \cdots \ge Q_K$. An active learner allowed exactly $t < N$ distinct queries has randomized minimax risk

$$R^*_{\mathrm{active}}(t) = \frac{\sum_{k>\min(t,K)} Q_k}{2(N-t)}.$$

Query one node from each of the $\min(t, K)$ largest blocks, then arbitrary unused nodes if necessary. Every remaining bit is fair under the product prior; no adaptive query can reveal an unqueried independent bit or remove more unqueried mass than this choice. This proves both bounds and gives zero risk once K queries are available. The passive label cost is therefore protocol specific, not an inherent cost under active querying.

For partial canonicalization, fix and announce a set J of blocks independently of their bits and of S, and complete precisely those blocks' missing bridges. The layout and designated bridges are known. Unselected bridges still reveal their bits. The exact randomized risk becomes

$$\frac{1}{2u} \sum_{k \in J} Q_k \frac{\binom{N-Q_k}{m}}{\binom{N}{m}}.$$

At most $|J|$ additions are used. The formula assumes J is announced and chosen independently of hidden bits and the sample; hidden bit-dependent selection or conditioning on realized edit counts can couple the posterior.

W Matched controls and executable evidence

The original experiments use NumPy/SciPy, and the maintained-operator replication uses PyTorch/PyG. The reported implementation checks compare gradients, forward predictions, and metric aggregation. All datasets are generated by the specified graph constructions.

W.1 Sampling, estimands, and matched objectives

Each of 12 independent seeds draws fair block bits and K uniformly sampled active vertices among KQ nodes; two initializations share that instance. An additional $F/8$ labels are drawn from F known-zero isolated fillers. The same active instance and initial parameters are used for $F = 0, 192, 448$. When filler BCE is masked, those filler labels still count as observed and are excluded from the test denominator. This is explicit stratified sampling. Its conditional coverage reference is missed active test mass divided by $2u$; its population expectation is $KQ\binom{KQ-Q}{K}/[2u\binom{KQ}{K}]$. Paired clean runs use the same target and sample but remove the zero-bit bridges. The clean structural decoder is tested against all clean targets.

Each run has equal weight. For 12 seed means z_j, each averaging its two initializations, $\mathrm{SE} = s_z/\sqrt{12}$, with $s_z^2 = \sum_j (z_j - \bar{z})^2/11$. Intervals use $t_{11,0.975}$. For the original 20 seed groups, the same formula uses 20 and 19, with each cluster averaging the three label counts sharing its seed index. This grouping defines the uncertainty unit for paired analyses; it does not make the repeated interventions independent replications. Pooled conditional rates are descriptive ratios of errors to covered test nodes; node-wise binomial intervals are not used. These intervals are marginal descriptive uncertainty, not multiplicity-adjusted discoveries.

Training and test classification use the exact comparison $f > 0$, with no numerical tolerance. Loss is mean binary cross-entropy over the stated objective indices. Strict zero-logit negatives can be classification-correct at BCE $\log 2$ and signed margin zero. Raw outputs include class-specific active errors, minimum signed margins, BCE, and threshold sensitivity at $\pm 10^{-6}$. No held-out labels affect optimization. Bias-free fillers contribute no parameter gradient; biased fillers can do so. The masked and no-filler active logits agree exactly in the executed comparisons, isolating that objective effect.

W.2 Architecture and feature controls

GCN/SGC are crossed with biases instead of comparing only bias-free versions against biased SAGE. SAGE has independent full-neighbor mean and self matrices; GCN uses symmetric self-loop normalization. Hidden width is 32, depth is three at $Q = 8$, and baseline optimization uses 200 Adam steps at learning rate .02. The additional SAGE controls use widths 16 and 64, 1,000 training steps, an L2 penalty of 0.001, or plain SGD at learning rate 0.05 for 1,000 steps. The latter is a fixed sensitivity setting, not a tuned validation recipe. Width remains at least K for $K = 4, 8, 16$. Varying $Q = 4, 8, 16$ changes the depth to $2, 3, 4$; the path control uses depth seven. Thus these geometry rows do not isolate a causal depth effect. Rank-one projection defects of preactivations are supplied by layer; small matrix rank alone does not ensure positive score orientation.

Feature rotations are fixed orthogonal matrices shared across worlds. The floor control has $\delta = .05$ at every nonpayload active node and amplitude 1 at each payload. The perturbed control adds bit-independent full-rank rows with norms $\eta = \delta/(4\sqrt{K})$ and rescales features by $1 + \eta$. Fillers remain zero. Proposition V.2 applies to row-normalized teacher propagation; the trained GCN uses its declared symmetric normalization. This distinction is maintained. The new theory does not assert that every trained hypothesis class fits every perturbed world.

Table 31: All matched controls. Covered contribution is the equal-run mean; SE clusters the two initializations within each of 12 samples. Conditional rates pool covered nodes descriptively. Perfect fits count zero classification error on all observed labels, including masked fillers.

Condition	Model	Bias	Covered	SE	Conditional	Fits
base	SGC	off	0.0000	0.0000	0.0%	24/24
base	SGC	on	0.0859	0.0067	53.5%	0/24
base	GCN	off	0.0000	0.0000	0.0%	24/24
base	GCN	on	0.0082	0.0028	5.1%	24/24
base	SAGE	off	0.0283	0.0048	17.6%	24/24
base	SAGE	on	0.0445	0.0073	27.7%	24/24
no fillers	SGC	off	0.0000	0.0000	0.0%	24/24
no fillers	SGC	on	0.1176	0.0261	18.3%	6/24
no fillers	GCN	off	0.0000	0.0000	0.0%	24/24
no fillers	GCN	on	0.0000	0.0000	0.0%	24/24
no fillers	SAGE	off	0.1131	0.0193	17.6%	24/24
no fillers	SAGE	on	0.0893	0.0281	13.9%	24/24
more fillers	SGC	off	0.0000	0.0000	0.0%	24/24
more fillers	SGC	on	0.0430	0.0033	53.5%	0/24
more fillers	GCN	off	0.0000	0.0000	0.0%	24/24
more fillers	GCN	on	0.0060	0.0015	7.5%	23/24
more fillers	SAGE	off	0.0141	0.0024	17.6%	24/24
more fillers	SAGE	on	0.0253	0.0032	31.5%	24/24
mask fillers	SGC	off	0.0000	0.0000	0.0%	24/24
mask fillers	SGC	on	0.0294	0.0065	18.3%	0/24
mask fillers	GCN	off	0.0000	0.0000	0.0%	24/24
mask fillers	GCN	on	0.0000	0.0000	0.0%	11/24
mask fillers	SAGE	off	0.0283	0.0048	17.6%	24/24
mask fillers	SAGE	on	0.0223	0.0070	13.9%	18/24
mask more	SGC	off	0.0000	0.0000	0.0%	24/24
mask more	SGC	on	0.0147	0.0033	18.3%	0/24
mask more	GCN	off	0.0000	0.0000	0.0%	24/24
mask more	GCN	on	0.0000	0.0000	0.0%	11/24
mask more	SAGE	off	0.0141	0.0024	17.6%	24/24
mask more	SAGE	on	0.0112	0.0035	13.9%	18/24
clean	SGC	off	0.0000	0.0000	0.0%	24/24
clean	SGC	on	0.0859	0.0067	53.5%	0/24

Continued on next page

Table 31 (continued)

Condition	Model	Bias	Covered	SE	Conditional	Fits
clean	GCN	off	0.0086	0.0010	5.3%	24/24
clean	GCN	on	0.0195	0.0029	12.2%	24/24
clean	SAGE	off	0.0069	0.0015	4.3%	24/24
clean	SAGE	on	0.0417	0.0057	25.9%	24/24
rotation	SGC	off	0.0286	0.0072	17.8%	8/24
rotation	SGC	on	0.0804	0.0068	50.0%	0/24
rotation	GCN	off	0.0000	0.0000	0.0%	24/24
rotation	GCN	on	0.0171	0.0036	10.6%	24/24
rotation	SAGE	off	0.0298	0.0055	18.5%	24/24
rotation	SAGE	on	0.0495	0.0079	30.8%	24/24
block feature floor	SGC	off	0.0000	0.0000	0.0%	24/24
block feature floor	SGC	on	0.0859	0.0067	53.5%	0/24
block feature floor	GCN	off	0.0000	0.0000	0.0%	24/24
block feature floor	GCN	on	0.0022	0.0016	1.4%	24/24
block feature floor	SAGE	off	0.0000	0.0000	0.0%	24/24
block feature floor	SAGE	on	0.0002	0.0002	0.1%	24/24
perturbed block floor	SGC	off	0.0000	0.0000	0.0%	24/24
perturbed block floor	SGC	on	0.0859	0.0067	53.5%	0/24
perturbed block floor	GCN	off	0.0000	0.0000	0.0%	24/24
perturbed block floor	GCN	on	0.0022	0.0012	1.4%	24/24
perturbed block floor	SAGE	off	0.0000	0.0000	0.0%	24/24
perturbed block floor	SAGE	on	0.0017	0.0010	1.0%	24/24
K4	SGC	off	0.0000	0.0000	0.0%	24/24
K4	SGC	on	0.0818	0.0124	50.9%	0/24
K4	GCN	off	0.0000	0.0000	0.0%	24/24
K4	GCN	on	0.0231	0.0078	14.4%	24/24
K4	SAGE	off	0.0394	0.0084	24.5%	24/24
K4	SAGE	on	0.0543	0.0073	33.8%	24/24
K16	SGC	off	0.0000	0.0000	0.0%	24/24
K16	SGC	on	0.0902	0.0048	57.7%	0/24
K16	GCN	off	0.0000	0.0000	0.0%	24/24
K16	GCN	on	0.0108	0.0035	6.9%	24/24
K16	SAGE	off	0.0166	0.0053	10.6%	24/24
K16	SAGE	on	0.0301	0.0073	19.3%	24/24
Q4	SGC	off	0.0000	0.0000	0.0%	24/24
Q4	SGC	on	0.0540	0.0047	38.9%	0/24
Q4	GCN	off	0.0000	0.0000	0.0%	24/24
Q4	GCN	on	0.0031	0.0021	2.2%	24/24
Q4	SAGE	off	0.0258	0.0051	18.6%	24/24
Q4	SAGE	on	0.0471	0.0064	33.9%	24/24
Q16	SGC	off	0.0000	0.0000	0.0%	24/24
Q16	SGC	on	0.0921	0.0060	54.3%	0/24
Q16	GCN	off	0.0000	0.0000	0.0%	24/24
Q16	GCN	on	0.0106	0.0038	6.2%	24/24
Q16	SAGE	off	0.0334	0.0054	19.7%	24/24
Q16	SAGE	on	0.0576	0.0065	33.9%	24/24
path	SGC	off	0.0000	0.0000	0.0%	24/24
path	SGC	on	0.0859	0.0067	53.5%	0/24
path	GCN	off	0.0000	0.0000	0.0%	24/24
path	GCN	on	0.0138	0.0034	8.6%	22/24
path	SAGE	off	0.0478	0.0082	29.7%	24/24
path	SAGE	on	0.0774	0.0088	48.1%	24/24
width16	SAGE	on	0.0359	0.0062	22.3%	24/24
width64	SAGE	on	0.0541	0.0080	33.7%	24/24
long	SAGE	on	0.0433	0.0077	27.0%	24/24

Continued on next page

Table 31 (continued)

Condition	Model	Bias	Covered	SE	Conditional	Fits
regularized	SAGE	on	0.0238	0.0068	14.8%	24/24
sgd	SAGE	on	0.0606	0.0062	37.7%	5/24

W.3 Information baselines and paired clean tasks

The component decoder uses connected components of the supplied released graph, not ground-truth block membership. Harmonic propagation (Zhu et al., 2003) solves the graph Dirichlet system on each component containing labels, with those labels clamped. On an unobserved component it uses probability $1/2$; known-zero isolated fillers receive zero. Loss integrates guessing randomness. Component correction overwrites SAGE's hard thresholded predictions on components with an observed common label; its unhit hard decisions are retained and known-zero fillers are set to zero. Sigmoid probabilities are not used as randomized decisions. Appendix A reports the corrected whole-test summary. All covered errors disappear under these repairs on the released family. The paired clean runs and zero-error structural decoder separate attack-induced information loss from training-dependent transfer error. The latter can occur on both clean and released inputs.

W.4 Original off-family boundary tests

The original shared-coordinate and attached-path conditions are retained below, separately from the matched realizable controls. The LP checks homogeneous linear readouts of the specified propagated features; nonlinear network realizability is a separate question. The shared-feature WL equivalence calculation can imply a stronger architecture obstruction. The probability $1/64$ that eight independent bits agree within the two groups of four explains why approximately one of 60 shared samples is linearly feasible.

Table 32: Original five-condition study: equal-run mean test error; parentheses are seed-cluster SEs. Reference is conditional on the sampled splits. The attached-path and combined rows use an oracle partition; baseline, shared-feature, and deep rows use the observable component partition.

Condition	Reference	SGC	GCN	SAGE
Baseline	0.0451	0.0398 (0.0037)	0.0475 (0.0046)	0.0750 (0.0042)
Shared	0.0451	0.0921 (0.0041)	0.0910 (0.0042)	0.1017 (0.0042)
Deep	0.0451	0.0398 (0.0037)	0.0458 (0.0041)	0.0799 (0.0043)
Attached paths	0.0451	0.0821 (0.0049)	0.0965 (0.0047)	0.0970 (0.0056)
Combined	0.0451	0.1571 (0.0105)	0.1571 (0.0105)	0.1277 (0.0061)

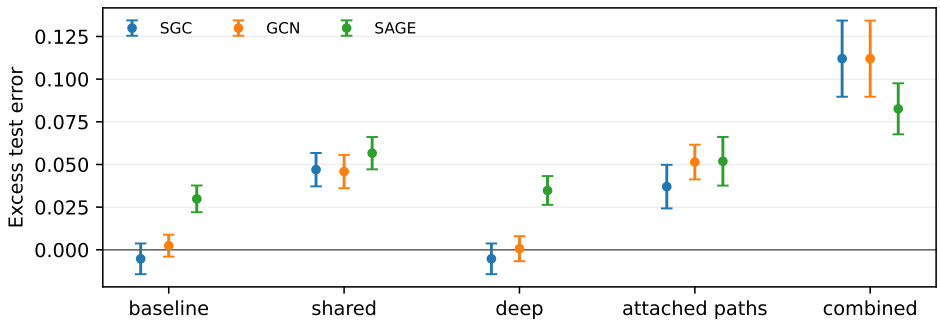

Figure 7: Architecture and graph-structure stress tests. Points show mean excess test-error fractions over the sample-conditional partition reference; bars show two standard errors across 20 seed groups, averaging three label counts per seed. Baseline and deep retain the original family. Shared features and attached paths modify representation settings; the attached-path graph has eight components and uses an oracle partition reference. Errors can reflect representation constraints as well as training.

W.5 Exact capacity optimization

For each integer cap $Q_{\max} \geq 2$, we maximize the homogeneous-family risk in Theorem 4.1 over $2 \leq Q \leq \min\{Q_{\max}, N\}$, including the packing factor $\min\{B, d, \lfloor N/Q \rfloor\}$. Figure 8 shows the optimized risk and the smallest maximizing block size at $N = 512$ for three choices of (b, m). Varying the cap changes the admissible construction family.

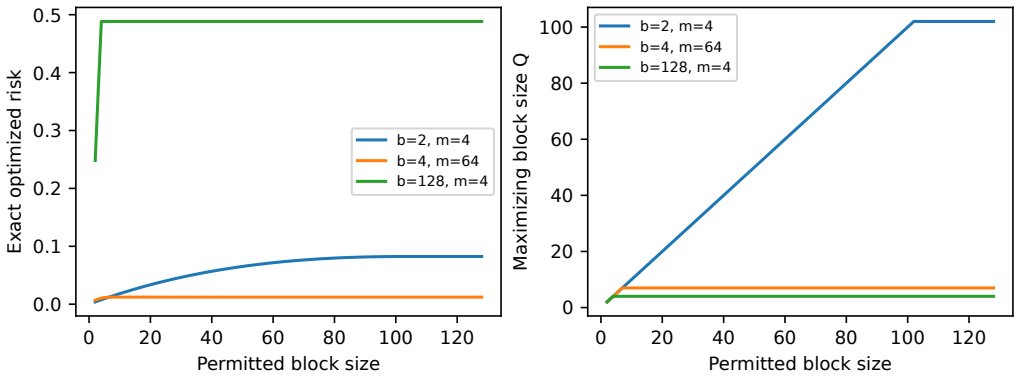

Figure 8: Exact homogeneous-family optimization at $N = 512$, including packing. Left: risk versus permitted block size. Right: the smallest maximizing size. Changing this cap changes the constructed family, not prediction depth on one fixed learning problem. All three phase regimes are represented.

W.6 Maintained-library and worst-world checks

PyTorch 2.14.0+cpu and PyTorch Geometric 2.8.0.post1 execute the declared SAGEConv, GCNConv, and SGConv updates. Initial matrices are copied from the NumPy initialization to avoid changing two factors at once. Six forward and gradient comparisons pass before fitting. All 144 matched final predictions agree, with largest logit discrepancy below 3×10^{-12}. The additional 24 normalized-SAGE and 24 GIN runs are specified in the separate protocol. GIN uses fixed $\epsilon = 0$, one bias-free linear map per layer, and hidden ReLU; its initialization is the library default. Normalization is applied only to SAGE's two hidden outputs, not its scalar final logit.

A separate fixed-split diagnostic trains bias-free SGC on all 256 bit worlds. Its world-average loss equals the fair coverage reference, but its all-one-world loss is twice that reference. This does not enumerate the worst-world behavior of all trained architectures across all splits. It verifies the specific default rule distinction and prevents describing prior-average agreement as minimax optimality. These 256 fits are separate from the 2,136 matched-control fits.

W.7 Closest-work comparison and precision of bounds

The comparison separates established inference tools (Tolstikhin & Lopez-Paz, 2016; Cesa-Bianchi et al., 2012; Tarlow et al., 2012; Foster et al., 2018) from the legal graph constructions and finite-risk results analyzed here. The relational bounds give uniform constant-factor control, with factors 113 or $2e^3$ in the stated regimes. A fair-guess alternative clips the minimax upper bound at $1/2$; the fixed proof decoder retains its own risk bound. Exact finite-game certificates and sampled decoder comparisons answer distinct numerical questions. The posterior classifier evaluates a specified cardinality prior, whereas the explicit decoder supplies the stated worst-risk upper bound. Representative-mask simulations report conditional sample means, with exact worst-world verification confined to the enumerated finite games.

For $r = 16$, $B = 1$, and $m = 0$, Appendix L.3 gives the exact minimax risk $15/152 \approx .09868$, while the general lower bound is $15/304 \approx .04934$. The universal multiplicative upper bound is loose at this endpoint; exact finite formulas and instance-specific decoder bounds provide the sharper numerical reference.

Table 33: Established ingredients and the paper's corresponding results.

Comparison	Established ingredient	Result established here
Realizable transduction	Dimension/sample lower bounds	Legal adjacency erasure, exact parsed coverage, and decision-class refinements
Signed prediction	Path products and balanced labels	Legal pre-sampling additions, a hidden cardinality mask, and normalized query risk
Cardinality inference	Efficient exact constrained counting	Posterior calculation for the released mask class and its finite-risk guarantee
Sparse pairwise recovery	Cut-based ambiguity and geometry	Fixed-size noiseless query sampling and the uniform half-component estimate
Training symmetry and dynamics	Symmetry lower bounds and kernel-mediated dynamics (Abbe & Boix-Adsera, 2022; Lahoti et al., 2024; Yang et al., 2024)	A connected finite-width one-aggregation training bound and an exact same-class linear optimizer intervention