

Dependence Is Not Advantage: What Randomizing the Graph Does Not Measure

Prahas Duggireddy 

Third Wheel

Irvine, USA

prahasduggireddy@thirdwheel.ai

Abstract

Graph ablation studies often interpret a performance drop after edge randomization as evidence that graph structure improves prediction. These are different comparisons: randomization measures how a fixed procedure depends on the observed edge arrangement, whereas improvement requires comparison with a specified graph-free baseline. Let $D = W(f, G) - W(f, N_1)$ denote dependence on the edge arrangement and $A = A(f; c)$ advantage over a comparator with the same node content and no graph. With survival ratio $s = W(f, N_1)/W(f, G)$, the relative drop is $1 - s$. For an enumerated scorer class, we derive the tightest monotone upper envelope of advantage as a function of this drop and characterize the penalty created by inverted scorer orderings. For scorers whose positive scores require a common neighbor, we also bound null performance using the surviving common-neighbor mass. On ogbl-collab, the 12 arrangement-reading scorers lose between 89.52% and 99.90% of their intact Hits@50 under the degree-preserving null, while their advantages range from -0.3147 to 0.2208 . The modularity-matrix spectral scorer loses 91.70% yet has advantage -0.0531 against the content-only comparator; GraphSAGE loses 34.25% and has advantage 0.1423 (95% CI $[0.1309, 0.1537]$). The same contrasts disagree across five additional public benchmarks. Finally, we treat a graph null model as a distribution over a constrained graph space and show why the support of its sampling algorithm must be checked separately. The results motivate reporting the intact score, graph-free baseline, survival ratio, and sampler diagnostics together. They concern evaluation validity, not downstream recommendation benefit.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Information systems** → **Evaluation of retrieval results**; • **Mathematics of computing** → *Random graphs*.

Keywords

graph neural networks, link prediction, graph ablations, graph dependence, graph advantage, evaluation validity, graph null models, degree-preserving randomization

1 Introduction

Graph-learning studies evaluate models under perturbed or randomized edges [7, 8]. Here we retrain under each graph condition and call the resulting performance drop the ablation. This tests whether the model uses the particular arrangement of edges. It does not, by itself, test whether that arrangement improves prediction relative to the same node content without a graph. That stronger claim requires a named no-graph comparator.

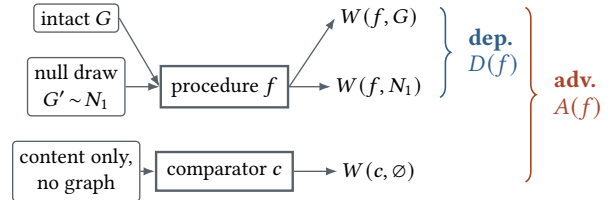


Figure 1: The two contrasts and the arm each is taken against. The same graph procedure f is evaluated on G and under the degree-preserving law N_1 ; condition-specific fitting and selection remain part of that fixed procedure definition. The no-graph condition is occupied by a distinct comparator c . Both braces share $W(f, G)$, but dependence subtracts $W(f, N_1)$ and advantage subtracts $W(c, \emptyset)$. Measured values for both contrasts are in Table 1 and Appendix I.

The two comparisons answer different questions. Randomizing the graph measures how far a scorer’s output *depends* on the arrangement of edges, on which node is linked to which, as opposed to how many edges each node has. The conclusion usually attached to that drop is stronger: that the arrangement improves performance beyond what the same node content supplies. Writing $W(f, \cdot)$ for the performance of a scorer f under an information condition, the ablation measures

$$D(f) = W(f, G) - W(f, N_1),$$

the drop from the intact graph G to a degree-preserving null draw, while the proposition asserted is about

$$A(f; c) = W(f, G) - W(c, \emptyset),$$

the gain over a comparator c that receives the same node content and no graph. These are differences against two different references, summarized in Figure 1.

To test the stronger claim, evaluate the graph-based scorer against a comparator trained on the same node content without access to the graph. This directly measures advantage over that named arm. Attributing the difference specifically to information additionally requires matched capacity and optimization, which our trained arms do not fully provide.

The distinction is algebraic. Both quantities subtract something from $W(f, G)$, so write $s(f) = W(f, N_1)/W(f, G)$ for the fraction of performance the null leaves standing, the *survival ratio*. Then $D(f) = W(f, G)(1 - s(f))$. The ablation adds one quantity to the intact evaluation: s . The reported relative drop is $1 - s$, whereas advantage does not involve s . Theorem 3.10 gives the least non-decreasing upper envelope obtainable from the drop and quantifies the unavoidable slack caused by an inverted pair.

The usefulness of the ablation depends on how much the survival ratio varies. If different scorers survive randomization to different degrees, an ablation is a coarse but real instrument; if s is pinned near zero across the scorers one might use, the reported relative drop is a constant and the absolute drop is intact performance under another name. Over the 15 scorers of Table 1 it is pinned: across the 12 that read which node is linked to which, the relative drop is confined to a band 10.38 percentage points wide while advantage crosses zero. The ablation is not resolving another quantity at this scale; within this class it largely re-reports the intact score. The most direct example is the spectral embedding of the modularity matrix, which loses 91.70% of its performance to randomization and still scores -0.0531 against a content-only comparator on identical node features: the ablation reports recoverable signal for a scorer that is worse than not using the graph at all.

This behavior follows from the benchmark’s co-occurrence structure, not from a particular scorer. Degree-preserving randomization on ogbl-collab does not gently perturb local co-occurrence; it removes it, leaving 3.80% of the co-occurrence the evaluation positives rest on. Combined with an intact-performance floor, this bounds how much relative drops can vary across scorers reading only shared neighbours. The diagnostic requires null draws but no training, so it can be run before the ablation rather than used to explain a result afterwards. Over 5 benchmarks it runs from 0.0380 to 1.0002, so whether an ablation can discriminate at all is knowable before it is run.

The same separation appears when we vary the benchmark instead of the scorer. Cora and CiteSeer lose 94.47% and 94.49% of Hits@20 to the same null, 0.02 points apart, while the arrangement is worth 2.3 times as much on the first. Across benchmarks the drop’s rank agreement with advantage runs from $+0.90$ to -0.60 , so not even its sign transfers. On ogbl-ddi dependence is measurable while a nontrivial content-matched advantage is undefined rather than small because the benchmark provides no node content for such an arm.

The diagnostic is a property of the null, not of the scorer. More generally, a graph null model specifies a distribution over a constrained graph space, while a particular sampler and budget may place mass on only part of that space. Separating these objects yields distinct checks on constraint preservation, mobility, and support adequacy.

Contributions.

- (1) **A formal separation of graph dependence and graph advantage** (Section 3), including the tightest monotone upper envelope, an inversion penalty, and a computable selectivity threshold.
- (2) **An empirical characterization of the separation** (Sections 5.2 and 5.4) across scorer families on ogbl-collab and across five additional public benchmarks.
- (3) **A prospective diagnostic for the ablation’s informative range** (Section 4.3), based on co-occurrence survival and requiring null draws but no model training.
- (4) **A distributional treatment of graph null models** (Section 4) that separates a constrained graph space from the distribution and support induced by a sampling algorithm.

Scope. We study Hits@ K on 5 public link-prediction benchmarks under one split each, and NDCG@8 on one typed synthetic generator (Appendix R). The measurements are designed to establish existence and instability, not prevalence: how often a published ablation would change interpretation under a content comparator remains open. The claims concern the validity of an evaluation contrast, not whether graph structure helps in general. Section 6 summarizes the measurements and sampler checks needed to support a graph-effect claim.

2 Related work

Graph dependence and graph advantage. Whether a measured quantity supports the construct a claim is about is the subject of measurement modelling [16]; for one such pair we give the exact algebra relating them. A natural comparator for this question is a structure-agnostic baseline: a model given the same node features but no adjacency. It is competitive far more often than reported, and omitting it distorts whole comparisons [9, 27]; on many common benchmarks graph learning turns out not to be necessary [17, 33]. Our comparator arm is that baseline. Our question is about the *other* contrast: what a randomization ablation identifies when the baseline is not run. The closest theoretical result is that a GNN is not guaranteed to converge on a graph-ignoring solution even with unlimited data when the target does not use the graph [4]. That is a pathology of trained models. Ours is a property of the evaluation contrast, and a checkable one: our sign-changing example is an eigendecomposition, with no gradient descent in it. The nearest empirical result is that GNNs on a protein interaction network do no better than uninformed models on cancer subtype classification, and are largely unchanged when the graph is heavily perturbed [8]: low dependence with low advantage, arrived at by running both comparisons. Why the two cannot be substituted for each other is what we add. The ablation contributes only s , advantage does not involve it, and where local co-occurrence does not survive randomization s has no room to vary.

Benchmark validity in link prediction. Link-prediction evaluation is distorted by easy negative sampling, inconsistent splits and understated heuristic baselines, with a corrected protocol proposed [18], and a position paper argues the field is losing relevance to poor benchmarks [5]. The same re-evaluations have landed in knowledge-graph completion [2, 28] and in recommendation [10, 25]. Those fix the negative-sampling instrument and the reporting culture. Whether the arrangement carries recoverable signal at all is orthogonal to both, and we can show the separation instead of asserting it: s_π is invariant to the negative protocol (Section 4.3). That a 2003 heuristic [1, 19] outscores the trained model here is consistent with that literature (Appendix L). Our co-occurrence measure belongs with dataset-level diagnostics computed before any model is fit [7], and is one number rather than a battery.

Null-model ladders and sampler validity. The dk -series gives a converging hierarchy of null ensembles with exact constraints at each rung [23], and the configuration-model literature catalogues how much the choice of ensemble matters [12]. Our ladder is a direct descendant. We ask something prior of a rung: whether it

can differ from the rung below it at all on the units the evaluation aggregates over, checked before any draw is made. Connected degree-preserving edge-swap sampling has known pitfalls and best practice [32], and swap-graph connectivity for a fixed degree sequence is classical [30]. That prior art targets *connectedness*; our constraint is the exact component partition. The bias we document is a protected-forest retention floor that standard diagnostics cannot distinguish from insufficient mixing, and we did not find it described.

Selectivity. Requiring a contrast to fail an appropriate control before it is credited with finding structure is long-standing practice, and a control earns that role through its own properties, not through the result it returns [20]. The closest instance in machine learning is the control task from probing [14]: a probe that also fits random labels has not found linguistic structure. That gate is on the probe. Definition 3.9 puts the graph-side analogue on the perturbation ensemble and the contrast, which is what lets it be evaluated by counting scorers instead of by fitting anything.

Synthetic benchmarks and claims from absence. Synthetic graph populations expose model comparisons a handful of real datasets cannot [24, 31], but do not by themselves show that a generated regime contains recoverable signal or validate their own perturbations. To our knowledge, three constructions below have not appeared in this form: the decomposition of Section 3.3, a control task required of a perturbation ensemble instead of a model, and our sign-changing scorer’s composition of the modularity matrix [21, 22] with the adjacency spectral embedding [3, 29]. If any has been stated before, the measurement and diagnostic results do not depend on priority.

3 The algebra of a graph ablation

The ablation drop and the advantage subtract different references from the same intact score. Perturbation-based variable importance likewise measures a predictor’s reliance on information by comparing performance before and after a stated intervention [11]. The graph setting adds a second, non-perturbative reference: a graph-free comparator. We formalize the relationship between these references and derive four consequences that transfer to any benchmark.

3.1 Conditions, arms, and two classes of scorer

Let $G = (V, E)$ be an observed simple undirected graph with node content fixed across conditions. A scoring *procedure* f includes any condition-specific fitting, validation selection, and internal randomness U ; write $M(f; H, U)$ for the resulting candidate-set metric on graph H , and $W(f, H) = \mathbb{E}_U M(f; H, U)$. For a probability law N on graphs,

$$W(f, N) := \mathbb{E}_{H \sim N, U} M(f; H, U). \quad (1)$$

Deterministic procedures are the degenerate case. Thus W is a protocol-level mean, not a single run. $\mathcal{F}$ denotes a class of procedures, and differences are read against a resolution scale $\varepsilon = 0.0147$ (Appendix K); Table 3 collects the notation.

The two ablated conditions are sampling distributions. Specifically, $N_1 = P_{S_1, b_1}$ is induced by the degree-preserving sampler at its fixed budget, and $N_2 = P_{S_2, b_2}$ additionally preserves G ’s exact

connected-component partition. Every graph in the support of N_1 has the same *labeled degree vector* $d_H = (\deg_H(v))_{v \in V}$ as G , the same unique-simple-edge count, no self-loops or duplicates, and no frozen evaluation pair introduced. The support of N_2 also preserves the full component-label partition, not merely its count. Section 4 separates these distributions from the constrained graph spaces named by the protocol.

The third condition, written $\emptyset$, is the edgeless graph on the same node set (same nodes, same content, no edges), and the arm occupying it is a *comparator* and not an absence. We instantiate all three on ogbl-collab, a collaboration graph with 235,868 nodes, 967,632 unique simple training edges and 621 connected components whose giant component holds 0.9873 of the nodes, evaluated over 106,413 candidate pairs [15] at the published split, held fixed across every arm (Appendix M). There c is a multilayer perceptron over the same node features, trained with the same objective, candidate pools, optimizer settings and an identical 131,841-parameter prediction head, differing from the graph arm only in receiving no message passing. Writing its performance as $W(c, \emptyset)$ instead of as a baseline constant keeps two facts visible: advantage is measured against a named arm, and that arm has capacity, 164,608 encoder parameters against the graph arm’s 328,448, so c is *content-matched*, *not capacity-matched*, with the consequence stated in Section 6. The untrained scorers have no capacity to match. For them, advantage compares an untrained graph scorer with a trained content-only arm and should be interpreted only as that arm contrast.

We call the intervention *degree-preserving randomization*, not “rewiring”, which in the graph-learning literature means editing a graph to relieve over-squashing. Two classes of scorer behave differently under it for reasons unrelated to how well they score, so we separate them by what they read.

Definition 3.1 (Degree-determined and arrangement-reading). A procedure f is *degree-determined* if, for every realization of its internal randomness, $f(H; U) = f(H'; U)$ whenever $d_H(v) = d_{H'}(v)$ for every labeled node v . Otherwise f is *arrangement-reading*.

Membership is settled by reading a scorer’s definition, never by running it: preferential attachment is degree-determined, its score being a function of two degrees; common neighbours is arrangement-reading, since two graphs with one degree sequence can disagree about whether a pair has a neighbour in common. Every claim below quantified over a family of scorers is quantified over one of these classes, so a reader can check what a claim covers without consulting our list.

3.2 Three population quantities and their estimates

Definition 3.2 (Dependence). For a null ensemble N , $D(f; N) = W(f, G) - W(f, N)$: positive when the procedure performs better on the observed arrangement than under the stated sampling distribution.

Definition 3.3 (Advantage). For a comparator arm c , $A(f; c) = W(f, G) - W(c, \emptyset)$: positive exactly when the graph procedure outperforms that named no-graph arm.

Definition 3.4 (Survival ratio). For a null ensemble N and any scorer with $W(f, G) > 0$, $s(f; N) = W(f, N)/W(f, G)$: the fraction of protocol-level performance the law leaves standing.

We write $s(f) = s(f; N_1)$ and otherwise drop indices at the paper’s fixed law and comparator. Unhatted W, D, A, s, R denote the protocol-level means above; hats denote finite-replicate plug-in estimates. The tables pair each replicate’s intact score with its null draw. All identities below hold for the population quantities and for coherently paired plug-in means. A graph ablation reports D ; a comparator-relative graph-effect claim asserts A .

3.3 Four lemmas

LEMMA 3.5 (FACTORIZATION). *For any scorer with $W(f, G) > 0$, writing $R(f) = D(f)/W(f, G)$ for the relative drop,*

$$D(f) = W(f, G)(1 - s(f)), \quad R(f) = 1 - s(f).$$

Hence the pair $(W(f, G), s(f))$ determines both readings of the ablation drop, and its first coordinate is measured without drawing a null.

PROOF. Substitute Definitions 3.2 and 3.4, and divide. $\square$

An ablation therefore supplies exactly one quantity the intact evaluation does not, namely $s(f)$, and by Definition 3.3 advantage does not involve it. The relative drop a paper reports is $R = 1 - s$.

LEMMA 3.6 (EXACT DISCREPANCY). *For every scorer,*

$$D(f) = A(f) + W(c, \emptyset) - W(f, N_1),$$

$$D(f) - (A(f) + W(c, \emptyset)) = -W(f, N_1).$$

Thus the discrepancy is exactly null performance. When $W(f, G) > 0$, it is equivalently $s(f)W(f, G)$. Its class-wide magnitude is $\sup_{f \in \mathcal{F}} |W(f, N_1)|$, which reduces to $\sup_{f \in \mathcal{F}} W(f, N_1)$ for a non-negative metric such as Hits@K.

PROOF. Definition 3.3 gives $A(f) + W(c, \emptyset) = W(f, G)$; subtract Lemma 3.5. $\square$

It holds scorer by scorer and predicts the agreement instead of observing it, which the regression of D on A cannot do.

LEMMA 3.7 (DEGREE INVARIANCE). *If f is degree-determined, then $W(f, H) = W(f, G)$ for every graph H in the support of a degree-preserving null law N . Hence, if $W(f, G) > 0$, then $s(f; N) = 1$ and $D(f; N) = 0$ for every such law.*

PROOF. Every draw shares G ’s labeled degree vector and node content, so f ’s output is unchanged and so is its performance under the fixed protocol. $\square$

A departure from this equality reveals an error in at least one of four places: labeled-degree preservation, the scorer implementation, the evaluation protocol, or numerical estimation. The equality is therefore a useful consistency check, although it cannot detect invariant violations outside the evaluated pairs. Section 4 validates the graph invariants directly, and Section 5.2 reports the degree-based check separately.

LEMMA 3.8 (SEPARATION OF ARGUMENTS). *As functions of the triple $(W(f, G), s(f), W(c, \emptyset))$, the relative drop $R(f)$ depends on $s(f)$ alone, and the advantage $A(f; c)$ on $W(f, G)$ and $W(c, \emptyset)$ alone.*

PROOF. Immediate from Lemma 3.5 and Definition 3.3. $\square$

3.4 What the drop alone does not order

Lemma 3.5 leaves two ways to read the drop and they fail differently. Read *absolutely*, D tracks A closely whenever s is small. But small s is exactly the condition under which $D \approx W(f, G)$, so the ablation agrees with a quantity that was in hand before the null was drawn. That is redundancy, and Lemma 3.6 says by how much. Read *relatively*, the drop can reverse the ordering of advantage.

Following the logic of negative controls and control tasks [14, 20], we ask when the contrast is selective for positive advantage.

Definition 3.9 (Selectivity and the selectivity threshold). A contrast C is *selective* for a property P at threshold τ over a class $\mathcal{F}$ if $C(f) > \tau$ implies $P(f)$ for every $f \in \mathcal{F}$. The *selectivity threshold* is

$$\tau^*(C, P, \mathcal{F}) = \sup \{ C(f) : f \in \mathcal{F}, \neg P(f) \},$$

with the supremum of an empty set taken as $-\infty$, so that C is selective for P at τ exactly when $\tau \geq \tau^*$. The *yield* at τ is the number of $f \in \mathcal{F}$ with $C(f) > \tau$.

A thresholded contrast is informative only when it is selective for the property of interest and retains non-negligible yield. Both quantities are defined on the enumerated scorer class and require neither a correlation nor a sampling frame.

THEOREM 3.10 (MONOTONE ENVELOPE AND INVERSION PENALTY). *Let $\mathcal{R}_{\mathcal{F}} = \{R(f) : f \in \mathcal{F}\}$ be the attained relative drops. For $r \in \mathcal{R}_{\mathcal{F}}$, define*

$$U_{\mathcal{F}}^*(r) = \sup \{ A(f; c) : f \in \mathcal{F}, R(f) \leq r \},$$

Then $U_{\mathcal{F}}^$ is the pointwise least non-decreasing function on $\mathcal{R}_{\mathcal{F}}$ that upper-bounds advantage. Moreover, if $R(f_1) \geq R(f_2)$ but $A(f_1; c) < A(f_2; c)$, every non-decreasing upper bound has slack at least $\delta = A(f_2; c) - A(f_1; c)$ at f_1 . If also $A(f_1; c) \leq 0 < A(f_2; c)$, then $\tau^*(R, A(\cdot; c), \mathcal{F}) \geq R(f_1)$, so no selective threshold identifies f_2 as having positive advantage.*

PROOF. The defining sets are nested in r , so $U_{\mathcal{F}}^*$ is non-decreasing and majorizes every attained point. If φ is any other non-decreasing majorant, then for every f with $R(f) \leq r$, $A(f; c) \leq \varphi(R(f)) \leq \varphi(r)$; taking the supremum gives $U_{\mathcal{F}}^*(r) \leq \varphi(r)$. At $r = R(f_1)$, the inverted scorer f_2 is in the defining set, which gives the slack δ . The selectivity claim follows because f_1 lacks positive advantage and therefore enters the supremum in Definition 3.9. $\square$

The finite table contains an estimated inversion: Structural role ($\widehat{R} = 98.63\%$, $\widehat{A} = -0.3147$) against Length-3 paths, normalised ($\widehat{R} = 93.38\%$, $\widehat{A} = 0.2208$), so the estimated inversion has gap $\widehat{\delta} = 0.5355$, which is 36.4% and 100% of the advantage span the class attains. The same inversion remains in any larger class containing these two scorers. Advantage can still be bounded, but the inversion prevents the relative drop from tightening that bound.

Dropping monotonicity yields a sharper conclusion under a stronger hypothesis on the whole class, stated and proved in Appendix B. Across both results, the two contrasts use different zero references. Advantage is zero at the measured comparator arm, and the evaluated scorers fall on both sides of it. Dependence is zero when null and intact performance agree; every measured arrangement-reading scorer has positive dependence under N_1 .

4 Null models as distributions over constrained graph spaces

An ablation specifies both a scorer and a graph distribution. Yet experimental descriptions often name only a constraint—for example, “degree preserving”—and leave the distribution induced by the sampling algorithm implicit. This distinction matters because two algorithms can preserve the same invariants while assigning mass to different subsets of the admissible graphs. This separation between constraints, ensembles, and sampling algorithms is inherited from the graph null-model literature [12, 23]; our focus is the diagnostic consequence of support restrictions.

4.1 Graph space, sampling distribution, and support

Definition 4.1 (Constrained graph space). For a set $\mathcal{I}$ of graph invariants, define $\mathcal{G}_{\mathcal{I}} = \{G' \text{ on } V : \iota(G') = \iota(G) \text{ for all } \iota \in \mathcal{I}\}$. This is a graph space, not a probability distribution; in particular, uniformity over $\mathcal{G}_{\mathcal{I}}$ is not implied. The two null conditions specify spaces $\mathcal{G}_{I_1}$ and $\mathcal{G}_{I_2}$.

Definition 4.2 (Sampler-induced distribution). A sampling algorithm S , including its moves, acceptance rule, and schedule, induces a distribution $P_{S,b}$ after budget b . Its support is $\text{supp}(P_{S,b})$, and write $\mathcal{G}_S^\infty = \bigcup_{b \geq 0} \text{supp}(P_{S,b})$ for the graphs reachable at some budget. In our experiments, $N_1 = P_{S_1, b_1}$ and $N_2 = P_{S_2, b_2}$ for fixed budgets b_1, b_2 .

Figure 3 distinguishes the constrained graph space from the support of the implemented sampler. Their relationship has three separate aspects.

Constraint preservation. The inclusion $\text{supp}(P_{S,b}) \subseteq \mathcal{G}_{\mathcal{I}}$ requires every draw to satisfy the stated invariants. We check these invariants directly for every retained draw. Lemma 3.7 supplies an additional consistency check for labeled-degree preservation, but cannot replace direct validation.

Sampler mobility. Write $F_S = \{e \in E : e \in E(H) \text{ for every } H \in \mathcal{G}_S^\infty\}$ for the edges no budget can move, and $F_{\mathcal{I}} = \{e \in E : e \in E(H) \text{ for every } H \in \mathcal{G}_{\mathcal{I}}\}$ for the edges forced by the constraints. If the sampler preserves the constraints at every budget, then $F_{\mathcal{I}} \subseteq F_S$.

Definition 4.3 (Algorithmically fixed edges). $F_S \setminus F_{\mathcal{I}}$ is the set of edges fixed by the sampling algorithm but not by the graph constraints. The sampler has full mobility for $\mathcal{I}$ when this set is empty.

LEMMA 4.4 (RETENTION FLOOR). *Writing E' for the edge set of a draw, every $G' \in \mathcal{G}_S^\infty$ satisfies $|E \cap E'| \geq |F_S|$, so the adjacency overlap between G and any draw at any budget is at least $|F_S|/|E|$.*

PROOF. The immobile edges lie in every member of $\mathcal{G}_S^\infty$. $\square$

High adjacency overlap can therefore reflect either insufficient mixing or edges that the algorithm cannot move. Increasing the sampling budget can address only the former.

Perturbation adequacy. Let $\psi(H)$ measure the perturbation delivered by a draw and let $\mathcal{B}$ be a prespecified acceptable range. The sampler is adequate for this diagnostic only if some $H \in \text{supp}(P_{S,b})$

satisfies $\psi(H) \in \mathcal{B}$ whenever some $H \in \mathcal{G}_{\mathcal{I}}$ does. Otherwise the experiment confounds the invariant being tested with restrictions imposed by the sampling algorithm.

4.2 Sampler support and design sensitivity

Definition 4.5 (Support adequacy). A sampler has adequate support for diagnostic ψ and range $\mathcal{B}$ if

$$(\exists H \in \mathcal{G}_{\mathcal{I}} : \psi(H) \in \mathcal{B}) \implies (\exists H \in \text{supp}(P_{S,b}) : \psi(H) \in \mathcal{B}).$$

PROPOSITION 4.6 (RESOLUTION BOUND). *Suppose a metric decomposes over n evaluation units as*

$$W = n^{-1} \sum_{i=1}^n m_i,$$

where $m_i \in [a, a+B]$, and two conditions can differ only on a known set J . Then $|W_1 - W_2| \leq B|J|/n$.

PROOF. Terms outside J cancel, and each remaining difference has magnitude at most B . $\square$

Application to the two null conditions. Every retained N_1 draw satisfies the stated invariants and retains only 0.000693 of the observed adjacency. The degree-determined scorers also behave as Lemma 3.7 predicts (Section 5.2). For N_2 , the reported 0.9772 share of both-in-giant evaluation pairs does not identify the units whose Hits@50 contribution can change. Proposition 4.6 therefore cannot be applied; moreover, the proposed ceiling 0.0228 exceeds the achieved half-width 0.0111. We draw no conclusion about the sensitivity of this contrast from that statistic.

The supplied partition-preserving sampler S_{forest} does fail the prespecified support-adequacy criterion. It fixes at least 2.2283% of the original edges, above the allowed overlap ceiling $100 \times 0.020693\%$, whereas the constrained graph space contains a leaf-exchange draw that retains 0.6597% and satisfies the criterion. Increasing the sampling budget cannot move algorithmically fixed edges (Appendices C and E).

4.3 A prospective co-occurrence diagnostic

Mobility and support adequacy address nulls that perturb the graph too little. The opposite problem arises when a null removes nearly all of the signal used by an entire scorer class, forcing their relative drops into a narrow range. This behavior is also a property of the graph and null distribution and can be assessed in advance.

Definition 4.7 (Co-occurrence survival). Write $\pi_+(H)$ for the fraction of positive evaluation pairs with at least one common neighbour in H . For a null ensemble N , the co-occurrence survival is $s_\pi = q_N/\pi_+(G)$, where $q_N := \mathbb{E}_{G' \sim N} \pi_+(G')$, written bare at the paper’s fixed null N_1 . A scorer is *strictly co-occurrence-supported* if, for every realization of its internal randomness, it assigns one value β_0 to every candidate pair with no common neighbour and a value strictly above β_0 to every pair with at least one.

Positives only: a Hits@K threshold is set by the negatives, so a figure pooling both bounds nothing. Common neighbours, its inverse-log-degree and resource-allocation weightings and the Jaccard and Salton normalizations all qualify, each a sum of positive terms over shared neighbours that is empty when there are none.

LEMMA 4.8 (CO-OCCURRENCE CEILING). *Let Hits@K count positives scoring strictly above the Kth highest negative score, over a candidate pool holding more than K negatives, as in every pool used here. Then $W(f, H) \leq \pi_+(H)$ for every strictly co-occurrence-supported f and every graph H . For any such scorer with $W(f, G) > 0$,*

$$s(f; N) \leq \frac{q_N}{W(f, G)} = \frac{s_\pi}{q_f}, \quad q_f := \frac{W(f, G)}{\pi_+(G)}. \quad (2)$$

Consequently, for any subclass $\mathcal{F}_0$ satisfying $W(f, G) \geq w_0 > 0$ and nonnegative Hits@K,

$$1 - \frac{q_N}{w_0} \leq R(f) \leq 1, \quad \text{diam}\{R(f) : f \in \mathcal{F}_0\} \leq \frac{q_N}{w_0}. \quad (3)$$

PROOF. If at least K negatives have a common neighbour the threshold exceeds β_0 ; if fewer do, it equals β_0 . Either way a positive with no common neighbour scores β_0 and is not counted. Taking expectations over $H \sim N$ and U gives $W(f, N) \leq q_N$; positivity of $W(f, G)$ also implies $\pi_+(G) > 0$, so dividing gives (2). On $\mathcal{F}_0$, $0 \leq s(f; N) \leq q_N/w_0$, so $R(f) = 1 - s(f; N)$ lies in the interval in Equation (3), whose width is q_N/w_0 . $\square$

The scorer-specific ceiling is s_π/q_f , not s_π alone. Equation (3) is the class-level saturation result: with a prespecified intact-performance floor w_0 , q_N/w_0 bounds how much the relative drop can distinguish within the whole class.

On ogbl-collab, over the 10 retained draws, $\pi_+(\cdot)$ falls from 0.5454 to 0.0207, so $s_\pi = 0.0380$. Across the 5 strictly co-occurrence-supported scorers of Table 1 the ceiling runs from 0.0396 to 0.0497, against measured survival ratios of 0.0010 to 0.0205. The other seven arrangement-reading rows lie outside the class and nothing here constrains them: the table’s largest survival ratio, 0.1048, is the landmark distance upper bound, which separates pairs having no common neighbour. Local co-occurrence is not perturbed here. It is removed.

Using the diagnostic prospectively. q_N needs the intact graph, evaluation positives, and null draws but no model training. Before running an ablation, declare either an intact-performance floor w_0 or the smallest class difference worth resolving, then compare it with q_N/w_0 . The graph-only ratio s_π remains useful triage, but without q_f or w_0 it is not itself a performance ceiling. Alongside the mobility and support checks, the resulting bound assesses the informative range from both sides: the null must deliver the intended perturbation while retaining enough variation to distinguish the class.

Behavior across benchmarks. Table 2 reports s_π on 5 benchmarks, spanning 0.0380 to 1.0002, and the ceiling held on 25 of 25 benchmark-scorer cells across every draw. On the 3 with $s_\pi \leq 0.0684$, the five measured scorers occupy bands at most 1.9 points wide. Retrospectively setting w_0 to each benchmark’s minimum intact score makes the theorem’s widest corresponding class bound 12.9 points; PubMed’s is 59.1 points. These are theorem envelopes, not uncertainty intervals. ogbl-ddi sits at the opposite extreme. At mean degree $\langle d \rangle = 501$ over 4,267 nodes a degree-preserving draw preserves essentially all co-occurrence, $s_\pi = 1.0002$, and the lemma constrains nothing. Cheaper still is $\hat{s}_\pi$, which predicts s_π from the degree sequence with no draw taken. It matches to 1.0% on the 2 benchmarks above 10^4 nodes and

over-states it on the two smallest, erring toward spending a draw (Appendix G).

Invariance to the negative protocol. s_π and $\pi_+(G)$ are functions of the graph and the positives alone. Sampling negatives to share a common neighbour [18] therefore leaves both unchanged to every displayed digit while every scorer’s intact performance falls, and the bound grows looser, not tighter (Appendix G).

5 What the ablation can distinguish

Lemma 3.5 leaves one quantity to measure: the range s occupies, over the scorers a practitioner might use, and then over the benchmarks they might run them on.

5.1 The scorer family

15 scorers are evaluated on ogbl-collab under the three conditions of Section 3.1, in three groups fixed by Definition 3.1. **Arrangement-reading scorers** (12 of them) score a pair from which nodes are linked to which: seven local, among them common neighbours and its inverse-log-degree weighting [1], and five reading beyond the neighbourhood. Every scorer we implemented appears in Table 1, four of them well below the intact performance anyone would propose in practice. Including them makes the class boundary a measured variable instead of a choice of ours, and Section 5.2 reports what removing those four costs every threshold below. **Degree-determined scorers** score a pair from the two endpoint degrees alone (preferential attachment, and a gradient-boosted model on the degree pair as an empirical estimate of how far degree alone can go here), and Lemma 3.7 predicts $s = 1$, $D = 0$ for both. **The trained arm** reads both: GraphSAGE [13] is the only scorer here that sees node content as well as structure, and our replication gives 0.4872 mean test Hits@50 against the published 0.481 ± 0.0081 . The row carrying the sign change is a **modularity-matrix spectral embedding**: the adjacency spectral embedding [3, 29] of Newman’s modularity matrix $B = A - dd^T/2m$ [21], scored by inner product, where A is the adjacency matrix and d the degree vector of G . Its rank is selected on validation within each condition, so no arm carries a rank tuned on another’s graph (Appendix H).

5.2 Selectivity of the relative drop

Table 1 gives $W(f, G)$, s , the relative drop and advantage for every scorer. Across the 12 arrangement-reading scorers s runs from 0.0010 to 0.1048, so the relative drop runs from 89.52% to 99.90%, a band 10.38 points wide (Figure 2). Over the same scorers intact performance spans 33-fold from 0.0167 to 0.5522, and advantage runs from -0.3147 to $+0.2208$, 36 resolution scales, with 5 of the 12 below zero.

Definition 3.9 turns that into a threshold. Over this class $\tau^* = 99.25\%$, attained by Louvain block affinity at advantage -0.3033 . Any lower threshold selects at least one scorer with nonpositive advantage. At τ^* , the yield is 3 of 12; among the 7 scorers with positive advantage, 4 fall below the threshold.

The threshold belongs to the class, not to the contrast. A supremum over a class moves when the class does. The value above is therefore portable only as far as the class it was computed on, and Louvain block affinity, which attains it, lies among the weakest scorers by

Table 1: Every scorer we evaluated on ogbl-collab, Hits@50, intact and under the retained degree-preserving null draws. s is the share of performance the null leaves standing, $R = D/W(f, G) = 1 - s$ the relative drop a paper reports, A against the content-only comparator, which scores 0.3314. *Inclusion rule: every scorer we implemented, nothing dropped after being computed. No intervals; reconciliation, and what removing the four weakest rows does to every threshold: Appendix F.*

Scorer	$W(f, G)$	$s(f)$	$R(f)$	A
<i>Arrangement-reading</i>				
Length-3 paths, normalised	0.5522	0.0662	93.38%	+0.2208
Length-3 paths	0.5328	0.0798	92.02%	+0.2014
Adamic-Adar	0.5240	0.0205	97.95%	+0.1926
Resource allocation	0.5235	0.0200	98.00%	+0.1921
Salton cosine	0.4888	0.0010	99.90%	+0.1574
Jaccard	0.4869	0.0010	99.90%	+0.1555
Common neighbours	0.4170	0.0052	99.48%	+0.0856
Modularity-matrix spectral	0.2783	0.0830	91.70%	-0.0531
Landmark sketch	0.0788	0.0090	99.10%	-0.2526
Landmark distance upper bound	0.0541	0.1048	89.52%	-0.2773
Louvain block affinity	0.0281	0.0075	99.25%	-0.3033
Structural role	0.0167	0.0137	98.63%	-0.3147
<i>Degree-determined</i>				
Gradient-boosted degree pair (oracle)	0.1245	1.0000	0.00%	-0.2069
Preferential attachment	0.0804	1.0000	0.00%	-0.2510
<i>Arrangement-reading, and reads node content</i>				
GraphSAGE	0.4737	0.6575	34.25%	+0.1423

Table 2: Public link-prediction benchmarks ordered by co-occurrence survival s_π , which $\hat{s}_\pi$ predicts from the degree sequence with no null drawn; n is the node count and $\langle d \rangle$ the mean degree. ogbl-ppa carries $\hat{s}_\pi$ only, so the measured columns cover 5. AA is the relative drop on Adamic-Adar, *band* its width across the 5 bounded scorers (a narrow band means the ablation returns nearly one number whatever scorer it is run on), and ρ the drop’s rank agreement with advantage over those same scorers, which changes sign across the rows. Read Cora against CiteSeer, matched on metric, K and negative protocol. An absent A is undefined, not unmeasured: † no node features, ‡ one-hot species labels. Per-scorer values: Appendix G.

Benchmark	n	$\langle d \rangle$	co-occurrence		relative drop R			
			$\hat{s}_\pi$	s_π	AA %	band	ρ	A
ogbl-collab	235,868	8.2	0.038	0.038	98.0	1.9	-0.60	0.193
CiteSeer	3,327	2.3	0.053	0.044	94.5	0.9	+0.90	0.087
Cora	2,708	3.3	0.098	0.068	94.5	1.8	-0.05	0.201
PubMed	19,717	3.8	0.095	0.095	73.7	24.8	+0.70	0.041
ogbl-ddi	4,267	500.5	1.000	1.000	79.5	41.0	-	$n/a^\dagger$
ogbl-ppa	576,289	73.7	0.196	-	-	-	-	$n/a^\ddagger$

intact performance. That makes the class boundary the parameter to vary. Raising an inclusion floor on $W(f, G)$, the one axis available without consulting advantage or the null, drops τ^* to 91.70% once the 4 weakest rows go. That is a 7.55-point swing across a 10.38-point band caused solely by changing the class boundary; the two thresholds are not transferable (Appendix F). Restricted to the 7

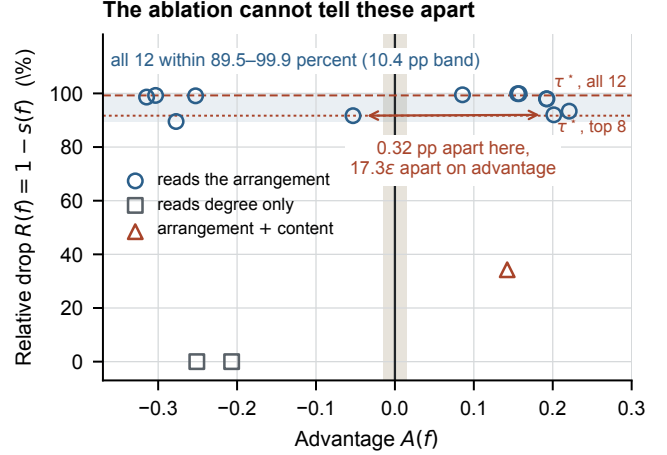


Figure 2: Table 1 plotted, marker shape for what the scorer reads. The vertical axis is what a paper reports when it says performance fell by some percentage: all 12 arrangement-reading scorers fall inside a 10.38-point band (shaded) while their advantages cross zero. Vertical shading is $\pm\epsilon = 0.0147$; the marked pair is 0.32 points apart vertically and 17.3ϵ apart horizontally. The two rust lines are one selectivity threshold over two classes (99.25% over all 12, 91.70% over the 8 that outscore degree alone), differing in which scorers are counted and in no measurement. The band is a property of this enumerated class, not an estimate for a population of scorers.

with positive advantage the contrast has nothing left to reject, so a full yield there is automatic. What can still be asked is whether the drop *orders* that class as advantage does. It nearly reverses it, ranking 17 of 21 pairs oppositely ($\rho = -0.82$). Narrowing the class to the scorers worth using leaves the drop less to distinguish, not more.

The absolute drop behaves as Lemma 3.6 says it must (Figure 4): scorer by scorer $|D - (A + W(c, \emptyset))| = sW$, at most 0.042524 or 2.89ϵ over the 12 across a 33-fold span of W . The drop is therefore advantage plus the constant $W(c, \emptyset)$ to within the benchmark’s own resolution, and that constant is measured without randomizing anything. With $W(c, \emptyset)$ constant, $\text{Pearson}(D, A) = \text{Pearson}(D, W(f, G))$ identically, both 0.9982 here, so correlating the ablation drop with the value of the graph is correlating it with the unablated score.

Degree-based consistency check. Both degree-determined scorers report $s = 1.0000$ and $D = 0$ to every printed digit, as Lemma 3.7 requires. A departure would reveal a violation of degree preservation or another part of the evaluation pipeline, but agreement cannot detect every possible invariant violation. Direct validation of each draw establishes constraint preservation for the sampled distribution.

5.3 Advantages of the retrained scorers

The three scorers for which an arm was retrained (Appendix I) all depend strongly on the arrangement: Adamic-Adar loses 97.95% of

its intact Hits@50, the spectral scorer 91.70%, GraphSAGE 34.25%, every interval excluding zero in all 10 replicates. An ablation-only report would call all three evidence of recoverable signal. Two are correctly described that way and the same protocol says so: GraphSAGE’s advantage is 0.1423 (CI [0.1309, 0.1537]) and Adamic–Adar’s 0.1926 (CI [0.1838, 0.2015]), lower bounds exceeding $\varepsilon = 0.0147$ by roughly an order of magnitude and stable when ε is halved or doubled (Appendices K and L). Under this protocol, GraphSAGE and Adamic–Adar outperform the named comparator; the spectral scorer does not. Attribution specifically to arrangement information remains limited by comparator match.

By contrast, the spectral scorer has advantage -0.0531 (CI $[-0.0620, -0.0443]$, 0/10 positive). We retain this row because it is an ordinary scorer in the evaluated class, not a specially constructed counterexample. Being untrained is not the explanation: 7 untrained scorers clear the same comparator and 6 outrank the trained graph arm. Two readings of the sign stay open, so the claim is comparator-relative in the way Definition 3.3 states (Appendix J).

5.4 The same two contrasts, across benchmarks

Table 2 varies the benchmark instead of the scorer, over the 6 non-learned scorers definable everywhere, with the comparator retrained per benchmark at its own K .

The dissociation survives the change of axis. Cora and CiteSeer lose 94.47% and 94.49% of Hits@20 to the same null, 0.02 percentage points apart, closer than any two scorers in Table 1, while the arrangement is worth 0.2009 on the first and 0.0870 on the second, a factor of 2.3. Of the 6 benchmark pairs 2 are ranked oppositely by the two contrasts; we report the count and not a correlation, which 4 points do not support.

Inside each benchmark the ordering question has an answer and a different sign each time. Over the 5 scorers defined everywhere, the drop’s rank agreement with advantage runs from $+0.90$ on CiteSeer to -0.60 on ogbl-collab (ρ , Table 2), negative on 2 of 4. An ordering calibrated on one of these graphs inverts on another, so the reversal of Section 5.2 is not a property of ogbl-collab. What does not survive the change of benchmark is the sign. The selectivity threshold is zero on all 4, with yield 5 of 6, and this too owes nothing to the measurement. The one scorer lacking positive advantage is preferential attachment, whose drop is zero by Lemma 3.7: the threshold separates exactly what Definition 3.1 separates by reading definitions.

ogbl-ddi separates the two contrasts at the benchmark level. Its dependence is measurable and large; a nontrivial content-matched advantage is undefined rather than small because the benchmark provides no node features for such an arm. The ablation still reports a number where that particular comparator question cannot be posed.

6 Discussion

Interpreting the two contrasts. Advantage and dependence answer different questions. Advantage compares two information conditions: a scorer with access to the graph and a named graph-free baseline. Dependence holds the scorer fixed and changes the edge arrangement through a sampling distribution. A large dependence estimate can therefore be scientifically useful—it shows that

a procedure reads the observed arrangement—without establishing that the procedure improves on information available without the graph.

Reporting a graph-effect claim. A complete report includes:

- (1) the intact score $W(f, G)$;
- (2) advantage $A(f; c) = W(f, G) - W(c, \emptyset)$ over a named graph-free comparator with the same node content, together with any differences in capacity or optimization;
- (3) survival $s(f) = W(f, N_1)/W(f, G)$, or equivalently the relative drop $R(f) = 1 - s(f)$, with the null distribution specified; and
- (4) checks that the sampler preserves the stated constraints and reaches an adequate perturbation. For co-occurrence-supported scorers, q_N/w_0 provides a prospective bound once the intact-score floor w_0 is declared.

The first two quantities address improvement over graph-free information; the third measures reliance on the sampled edge arrangement; the fourth determines whether the implemented null supports that interpretation.

A discriminating case. A scorer with high advantage and low dependence would show that the two contrasts can agree on a useful separation. None of the evaluated scorers occupies this region. Survival under the degree-preserving null requires using information that the null retains, of which degree is the most substantial here. A gradient-boosted model on the degree pair alone reaches 0.1245 against the comparator’s 0.3314, placing that region near -0.2069 advantage on ogbl-collab. Among the additional benchmarks, PubMed—with a 24.8-point drop band and $\rho = +0.70$ —is the clearest candidate for a learned follow-up.

What the evidence establishes. The analysis supports three claims. First, a large relative drop can coexist with negative advantage; the modularity-matrix scorer provides such a case. Second, relative drop does not preserve the ordering of advantage within the enumerated scorer class, and the ordering also changes across benchmarks (Sections 5.2 and 5.4). Third, for strictly co-occurrence-supported scorers, Lemma 4.8 bounds null performance scorer by scorer: the bound holds in 25 of 25 benchmark–scorer cells across all draws, while q_N/w_0 extends it to a declared class. These statements concern existence, ordering, and an analytical bound. They do not estimate how prevalent the mismatch is among graph-learning methods in general.

Scope and comparability. The trained models are evaluated only on ogbl-collab; the cross-benchmark comparison uses non-learned scorers. We therefore do not extrapolate the measured ordering to learned models on the other benchmarks. Each benchmark retains its published metric, K , and negative-sampling protocol (Appendix G), which preserves within-benchmark comparability but makes cross-benchmark magnitudes descriptive rather than commensurate. The content-only comparator c is matched on node content and prediction target but not on capacity or optimization (Section 3.1). Its positive and negative differences are thus comparator-relative; parameter count alone does not determine the direction of any remaining mismatch.

Uncertainty. The 10 replicates concentrate precision on the primary contrasts, whose effects exceed ε by an order of magnitude (Section 5.3). Enumerating the scorer class removes uncertainty about which scorers were included, but finite null draws and stochastic training still affect the estimated scores and selectivity threshold. Three secondary contrasts have half-widths above the prespecified 95% target 0.00735 and should not be interpreted at that resolution. The co-occurrence diagnostic is a property of the graph, positive evaluation pairs, and specified null distribution; its estimate nevertheless retains finite-draw uncertainty.

7 Conclusion

Edge randomization measures dependence on an observed arrangement, not improvement over graph-free information. On ogbl-collab, the modularity-matrix spectral scorer loses 91.70% under the degree-preserving null while remaining below the content-only comparator, whereas GraphSAGE has positive advantage under the same evaluation protocol. The discrepancy persists across a broader scorer class and changes across benchmarks. Treating the null as a distribution over a constrained graph space further shows why invariant preservation, sampler support, and perturbation strength must be reported alongside the drop. Together, the intact score, named graph-free baseline, survival ratio, and sampler diagnostics separate what a model uses from whether that information improves prediction.

Acknowledgments

I thank Bryan Tam, founder and CEO of Third Wheel, for the curated corpus of professional profiles behind the testbed of Appendix P, and for bringing me onto the team.

References

- [1] Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social Networks* 25 (2003), 211–230. Issue 3. doi:10.1016/s0378-8733(03)00009-1
- [2] Farahnaz Akrami, Mohammed Samiul Saeef, Qingheng Zhang, Wei Hu, and Chengkai Li. 2020. Realistic re-evaluation of knowledge graph completion methods: An experimental study. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2020-06-11). ACM, 1995–2010. doi:10.1145/3318464.3380599
- [3] Avanti Athreya, Donniell E Fishkind, Keith Levin, Vince Lyzinski, Youngser Park, Yichen Qin, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. 2017. Statistical inference on random dot product graphs: a survey. *arXiv [stat.ME]* (2017), 1–92. Issue 226. doi:10.48550/arXiv.1709.05454
- [4] Maya Bechler-Speicher, Ido Amos, Ran Gilad-Bachrach, and Amir Globerson. 2023. Graph Neural Networks use graphs when they shouldn't. *arXiv [cs.LG]* (2023), 3284–3304. doi:10.48550/arXiv.2309.04332
- [5] Maya Bechler-Speicher, Ben Finkelshtein, Fabrizio Frasca, Luis Müller, Jan Tönshoff, Antoine Siraudin, Viktor Zaverkin, Michael M Bronstein, Mathias Niepert, Bryan Perozzi, Mikhail Galkin, and Christopher Morris. 2025. Position: Graph learning will lose relevance due to poor benchmarks. *arXiv [cs.LG]* (2025). doi:10.48550/arXiv.2502.14546
- [6] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitry Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. 2021. Accounting for variance in machine learning benchmarks. *arXiv [cs.LG]* (2021). doi:10.48550/arXiv.2103.03098
- [7] Corinna Coupette, Jeremy Wayland, Emily Simons, and Bastian Rieck. 2025. No metric to rule them all: Toward principled evaluations of graph-learning datasets. *arXiv [cs.LG]* (2025), 11405–11434. doi:10.48550/arXiv.2502.02379
- [8] Kutalmış Coşkun, Ivo Kavisanzcki, Amin Mirzaei, Tom Siegl, Bjarne C Hiller, Stefan Lüdtkke, and Martin Becker. 2025. Informed, but not always improved: Challenging the benefit of background knowledge in GNNs. *arXiv [cs.LG]* (2025). doi:10.48550/arXiv.2505.11023
- [9] Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. 2019. A fair comparison of graph neural networks for graph classification. *arXiv [cs.LG]* (2019). doi:10.48550/arXiv.1912.09893
- [10] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches: a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems* (New York, NY, USA, 2019-09-10). ACM, 101–109. doi:10.1145/3298689.3347058
- [11] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20 (2019), 1–81. Issue 177. https://jmlr.org/papers/v20/18-760.html
- [12] Bailey K Fosdick, Daniel B Larremore, Joel Nishimura, and Johan Ugander. 2018. Configuring random graph models with fixed degree sequences. *SIAM Review. Society for Industrial and Applied Mathematics* 60 (2018), 315–355. Issue 2. doi:10.1137/16m1087175
- [13] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *arXiv [cs.LG]* (2017). doi:10.48550/arXiv.1706.02216
- [14] John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Stroudsburg, PA, USA, 2019). Association for Computational Linguistics, 2733–2743. doi:10.18653/v1/d19-1275
- [15] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for machine learning on graphs. *arXiv [cs.LG]* (2020), 22118–22133. doi:10.48550/arXiv.2005.00687
- [16] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021-03-03). ACM, 375–385. doi:10.1145/3442188.3445901
- [17] Isay Katsman, Ethan Lou, and Anna Gilbert. 2024. Revisiting the necessity of graph learning and common graph benchmarks. *arXiv [cs.LG]* (2024). doi:10.48550/arXiv.2412.06173
- [18] Juanhui Li, Harry Shomer, Haitao Mao, Shenglai Zeng, Yao Ma, Neil Shah, Jiliang Tang, and Dawei Yin. 2023. Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking. In *Advances in Neural Information Processing Systems 36* (San Diego, California, USA, 2023). Neural Information Processing Systems Foundation, Inc. (NeurIPS), 3853–3866. doi:10.52202/075280-0169
- [19] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58 (2007), 1019–1031. Issue 7. doi:10.1002/asi.20591
- [20] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. 2010. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 21, 3 (2010), 383–388. doi:10.1097/EDE.0b013e3181d61eeb
- [21] M E J Newman. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics* 74 (2006), 036104. Issue 3 Pt 2. doi:10.1103/PhysRevE.74.036104
- [22] M E J Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America* 103 (2006), 8577–8582. Issue 23. doi:10.1073/pnas.0601602103
- [23] Chiara Orsini, Marija M Dankulov, Pol Colomer-de Simón, Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Kevin E Bassler, Zoltán Toroczka, Marián Boguñá, Guido Caldarelli, Santo Fortunato, and Dmitri Krioukov. 2015. Quantifying randomness in real networks. *Nature Communications* 6 (2015), 8627. Issue 1. doi:10.1038/ncomms9627
- [24] John Palowitch, Anton Tsitsulin, Brandon Mayer, and Bryan Perozzi. 2022. GraphWorld: Fake graphs bring real insights for GNNs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2022-08-14). ACM, 3691–3701. doi:10.1145/3534678.3539203
- [25] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural collaborative filtering vs. Matrix factorization revisited. In *Fourteenth ACM Conference on Recommender Systems* (New York, NY, USA, 2020-09-22). ACM, 240–248. doi:10.1145/3383313.3412488
- [26] Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E Priebe. 2022. A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 84 (2022), 1446–1473. Issue 4. doi:10.1111/rssb.12509
- [27] Oleksandr Shchur, Maximilian Mummé, Aleksandar Bojchevski, and Stephan Günnemann. 2018. Pitfalls of graph neural network evaluation. *arXiv [cs.LG]* (2018). doi:10.48550/arXiv.1811.05868
- [28] Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang. 2020. A re-evaluation of knowledge graph completion methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA, USA, 2020). Association for Computational Linguistics, 5516–5522. doi:10.18653/v1/2020.acl-main.489
- [29] Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. 2012. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* 107 (2012), 1119–1128. Issue 499. doi:10.1080/01621459.2012.699795
- [30] R Taylor. 1981. *Constrained switchings in graphs*. Lecture Notes in Mathematics, Vol. 884. Springer Berlin Heidelberg, 314–336. doi:10.1007/bfb0091828
- [31] Louis Van Langendonck, Guillermo Bernárdez, Nina Miolane, and Pere Barlet-Ros. 2025. GraphUniverse: Synthetic graph generation for evaluating inductive generalization. *arXiv [cs.LG]* (2025). doi:10.48550/arXiv.2509.21097
- [32] Fabien Viger and Matthieu Latapy. 2005. *Efficient and simple generation of random simple connected graphs with prescribed degree sequence*. Lecture Notes in Computer Science, Vol. 3595. Springer Berlin Heidelberg, 440–449. doi:10.1007/11533719_45
- [33] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza, Jr, Christopher Fifty, Tao Yu, and Kilian Q Weinberger. 2019. Simplifying Graph Convolutional Networks. *arXiv [cs.LG]* (2019), 6861–6871. doi:10.48550/arXiv.1902.07153

Guide to the appendices

Core notation is collected in Table 3; symbols local to one definition or proof are defined at first use.

Units and provenance

- A the replicate unit; what no interval covers
- R what every number is traceable to

The algebra, and the null’s validity theory

- B Theorem 3.10’s limiting form
- C the retention floor as a corollary
- D specification against mechanism, drawn
- E three samplers; which one is rejected

The ogbl-collab measurements

- F how the scorer family was assembled
- G every cell behind the cross-benchmark table
- H the eigenpair restriction is not the cause
- I per-replicate intervals, three retrained arms
- J interpretations not supported by the measurements
- K stability across a fourfold range of ε
- L why the heuristic’s numbers can be relied on

Design, variance, and the controlled testbed

- M training conditions, seeds, preregistration
- N what each interval resamples
- O three failures and one unresolved design-power diagnostic
- P the controlled testbed and its access boundary
- Q confirmatory topology contrasts

A Units, intervals, and what is not represented

On ogbl-collab the replicate unit is nested: u_k pairs training seed k with one N_1 draw and one N_2 draw, giving $n = 10$ replicates over seeds 21–30. Null draws within a replicate are not independent observations. Intervals are paired Student- t at $df = n - 1$, two-sided 95%.

Which sources of variation enter an interval differs by contrast, and one case would otherwise look like a defect: both non-learned scorers are deterministic on the intact graph, so their advantage contrasts inherit their entire replicate variance from the comparator arm, and their dependence contrasts vary only through the null draw. Narrow intervals there reflect a low-variance object rather than extra evidence. Appendix N gives the decomposition per contrast.

On the controlled testbed the metric is NDCG@8 and the unit is the generator world. The confirmatory experiment aggregates the mean case-level paired difference within a world and then resamples worlds (30 worlds, percentile bootstrap). The multi-arm comparison resamples generator seeds and cases within seeds (20 seeds, hierarchical bootstrap reported as primary because it is the only interval that propagates both levels).

Not represented in any interval anywhere in this work: split variation, which is held at a fixed seed throughout; hyperparameter variation; and, for the visible-bridge experiment, generator variation, whose intervals are case-level at five fixed generator seeds. Seed variation alone understates benchmark variance [6], so these intervals are narrower than a statement about the benchmark would warrant and we read them only as statements about this split.

B A conditional non-identifiability result

Theorem 3.10 constructs the exact monotone upper envelope and reports what an inverted pair costs it. Dropping monotonicity requires the stronger, explicitly conditional hypothesis below. The empirical table does not verify that hypothesis.

Definition B.1 (Variation independence). Two functionals on a class are *variation independent* over it if their joint range is the product of their individual ranges: every combination of separately attainable values is jointly attainable. The condition concerns attainable values and not probability.

PROPOSITION B.2. *Let $\mathcal{F}$ be a class of scorers with $W(f, G) \in (0, 1]$ and $s(f) \in [0, 1]$ on which s and $W(\cdot, G)$ are variation independent, and let $\mathcal{A} = \{A(f; c) : f \in \mathcal{F}\}$. Then (i) R and $A(\cdot; c)$ are variation independent over $\mathcal{F}$; (ii) any φ with $A(f; c) \leq \varphi(R(f))$ throughout satisfies $\varphi(r) \geq \sup \mathcal{A}$ at every attainable r , and symmetrically for lower bounds, so the sharpest bound on advantage expressible as a function of the relative drop is constant; and (iii) if $\inf \mathcal{A} < 0 < \sup \mathcal{A}$ then $\tau^*(R, A(\cdot; c) > 0, \mathcal{F}) = \sup R(\mathcal{F})$ and its yield is zero.*

PROOF. Advantage is strictly increasing affine in $W(f, G)$ and R strictly decreasing affine in $s(f)$, so $(R, A(\cdot; c))$ is the image of $(s, W(\cdot, G))$ under a coordinatewise strictly monotone bijection, which preserves variation independence: that is (i), and it restates Lemma 3.8. For (ii), fix an attainable r ; by (i) the advantages attained where $R = r$ are all of $\mathcal{A}$, so $\varphi(r) \geq \sup \mathcal{A}$ irrespective of r . For (iii), $\inf \mathcal{A} < 0$ supplies at each attainable r some scorer with $A(f; c) \leq 0$, so the supremum runs over all of $R(\mathcal{F})$, and no member exceeds it. $\square$

Two remarks on the hypothesis, since it is the part of the statement a reader is entitled to test. It is not a condition on cardinality. Four scorers whose $(s, W(f, G))$ values are $\{s_1, s_2\} \times \{w_1, w_2\}$ have exactly that product as their joint range, satisfy Definition B.1, and make the sharpest bound constant at $w_2 - W(c, \emptyset)$; what the condition asks is that the survival ratio not determine the intact score. Table 1 contains a local pair with that behavior (Jaccard and Salton cosine share survival 0.0010 but have intact scores 0.4869 and 0.4888); this does not establish the full product-range condition. The body theorem avoids that assumption. Second, confining s to $[0, 1]$ is what leaves the product unobstructed: $s(f)W(f, G) = W(f, N_1) \leq 1$ would otherwise tie the coordinates together near the top of the range. No draw in this study helps a scorer, so the restriction costs nothing here.

C The immovable-edge floor is a corollary

Let the move be the double-edge swap: from edges $\{p_1, p_2\}$ and $\{q_1, q_2\}$, propose $\{p_1, q_1\}$ and $\{p_2, q_2\}$, rejecting on a self-loop or if either proposed edge already exists. Under N_2 both edges must lie in one component of the original graph and the component partition must be preserved. Call a component *rigid* if no accepted swap has both its edges inside it.

Taylor shows that any two connected labeled simple realizations of a fixed labeled degree vector are connected by standard double-edge swaps whose intermediates remain connected [30]. Unique labeled realizability is therefore a sufficient condition for rigidity: stars, triangles, and complete components are immediate examples.

Table 3: Core symbols, in dependency order. *Type* separates the observed graph from the sampling distributions that ablate it, the conditions a scorer is evaluated under, and the two estimands. The estimands differ only in what they subtract.

Symbol	Type	Meaning
G, H	graph	G is the intact observed graph, simple and undirected, on node set V with node content X held fixed across every condition; H denotes a generic graph on the same node set
A, d, m	structure	the adjacency matrix of G , its degree vector, and its edge count $m = E = 967,632$ unique simple training edges. A is bold to keep it apart from advantage A
N	distribution	a generic null ensemble; $H \sim N$ denotes a graph draw
N_1	distribution	P_{S_1, b_1} : the distribution induced by the degree-preserving sampler at its fixed budget; its support preserves the labeled degree vector, unique-simple-edge count, and evaluation-pair constraints
N_2	distribution	P_{S_2, b_2} : the distribution induced by the component-preserving sampler; its support also preserves G 's full component-label partition
$\emptyset$	condition	the same nodes and node content with no edges
$f, \mathcal{F}$	procedure	a scoring procedure, including condition-specific fitting, selection, and internal randomness U ; $\mathcal{F}$ is a class of procedures
$W(f, \cdot)$	scalar	protocol-level mean performance; $W(f, H) = \mathbb{E}_U M(f; H, U)$ and $W(f, N) = \mathbb{E}_{H \sim N, U} M(f; H, U)$. Hats denote finite plug-in estimates
c	arm	the content-only comparator: same features, objective, pools and prediction head; no message passing
$D(f)$	estimand	dependence, $W(f, G) - W(f, N_1)$; written $D(f; N)$ where the ensemble is at issue
$A(f)$	estimand	advantage, $W(f, G) - W(c, \emptyset)$; written $A(f; c)$ where the comparator is at issue
$s(f), s(f; N)$	ratio	survival ratio, $W(f, N)/W(f, G)$; $s(f)$ fixes $N = N_1$ and is the one quantity an ablation adds. Written s rather than ρ because this paper prints rank correlations. Nothing forbids $s > 1$, a scorer doing <i>better</i> on a randomized graph, but no draw here produces one
$R(f)$	ratio	relative drop, $D(f)/W(f, G) = 1 - s(f)$ (Lem. 3.5); the number a graph ablation reports
$\pi_+(H), q_N$	ratios	share of evaluation <i>positives</i> with a common neighbour in H , and its expectation $q_N = \mathbb{E}_{H \sim N} \pi_+(H)$; $\pi_+(G) = 0.5454$
q_f, w_0	scalars	intact co-occurrence efficiency $q_f = W(f, G)/\pi_+(G)$ and a prespecified lower bound $W(f, G) \geq w_0 > 0$ for a scorer class
$s_\pi, \hat{s}_\pi$	ratios	co-occurrence survival, $q_{N_1}/\pi_+(G) = 0.0380$; the scorer-specific ceiling is s_π/q_f , and q_N/w_0 bounds a class (Lem. 4.8). $\hat{s}_\pi$ predicts the graph ratio before a draw
τ^*	threshold	selectivity threshold of a contrast for a property over a class (Def. 3.9)
$\mathcal{G}_I$	graph space	graphs admitted by the stated invariants; the space alone carries no probability distribution
$S, S_{\text{forest}}, S_{\text{leaf}}$	samplers	a generic sampling algorithm, the retained protected-forest sampler, and its leaf-exchange diagnostic variant
$P_{S, b}, \text{supp}(P_{S, b})$	distribution, set	distribution induced by sampler S at budget b and its support; $\mathcal{G}_S^\infty$ is the union of these supports over budgets
F_S	edge set	edges no budget of S moves; F_I those forced by the graph constraints
ε	constant	benchmark-relative resolution scale, 0.0147; not an application-level minimum useful effect

We use only this sufficient direction and do not enumerate the full component-preserving realization graph.

Our enumeration identifies 3,021 edges as a constraint-forced lower bound. The unique-realizability argument explains why a nonzero floor is expected in a co-authorship graph whose small components often arise as paper cliques; it does not upgrade that enumerated lower bound to a complete characterization. The asymmetry between the rungs is still exact: N_1 permits cross-component swaps, so a locally rigid edge can move by pairing with an edge elsewhere, whereas N_2 forbids that route. The much larger 21,562-edge protected forest belongs to S_{forest} 's mechanism, not to the target.

D The specification and the mechanism

The figure makes explicit the containment relation between the constrained graph space, the sampling distribution's support, and the fixed-edge set. Definitions 4.1 and 4.2 and the conditions in Section 4 do not depend on the drawing.

E Sampler variants

Three samplers were compared at 10 attempted swaps per edge, averaged over three diagnostic draws whose seeds are disjoint from every inferential draw. The degree-preserving sampler retains 0.000693 of the original adjacency and does not preserve the component partition, which it is not asked to. The supplied partition-preserving sampler S_{forest} retains 0.029143, preserves the partition on every draw, and moves none of the 21,562 edges in its protected forest. The leaf-exchange variant S_{leaf} retains 0.6597% and moves 7.1-fold more of that protected mass.

We evaluate support adequacy with a perturbation-matching statistic that the two partition-preserving samplers must hold below 0.02: S_{forest} attains 0.028450, above the threshold, and S_{leaf} attains 0.005904, below it. This is the quantitative form of the reading in Section 4, where Lemma 4.4 bounds S_{forest} 's achievable perturbation by its retention floor irrespective of the mixing budget, so no longer chain repairs it. A fourth variant, leaf-reattach, gives an almost identical statistic, but we exclude it from the formal comparison because its implementation is unavailable for reproduction.

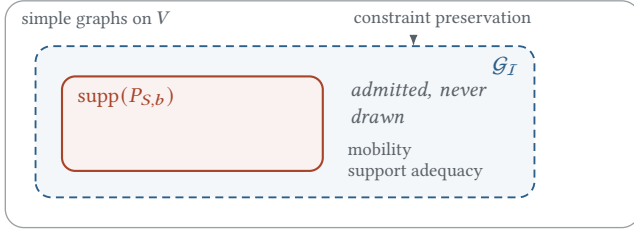


Figure 3: A constrained graph space and an implemented sampling distribution are distinct objects. The dashed region $\mathcal{G}_I$ contains the graphs admitted by the stated invariants; the solid region $\text{supp}(P_{S,b})$ is the support of the distribution produced after budget b . Constraint preservation requires containment. The gap is governed by sampler mobility and support adequacy: an algorithm may leave edges fixed even when the graph constraints do not. Not to scale; area does not encode probability mass.

The both-in-giant diagnostic does not identify a set of Hits@50 units that cannot change, so Proposition 4.6 does not yield a resolution bound for the partition condition. Separately, S_{forest} fails the support-adequacy condition in Definition 4.5.

F Scorer-family construction and sensitivity

Table 1 is the evidence behind Section 5.2. Two provenance details are not visible in the table. First, the recomputed quantities and their scope. The 12 arrangement-reading scorers and the two degree-only scorers were evaluated on the same intact graph and the same retained N_1 draws as the main run, on the same split, but no arm was retrained for them: a non-learned scorer requires no training, and advantage requires only the intact score and the comparator mean that the main run already recorded. This is why they carry no intervals of their own. Adamic-Adar and the modularity-matrix spectral embedding appear in both the main run and this table, computed by two independent implementations on the same objects, and agree on all three of $W(f, G)$, D and A to within 5×10^{-6} , which supports use of the independently recomputed table.

The table is not a sample of scorers. The scorers were chosen to span a wide range of $W(f, G)$ because Lemma 3.5 is a claim about a range; the four weakest rows are included to make that range visible. Neither of the two statements Section 5.2 rests on requires a sampling frame: Lemma 3.6 bounds $|D - (A + W(c, \emptyset))|$ for each scorer separately, and the selectivity threshold τ^* is a supremum over a class that this table enumerates in full. We report the regression as a descriptive summary of this enumerated class, not as a population estimate. The wide span allows us to state whether the survival ratio remains small across the class.

The class boundary, varied. Because τ^* is a supremum over a class, it is a joint property of the contrast and of where the analyst draws the class boundary, which makes the boundary the first parameter to vary. Every nested class obtained by raising an inclusion floor on $W(f, G)$ is in Table 4. The floor is the right axis to vary because it is the only one available before the question is answered: it consults neither advantage, nor the null, nor a threshold, so a

Table 4: τ^* against the class boundary, on the same 12 measurements throughout. Each row keeps every arrangement-reading scorer above an inclusion floor on $W(f, G)$ and is named by the weakest scorer it still contains. *band* is the width in points of the relative-drop interval the class occupies, *yield* the number certified at that class’s own threshold, and ρ the rank agreement between the drop and advantage within it. Rows marked vacuous have no member with non-positive advantage, so $\tau^* = -\infty$ by the empty-supremum convention of Definition 3.9 and the contrast certifies everything without having rejected anything.

$ \mathcal{F} $	weakest member kept	band	τ^*	yield	ρ
12	Structural role	10.4	99.25%	3/12	-0.13
11	Louvain block	10.4	99.25%	3/11	-0.07
10	Landmark dist. UB	10.4	99.10%	3/10	+0.05
9	Landmark sketch	8.2	99.10%	3/9	-0.30
8	Modularity spectral	8.2	91.70%	7/8	-0.21
7	Common nbrs	7.9	$-\infty$	7/7 (vacuous)	-0.82
6	Jaccard	7.9	$-\infty$	6/6 (vacuous)	-0.89
5	Salton	7.9	$-\infty$	5/5 (vacuous)	-0.90
4	Resource alloc.	6.0	$-\infty$	4/4 (vacuous)	-0.80
3	Adamic-Adar	5.9	$-\infty$	3/3 (vacuous)	-0.50

boundary drawn on it cannot have been drawn to produce a desired conclusion. The table has three implications.

The threshold is flat and then steps. Removing the 3 weakest rows moves it by less than a point, from 99.25% to 99.10%; the fourth takes it to 91.70%, 7.55 points below where it started; at $|\mathcal{F}| = 7$ it is $-\infty$. No drop threshold transfers across these class definitions: the curve appears stable until the row that changes it. Neither of the two named values is therefore portable without stating the class. Applied to all 12, the 8-scorer threshold 91.70% admits 10 scorers of which 3 have non-positive advantage; applied to the 8, the 12-scorer threshold 99.25% certifies 3 of the 7 positives it exists to find. The measurements are identical in both directions; only the class changes. The *false* count a reader might look for is zero in every row of the table by construction: Definition 3.9 makes τ^* the supremum over the members lacking the property, so no class can certify one of its own negatives. That is exactly why the threshold, and not an error rate computed inside the class that defined it, is the quantity that has to be reported.

The yield is not evidence where the class has nothing to reject. From $|\mathcal{F}| = 7$ down, every member has positive advantage, τ^* is the supremum of the empty set, and the full yield follows from arithmetic. A full yield on such a class records the composition of the scorer list rather than a property of the contrast.

What survives the restriction is the ordering, and it points the wrong way. Over the 7 scorers with positive advantage the drop ranks 17 of 21 pairs opposite to advantage, $\rho = -0.82$: the largest drop in that class belongs to Salton cosine, 5th of 7 by advantage, while Length-3 paths, normalised is first by advantage and 6th by drop. The coefficient describes this enumerated class and estimates nothing about a population (Appendix J); the pair count is the statement the design supports, and it is the one the body quotes. The identity is checked rather than assumed: $|D - (A + W(c, \emptyset))|$ and $sW(f, G)$ agree to within 10^{-9} on every arrangement-reading row. The largest value of either is 0.042524, attained by Length-3

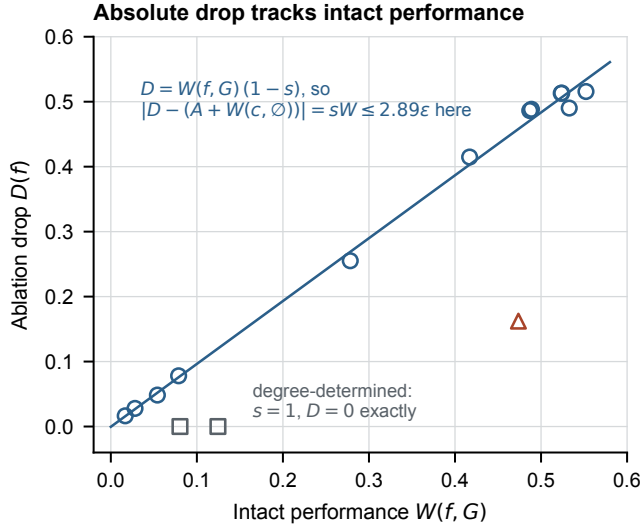


Figure 4: The absolute ablation drop against intact performance, same scorers and markers as Figure 2. By Lemma 3.6 the drop differs from advantage plus a constant by exactly $sW(f, G)$, at most 2.89ϵ here across a 33-fold span, which is why the points track a line, and why no fit is needed to say so. Degree-determined rows sit at $D = 0$ exactly (Lemma 3.7); the learned arm sits off the line because its survival ratio is not small. The line is drawn to make the alignment visible; the bound it illustrates holds for each point separately.

paths. From the two measured ranges alone, without pairing any particular s with any particular $W(f, G)$, the same discrepancy is bounded a priori by Lemma 3.6’s $s_{\max}W_{\max} = 0.0579 = 3.94\epsilon$, weaker than the observed maximum, and available from Table 1 without the per-scorer pairing.

G Cross-benchmark co-occurrence diagnostic

Table 2 provides a different kind of evidence from Table 1. The scorer sweep varies the scorer under one fixed protocol, so its rows are commensurable by construction. The cross-benchmark rows vary the benchmark, and benchmarks differ in more than the graph.

Figure 5 shows the relative-drop range permitted by each benchmark’s co-occurrence bound, with the bounded class’s measured drops inside it. Figure 6 compares the degree-sequence estimate $\hat{s}_\pi$ with the measured s_π ; its error is in the direction that recommends an unnecessary draw.

What differs across the rows. Each benchmark keeps its own published metric and K : Hits@50 on ogbl-collab, Hits@20 on the other four. Negatives follow each benchmark’s own convention: the published fixed set on the OGB graphs, per-positive candidate pools on Cora, CiteSeer and PubMed. Neither choice is ours to make and neither is neutral, which is why the comparison the argument rests on is Cora against CiteSeer: same metric, same K , same negative convention, same generator of node content, and a s_π that differs by less than a factor of two. Rows that cross those lines support ordinal statements and are read no more strongly.

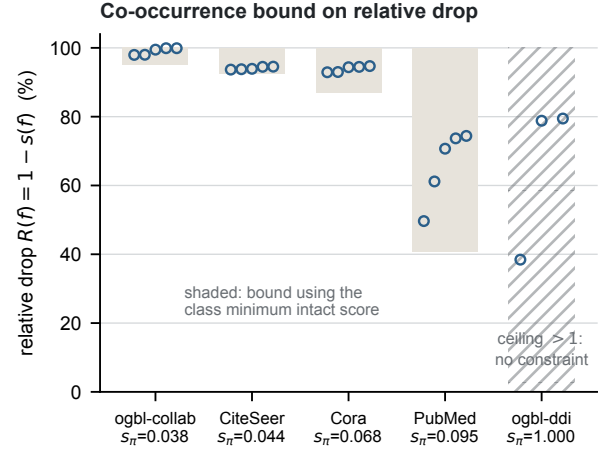


Figure 5: Range permitted by the co-occurrence bound. Each open circle is one of the 5 strictly co-occurrence-supported scorers at its measured relative drop; each shaded region runs from the smallest drop Lemma 4.8 permits *any* member of the observed class up to 100%. The displayed envelopes insert the null co-occurrence mass and each observed class’s minimum intact score, and are therefore retrospective. Benchmarks are ordered by s_π . On the three with $s_\pi \leq 0.0684$ the measured drops span at most 1.9 percentage points, while the corresponding theorem envelopes are at most 12.9 points wide; on ogbl-ddi the ceiling exceeds one and constrains nothing. Every point falls inside its own region, 25/25 across all draws. Envelope width is an analytical range, not evidence of advantage, which is a separate axis and is not shown here.

What does not differ. s_π and $\pi_+(G)$ are functions of the graph, the null and the evaluation positives, so they are invariant to the negative protocol. Re-running ogbl-collab against 500 hard candidates per positive [18] reproduces both to the displayed precision, while every scorer’s $W(f, G)$ and s changes (Section 4.3). This invariance makes s_π suitable as a dataset-level diagnostic.

The ρ column. ρ is the rank agreement between the relative drop and advantage over the 5 bounded scorers on each benchmark, computed on 5 points and reported per benchmark rather than pooled. Its range across the 4 is -0.60 to $+0.90$; the sign is not a constant of the contrast. The bands those coefficients rank inside provide necessary context, 0.9 points at the narrowest and 24.8 at the widest, because a rank computed inside a band a point or two wide is ordering differences the benchmark can barely resolve, which is one reason the sign moves. Whether band width determines the sign is more than 4 benchmarks can support; the diagnostic’s claim (Section 4.3) is about the band and not about the rank.

Per-scorer values. Table 5 gives every cell behind Table 2. Two entries are absent rather than zero: on ogbl-ddi the Jaccard and Salton normalizations score 2 positives above threshold on the intact graph, so s is 0/0 there and no survival ratio exists to report. The degree control is marked; Lemma 4.8 does not cover it. We include it to make the one family member outside the bound explicit.

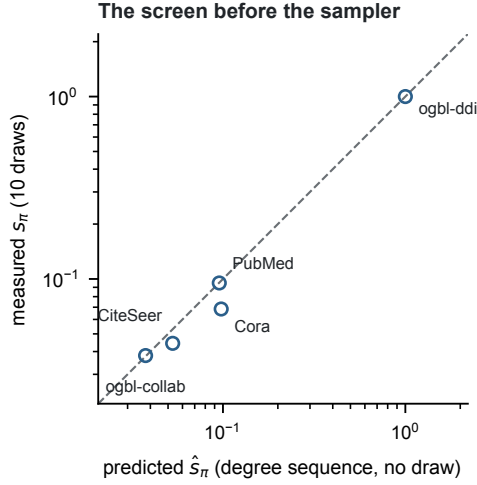


Figure 6: Degree-sequence approximation to the measured diagnostic. $\hat{s}_\pi$ is computed from the degree sequence and the evaluation positives with no null draw; s_π is measured over 10 draws. Both values per benchmark are in Table 2. On the 2 benchmarks with more than 10^4 nodes the two agree to within 1.0%; on the two smallest and sparsest, $\hat{s}_\pi$ over-states survival by up to 1.43 $\times$. Over-stating survival over-states how much room the ablation has, so the error is in the direction that wastes a draw rather than the direction that suppresses an informative experiment, which is why $\hat{s}_\pi$ is offered as a triage step before s_π and not as a replacement for it. Five points: this establishes the sign of the error and its scale on these graphs, not a calibration curve.

H A check on the spectral restriction

Restricting the embedding to positive eigenpairs assumes the latent-position model is positive semi-definite, which the generalised random dot product graph relaxes [26]. Including negative eigenpairs does not change the conclusion for this scorer. At rank 512, twice the selected rank and split evenly between positive and negative eigenpairs, the indefinite embedding scores 0.2609 on the intact graph, 1.18 ϵ below the positive-eigenpair scorer on twice the spectral budget, for an advantage of -0.0705 , more negative than the -0.0531 reported in the body. The negative-eigenpair block alone scores 0.1673, advantage -0.1640 . At a smaller rank the ordering between the two reverses by 0.25ϵ , inside the resolution scale and therefore not a difference we would report in either direction. These are fixed-rank probes rather than scorers selected under the protocol of Section 5, so they answer only whether the restriction is what produces the negative advantage. It is not.

I Per-replicate intervals for the retrained scorers

Table 6 and Figure 7 give the three scorers for which an arm was retrained, and therefore the three that carry per-replicate intervals; the remaining rows of Table 1 are point estimates for the reason given in Appendix F. The values are quoted in Section 5.3.

Table 5: Every cell behind Table 2, at each benchmark’s own K . Ceiling is Lemma 4.8’s bound on s . Ceilings above one are vacuous, which is the ogbl-ddi case. † degree control: outside the bound’s scope by construction, and reported so its exclusion is checkable.

Benchmark	Scorer	$W(f, G)$	$s(f)$	ceiling	A
ogbl-collab	Common nbrs	0.4170	0.0052	0.0497	0.0856
	Adamic-Adar	0.5240	0.0205	0.0396	0.1926
	Resource alloc.	0.5235	0.0200	0.0396	0.1921
	Jaccard	0.4869	0.0010	0.0426	0.1555
	Salton	0.4888	0.0010	0.0424	0.1574
	Pref. attach. †	0.0804	1.0000	0.2581	-0.2510
CiteSeer	Common nbrs	0.2088	0.0621	0.0747	0.0365
	Adamic-Adar	0.2593	0.0551	0.0602	0.0870
	Resource alloc.	0.2615	0.0546	0.0597	0.0892
	Jaccard	0.2418	0.0609	0.0645	0.0695
	Salton	0.2330	0.0632	0.0670	0.0607
	Pref. attach. †	0.0242	1.0000	0.6455	-0.1481
Cora	Common nbrs	0.2258	0.0529	0.1294	0.1213
	Adamic-Adar	0.3055	0.0553	0.0957	0.2009
	Resource alloc.	0.3036	0.0563	0.0963	0.1991
	Jaccard	0.2676	0.0709	0.1092	0.1630
	Salton	0.2676	0.0702	0.1092	0.1630
	Pref. attach. †	0.0190	1.0000	1.5400	-0.0856
PubMed	Common nbrs	0.0589	0.2559	0.4510	0.0162
	Adamic-Adar	0.0837	0.2631	0.3173	0.0410
	Resource alloc.	0.0754	0.2931	0.3524	0.0326
	Jaccard	0.0580	0.3883	0.4580	0.0153
	Salton	0.0449	0.5035	0.5915	0.0022
	Pref. attach. †	0.0196	1.0000	1.3529	-0.0231
ogbl-ddi	Common nbrs	0.1773	0.2118	5.6396	–
	Adamic-Adar	0.1861	0.2055	5.3735	–
	Resource alloc.	0.0623	0.6157	16.0577	–
	Jaccard	0.0000	–	–	–
	Salton	0.0000	–	–	–
	Pref. attach. †	0.0393	1.0000	25.4668	–

J What the measurements do not establish

The spectral scorer is an ordinary combination of two standard constructions, with eleven other scorers in the same ablation band; it is not a constructed counterexample. Its negative advantage is measured against a trained arm, however, while the scorer itself trains nothing. The result is therefore consistent with the scorer reading real arrangement information through a fixed inner product less efficiently than c reads content. Lack of training alone does not explain the result: 7 other untrained scorers clear the same comparator and 6 outrank the trained graph arm. Nor is the result specific to the selected spectral rank: rank is chosen on validation within each condition, and adding negative eigenpairs makes the advantage more negative (Appendix H). A trained readout on the embedding at the comparator’s budget would distinguish the remaining explanations; we did not evaluate one, so the claim remains comparator-relative as Definition 3.3 requires.

Dependence and advantage are correlated here for the algebraic reason in Lemma 3.6. The *absolute* drop has correlation 0.9982 with advantage over the 12 arrangement-reading scorers, but this is

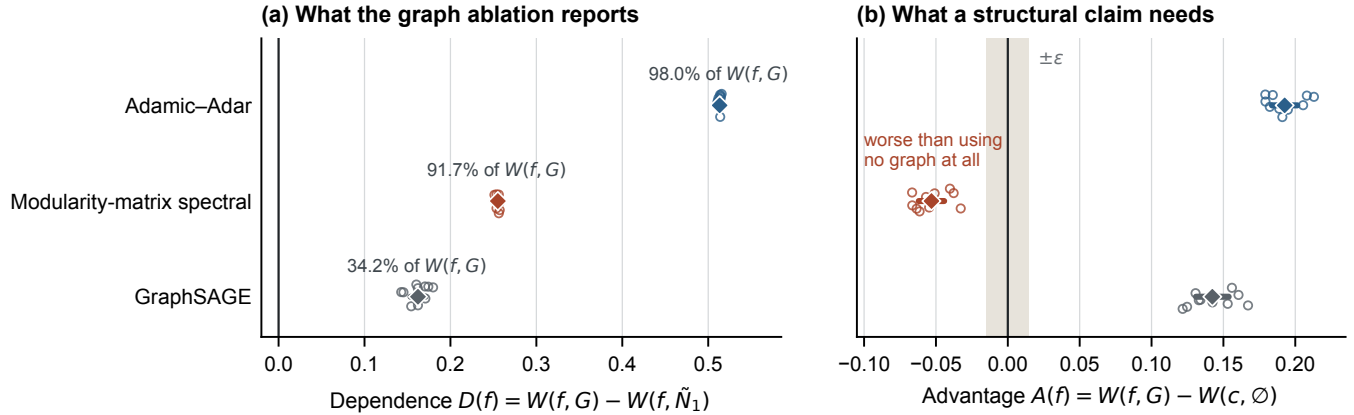


Figure 7: The two contrasts measured, ogbl-collab, Hits@50. Rows are the three scorers of Table 6; colour identifies the scorer and is constant across panels, so no hue changes meaning between them. Open circles are the 10 paired per-replicate differences, the filled diamond their mean, the bar the two-sided 95% paired Student- t interval at $df = 10 - 1$; each circle is one replicate, not one evaluation pair. (a) What an ablation reports: all three scorers are strongly dependent, annotated as a share of intact performance. The two non-learned intervals are narrower than the diamond that covers them because those scorers are deterministic on the intact graph, so only the null draw varies: a low-variance object, not extra evidence. (b) What a structural claim asserts, against a content-only comparator on identical features. The band is $\pm \varepsilon = 0.0147$, a benchmark-relative resolution scale and not an application-level minimum useful effect. The comparator is trained and the spectral scorer is not, so panel (b) shows arrangement information failing to beat trained content, not arrangement information redundant with node content (Section 5.3); and three scorers illustrate the dissociation, not a rule for how D and A order scorers.

Table 6: Dependence and advantage on ogbl-collab, Hits@50, $n = 10$ paired replicates over training seeds 21–30; bracketed rows are two-sided 95% paired Student- t intervals at $df = 10 - 1$. D is what a graph ablation reports, against a degree-preserving null draw; A is measured against a content-only comparator on identical node features, which is content-matched and *not* capacity-matched, with the mismatch favouring the graph arm (Section 3.1). Arm means are in Section 5.3. The non-learned scorers are deterministic on the intact graph, so their A intervals inherit their variance from the comparator arm alone (Appendix A). These are the three scorers for which an arm was retrained and intervals therefore exist; the remaining 15–3 are in Table 1 as point estimates. Every scorer here has large D ; one has negative A ; what the full family supports is in Section 5.2.

Scorer	D (% of $W(f, G)$)	A
Adamic-Adar	+0.5133 (97.95%) [0.5122, 0.5144]	+0.1926 [0.1838, 0.2015]
Modularity-matrix spectral	+0.2552 (91.70%) [0.2538, 0.2565]	−0.0531 [−0.0620, −0.0443]
GraphSAGE	+0.1622 (34.25%) [0.1535, 0.1709]	+0.1423 [0.1309, 0.1537]
Content only (no graph)	—	0

induced by their shared intact score and is not independent evidence. The non-ordering result instead concerns survival, and hence

relative drop. The inverted pair in Table 1 satisfies the empirical condition in Theorem 3.10.

The saturation is a statement about an enumerated class, the 12 scorers listed in full under a stated inclusion rule, and not about a population; why neither result of Section 5.2 needs a sampling frame is set out in Appendix F.

The large dependence estimate does not quantify how much recoverable spectral signal ogbl-collab contains. Its value remains compatible with preserved degree statistics, incomplete null-chain mixing, other preserved structural summaries, and candidate-set composition.

Separately, negative advantage does not establish that the arrangement information is redundant with node content. Here “worse than using no graph” means worse than the specific trained content-only model with 164,608 encoder and 131,841 predictor parameters. Section 5.3 reports the within-table comparisons that narrow this ambiguity, but the present protocol does not distinguish redundancy from a less effective fixed readout. This is the relevant limitation of the measurement.

K Sensitivity to the resolution scale

The resolution scale’s recorded derivation is $0.0147 = 0.1 \times (0.4812 - 0.3342)$, ten percent of a gap between a published reference value and a prior three-seed comparator mean. It is therefore calibrated on the same type of comparator gap evaluated here. It is also close to one replicate standard deviation of the primary contrasts (their mean per-replicate standard deviation is 0.015385, so ε is 0.9555 of it), which is an observation about the value rather than about its provenance. Neither reading is necessary for the conclusions, because Table 7 evaluates each result across a fourfold range of ε

Table 7: Every ogbl-collab verdict recomputed at $0.5\times$, $1\times$ and $2\times$ the frozen resolution scale $\varepsilon = 0.0147$, from the retained per-replicate arrays. A verdict is $> +\varepsilon$ when the interval’s lower bound exceeds $+\varepsilon$, $< -\varepsilon$ when its upper bound is below $-\varepsilon$, within $\pm\varepsilon$ when the whole interval lies inside the band, and unresolved otherwise. Two contrasts change verdict across the range, and they are the two the paper reports as unresolved; no claim made in this paper depends on the value of the constant.

Quantity	$0.5 \times \varepsilon$	ε	$2 \times \varepsilon$
D , GraphSAGE	$> +\varepsilon$	$> +\varepsilon$	$> +\varepsilon$
D , GraphSAGE (N_2 rung)	$> +\varepsilon$	$> +\varepsilon$	$> +\varepsilon$
A , GraphSAGE	$> +\varepsilon$	$> +\varepsilon$	$> +\varepsilon$
D , Adamic-Adar	$> +\varepsilon$	$> +\varepsilon$	$> +\varepsilon$
A , Adamic-Adar	$> +\varepsilon$	$> +\varepsilon$	$> +\varepsilon$
D , spectral	$> +\varepsilon$	$> +\varepsilon$	$> +\varepsilon$
A , spectral	$< -\varepsilon$	$< -\varepsilon$	$< -\varepsilon$
GraphSAGE – Adamic-Adar	$< -\varepsilon$	$< -\varepsilon$	$< -\varepsilon$
randomized – content [†]	$< -\varepsilon$	unresolved	unresolved
rung difference [†]	unresolved	within $\pm\varepsilon$	within $\pm\varepsilon$

[†] verdict is not stable across the range.

and none changes. The sensitivity table allows the threshold to be stated without making the conclusion depend on the calibration constant.

L Precision of the heuristic contrast

Adamic-Adar outscores the trained model: the trained-minus-heuristic contrast is -0.050327 (CI $[-0.0573, -0.0434]$). Prior work also reports strong heuristic performance [18]; here the relevant property is the contrast’s precision. Its half-width is 0.006954 , below the 0.00735 target, which makes it the only non-deterministic contrast in the study meeting the protocol’s prospective precision target. When Adamic-Adar is used in Section 5.3 as the scorer whose dependence and advantage are both large, the credibility of that measurement therefore rests on this study’s best-estimated quantity. We use that precision only for the specific contrast in Section 5.3.

M Training conditions and study provenance

All four arms share one split over train, validation and test positives, one fixed negative set and one set of node features, and all four were trained with identical hyperparameters on all ten seeds: three layers, 256 hidden channels, dropout 0, learning rate 10^{-3} , batch size 65536, 400 epochs with per-epoch validation selection. The arms therefore differ in their information condition and in message passing, not in their optimization. Encoder and predictor parameter counts were read off each trained model, which is why the capacity mismatch of Section 3.1 is a measured gap rather than one inferred from a pilot. Training seeds are 21–30, the extension seeds named in Section 6 are 31–40, and the seeds used for the diagnostic draws of Appendix E are disjoint from every draw entering an interval.

Three limitations qualify the findings above. The study was not preregistered, was partially informed by prior analyses, and its original analysis plan is unavailable. The resolution scale ε was

derived from three earlier training seeds, disjoint from the ten used here, so it is calibrated on this benchmark but not on this sample, which is why every conclusion is evaluated across a fourfold range of it (Appendix K). And the 0.02 perturbation-matching threshold of Appendix E has no documented derivation; no claim in the paper turns on its exact value, only on Lemma 4.4, which bounds S_{forest} ’s perturbation without reference to it.

N What the intervals resample

Each interval in this work resamples a different object. Table 8 identifies that object and the resulting interpretation. Two rows require particular care.

Not represented in any interval anywhere: split variation, held at a fixed seed; hyperparameter variation; and generator variation for the visible-bridge experiment, whose intervals are case-level at five fixed generator seeds. Secondary contrasts are not adjusted for multiplicity, and the primary contrast, primary slice and primary k were declared in the protocols before analysis.

O Failure modes and diagnostics

Table 9 organizes four checks around the intervention, the contrast, and the compared arms. Support adequacy (Definition 4.5) tests whether the sampler can deliver the intended perturbation; the selectivity threshold (Definition 3.9) tests whether the contrast separates the property of interest. Arm comparability is a property of the two protocols themselves: an anchor offered as a ceiling must share the observation contract of the arms it bounds. For the partition condition, the available statistic does not identify the affected evaluation units required by Proposition 4.6, so it provides no resolution bound.

P The controlled testbed

The testbed is a typed co-affiliation graph built from a curated corpus of professional profiles, in which every edge traces to a named profile field and the relevance labels are held apart from graph construction. Two regimes are held fixed: in the *hidden-dominant* regime much of the relevance-generating state is available only to the evaluator, and in the *visible-bridge* regime the relevance rule is materialized in the typed structure and withheld from the text-only view, so recoverability can be checked before model results are interpreted. The profile-level inputs sit under a restricted access boundary. The public benchmark therefore carries the reproducible empirical analysis, while this testbed is reported as a separate controlled case. The testbed informed the diagnostic design; the primary evidence and conclusions come from the public benchmark analysis presented first.

Contract-matched results. In the hidden-dominant regime a linear typed-incidence scorer, matched to the learned family under the same statically enforced observation contract, beats the content-only arm by 0.0292 (hierarchical bootstrap CI $[0.0209, 0.0380]$, 20/20 generator seeds positive; the seed-level interval $[0.0229, 0.0356]$ agrees). Both quantities are positive, so this regime’s signal is reachable under the contract the arms share. The trained model does not reach it: the best of six GraphSAGE configurations scores -0.0184 against that scorer (CI $[-0.0286, -0.0090]$), and all six lose. In the visible-bridge regime the same scorer’s advantage is 0.0000 (CI

Table 8: What each interval resamples. Two rows would otherwise read as defects: the non-learned A contrasts report an identical dispersion because it comes entirely from the comparator arm, and the non-learned D contrasts are narrow because a deterministic scorer on a fixed graph leaves only the null draw to vary. Narrowness here reflects a low-variance object, not extra evidence.

Contrast	Varies across replicates	Consequence
D , GraphSAGE	training seed on both arms and the null draw jointly	confounds training-seed with draw-to-draw variance; not separable at $n = 10$
A , GraphSAGE	training seed on both arms	the cleanest of the three
A , non-learned	the comparator arm only	the scorer is deterministic on the intact graph, so both non-learned A contrasts report an identical per-replicate standard deviation
D , non-learned	the null draw only	no training is involved; these are the only contrasts meeting the precision target, and the small dispersion reflects a low-variance object
rung difference	training seed plus two draws	shares replicate structure with both D contrasts, so its variance is not independent of theirs

Table 9: Diagnostic checks and their implications for the reported contrasts.

Check	Finding and implication
Design sensitivity <i>resolution bound</i>	For the partition condition, the 0.9772 both-in-giant share does not identify which Hits@50 units can change. Proposition 4.6 therefore does not apply; the proposed ceiling 0.0228 also exceeds the achieved half-width 0.0111.
Sampler support <i>perturbation check</i>	S_{forest} fixes 2.2283% of edges, compared with the constraint-forced lower bound 0.3122%, and cannot meet the overlap threshold 0.020693 at any sampling budget. A leaf-exchange sampler shows that the constrained graph space does admit a qualifying perturbation.
Contrast selectivity <i>selectivity</i>	Dependence is 91.70% for a scorer whose advantage is -0.0531 ; selectivity threshold $\tau^* = 99.25\%$ with yield 3/12 (§5.2). Dependence should therefore be reported alongside a graph-free comparison rather than interpreted as advantage.
Comparator alignment <i>comparability</i>	An anchor at 375 warm cells with no hold-out, offered as a ceiling for arms run at 241 cases under an inductive split (§P). The anchor must be recomputed under the same observation and holdout protocol before it can bound those arms.

$[-0.0051, 0.0049]$). A contract-matched scorer reports reachable signal in one regime and none in the other; the second null retains enough support to be informative under the stated protocol (Figure 8).

The withdrawn anchor, in full. Appendix O states the comparability failure. The numbers behind it: an evaluator oracle reaches 0.9915 NDCG@8 in the visible-bridge regime and a structural anchor that reads each candidate’s own target-bearing edge reaches 0.7053, while the text ladder sits at chance; the learned arms scored 0.0195. The anchor was computed under a warm regime with no holdout over 375 evaluation cells, and the learned arms ran under

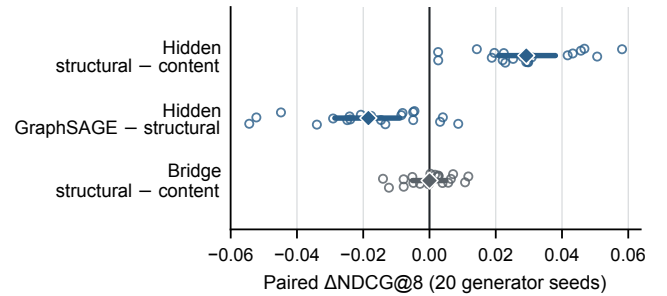


Figure 8: Controlled testbed, NDCG@8, unit = generator seed (20 seeds). Open circles are per-seed paired differences, the filled diamond their mean, the bar a 95% hierarchical bootstrap interval resampling seeds and cases within seeds; colour repeats the regime already named in each row label and carries nothing further. A contract-matched structural scorer reports reachable signal in the hidden-dominant regime and none in the visible-bridge regime; the trained model reaches neither. The bridge contrast has enough sensitivity to detect 0.0047, a fifth of the hidden-regime effect. Relevance is evaluator-constructed, so the contrast measures signal recovery under a known generator, not recommendation quality. This controlled result is supplementary; no primary claim depends on it.

an active inductive split retaining 70% of candidates over 241 cases, so a comparison between them crosses two populations under two observation contracts in either direction.

What remains open. For held-out candidates the exact target block is absent from the declared InferenceView; the contract-matched ceiling uses only allowlisted declared-entity relations. Its implementation is identifier-equivariant: under any bijective relabeling of opaque candidate identifiers, with allowed attributes and relations carried along, its scores relabel but do not change. The test suite checks that property. This is an interface guarantee for the audited scorer, not an information-theoretic claim that every

conceivable program lacks the generator seed or could not recompute the identifier hash. For the roughly 70% held in, the target is present in the graph the model fits on, and whether an arbitrary 20-pair matching over 40 block representations is expressible by the decoder in use and learnable from the available supervision is a question about a bilinear form that we have not settled. It is not closed by the structural scorer’s null in this regime, because a similarity-shaped scorer returning zero against a disassortative target is equally consistent with no reachable signal and with no similarity-shaped method reaching it.

Q Confirmatory topology contrasts

The confirmatory experiment in the hidden-dominant regime ran over 30 generator worlds (1000–1029) at a fixed split seed, with the primary contrast, primary slice and primary k declared in the protocol before analysis. On the primary full slice the randomization contrast is 0.0068 (CI [0.0041, 0.0098], 22/30 worlds positive) and the content contrast is 0.0079 (CI [0.0041, 0.0117], 22/30). On the secondary non-obvious slice both intervals contain zero: -0.0029 (CI $[-0.0081, 0.0022]$) and 0.0040 (CI $[-0.0026, 0.0104]$).

We keep the non-obvious slice out of the primary claim because the minimum detectable effect at this sample size is 0.0077, larger than the 0.0068 effect detected on the full slice. A design that could not have detected an effect of that size has not shown that no effect is present. Figure 9 shows the world-level dispersion behind each interval, which the intervals alone do not convey and which pre-empts the reading that a mean is carried by a few worlds.

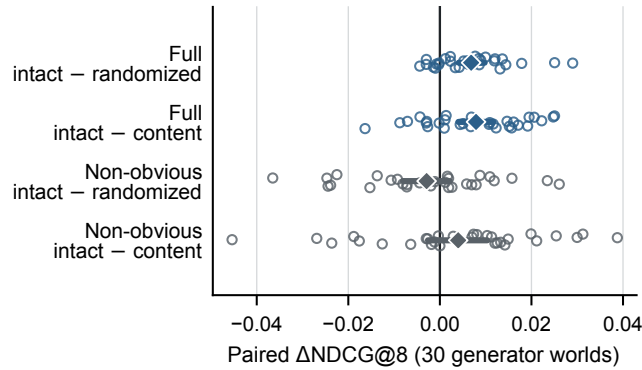


Figure 9: Hidden-dominant regime, NDCG@8, 30 generator worlds. Open circles are per-world means of the paired case-level difference, the filled diamond their mean, the bar a percentile bootstrap over worlds. Blue is the primary full slice (22/30 worlds positive on the topology contrast), grey the secondary non-obvious slice. The non-obvious rows are *not* evidence that typed structure fails there: the minimum detectable effect at this sample size is 0.0077, larger than the 0.0068 effect measured on the full slice. Primary contrast, primary slice and primary k were declared in the protocol before analysis; the remaining contrasts are secondary and not adjusted for multiplicity.

R Reproducibility and access

The ogbl-collab side rests entirely on public data: the benchmark, its published split and its fixed negatives. Every quantity in this paper is traceable to the run that produced it, and every contrast, recomputed from the per-replicate values, agrees with the summary it is quoted from to within 10^{-12} for all thirteen contrasts. The controlled testbed’s profile-level inputs sit under a restricted access boundary, so what stands behind Appendix P is its summaries and protocol files, not the inputs themselves.