A STRUCTURAL AUDIT OF MISSING PROPAGATED EVIDENCE IN TEMPORAL LINK PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Message-passing graph neural networks are the standard architecture for temporal link prediction, yet practitioners repeatedly observe that feature-engineered gradient-boosted trees outperform carefully tuned GNNs on many graphs. Current practice offers expressive-power guarantees that characterize what a model class can represent given suitable input, but no training-free diagnostic for whether a benchmark's queries actually supply the connectivity that message passing needs. We introduce a structural audit computed from the history graph alone that flags queries whose target lies outside the source's receptive field or is indistinguishable from a reachable negative under Weisfeiler-Leman refinement. We prove the flag identifies queries where no source-specific signal reaches the target representation, yet show it cannot upper-bound model accuracy because free embeddings achieve perfect ranking on any consistent query set. The audit therefore measures available propagated evidence rather than representational capacity. On two graphs where trained models solve the unflagged set, every architecture from local aggregation to temporal memory collapses toward random on flagged queries, with gaps surviving degree and depth controls. Controlled feature ablation reveals hard-query performance scales with feature count rather than model architecture: expanding features raises one model's hard-set score by 0.424 MRR while cross-model gaps vanish when features are matched. Where the class is weak throughout, the gap reflects degree confounding rather than the mechanism. We state applicability conditions, verify them on held-out graphs, and show that on the one graph meeting all three, a non-message-passing baseline concentrates its advantage exactly on the flagged queries. A graded hop score derived from the audit ranks per-query benefit of source conditioning at Spearman 0.65 without training.

1 INTRODUCTION

Temporal link prediction asks which node a source will interact with next, given a history of timestamped edges. It underlies recommendation, knowledge-base completion, dynamic social modelling and fraud detection, and its default toolkit is the message-passing graph neural network.

A recurring and under-explained difficulty confronts practitioners: on many deployed graphs a feature-engineered gradient-boosted tree outperforms carefully tuned GNNs. We see it on one of ours, where such a ranker beats every architecture we test, including the path-aware NBFNet, by a margin that persists across depth and width; controlled feature ablation traces the gap to feature richness rather than to model class, since a matched interface removes the difference between the tree and the source-conditioned GNN. Both responses, tuning the GNN further or engineering features, are chosen only after paying the full cost of training: no cheap test quantifies what share of a benchmark's evaluation set carries source-to-target propagated evidence, or which subsets are structurally deprived of it.

Current practice has expressive-power theory but lacks a way to inspect a benchmark before committing to an architecture. Expressiveness results characterize what a model class can represent given the right input, not whether a particular evaluation set supplies it; structural analyses that do relate graph statistics to accuracy run after training and correlate at dataset level, so they cannot identify the affected queries or the reason for their failure. We therefore take a measurement-first stance and study a training-free structural diagnostic for the local message-passing class. A depth-L

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

GNN scores a candidate c for source s through a decoder $f(\boldsymbol{h}_s, \boldsymbol{h}_c)$ in which each $\boldsymbol{h}_v$ is computed by L rounds of aggregation, so it depends only on the depth-L computation neighbourhood at v. Two structural facts follow, with different scopes. If the target p lies outside s's L-hop ball then $\boldsymbol{h}_p$ receives no information propagated from s, which holds for every model in the class. If p shares its L-round Weisfeiler–Leman (WL) colour with a reachable distractor, the two are assigned identical embeddings, which holds only for the anonymous subclass, since a free embedding table separates them. A query where either holds is flagged (Definition 1, Figure 1), and the resulting fraction hard is computed in seconds. Neither fact says the target cannot be ranked first: the decoder still sees $\boldsymbol{h}_s$ and may exploit source–target compatibility. The propagated evidence is absent; its contribution to the observed failure and the conditions under which it matters are evaluated on ten temporal graphs spanning six domains, eight of them public. The resulting boundary is sharp.

Our contributions are threefold:

- **An audit whose predicted failures are real.** We formalize flagged queries and prove the per-query soundness statement above. On the two graphs where the class solves the unflagged complement, four theorem-covered static architectures drop sharply on flagged queries while performing well on the complement; a temporal memory-plus-attention network shows the same empirical pattern as a stress test outside the proposition; the separation is seed-stable and survives matching on target degree, source degree and depth, source-clustered intervals and an eight-fold widening (§4–§5.3). A controlled ablation across three feature sets and three model classes shows within-model gains from feature expansion exceed cross-model gaps at matched features, so feature richness rather than architecture drives hard-query recovery on the one graph where a break is warranted.
- **Audit scope and a no-ceiling theorem.** With free node embeddings the class can achieve a perfect ranking on any pairwise-consistent query set, irrespective of the flagged fraction (§4.2); the audit therefore measures available propagated evidence, and its realized cost is the object of study.
- **Applicability conditions and a graded pre-training decision rule.** The audit speaks only where reachability discriminates and the class solves the complement; recommending a different model class requires a third condition, met on one graph and explained by a controlled public construction. In our fixed-radius experiments, minimum reachable hop separates shell from off-shell queries and predicts which tested intervention helps. On D1, the resulting score $s(q) = 1/h(q)$ ranks the per-query benefit of source-conditioning at Spearman 0.65: the mean gain is 0.42 MRR across cold-start queries versus 0.12 on flagged ones. Its inversion on low-signal graphs bounds the rule to graphs with usable side-features (§5.4, App. F).

2 Related Work

Expressive power of message passing. Standard message-passing GNNs are at most as discriminative as the 1-dimensional Weisfeiler–Leman test (Xu et al., 2019; Morris et al., 2019), and for link prediction node embeddings fed to an inner-product decoder cannot represent pairwise structure such as the number of paths between endpoints, while endpoint-conditioned labeling exceeds 1-WL (Zhang & Chen, 2018; Zhang et al., 2021; Wang et al., 2025); NBFNet (Zhu et al., 2021) realizes this via a learned Bellman–Ford recursion, with a complementary logical characterization by Barceló et al. (2020) and depth/width capacity bounds by Loukas (2020b;a). Closest in spirit, Lachi et al. (2025) place link-representation models in a hierarchy where structural-feature methods dominate pure message passing at every radius and depth, and test whether it matters in practice with a per-dataset symmetry score against aggregate MRR; we ask that question at query resolution, on the reachability axis rather than endpoint symmetry, and §5.4 finds the two resolutions disagree. This literature builds more expressive models; we instead quantify, before training and per graph, how much of a given task lies beyond the class, and find even NBFNet limited on the flagged queries, because the dominant barrier here is the receptive field, not WL indistinguishability.

Temporal models, and structural explanations that come after training. Time-aware message passing for dynamic graphs (TGAT (Xu et al., 2020), TGN (Rossi et al., 2020), DyRep (Trivedi et al., 2019)) and scalable structure-aware predictors such as BUDDY/ELPH (Chamberlain et al., 2023) and Neural Common Neighbor (Wang et al., 2024) are the practical state of the art. Sequence-based

temporal predictors (CAW (Wang et al., 2021), GraphMixer (Cong et al., 2023), DyGFormer (Yu et al., 2023)) replace propagation with encoders over each endpoint's recent one-hop interactions and pairwise features reaching at most two hops, so a target flagged at $L{=}2$ lies beyond their structural evidence too; we do not train them (App. U). TGN's recurrent memory is not confined to an L-hop ball, so Proposition 1 does not cover it. We train it alongside the static backbones only as an empirical stress test (§5.2). A separate line relates dataset structure to realized accuracy after training, closest being Jiang & Deng (2025), which resamples subgraphs of FB15k-237, retrains, and correlates structural statistics with accuracy; ours is computed before training, concerns a model class, not a fitted model, and is anchored to a mechanism rather than to correlations. Over-smoothing (Oono & Suzuki, 2020; Cai & Wang, 2020; Rusch et al., 2023) enters only to explain the failure of increasing L to rescue flagged queries (App. I). App. V extends this discussion, including the independent temporal-receptive-field notion of Karmim et al. (2024).

3 Problem Setup

Temporal link prediction and evaluation. A temporal graph is a stream of timestamped edges $\{(s_i, d_i, t_i)\}$. A query is a triple $(s, t, \mathcal{C})$: a source s, a time t, and a candidate set $\mathcal{C}$ of size K containing the true target p and $K-1$ negatives ($K = 100$ throughout). A model scores each candidate c as $f(\boldsymbol{h}_s, \boldsymbol{h}_c)$, and we report the mean reciprocal rank (MRR) of the true target under the tie-aware convention,

$$\mathrm{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathrm{rr}(q), \qquad \mathrm{rr}(q) = \frac{2}{2g_q + m_q + 1}, \tag{1}$$

with g_q the number of candidates scored strictly above the target and m_q the number tied with it including the target itself, so m candidates tied at the top give $2/(m{+}1)$ and a unique top rank gives 1. The convention matters: Proposition 1(ii) caps tied scores rather than forbidding them. The random-guessing baseline is $\mathrm{MRR}_{\mathrm{rand}} = \frac{1}{K}\sum_{r=1}^{K} \frac{1}{r} \approx \frac{\ln K + \gamma}{K} \approx 0.052$ for $K = 100$.

Leak-free evaluation. Temporal leakage is a notorious source of inflated numbers. We construct a strict split: the history graph consists of all edges before a cutoff time; the message-passing graph and history-derived features use only those edges. Query edges are drawn from a later window and, by construction, never appear in history. Four D1 XGBoost features are cross-query aggregates computed from the evaluation candidate batch and are therefore transductive; we identify that interface below. We verify 0 leaked (s, d, t) triples on every dataset (App. G).

Benchmarks and protocol. Our leak-free protocol follows the standard set by recent dynamic-link-prediction benchmarks (Huang et al., 2023; Kumar et al., 2019), and our graphs come from those sources and SNAP (Leskovec & Krevl, 2014). §5.3 reports a negative-sampling asymmetry we found in them.

Message-passing GNNs. A depth-L GNN computes

$$\boldsymbol{h}_v^{(\ell)} = \phi^{(\ell)}\Big(\boldsymbol{h}_v^{(\ell-1)},\ \mathrm{agg}\big\{\boldsymbol{h}_u^{(\ell-1)} : u \in \mathcal{N}(v)\big\}\Big), \qquad \ell = 1, \dots, L, \tag{2}$$

then scores a candidate via a decoder $f(\boldsymbol{h}_s^{(L)}, \boldsymbol{h}_c^{(L)})$. Crucially, $\boldsymbol{h}_v^{(L)}$ is a function only of the depth-L rooted subtree at v; equivalently, of v's L-round WL colour and the features within $\mathcal{B}_L(v)$, the L-hop ball around v. We consider four architectures spanning the inductive-bias spectrum: GCN (Kipf & Welling, 2017) and GraphSAGE (Hamilton et al., 2017) (local mean/max aggregation), APPNP (Gasteiger et al., 2019) (personalized-PageRank global mixing), and NBFNet (Zhu et al., 2021) (source-conditioned, path-aware; provably beyond 1-WL). All share an identical decoder, feature interface, training loop, and seeds, so only the encoder differs.

Throughout, the flag is a per-query structural statement computable before training: on a flagged query, nothing injected at the source reaches the target's representation (Proposition 1(i)). It is not a numerical bound on MRR (§4.2), so a flagged query may still be ranked correctly; and it is an evaluation-set instrument rather than a prediction-time router, since deciding whether $p \in \mathcal{B}_L(s)$ requires knowing p (App. N). We measure the flag's cost in realized MRR and test the conditions under which that cost is attributable to the mechanism and not to a confound.

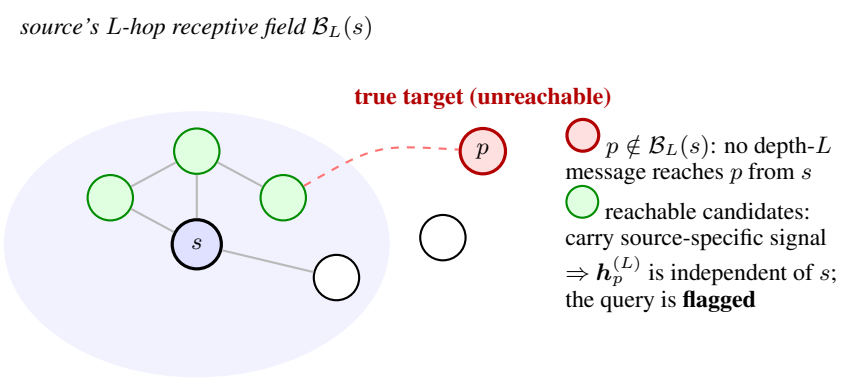

Figure 1: The receptive-field mechanism. A depth-L message-passing GNN can only inject source-specific signal into candidates within s's L-hop ball $\mathcal{B}_L(s)$ (shaded). When the true target p lies outside the ball, no evidence reaches $\boldsymbol{h}_p^{(L)}$ by propagation from s, and the query is flagged. The decoder still sees $\boldsymbol{h}_s$ and may use it; what is missing is the propagated source evidence in the target representation. The test runs before training, in seconds, from the graph alone.

4 The Training-Free Diagnostic

4.1 Flagged queries

Fix message-passing depth L. Let $\mathcal{B}_L(s)$ be the set of nodes within L hops of s in the (undirected) history graph, and let $\mathrm{col}_L(v)$ be the depth-L WL colour of v under constant initialization (pure structural role, no node identity).

Definition 1 (Reachability-flagged and collision-flagged)**.** A query $(s, p, \mathcal{C})$ at depth L is reachability-flagged if $p \notin \mathcal{B}_L(s)$, and collision-flagged if there is a negative $c \in \mathcal{C} \cap \mathcal{B}_L(s)$ with $\mathrm{col}_L(c) = \mathrm{col}_L(p)$. It is flagged if either holds, and we write hard for the flagged fraction. Here $\mathcal{B}_L$ and col_L are computed on the deduplicated, undirected simple history graph with constant initial WL colours; edge multiplicity, direction, labels, and attributes are not part of the collision clause.

These are separate objects: an information-flow statement about the whole local class and an expressiveness statement about the anonymous subclass (Proposition 1), so we keep the names distinct. We report their union because hard is the quantity a practitioner would compute; the second clause contributes $< 0.4\%$ of queries on nine of ten graphs. Both clauses are conservative: reachability is topological, exact for the static encoders that propagate on this graph and an over-count of information flow for time-respecting models, whose flagged fraction is thus under-stated; and an anonymous model ties p with a colour-tied negative even when it is unreachable, so the collision term under-counts such ties (App. B).

Proposition 1 (Source-invariance, and a tie for the anonymous class)**.** Let a depth-L message-passing GNN score candidate c for source s as $f(\boldsymbol{h}_s^{(L)}, \boldsymbol{h}_c^{(L)})$, where $\boldsymbol{h}_v^{(L)}$ is a function of the depth-L computation neighbourhood at v. (i) (any such model.) If $p \notin \mathcal{B}_L(s)$ then $\boldsymbol{h}_p^{(L)}$ is invariant to every perturbation of the features, identity or embedding of s: the target's representation carries no information injected at the source. The score may still depend on s through its other argument. (ii) (the anonymous simple-graph class $\mathcal{M}_L^{\mathrm{anon}}$ only.) If in addition every $\boldsymbol{h}_v^{(0)}$ is determined by $\mathrm{col}_0(v)$, and p shares its colour with a reachable negative c, then $\boldsymbol{h}_p^{(L)} = \boldsymbol{h}_c^{(L)}$, so $f(\boldsymbol{h}_s^{(L)}, \boldsymbol{h}_p^{(L)}) = f(\boldsymbol{h}_s^{(L)}, \boldsymbol{h}_c^{(L)})$ and no decoder ranks p strictly above c; under the tie-aware convention $\mathrm{rr} \le 2/3$ (and with r tied negatives, $\mathrm{rr} \le 2/(r+2)$).

Proof. For (i), L-round aggregation propagates information at most L hops; if $p \notin \mathcal{B}_L(s)$ no message originating at s reaches p within L rounds, so $\boldsymbol{h}_p^{(L)}$ is measurable with respect to $\mathcal{B}_L(p)$, which excludes s. For (ii), with $\boldsymbol{h}^{(0)}$ determined by the initial colouring, depth-L message passing is at most as discriminative as the L-round WL test (Xu et al., 2019; Morris et al., 2019), so equal colours imply equal embeddings and equal decoder scores. □

Scope of clause (i). Clause (i) does not imply that the target cannot be ranked first: the score still depends on s through $\boldsymbol{h}_s^{(L)}$, and a free embedding table can rank p first for reasons unrelated to s (§4.2). The missing signal is evidence propagated from this source into the candidate representation; its effect on accuracy is empirical (§5.2–§5.3).

The clauses differ in scope, which bounds the experimental claims. Clause (i) constrains information flow and holds for every model whose $\boldsymbol{h}_v^{(L)}$ is a function of $\mathcal{B}_L(v)$: the four static encoders and the source-conditioned NBFNet. Clause (ii) constrains expressiveness and holds only for $\mathcal{M}_L^{\mathrm{anon}}$; every architecture we train carries a free embedding table or is beyond 1-WL, so it does not bind them, and hard is dominated by clause (i) throughout: a receptive-field diagnostic computed in seconds by breadth-first reach and WL refinement. The exception is LASTFM at $L{=}1$, added precisely because it activates the collision term, which App. B tests.

TGN is outside Proposition 1(i): its GRU memory can propagate along chains of temporally ordered events rather than only along L history-graph edges. We therefore include it as an empirical stress test, not as an instance of the theorem; it has a mechanism for recovering the information indicated as missing by the flag (App. H).

4.2 Reachability cannot upper-bound achievable MRR

hard does not induce a numeric ceiling of the form $\mathrm{MRR} \le (1 - \mathrm{hard}) + \mathrm{hard} \cdot \epsilon$; this section proves it. Write $\mathcal{M}_L^{\mathrm{free}}$ for the depth-L class with freely learnable $\boldsymbol{h}_v^{(0)}$, the standard implementation since essentially every deployed link predictor contains an embedding table, and $\mathcal{M}_L^{\mathrm{anon}}$ for the class where $\boldsymbol{h}_v^{(0)}$ is a given feature, constant if the graph is featureless.

Proposition 2 (No non-trivial hard-derived bound under pairwise consistency)**.** Let $\mathcal{Q}$ be any finite pairwise consistent query set: for every pair (s, c), node c is the target of all source-s queries whose candidate set contains it, or of none. Then $\sup_{M \in \mathcal{M}_L^{\mathrm{free}}} \mathrm{MRR}(M; \mathcal{Q}) = 1$ for every $L \ge 1$ and every history graph, independently of hard. The same holds for $\mathcal{M}_L^{\mathrm{anon}}$ on any graph whose L-round WL colours separate the nodes appearing in $\mathcal{Q}$.

The proof (App. E) is a one-line construction: a message-passing layer may ignore its aggregate, so one-hot inputs survive L rounds unchanged and the decoder sees (s, c) exactly.

The second half bites in practice: WL collisions are near-zero on nine of our ten graphs (§5.2), so Definition 1 supplies no non-trivial class-supremum bound there. Any training-free "ceiling" derived from this history flag must therefore measure something other than the class supremum.

Definition 1 instead measures the propagated evidence available to the computation. On a flagged query $\boldsymbol{h}_p$ carries nothing routed from s (Proposition 1(i)), so any remaining signal must come from source–target compatibility learned from the training distribution rather than from structure connecting this pair. Our leak-free protocol (§3) makes queries future edges absent from history, which closes that route empirically, not by corollary. This makes §5.2–§5.3 and App. J load-bearing.

Depth matching. Proposition 1 is stated for a fixed L, so we train every architecture at $L \in \{2, 3, 4\}$ and score each against the diagnostic at its own depth. Deepening changes the partition as expected but does not rescue the queries inside it: on D1 the hard fraction shrinks with L while depth-matched hard-set MRR falls alongside it, so depth is a trade, not a rescue (App. I).

5 Experiments

5.1 Datasets and protocol

We evaluate on ten temporal graphs spanning six domains: recommendation, review, Q&A, signed trust, social messaging and institutional email, sized 2k to 674k nodes and including one graph with 172-dimensional edge features (Tables 7 and 10 give sources, sizes and each graph's role). Eight are public, from SNAP (Leskovec & Krevl, 2014), JODIE (Kumar et al., 2019) and TGB (Huang et al., 2023), and carry every claim except the class-specificity break; D1/D2 come from a machine-learning competition and are not redistributable. All splits are leak-free (App. G), and §5.3 reports

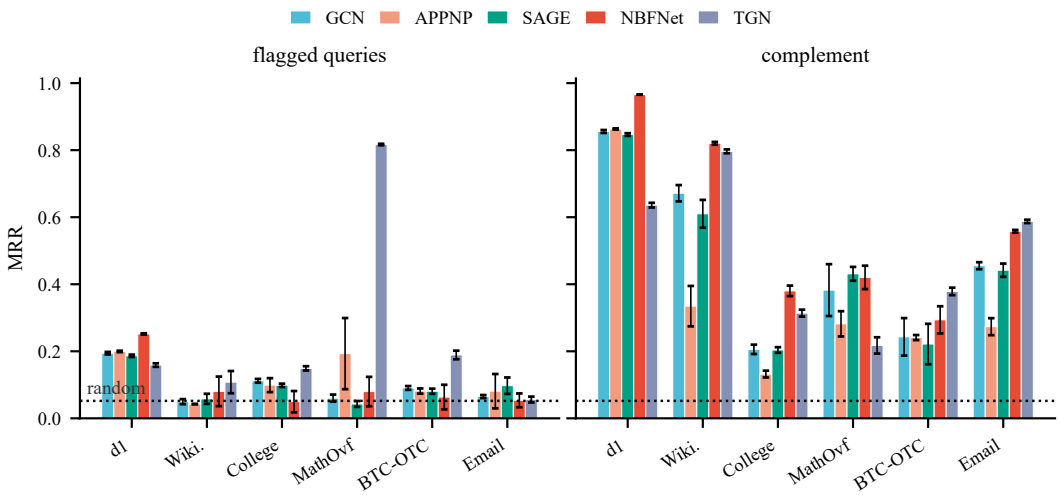

Figure 2: MRR on flagged queries and on the complement, per graph and architecture (mean±sd over seeds); the dotted line is the random baseline 0.052. On the two high-signal graphs, D1 and WIKIPEDIA, every architecture moves sharply toward that line on the flagged set and remains far above it on the complement. The low-signal panels are descriptive rather than headline evidence; EMAIL-EU has only 35 flagged queries. Same source as Table 1.

Table 1: MRR on flagged queries → MRR on the complement, mean over seeds $\{11, 23, 37\}$; per-seed deviations in App. J. The random baseline is 0.052 and the partition is computed before training, identically for every seed and architecture. EMAIL-EU's hard set (n=35) is too small to interpret. The one cell that does not fall is TGN on MATHOVERFLOW, a negative-sampling artefact diagnosed in §5.3.

Graph	Domain	n_{hard}	hard frac	GCN hard→OK	APPNP hard→OK	SAGE hard→OK	NBFNet hard→OK	TGN hard→OK
EMAIL-EU	dense social	35	0.006	0.065→0.455	0.081→0.273	0.097→0.442	0.054→0.558	0.056→0.587
D1	user–item	5955	0.177	0.194→0.856	0.200→0.863	0.186→0.847	0.252→0.966	0.159→0.636
WIKIPEDIA	bipartite+feat	558	0.093	0.050→0.672	0.042→0.335	0.058→0.610	0.080→0.820	0.108→0.796
COLLEGEMSG	sparse social	2137	0.356	0.112→0.206	0.099→0.132	0.098→0.204	0.050→0.380	0.149→0.314
MATHOVERFLOW	Q&A	2184	0.364	0.060→0.382	0.193→0.282	0.043→0.431	0.080→0.420	0.817→0.217
BITCOIN-OTC	trust	2002	0.550	0.091→0.243	0.081→0.241	0.081→0.222	0.063→0.294	0.189→0.379

a negative-sampling asymmetry we found in the public builds and corrected. We train four static encoders (GCN, GraphSAGE, APPNP, NBFNet) under a shared protocol; on D1 they receive the same 26 core features; every other graph, WIKIPEDIA included, runs graph-only, since its 172-dimensional edge features describe edits already made and carry no signal about the next one (the one XGBoost arm that used them is omitted as invalid). We call D1 and WIKIPEDIA high-signal, since on both the best encoders solve the unflagged complement (Table 1), and the remaining graphs low-signal. TGN is an additional empirical stress test outside Proposition 1 (App. H).

5.2 THE PARTITION IS REAL

Figure 2 and Table 1 report MRR on flagged versus complementary ("OK") queries over three seeds, for five encoders spanning local aggregation, global mixing, source-conditioned path awareness and temporal memory. The direction is the same in 29 of 30 graph–architecture pairs; the interpretation differs by graph, and the single exception, TGN on MATHOVERFLOW, diagnoses a flaw in the benchmark rather than in the audit (§5.3). This first table uses a fixed benchmark-audit depth (L=2 on non-bipartite graphs, L=3 on the bipartite graphs) and is descriptive; it is not an invocation of Proposition 1 for every trained architecture. The theorem-backed comparison for the four static encoders is the depth-matched sweep above and in App. I, where each model is scored at its own number of propagation rounds.

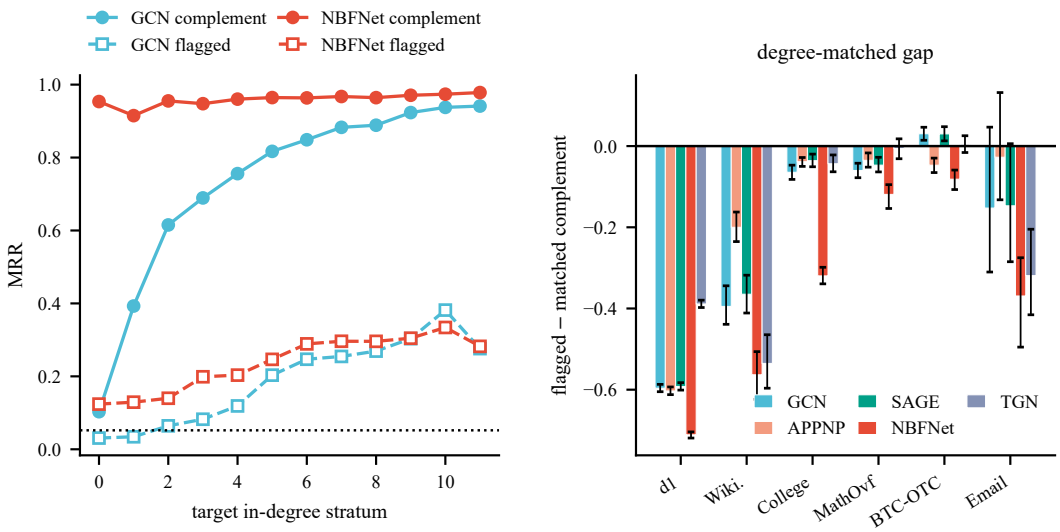

Figure 3: The degree control. Left: on D1, MRR by target history in-degree stratum, solid for OK queries and dashed for flagged; the curves never meet: the gap is present at every degree, not a popularity effect. Right: the degree-matched gap (hard minus reweighted OK, 95% bootstrap CI) per graph and architecture. High-signal graphs show a large negative gap for every architecture; on low-signal graphs the local architectures sit near zero and only the source-conditioned NBFNet retains an effect everywhere. Full numbers in Table 12.

On D1 and WIKIPEDIA the drop is large, seed-stable and not an artefact of one architecture: NBFNet follows the local models, and TGN does so as an empirical stress test. On the low-signal graphs it is present but seed-unstable and indistinguishable from target-degree composition for the local encoders (App. J, §5.3), so we treat them as descriptive. Nor do we rest anything on the graph-level correlation between hard-fraction and realized ceiling ($r = -0.88$, but $n = 5$ and Spearman only -0.60; Figure 6).

5.3 THE COLLAPSE IS NOT A DEGREE ARTEFACT ON HIGH-SIGNAL GRAPHS

Unreachable targets are disproportionately low-degree or newly-appearing nodes, and low-degree targets are hard to rank for reasons unrelated to the receptive field, so "flagged" could be a proxy for "rare node". The confound is real: on D1 the median target history in-degree is 16 on flagged queries against 52 on the complement, and every graph shows the same direction. We control for it by direct standardization, reweighting OK-query MRR onto the flagged set's target-in-degree distribution, alongside a conservative estimate restricted to strata holding at least 20 queries of each kind. One correction comes first: on the public graphs some flagged targets appear nowhere in history (in-degree 0, identifiable by novelty alone, App. C), so the control keeps only targets previously seen in the target role, raising coverage from 0.17–0.58 to 0.83–0.99 (Table 12); D1 has no such rows. Two outcomes follow (Figure 3).

On the two high-signal graphs the collapse is not a degree effect. D1 has coverage 1.00, so both estimates coincide: the gap runs from -0.389 (TGN) through -0.596 (GCN, 95% CI $[-0.605, -0.587]$) to -0.711 (NBFNet), and stratum by stratum it appears at every degree, including the most popular targets. In the top stratum (targets seen >1000 times, n_{OK}=912) GCN scores 0.276 on hard against 0.941 on OK. WIKIPEDIA agrees at coverage 0.84, every architecture between -0.200 and -0.563. Source-degree matching with source-clustered intervals moves the D1 estimates by under a hundredth and leaves every interval below zero (App. O).

On the low-signal graphs it is weaker but real, except on one. With the novelty correction, COLLEGEMSG (coverage 0.89) gives gaps of -0.036 to -0.320 and MATHOVERFLOW (0.83) gives -0.006 to -0.119: negative for every architecture but an order of magnitude smaller than on the high-signal graphs, as expected where the models have little signal (OK-set MRR 0.13–0.44).

Table 2: Class-specificity, all nine graphs with a non-degenerate flag. conc is the baseline's flagged-set edge divided by its complement edge; bold marks $\mathrm{conc} > 1.3$. Rows are hard-node-seen (over all rows WIKIPEDIA would read 3.25, see §5.3). Three seeds on both sides except EMAIL-EU and TGB-REVIEW (one baseline seed).

Graph	L	hard	best MP	best non-MP	raw	conc
D1	2	0.177	0.252	0.543	2.16	**2.13**
TGBL-REVIEW	3	0.561	0.164	0.300	1.83	**1.40**
LASTFM	1	0.360	0.071	0.108	1.52	1.10
MATHOVERFLOW	2	0.364	0.107	0.156	1.46	1.08
REDDIT	1	0.095	0.231	0.236	1.02	0.99
WIKIPEDIA	3	0.093	0.129	0.128	0.99	0.95
EMAIL-EU	2	0.006	0.132	0.142	1.08	0.92
BITCOIN-OTC	2	0.550	0.216	0.233	1.08	0.76
COLLEGEMSG	2	0.356	0.220	0.177	0.80	0.63

BITCOIN-OTC is the genuine exception: at coverage 0.99 the two local architectures come out positive (GCN and SAGE both $+0.030$) while the source-conditioned and global models stay negative (-0.047, -0.082), so there the raw gap is degree composition rather than reachability; §5.4 turns this into an applicability condition. Across all five, NBFNet, the only source-conditioned representation, has a negative degree-matched gap on every graph.

A negative-sampling asymmetry. Two of our builds contain a negative-sampling asymmetry that a memory-based model exposes: on MATHOVERFLOW TGN scores 0.817 on flagged against 0.217 on the complement, the one inversion in Table 1, as 80.2% of flagged targets never appear in history while no negative is novel. A flagged row is hard-node-seen when its target appears anywhere in history, hard-target-seen when it previously appears as target; each analysis names its cohort (App. C).

5.4 CLASS-SPECIFIC COLLAPSE AND ITS APPLICABILITY CONDITION

Feature width dominates architectural choice on flagged queries. On D1, where every model receives engineered features, expanding the feature set from 26 core covariates to the full 46 raises XGBoost from 0.120 to 0.544 MRR on flagged queries against a much smaller OK-set gain; the same direction holds for GCN and NBFNet. At the matched all-46 interface XGBoost and NBFNet sit within seed noise of each other, and both exceed GCN with 26 features by over three-fold. Feature richness therefore controls the recovery of flagged queries more than the choice of encoder (Table 3 in App. D).

Is the audit about GNNs, or are flagged queries hard for everyone? A baseline stronger everywhere would also be stronger on flagged queries, so the claim needs the advantage concentrated there: we report conc, its edge on flagged queries divided by its edge on the complement (Table 2). On D1 the GNN arms receive 26 core engineered features; XGBoost (Chen & Guestrin, 2016) receives those 26 plus 16 structural and four cross-query features, the latter using the evaluation candidate batch, so the comparison is transductive rather than history-only; with history-only covariates the tree's edge is confined to the shell (App. R), so the break rests on those dataset features rather than on the tree as such. XGBoost breaks the diagnostic exactly where the GNNs collapse (Figure 4): 0.543 on flagged queries, $2.16\times$ the best message-passing model (NBFNet, 0.252), while level with them on the complement (0.975 against 0.966), $\mathrm{conc} = 2.13$. Flagged queries there are not intrinsically unsolvable; the GNNs fail because their representation discards evidence a non-path model can use, a principled account of the advantage feature-engineered trees often have over GNNs on such graphs.

Across eight public graphs the advantage is not concentrated. Two near-misses show where the raw hard-set ratio misleads: TGB-REVIEW is a baseline stronger everywhere (App. A), while WIKIPEDIA's $\mathrm{conc} = 3.25$ comes from novel-target asymmetry (§5.3); all reported rows are therefore hard-node-seen.

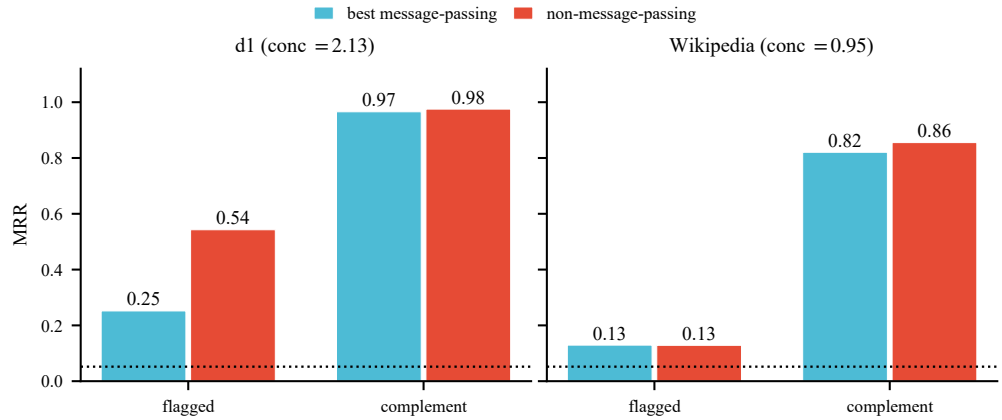

Figure 4: MRR on flagged and complementary queries. On D1, XGBoost (46 features, 26 shared) is 2.16× above the best message-passing model on flagged queries and level on the complement; on WIKIPEDIA, conc = 0.95. The invalid 172-feature arm is omitted.

The condition, isolated on public inputs. Rather than keep searching, we test the condition on unmodified public COLLEGEMSG, synthesizing only targets from a pairwise rule (probability α) or a walk-count rule (oracle separation 0.973 vs. 0.031). A plain inner product is learned by the MLP and lowers conc (1.15, 0.89, 0.84, 1.02 across α), whereas a high-frequency function with the same inputs restores the effect (9.40 vs. 2.144). Thus the covariate must be path-free and not recoverable by the class's readout, as reproduced on two public histories (App. Q).

Which remedy applies depends on the stratum, not on the flag. In our fixed-radius setup, shell targets first appear at $L+1$, so pair-conditioned path counts expose evidence just outside the audited radius (Srinivasan & Ribeiro, 2020; Zhang et al., 2021). BUDDY-lite wins this stratum on two of four measurable graphs, and adding path counts gives $+0.091$ (CI $[+0.078, +0.103]$; Table 9). Off-shell results likewise favour covariate-bearing arms; these are fixed-radius measurements, not consequences of Proposition 1.

An encoder-side intervention, with NBFNet's learned source indicator ablated in an otherwise identical model, ranks the strata identically without shared readout parameters (App. M). With all features and queries fixed, the indicator is worth $+0.4216$ MRR on cold-start queries but only $+0.1223$ on flagged ones (0.5814 against 0.1598, and 0.2517 against 0.1295, with and without it); permuting source identity reproduces $+0.4187$ against a degree baseline (Table 5, App. F). The residual is the decoder still seeing $\boldsymbol{h}_s$ although the target representation receives nothing from s.

Regimes in which the audit is informative. Table 4 (App. D) gives the per-graph verdict. Of four trained outcomes under the frozen conditions, three predictions hold and one is refuted (App. S); a model-class switch requires the third condition, met only on D1.

6 CONCLUSION

Our training-free audit flags depth-L receptive-field gaps and L-round WL collisions from the history graph alone. On the high-signal graphs every architecture collapses toward random on flagged queries while performing well on the complement, with degree-matched gaps of -0.39 to -0.71 MRR on D1 at every degree stratum; a feature-matched ablation attributes the residual tree-versus-GNN gap on that graph to feature width rather than to architecture; frozen applicability conditions predict three of four held-out outcomes; and where a third condition holds, a non-message-passing baseline recovers 2.16× the best GNN on exactly the flagged queries.

The audit should be read as an evidence diagnostic, not as a ceiling on model accuracy: the flag cannot bound the class supremum (Proposition 2), while the graded hop score ranks where source conditioning helps. Its reliable use therefore depends on the applicability conditions and on checking the evaluation split for the sampling asymmetries documented in the appendix.

AI USE STATEMENT

Generative AI tools were used for manuscript polishing and typesetting, organizing code and experimental results, and code standardization. The authors reviewed all AI-assisted work and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Definition 1 is deterministic; masks are cached per (graph, depth) and shared across architectures and seeds $\{11, 23, 37\}$. Data construction, splits, settings, results, the independent implementation, and claim scope are documented in Apps. G, H, J, T, and U; public and non-redistributable graph sources are identified there.

REFERENCES

Pablo Barceló, Egor V. Kostylev, Mikaël Monet, Jorge Pérez, Juan Reutter, and Juan Pablo Silva. The logical expressiveness of graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

Benjamin Paul Chamberlain, Sergey Shirobokov, Emanuele Rossi, Fabrizio Frasca, Thomas Markovich, Nils Hammerla, Michael M. Bronstein, and Max Hansmire. Graph neural networks for link prediction with subgraph sketching. In *International Conference on Learning Representations (ICLR)*, 2023.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

Weilin Cong, Si Zhang, Jian Kang, Baichuan Yuan, Hao Wu, Xin Zhou, Hanghang Tong, and Mehrdad Mahdavi. Do we really need complicated model architectures for temporal networks? In *International Conference on Learning Representations (ICLR)*, 2023.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations (ICLR)*, 2019.

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

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xiaobo Jiang and Yadong Deng. Understanding the mechanisms behind structural influences on link prediction: A case study on fb15k-237. *arXiv preprint arXiv:2505.01655*, 2025.

Yannis Karmim, Leshanshui Yang, Raphaël Fournier S'Niehotta, Clément Chatelain, Sébastien Adam, and Nicolas Thomé. Temporal receptive field in dynamic graph learning: A comprehensive analysis. *arXiv preprint arXiv:2407.12370*, 2024.

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

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

Veronica Lachi, Francesco Ferrini, Antonio Longa, Bruno Lepri, Andrea Passerini, and Manfred Jaeger. Bridging theory and practice in link representation with graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, 2014.

Andreas Loukas. How hard is to distinguish graphs with graph neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Andreas Loukas. What graph neural networks cannot learn: depth vs width. In *International Conference on Learning Representations (ICLR)*, 2020b.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *AAAI Conference on Artificial Intelligence*, 2019.

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

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. In *ICML Workshop on Graph Representation Learning*, 2020.

T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

Balasubramaniam Srinivasan and Bruno Ribeiro. On the equivalence between positional node embeddings and structural graph representations. In *International Conference on Learning Representations (ICLR)*, 2020.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. DyRep: Learning representations over dynamic graphs. In *International Conference on Learning Representations (ICLR)*, 2019.

Xiyuan Wang, Haotong Yang, and Muhan Zhang. Neural common neighbor with completion for link prediction. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiyuan Wang, Pan Li, and Muhan Zhang. Improving graph neural networks on multi-node tasks with the labeling trick. *Journal of Machine Learning Research*, 26(23):1–44, 2025.

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *International Conference on Learning Representations (ICLR)*, 2021.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Muhan Zhang, Pan Li, Yinglong Xia, Kai Wang, and Long Jin. Labeling trick: A theory of using graph neural networks for multi-node representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural bellman-ford networks: A general graph neural network framework for link prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

A THE BASELINE TRAINING PROTOCOL AND THE FIRST-PASS NULL RESULT

The class-specificity test compares a non-message-passing baseline against the message-passing models on flagged queries, so the baseline has to be trained competently or the comparison is meaningless in the direction that flatters our claim. Our first implementation was not, and the failure is easy to repeat, so we record it.

We sampled training positives uniformly from history edges (u, i, t) and drew negatives from the pool of history targets. Every training row therefore had the pair-seen indicator equal to 1 and the pair-count feature at least 1 for the positive, because the positive was a history edge. At evaluation the positive is a future edge, so those features are near zero: on TGB-REVIEW, the pair-seen indicator was 1.000 on training positives, 0.002 on evaluation positives, and exactly 0 on flagged ones. Gradient boosting duly put almost all of its gain on those two columns and learned "the answer is the item this source already interacted with", which is correct on the training distribution and useless on the flagged queries.

The fix is a nested split. Cut history at its 85th time percentile; compute the training rows' features from the earlier part only; draw training positives from the later part. Training rows then have the same feature distribution as evaluation rows, the positive is a future edge relative to the history its features were built from. This raises TGB-REVIEW from 0.240 to 0.456 overall and from 0.108 to 0.300 on hard-node-seen queries, and it is what made the concentration-ratio analysis of §5.4 meaningful: before the fix every public graph showed a spurious "break" of the uninteresting kind, where the baseline is bad everywhere and slightly less bad on the flagged set.

The general point is that a baseline whose features are computed from the same edges that supply its training labels will be mis-trained in a way aggregate metrics hide, and the bias runs towards the diagnostic looking more class-specific than it is.

B THE COLLISION CLAUSE ON A GRAPH THAT ACTIVATES IT

The seven graphs of Table 7 leave Definition 1's second clause untested, so we added one that activates it: on JODIE-LastFM at $L{=}1$ one WL round separates only 390 of 1980 nodes and 11.4% of queries have a reachable, colour-tied distractor, giving 570 collide-only rows disjoint from 1474 unreachable ones. For the anonymous class the clause predicts more than a collapse: colour-tied candidates receive identical embeddings, so one tied negative caps tie-aware reciprocal rank at $2/3$. The deployed architectures in Table 11 are outside that class and show a smaller drop than on unreachable rows. The reason is structural. The clause assumes $\boldsymbol{h}_v^{(L)}$ is determined by $\mathrm{col}_L(v)$, which none of our architectures satisfies: GCN, APPNP, SAGE and TGN carry a free node-embedding table, so colour-tied nodes still differ (the mechanism Proposition 2 exploits) and NBFNet is source-conditioned and beyond 1-WL, so a shared colour never implied a shared representation for it. Definition 1 also restricts the colliding negative to $\mathcal{B}_L(s)$, a restriction the anonymous class does not need: $\boldsymbol{h}_v^{(L)}$ is a function of $\mathrm{col}_L(v)$ alone, so a reachable target is scored identically to a colour-tied negative whether or not that negative is reachable. The restricted count is therefore a lower bound on the anonymous-class tie set; we keep it because it makes the flag a statement about the source's neighbourhood, and because the clause binds no deployed architecture either way.

The clauses therefore differ in scope, which the earlier graphs could not reveal. The receptive-field clause constrains information flow and holds for any depth-L message-passing model regardless of initialization or conditioning; the collision clause constrains expressiveness and holds only for the anonymous class of Proposition 2. Since deployed link predictors almost always include an embedding table, that clause is close to vacuous in practice, a stronger statement than "untested", and a further narrowing of the claims available to a training-free certificate here.

The same conclusion on a benchmark built to activate the clause. One organic graph is thin evidence about a whole clause, so we repeat the test on LR-EXP, the synthetic benchmark Lachi et al. (2025) introduce precisely to contain non-automorphic links whose endpoints share 1-WL colours. We reimplement their generator from its published specification rather than their code, so the check is independent: sample an Erdős–Rényi graph on $n \in \{5, \ldots, 17\}$ nodes with $p \sim U(0, 1)$, duplicate it, add inter-block edges with $p' \sim U(0, 1)$, and keep graphs containing a colour-

tied pair. The precondition is indeed active, 94.3% of retained graphs contain at least one reachable colour-tied pair, with mean WL collapse $1 - |\mathcal{C}|/n = 0.309$.

On 1276 decisive queries, a source and two reachable candidates sharing a WL colour, exactly one of which is the true target, we train two decoders that differ only in the representation assigned to a node. A WL-determined decoder, $\mathrm{MLP}([\mathrm{col}_L(s), \mathrm{col}_L(c)])$, satisfies the clause's premise; a free per-node table, $\mathrm{MLP}([\boldsymbol{E}_s, \boldsymbol{E}_c])$, does not. The former achieves $\mathrm{rr} > 2/3$ on 0.00% of the queries and ties exactly at $2/3$ on all of them, confirming both the cap and our implementation of it. The latter achieves $\mathrm{rr} > 2/3$ on 100%. The gap is not a matter of degree: on a benchmark designed to make the collision clause bite, the clause binds a class of models that includes essentially nothing anyone trains.

C THE BENCHMARK ASYMMETRY, IN FULL

Training a memory-based model surfaced a flaw in our splits that the static architectures had hidden: the one inversion in Table 1. On MATHOVERFLOW, TGN scores 0.817 on the flagged set against 0.217 on the complement, reproducibly across three seeds. The explanation is not that hard queries are easy: 80.2% of that graph's hard-query targets never appear in the history graph, while 0% of OK-query targets and negatives do, so the target is the only candidate with no history and TGN's memory, zero-initialized for an unseen node, is exactly a novelty detector. Its 0.817 tracks the 0.802 novel share almost exactly.

Definition 1 is not at fault, a node with no history edges is necessarily outside every L-hop ball, so novelty is a strict subset of unreachability and flagging such queries is correct. The fault is an asymmetry between how our builds sample negatives (always from nodes with a history) and positives (any future edge), which leaves the flagged set conflating two real but distinct reasons a model fails: no path, and never seen. We separate them throughout (Table 6 in Appendix D): hard-node-seen queries carry the receptive-field analysis, while hard-node-novel queries show only that negative sampling is degenerate there. On hard-node-seen rows the models remain below the OK set (WIKIPEDIA 0.072–0.129 against OK 0.335–0.820) and TGN's MATHOVERFLOW score falls from 0.995 to 0.095. D1 has no novel-target queries and is unaffected. The caution concerns split construction, not the published benchmarks: drawing negatives from historical targets while letting positives be any future edge hands a free signal to any novelty-sensitive model, the obvious construction, which we used, and which cost us two conclusions before we caught it (§5.4). The fix is cheap and TGB already applies it (asymmetry -0.0003 on TGBL-REVIEW), so we recommend reporting that asymmetry alongside any hard-query analysis on a self-built split.

The collision clause applies more narrowly. The seven graphs of Table 7 have near-zero WL collisions, so Definition 1's second clause went untested there; Appendix B constructs a setting that activates it (JODIE-LastFM at $L{=}1$, 570 collide-only rows) and gives the per-stratum numbers. The outcome limits the clause's practical reach: its premise, an embedding fully determined by the WL colour, fails for any model carrying a free node-embedding table and for any source-conditioned model, so the $\mathrm{rr} \leq 2/3$ prediction does not bind deployed architectures. In scope the two clauses part ways, information flow versus expressiveness, and only the former holds across all depth-L models. The operative flag is reachability throughout.

D ADDITIONAL TABLES AND FIGURES

For space, several supporting figures are collected here.

E PROOF OF PROPOSITION 2

Take $d = n$ and $\boldsymbol{h}_v^{(0)} = \boldsymbol{e}_v$ (one-hot). A message-passing layer is free to ignore its aggregate: choosing $\phi^{(\ell)}(\boldsymbol{h}_v, \mathrm{agg}) = \boldsymbol{h}_v$ at every layer gives $\boldsymbol{h}_v^{(L)} = \boldsymbol{e}_v$ for every v, so the decoder input $(\boldsymbol{h}_s^{(L)}, \boldsymbol{h}_c^{(L)})$ determines the pair (s, c) exactly and a sufficiently wide decoder realizes any function of (s, c). Pairwise consistency is exactly the condition for the target-indicator to be such a function: the decoder cannot see which candidate set a pair was drawn from, so it must return the same value for (s, c) at every occurrence, which the hypothesis guarantees is well defined. Realizing it ranks

Table 3: Feature-width ablation on D1, three seeds per row. Hard-query MRR scales with feature count within every model class, and within the all-46 interface XGBoost and NBFNet agree to within seed noise. shell and off report the two flagged sub-strata (minimum reachable hop $L{+}1$ against later or off-shell).

Model	Features	overall	hard	OK	shell	off
XGBoost	core26	0.690	0.120	0.812	0.153	0.090
	hist42	0.844	0.282	0.965	0.463	0.116
	all46	0.899	0.544	0.975	0.646	0.450
GCN	core26	0.716	0.170	0.834	0.226	0.118
	hist42	0.778	0.191	0.904	0.300	0.091
	all46	0.830	0.369	0.928	0.480	0.268
NBFNet	core26	0.839	0.251	0.966	0.454	0.066
	all46	0.898	0.552	0.972	0.638	0.473

Table 4: Conditions for an informative audit. C1 asks whether reachability discriminates and is checkable before training; C2 asks whether the class solves the OK set and needs one training run; conc. is the concentration ratio of §5.4. D2 is the sole C1 failure, by construction (Appendix G).

Graph	hard	C1	C2 (OK MRR)	deg-robust	conc.	verdict
D1	0.177	yes	yes (0.85–0.97)	yes	2.13	informative, class-specific
WIKIPEDIA	0.093	yes	yes (0.42–0.83)	yes	0.95	informative, info-theoretic
TGB-REVIEW	0.561	yes	yes (0.23–0.44)	no	1.40	uniformly stronger baseline
MATHOVERFLOW	0.364	yes	no (0.28–0.43)	weak	1.08	weak / degree-confounded
COLLEGEMSG	0.356	yes	no (0.13–0.38)	weak	0.63	weak / degree-confounded
BITCOIN-OTC	0.550	yes	no (0.22–0.29)	local: no	0.76	silent for local models
EMAIL-EU	0.006	marginal	yes	$n{=}35$	0.92	too few hard queries
D2	pinned	no	–	–	–	inapplicable

every target first and gives $\mathrm{MRR} = 1$. For $\mathcal{M}_L^{\mathrm{anon}}$ the encoder can no longer ignore its aggregate, since $\boldsymbol{h}_v^{(0)}$ is constant; let it instead realize L rounds of injective aggregation, which a sufficiently wide layer can (Xu et al., 2019), so that $\boldsymbol{h}_v^{(L)}$ is a one-hot encoding of $\mathrm{col}_L(v)$. If the L-round WL colours separate the nodes appearing in $\mathcal{Q}$ then col_L is injective on them, $\boldsymbol{h}_v^{(L)}$ again determines v, and the decoder construction above applies unchanged. □

Pairwise consistency is required. An earlier version of this proposition assumed only that no two queries share a source and candidate set while differing in their target, which is weaker and does not suffice. Our own query sets satisfy it and still violate the conclusion. On COLLEGEMSG, queries 1 and 635 share source 594 and have candidate sets differing in 90 of 100 entries, so the weaker hypothesis holds; but query 1's target 1124 is a negative of query 635 and query 635's target 320 is a negative of query 1. Ranking both targets first requires $f(\boldsymbol{h}_{594}, \boldsymbol{h}_{1124}) > f(\boldsymbol{h}_{594}, \boldsymbol{h}_{320})$ and the reverse simultaneously, which no function of (s, c) satisfies; the construction above yields precisely such a function, since $\boldsymbol{h}_v^{(L)} = \boldsymbol{e}_v$ carries no dependence on query time, so it cannot rank both first. Pairwise consistency rules this out and is the condition used in the proof; this finite counterexample has been checked mechanically.

The scope is worth stating precisely, because it does not rescue the diagnostic. Such a cycle caps MRR below 1 for any time-independent scorer, so a structural quantity computable from the query set alone, no BFS, no WL, does bound the achievable accuracy of the static architectures. It is a property of how the evaluation protocol samples negatives rather than of the history graph, it is uncorrelated with hard, and it does not constrain a time-aware encoder, whose $\boldsymbol{h}_s(t)$ takes a different value at each query time, except on conflicting pairs sharing a timestamp. Thus the reading of §4.2 is unchanged: no quantity derived from the history graph or from Definition 1 upper-bounds achievable MRR, but the unqualified form "no structural quantity can" is too strong, and we do not claim it.

Table 5: Two interventions at different points in the pipeline, on one cohort (D1 hard-target-seen, n=5878), three seeds each, query bootstrap. Encoder change: NBFNet minus its source-ablated control. Decoder change: the readout arms of §5.4. Both peak at $h{=}L{+}1$ and both favour the shell over the rest of the flagged set with intervals excluding zero.

		encoder: source-conditioning		decoder: pairwise readout	
Stratum	n	Δ	95% CI	Δ	95% CI
$h{=}L{+}1$ (shell)	2834	$+0.2807$	$[+0.2659, +0.2951]$	$+0.0414$	$[+0.0311, +0.0517]$
$h{=}L{+}2$	1644	-0.0454	$[-0.0519, -0.0391]$	-0.0756	$[-0.0844, -0.0669]$
$h{=}L{+}3$	759	-0.0047	$[-0.0105, +0.0007]$	-0.0298	$[-0.0402, -0.0195]$
$h{\geq}L{+}4$	309	$+0.0047$	$[-0.0046, +0.0142]$	-0.0018	$[-0.0171, +0.0129]$
censored	331	$+0.0019$	$[-0.0025, +0.0060]$	-0.0064	$[-0.0177, +0.0051]$
shell − rest		$+0.3057$	$[+0.2903, +0.3205]$	$+0.0905$	$[+0.0782, +0.1029]$

Table 6: The node-novelty split inside the flagged set. seen/novel: number of flagged queries whose target does and does not appear anywhere in the history graph. Each cell is MRR on hard-node-seen / hard-node-novel, averaged over three seeds. Node-novel rows are trivially identifiable because all negatives are drawn from nodes with a history, and TGN's zero-initialized memory detects exactly that, hence 0.995 on MATHOVERFLOW's novel rows against 0.095 on its seen rows. The receptive-field claim is read off the seen columns. D1 has no novel-target queries.

	n		MRR on hard-node-seen / hard-node-novel				
Graph	seen	novel	GCN	APPNP	SAGE	NBFNet	TGN
EMAIL-EU	15	20	0.096/0.042	0.092/0.073	0.132/0.071	0.068/0.043	0.078/0.039
D1	5955	0	0.194/–	0.200/–	0.186/–	0.252/–	0.159/–
WIKIPEDIA	190	368	0.107/0.020	0.072/0.027	0.100/0.036	0.119/0.061	0.129/0.097
COLLEGEMSG	1392	745	0.131/0.077	0.071/0.152	0.113/0.071	0.054/0.042	0.220/0.017
MATHOVERFLOW	433	1751	0.107/0.049	0.090/0.219	0.092/0.031	0.096/0.076	0.095/0.995
BITCOIN-OTC	812	1190	0.145/0.054	0.109/0.063	0.143/0.038	0.105/0.035	0.216/0.171

Three scope remarks apply to the proposition. It is a statement about the class supremum, not about the outcome of gradient descent. It presupposes a transductive identifier space, in which every node appearing in $\mathcal{Q}$ owns a row of the table; this holds for the benchmarks here, whose candidates carry benchmark identifiers, but not for a strictly inductive setting where unseen nodes arrive at test time, and there only the anonymous clause applies. And it does not claim the diagnostic is vacuous, only that its content is about available evidence rather than representational capacity. The empirical weight therefore lies in §5.2–§5.3.

A failed attempt at a structural ceiling. Proposition 2 needs one hypothesis (pairwise consistency) so a ceiling would have to exploit its violation. There is a natural route. The history graph is fixed, so for a static model $\boldsymbol{h}_s^{(L)}$ is one vector per source, and for an anonymous model it is a function of $\mathrm{col}_L(s)$ alone. On a flagged query $\boldsymbol{h}_s$ carries no information about which candidate is the target, so every query in a source-colour class is scored by one shared ordering of candidate colours. Where those queries have targets of different colours they compete for the top positions, and summing the resulting per-query limits gives a bound computable from BFS and WL before training, for any width and any decoder.

We report the attempt because it fails. Two earlier versions of it were arithmetically wrong in ways a validation script caught and re-reading did not: they substituted a greedy ordering for a maximum, so the quantity sat below the supremum it was meant to bound and an anonymous scorer fitted directly to the flagged queries exceeded it on six of seven graphs; a third omitted the tie convention and was violated on three. The lesson is about verification rather than about the bound, so we state it once here rather than reconstructing the sequence. The surviving bound is sound, the fitted optimum respects it on all seven graphs where we evaluate it (D1, WIKIPEDIA, COLLEGEMSG, MATHOVERFLOW, BITCOIN-OTC, LASTFM, TGB-REVIEW; not the seven of Table 7), and matches it to four

Table 7: The seven graphs carrying the main experiments; Table 10 gives the remaining three. hard is the flagged fraction at $L{=}2$ (computed before training). "bipartite" marks disjoint source/target node types.

Graph	Domain	#nodes	#edges	bipartite	hard
EMAIL-EU	dense social	986	332k	no	0.006
D1	user–item recsys	43k	626k	no	0.177
WIKIPEDIA[‡]	edits (+feat)	9.2k	157k	yes	0.093
COLLEGEMSG	messaging	1.9k	59k	no	0.356
MATHOVERFLOW	Q&A	24.8k	390k	no	0.364
BITCOIN-OTC	trust	5.9k	36k	no	0.550
D2	bipartite recsys	140k	2.0M	yes	degenerate

[‡]WIKIPEDIA is bipartite, so reachability is evaluated at odd depth $L{=}3$ (a source reaches items only at odd hops); all non-bipartite graphs use $L{=}2$.

Table 8: Distance strata among hard-target-seen queries flagged by the reachability clause; collide-only rows are excluded. Columns give the share first reached at $L{+}1$, $L{+}2$, or later within the stored BFS cap. A row is censored when the minimum reachable hop is recorded as $h_{\min} = -1$: farther than that cap or disconnected, not verified "never reachable." Fractions close by construction. Bipartite graphs place their first target shell at $L{+}2$; rows with fewer than 100 eligible queries are omitted. REDDIT is also omitted because the frozen artefacts do not separate its 13 residual rows into collide-only and censored cases.

Graph	L	$n_{\text{RF, target-seen}}$	$L{+}1$	$L{+}2$	$\geq L{+}3$	censored
COLLEGEMSG	2	1382	0.978	0.020	0.002	0.000
MATHOVERFLOW	2	402	0.935	0.052	0.000	0.012
BITCOIN-OTC	2	803	0.827	0.087	0.085	0.001
D1	2	5877	0.482	0.280	0.182	0.056
WIKIPEDIA	3	166	0.000	0.946	0.000	0.054
LASTFM	1	1543	0.000	1.000	0.000	0.000
TGBL-REVIEW	3	4072	0.000	0.000	0.000	1.000

decimals on WIKIPEDIA and MATHOVERFLOW, so it is tight for this class, but it is non-vacuous on one graph only (LASTFM, 0.784). It is also not new: it is the collision clause of Definition 1 applied per query, which Appendix B already shows to be near-vacuous once a model carries an embedding table.

The missing ingredient is coupling across queries, but extracting it appears to require maximizing over orderings of target colours, a linear ordering problem. We can now obtain a weaker form without solving that maximization, which the previous relaxations discarded. Within a source-colour class, any two distinct target colours must be placed in one of three relations by the shared preorder, and each relation forces a measurable loss on one side; taking the maximum over the three is a valid bound on the pair's contribution, and charging each query to at most one pair makes the tightest such bound a maximum-weight matching over target colours, polynomial, and computable from

Table 9: The pairwise-readout test on the three graphs with a large $L{+}1$ shell. Arms differ only in whether the readout sees src→cand path counts; Δ is arm B minus arm A, averaged over three seeds, with a 2000-fold bootstrap CI over queries. ratio is the $L{+}1$ gap divided by the OK gap: on these graphs it never exceeds 1, though the contrast they can support is shell-versus-OK rather than the shell-versus-off-shell one the mechanism predicts (§5.4).

	flagged, shell ($h{=}L{+}1$)		OK		
Graph	n	Δ (95% CI)	n	Δ	ratio
COLLEGEMSG	1359	$+0.0383$ $[+0.027, +0.050]$	3863	$+0.1820$	0.21
MATHOVERFLOW	398	-0.0129 $[-0.031, +0.004]$	3816	-0.0440	0.29
BITCOIN-OTC	666	-0.0307 $[-0.040, -0.022]$	1639	$+0.0160$	-1.91

Table 10: Three further public graphs, added to test claims the first six could not. TGBL-REVIEW (Huang et al., 2023) and REDDIT (Kumar et al., 2019) are featured consumption graphs, added to attempt the class-specificity break on redistributable data (§5.4); LASTFM (Kumar et al., 2019) activates the collision clause (Appendix B). All three are bipartite, so depths are odd. REDDIT is also the only public graph here with no novel-target queries at all, so the negative-sampling asymmetry of §5.3 is absent there by construction rather than by correction.

Graph	Domain	#nodes	#edges	L	hard
TGBL-REVIEW	product reviews	705k	4.9M	1	0.998
				3	0.561
LASTFM	music listening	2.0k	1.29M	1	0.360 (collide 0.114)
				3	0.008 (all novel)
REDDIT	posts to subreddits	11.0k	672k	1	0.095
				3	0.001 (degenerate)

Table 11: The collision clause on LASTFM at L=1, the one graph/depth in our study that activates it (Appendix B). Strata are disjoint: collide-only means the target is reachable but a reachable negative shares its 1-WL colour. The anonymous-class cap is $\mathrm{rr} \leq 2/3$; the trained architectures below carry richer node representations and are reported only to show the empirical stratum difficulty.

Model	MRR by stratum			
	collide-only n=570	unreach-only n=1474	both n=114	OK n=3842
GCN	0.099	0.057	0.050	0.114
APPNP	0.058	0.071	0.042	0.095
SAGE	0.097	0.058	0.040	0.109
NBFNet	0.120	0.122	0.052	0.227

BFS and WL alone. Unlike the per-query bound this can bind when $m_q = 1$ for every query, i.e. where Definition 1's collision clause is silent, since all it needs is two queries in one class whose distinct target colours appear in each other's candidate sets. On COLLEGEMSG it does: among the classes where every $m_q = 1$, five of the sixty largest have an optimum strictly below the per-query bound, by exactly 0.5, one query displaced from rank 1 to rank 2. Exhaustive enumeration over all preorders on the classes small enough to permit it (at most six target colours) finds no violation, and on every class where the matching bound binds it equals the enumerated optimum exactly, so this form of the bound is tight rather than merely sound.

We report it as a partial result rather than the calibrated ceiling. The graph-level number it yields is 0.987 against 0.996 for the per-query bound on COLLEGEMSG, far above the 0.963 an overfitted anonymous scorer actually reaches, so it does not constrain any model we train; and it is silent on MATHOVERFLOW, where no collision-free class binds at all. It settles the question the previous version left open, the cross-query constraint can be kept without solving the linear ordering problem, while leaving the gap between a tight class-level bound and a useful one where it was.

A third remark places it against the capacity literature. The construction needs $d = n$ and a decoder wide enough to realize an arbitrary function of (s, c), so it lives at the top of the capacity scale. Loukas (2020b) show that message-passing networks lose much of their power once the product of depth and width is restricted, and Loukas (2020a) sharpen this to a lower bound on communication capacity, a quantity combining depth, width, message size and global state, of $\Omega(n^2)$ to determine the isomorphism class of connected graphs and $\Omega(n)$ for trees, in expectation over an input distribution as well as in the worst case, and for anonymous and non-anonymous networks alike. Proposition 2 is therefore consistent with those bounds rather than in tension with them: it exhibits an escape that consumes exactly the capacity they show is required. Two consequences for reading this paper. Our encoders use d=16 and $L \in \{2, 3, 4\}$, far below that regime, so the supremum the proposition constructs is not one our models could reach; and the question the proposition closes, a calibrated ceiling for the unrestricted class, is narrower than it first appears, since a capacity-restricted ceiling is not ruled out and the cited bounds are evidence that one exists. We do

Table 12: Degree-matched control on hard-target-seen rows, all six graphs. hard/OK: raw MRR; gap: hard minus OK reweighted to the hard target-in-degree distribution (400-fold bootstrap CI). The right block restricts to strata with ≥ 20 queries on both sides; cov. is 1.00 on D1, 0.84 on WIKIPEDIA, and 0.83–0.99 on the low-signal graphs. EMAIL-EU's 35 hard queries make its row uninterpretable.

				all strata	overlapping strata	
Graph	Model	hard	OK	gap 95% CI	gap	cov.
EMAIL-EU	GCN	0.088	0.460	$-0.153\ [-0.310, +0.047]$	–	–
	APPNP	0.107	0.281	$-0.027\ [-0.132, +0.132]$	–	–
	SAGE	0.090	0.455	$-0.147\ [-0.285, +0.006]$	–	–
	NBFNet	0.092	0.563	$-0.369\ [-0.495, -0.275]$	–	–
	TGN	0.060	0.589	$-0.319\ [-0.416, -0.205]$	–	–
D1	GCN	0.199	0.860	$-0.596\ [-0.605, -0.587]$	-0.596	1.00
	APPNP	0.202	0.866	$-0.603\ [-0.612, -0.593]$	-0.603	1.00
	SAGE	0.183	0.848	$-0.592\ [-0.601, -0.583]$	-0.592	1.00
	NBFNet	0.250	0.966	$-0.711\ [-0.719, -0.704]$	-0.711	1.00
	TGN	0.165	0.643	$-0.389\ [-0.398, -0.380]$	-0.389	1.00
WIKIPEDIA	GCN	0.113	0.705	$-0.395\ [-0.439, -0.344]$	-0.423	0.84
	APPNP	0.090	0.419	$-0.200\ [-0.235, -0.162]$	-0.211	0.84
	SAGE	0.106	0.659	$-0.365\ [-0.411, -0.318]$	-0.396	0.84
	NBFNet	0.116	0.826	$-0.563\ [-0.624, -0.506]$	-0.594	0.84
	TGN	0.133	0.801	$-0.535\ [-0.596, -0.465]$	-0.551	0.84
COLLEGEMSG	GCN	0.124	0.218	$-0.064\ [-0.082, -0.047]$	-0.054	0.89
	APPNP	0.076	0.146	$-0.038\ [-0.050, -0.028]$	-0.040	0.89
	SAGE	0.122	0.215	$-0.036\ [-0.051, -0.020]$	-0.026	0.89
	NBFNet	0.027	0.359	$-0.320\ [-0.339, -0.299]$	-0.318	0.89
	TGN	0.231	0.315	$-0.043\ [-0.063, -0.022]$	-0.044	0.89
MATHOVERFLOW	GCN	0.096	0.438	$-0.060\ [-0.078, -0.042]$	-0.032	0.83
	APPNP	0.082	0.323	$-0.035\ [-0.052, -0.017]$	-0.013	0.83
	SAGE	0.098	0.430	$-0.047\ [-0.064, -0.028]$	-0.031	0.83
	NBFNet	0.057	0.398	$-0.119\ [-0.154, -0.095]$	-0.103	0.83
	TGN	0.083	0.185	$-0.006\ [-0.031, +0.018]$	-0.001	0.83
BITCOIN-OTC	GCN	0.142	0.184	$+0.030\ [+0.014, +0.046]$	$+0.029$	0.99
	APPNP	0.108	0.231	$-0.047\ [-0.065, -0.030]$	-0.046	0.99
	SAGE	0.134	0.165	$+0.030\ [+0.013, +0.048]$	$+0.029$	0.99
	NBFNet	0.070	0.237	$-0.082\ [-0.107, -0.059]$	-0.081	0.99
	TGN	0.215	0.379	$+0.004\ [-0.016, +0.025]$	$+0.003$	0.99

not attempt it here: those results bound graph distinguishability, not ranking accuracy on a query set, and bridging the two is open.

F A GRADED PREDICTOR OF SOURCE-CONDITIONING GAINS

The binary flag marks queries whose target receives nothing propagated from the source. A graded version of the same structural quantity says how much a source-conditioned architecture helps on the queries it can reach. Define the training-free score $s(q) = 1/h(q)$, where $h(q)$ is the minimum hop at which the target becomes reachable; when the stored BFS is censored before reaching it, we set $s(q) = 0$. We measure the per-query benefit of source-conditioning as NBFNet minus an otherwise identical variant with the source indicator ablated, averaged over three seeds.

Which query set this is evaluated on matters, and we report both. On D1's cold-start set (no candidate has any prior interaction with the source; 33.8% of queries) the benefit is monotone in the pre-training score, $+0.87$ at h=1, $+0.67$ at h=2, $+0.29$ at h=3, then -0.05 at h=4 and ≈ 0 beyond, including the censored bin (Figure 7). Spearman correlation is 0.651 ± 0.003, and selecting the top 10% by s gives mean benefit $+0.87$ against $+0.42$ over the full cold-start set (Figure 8). On the flagged set the same quantity is weaker, as it must be: that set excludes every target reachable within L=2, so the two most informative strata are gone by construction and only $h \geq 3$ remains. Spearman is 0.435 ± 0.013, with top-10% benefit $+0.25$ against $+0.12$.

Table 13: Depth-matched diagnostic: each model is trained at depth L and scored against the diagnostic computed at the same L, so Proposition 1 applies exactly. Columns are MRR on the depth-L flagged set. Raising L shrinks the hard fraction but lowers hard-set MRR, so the queries that survive at greater depth are harder rather than easier. EMAIL-EU at L=3 has 24 hard queries and is not interpretable.

Graph	L	hard	GCN	APPNP	SAGE	NBFNet
EMAIL-EU	3	0.004	0.101	0.226	0.147	0.079
D1	2	0.177	0.191	0.194	0.194	0.108
D1	3	0.092	0.131	0.128	0.124	0.065
D1	4	0.043	0.077	0.080	0.074	0.064
WIKIPEDIA	3	0.093	0.060	0.043	0.045	0.143
COLLEGEMSG	2	0.356	0.112	0.090	0.107	0.072
COLLEGEMSG	3	0.130	0.067	0.209	0.061	0.012
COLLEGEMSG	4	0.126	0.040	0.159	0.052	0.080
MATHOVERFLOW	2	0.364	0.074	0.121	0.075	0.093
MATHOVERFLOW	3	0.298	0.064	0.033	0.038	0.011
MATHOVERFLOW	4	0.294	0.020	0.188	0.019	0.044
BITCOIN-OTC	2	0.550	0.082	0.109	0.079	0.113
BITCOIN-OTC	3	0.371	0.055	0.081	0.033	0.016
BITCOIN-OTC	4	0.351	0.059	0.168	0.051	0.088

The gap between those two numbers gives the scope of this appendix: the graded score is a useful router over cold-start queries, where it separates a $+0.87$ stratum from a ≈ 0 one, and a much weaker one over exactly the queries the diagnostic flags. It is a statement about which queries to send to a path-aware model, not about the diagnostic's ceiling, which is why it sits here rather than in §5.

The control preserves a degree signal and is not a null model. Removing the source indicator does not leave a model with no signal. The propagation still runs, but its boundary is then uniform, so the first layer produces one vector shared by every node and the second turns it into a multiple of node degree: the arm becomes a degree ranker. Measured, it is close to one, correlating with target degree at Spearman 0.85–0.96 on the low-signal graphs and 0.33 on D1, against 0.17–0.71 for the arms that keep the indicator; a hand-built degree ranker scores 0.101 on COLLEGEMSG where that arm scores 0.124. So the benefit above is measured against a degree baseline. We therefore also ran the control this suggests, seeding propagation at another query's source under a fixed permutation, which keeps one seeded node and the same recursion and destroys only source identity. On D1, three seeds as elsewhere, it reproduces the effect: benefit $+0.4187$ with strata $+0.864$, $+0.664$, $+0.288$, -0.046 at h=1..4 and Spearman $+0.663$, against $+0.4216$ and $+0.651$ for the degree-baseline control.

The ordering does not survive on low-signal graphs, which bounds the claim. Under the permuted-source control the cold-start benefit is negative on COLLEGEMSG (-0.086 at h=2, -0.025 at h=3), BITCOIN-OTC (-0.080, -0.072) and MATHOVERFLOW (-0.105, -0.026), with Spearman -0.47 to -0.55 where D1 gives $+0.65$. Only the h=1 stratum keeps the predicted sign ($+0.313$ on MATHOVERFLOW). The reading is the one the degree ranker suggests: where no covariate predicts the target without a path, degree is the best remaining signal, and a model that spends its capacity on source-conditioned propagation gives up that signal on precisely the queries with no pair history. The graded score is therefore a router on graphs with usable side-features, the same condition §5.4 places on the binary flag, and we do not claim it beyond them.

G DATASET CONSTRUCTION AND LEAK-FREE SPLITS

For each temporal graph we sort edges by time and place a cutoff so that the last 15% of the timeline forms the query window. The history graph is all edges strictly before the cutoff; message passing, WL colours, reachability, and history-derived features are computed from history only. The four transductive D1 cross-query features are the stated exception. Queries are sampled from the query window with sources that appear in history; each query's candidate set is the true target plus 99

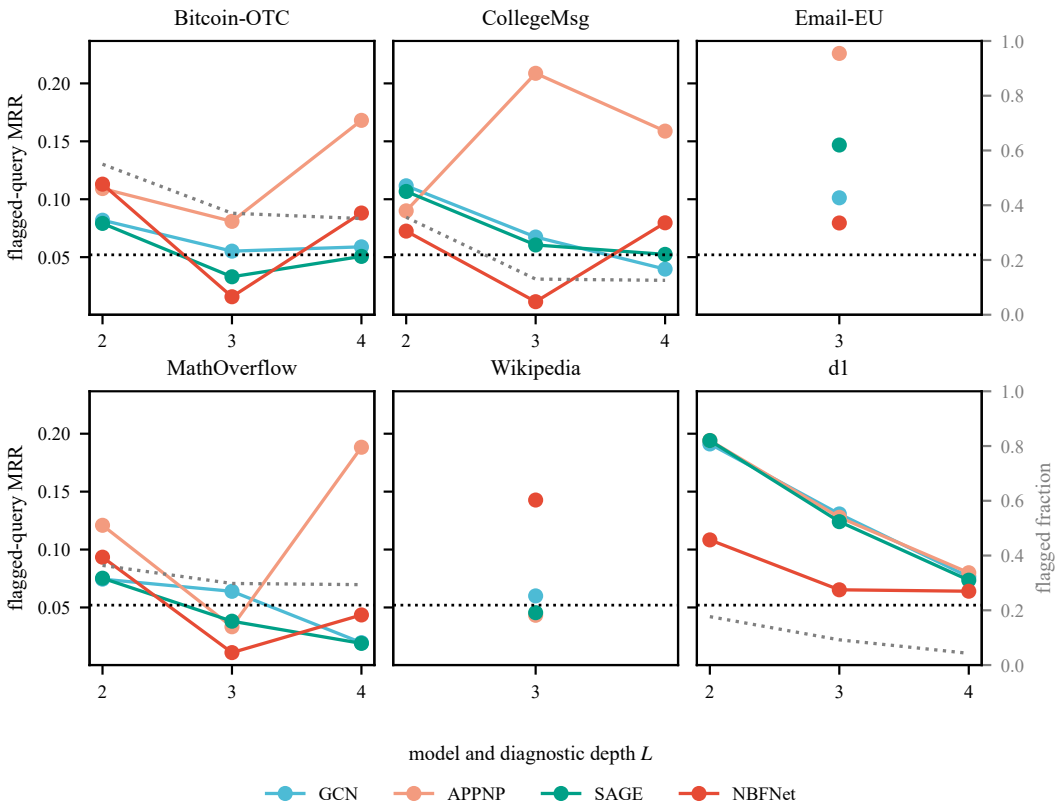

Figure 5: Depth-matched diagnostic. Solid lines: MRR on the depth-L flagged set for a model trained at that same L (left axis). Dotted grey: the hard fraction at that depth (right axis). Raising L shrinks the hard set while lowering hard-set MRR, so depth trades size for difficulty rather than rescuing the flagged queries.

negatives sampled from the pool of history targets, with the target placed at a uniformly random column. We verify 0 leaked (s, d, t) triples for every dataset. For the bipartite graphs (WIKIPEDIA, D2) we offset target ids into a disjoint range so that user id 5 and item id 5 are distinct nodes; omitting this silently merges the two node types (we caught this: without the offset the two id spaces collapse onto one another, so WIKIPEDIA's 7327 users and 984 items are counted as 7327 nodes rather than the correct 9227). D2 is the one graph on which the flag is degenerate: its construction pins hard at 1 for even L, where a bipartite source reaches no target, and at 0 for odd L, where every candidate is reachable, so condition 1 of Table 4 fails. This is not a bipartite artefact: WIKIPEDIA is bipartite too and discriminates cleanly at odd depth.

Negative-sampling artefact check. The overall fraction is stable, but membership in the flagged set can change. The collision term of Definition 1 depends on the candidate set, yet WL collisions are near-zero on nine of the ten graphs, so hard is dominated by the unreachability term (whether the true target lies within L hops of the source) which depends only on (s, p) and the history graph. On D1, hard $= 0.1768$ decomposes as unreachable 0.1768 and collision 3×10^{-5}, so the hard-fraction is effectively negative-independent.

The composition of the hard set is a different matter, and §5.3 reports the datasets on which our own builds get it wrong: on the five graphs constructed from public edge lists, negatives are drawn from nodes that appear as targets in history while positives come from a later window, so a positive can be a node with no history at all. Such a positive is necessarily unreachable (correctly flagged) but it is also the only candidate lacking a history, which is a second, path-free way to identify it. The hard-fraction is therefore still a property of the graph and the split, but on those graphs the flagged set mixes two mechanisms, and any per-model number computed on it inherits the mixture.

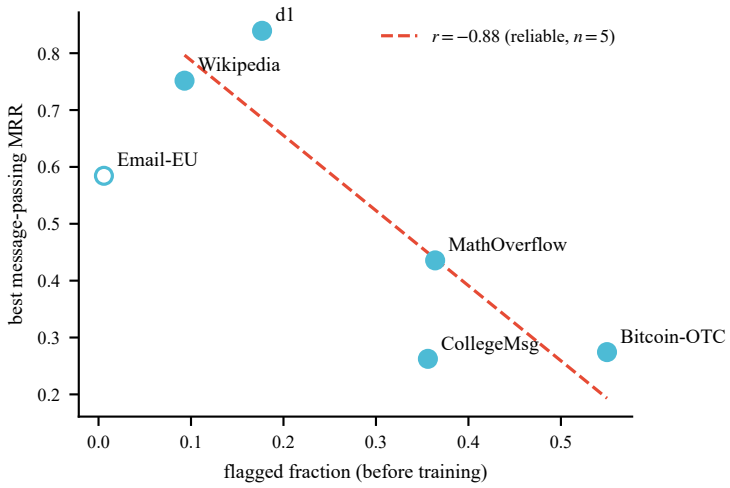

Figure 6: The flagged fraction, computed before training, predicts each graph's realized GNN ceiling (best overall MRR after training), $r = -0.88$ over the five graphs with a reliable hard set. Low hard-fraction graphs admit strong GNNs; high hard-fraction graphs cap them. The correlation might merely reflect density, since a denser graph is both easier and has fewer unreachable targets: on the same five graphs density proxies reach $|r|$ between 0.09 and 0.73 (mean edge multiplicity $+0.73$, average degree -0.43, edges per node -0.09) against -0.88 for the hard-fraction, so the ordering favours the diagnostic but not decisively at $n = 5$, where the Spearman coefficient is only -0.60. We therefore quote the correlation only as consistent with the per-query mechanism, which is where the evidence lives.

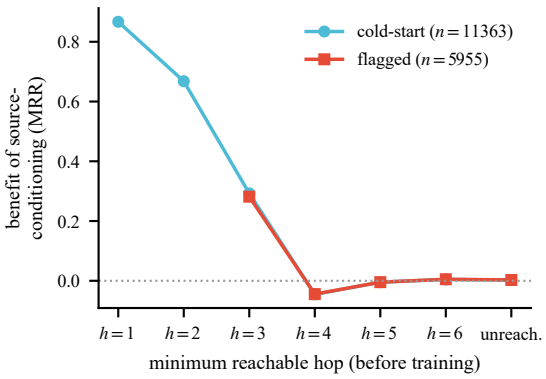

Figure 7: Source-conditioning benefit on D1, stratified by the training-free minimum reachable hop h: $+0.87$ at $h{=}1$ and $+0.67$ at $h{=}2$, then decaying with distance. The final bin is censored beyond the stored BFS cap, not verified unreachable.

Analyses in this paper accordingly restrict to hard queries whose target appears in history wherever the distinction matters, and Table 6 gives the split. D1 and the TGB-REVIEW build, whose negatives come from the full target vocabulary, do not have the problem: novel-positive and novel-negative shares agree to within 0.005 on TGB-REVIEW (0.052 vs 0.047).

H EXPERIMENTAL SETUP

Encoders. The four static architectures use hidden dimension 16 and $L{=}3$ layers unless a depth-matched comparison says otherwise (Appendix I). GCN and GraphSAGE aggregate locally, APPNP

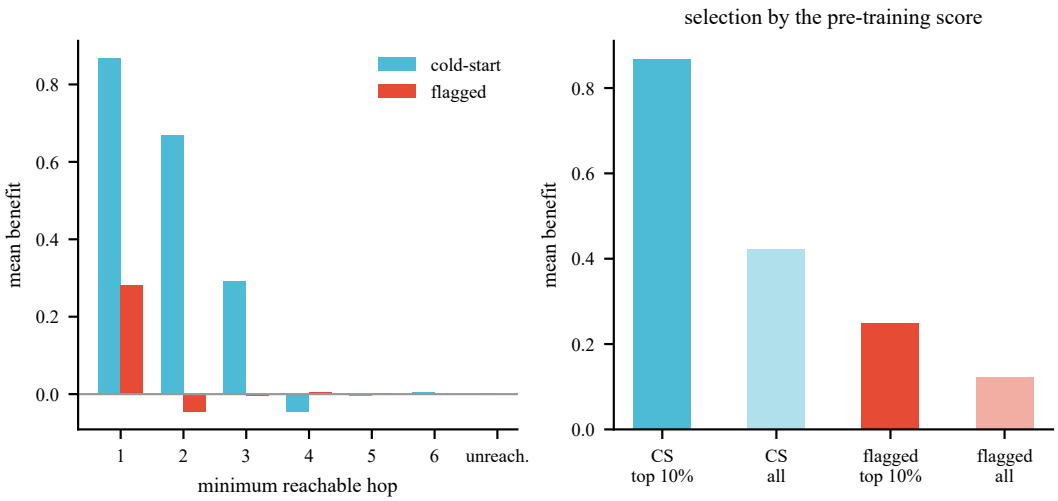

Figure 8: Left: training-free score $s = 1/h$ vs. measured per-query source-conditioning benefit on D1 (Spearman 0.651). Right: risk–coverage, top-scored queries concentrate the largest benefit, above the full cold-start cohort mean.

applies personalized-PageRank propagation, and NBFNet uses a generalized Bellman–Ford recursion with a learned source indicator propagated as batched sparse matrix multiplications, one per layer, independent of batch size. TGN uses hidden dimension 32, two layers of temporal graph attention over the 20 most recent neighbours strictly preceding the query time, a GRU memory updated on every interaction, and harmonic (Time2Vec) encodings of the elapsed time; its neighbour lookup is vectorized by offsetting each node's time-sorted adjacency slice so one binary search serves a whole batch. At evaluation the memory is rebuilt by replaying the full history once and then frozen, so query-window events never enter it and no test event can inform a concurrent or later query. All five share the readout MLP, the AdamW optimizer, a listwise softmax loss over the 100 candidates, and identical query sampling. The query seed is fixed, so only initialization and batch order vary with the reported seed $\in \{11, 23, 37\}$. On graph-only datasets no engineered features are supplied to any model. Static encoders train for 12 epochs on 60,000 sampled training queries in batches of 256 with AdamW (10 epochs on D1 in the depth and width sweeps); TGN trains for 8 epochs at learning rate 10^{-3} and weight decay 10^{-5}. XGBoost uses the pairwise ranking objective for 400 rounds at learning rate 0.05, maximum depth 6, minimum child weight 5, row and column subsampling 0.9 and 0.8, and L_2 penalty 1. Reported MRR is $1/(1 + |\{c : S_c > S_p\}|)$, which counts strictly better candidates and so treats a tie optimistically. We checked this is not a distinction that matters here: across the evaluation sets no row contains a score exactly tied with the positive, or tied within 10^{-6}, so the tie-aware convention gives the same MRR to four decimals. Ties are the one place it could have mattered, since Proposition 1(ii) predicts them; Appendix B reports that comparison directly.

TGN lies outside Proposition 1(i). The memory is updated at every interaction, so a node's state summarizes its past interactions and a later interaction carries that summary onward. Composed over many events this could in principle reach a target outside $\mathcal{B}_L(s)$, so the number of attention layers is not the model's full propagation radius and $\boldsymbol{h}_v^{(L)}$ is not a function of $\mathcal{B}_L(v)$ alone. Establishing L-locality for this architecture would require an argument about the memory recursion that we do not have, so we treat its collapse on flagged queries (§5.2) as evidence about a model the theorem does not cover.

Diagnostic depth. The diagnostic is reported at L=2 for the non-bipartite graphs and at L=3 for the bipartite WIKIPEDIA and TGB-REVIEW, where a source reaches targets only at odd depth; on TGB-REVIEW L=1 is degenerate (hard = 0.998) for the same reason. Appendix I reports every model against a diagnostic computed at its own depth.

Non-message-passing baselines. XGBoost with a pairwise ranking objective is trained on queries sampled from history; its history-derived features use only the corresponding history. It is run in nested feature sets so a break can be attributed: struct uses path and degree structure (2- and 3-hop path counts, common neighbours, endpoint degrees, pair history), pop adds popularity, recency and short/mid/long-window target counts, and on WIKIPEDIA a further variant adds the graph's 172-dimensional edge features compressed to 16 dimensions by SVD. On D1 the baseline uses that dataset's 46 engineered features: 26 core features shared with the GNN arms, 16 additional structural features, and four transductive cross-query aggregates computed from the evaluation candidate batch. BUDDY-lite implements subgraph sketching as exact distance-dependent neighbourhood counts $|\{v : d(s,v){=}i, d(c,v){=}j\}|$ for $i, j \in \{1, 2\}$ plus the capped shortest-path distance, deliberately excluding degree and popularity so that it isolates path evidence.

I THE DEPTH SQUEEZE: REACHABILITY VERSUS OVER-SMOOTHING

DEPTH-MATCHED SWEEP

Proposition 1 is stated for a fixed depth L, so a diagnostic computed at one depth says nothing about a model trained at another. This matters in both directions: it is a soundness requirement, and it also provides the natural rescue test: increase L until the target enters the receptive field. We match depth in the sense the proposition requires, the diagnostic's L equals the number of aggregation rounds, and claim nothing stronger. A GCN layer, an NBFNet Bellman–Ford round and an APPNP propagation round are not equivalent units of capacity, so this sweep isolates receptive-field radius rather than holding capacity fixed; the width experiment below is the capacity control. We therefore train the four static architectures at $L \in \{2, 3, 4\}$ and evaluate each model against the diagnostic computed at its own depth (Table 13, Figure 5).

Deepening refines the partition on schedule; the queries left inside gain nothing. On D1 the hard fraction falls from 0.177 at $L{=}2$ to 0.092 at $L{=}3$ to 0.043 at $L{=}4$: more targets become reachable. But MRR on the depth-matched hard set falls too: GCN $0.191 \to 0.131 \to 0.077$, and likewise for the other architectures, so the queries that remain flagged at larger L are harder still, not easier. This is the mechanism's prediction: raising L removes from the hard set exactly the queries whose targets sit at moderate distance, leaving a residue of genuinely distant ones, and it does not propagate source evidence into that residue. The same pattern holds on COLLEGEMSG ($0.356 \to 0.130$ hard fraction, GCN $0.112 \to 0.067$), MATHOVERFLOW and BITCOIN-OTC. Depth is therefore not a rescue but a trade: a smaller flagged set whose members are less tractable.

A capacity-side argument points the same way, and we treat it as secondary: the normalized Dirichlet energy of the propagation operator decays geometrically in L (0.084 at $L{=}1$ to 0.0017 at $L{=}6$ on D1) while reachability needs $L \approx 5$–6 to cover most targets, so the two never overlap usefully. We report it as a naive-operator rate, residual connections and normalization slow the decay, and rely on the depth-matched measurements above for the claim itself.

Table 14 gives the two curves of Figure 9 on D1: the fraction of cold-start targets within L hops of the source, and the normalized Dirichlet energy $\mathcal{E}(\mathbf{H}^{(L)})/\mathcal{E}(\mathbf{H}^{(0)})$ of the symmetric-normalized propagation operator applied to random node signals (a standard over-smoothing measure; Cai & Wang, 2020). Most targets only come within reach around $L \approx 5$–6, by which depth energy has decayed below 0.3%. This is an operator property independent of trained weights; residual connections and normalization can slow the decay, so we report it as the naive-operator rate and use it only to explain the failure of deeper GCNs on D1 (measured cold MRR $0.341 \to 0.357$ from 2 to 5 layers).

Table 14: Depth squeeze on D1: reachability rises, distinguishability (Dirichlet energy ratio) falls; they never coincide in a useful region.

L	1	2	3	4	5	6
target reachable	0.01	0.44	0.71	0.86	0.94	0.97
energy ratio	0.084	0.022	0.009	0.005	0.003	0.0017

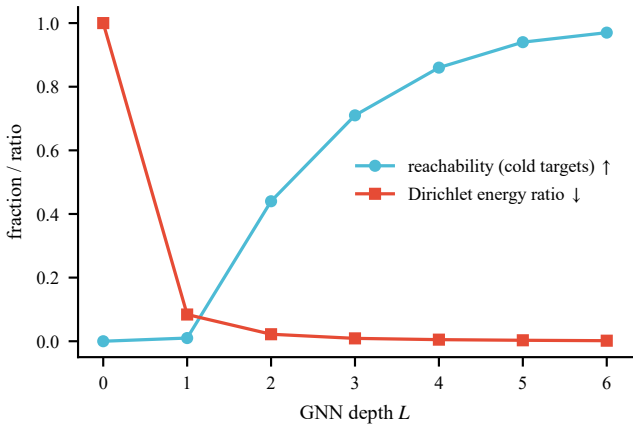

Figure 9: The depth squeeze on D1. Reachability of cold targets rises with depth L (blue), but the propagation operator's normalized Dirichlet energy, node distinguishability, decays geometrically (red). No depth both reaches distant targets and preserves distinguishability.

J PER-GRAPH MULTI-SEED TABLES

UNDER-TRAINING CHECK

Were the collapse a training artefact, seeds would scatter. On D1 they do not: $\mathrm{std} \leq 0.008$ for every architecture (GCN 0.194 ± 0.004, APPNP 0.200 ± 0.002, SAGE 0.186 ± 0.004, NBFNet 0.252 ± 0.002), and the smallest hard-vs-OK gap ($0.856 - 0.194 = 0.662$ for GCN) exceeds 100 seed standard deviations, so the separation is significant without a formal test; WIKIPEDIA's local models agree ($\mathrm{std} \leq 0.015$). On the low-signal graphs they do scatter, and we report it rather than average it away: NBFNet has $\mathrm{std} = 0.037$ on BITCOIN-OTC and 0.032 on COLLEGEMSG, and APPNP on MATHOVERFLOW is 0.193 ± 0.106, one seed reaches 0.27, five times the random baseline, so "collapses to near-random" is false there for some runs. This is the boundary the degree control finds independently (§5.3): with no usable side-features there is little signal anywhere, so which optimum a run lands in is close to arbitrary. The under-training objection is refuted where the diagnostic is informative; where it is not refuted, the diagnostic makes no claim.

Table 15 gives overall MRR, over all queries, not the partition, for every architecture, which is the quantity the hard-fraction correlation of §5.2 is computed against and the source of the D1 comparison described in §1 (0.899 for the non-message-passing baseline against 0.839 for the best of five encoders). It also shows that no encoder is a weak baseline on the graphs carrying the claim. Table 16 gives the full multi-seed flagged and OK MRR for D1 and BITCOIN-OTC (three seeds each). Standard deviations are below 0.008 except NBFNet on BITCOIN-OTC (0.037); every flagged mean is below its corresponding OK mean.

Table 15: Overall MRR (all queries), mean±std over seeds $\{11, 23, 37\}$, with the best non-message-passing baseline for reference. D1's 0.899 against 0.839 is the gap quoted in §1.

Graph	hard	GCN	APPNP	SAGE	NBFNet	TGN	best non-MP
EMAIL-EU	0.006	0.453±0.010	0.272±0.025	0.440±0.020	0.555±0.004	0.584±0.005	0.548 (XGB-path)
D1	0.177	0.739±0.004	0.746±0.002	0.730±0.004	0.839±0.000	0.552±0.006	0.899 (XGBoost)
WIKIPEDIA	0.093	0.614±0.023	0.307±0.055	0.559±0.036	0.752±0.008	0.732±0.006	0.767 (XGB-path)
COLLEGEMSG	0.356	0.172±0.009	0.120±0.013	0.166±0.007	0.263±0.021	0.255±0.006	0.300 (BUDDY)
MATHOVERFLOW	0.364	0.265±0.053	0.250±0.025	0.290±0.010	0.296±0.030	0.436±0.016	0.310 (BUDDY)
BITCOIN-OTC	0.550	0.159±0.024	0.153±0.001	0.144±0.028	0.167±0.035	0.274±0.009	0.195 (BUDDY)

K THE INFORMATION-THEORETIC REGIME ON LOW-SIGNAL GRAPHS

In our fixed-radius comparisons on graphs where no measured covariate predicts the target without a path, the flagged set is difficult across the tested model families rather than a boundary of one.

Table 16: Full multi-seed flagged / OK MRR (mean±std, seeds $\{11, 23, 37\}$).

	D1 (hard=0.177)		BITCOIN-OTC (hard=0.550)	
Model	hard	OK	hard	OK
GCN	0.194 ± 0.004	0.856	0.091 ± 0.006	0.243
APPNP	0.199 ± 0.002	0.863	0.082 ± 0.008	0.240
SAGE	0.186 ± 0.004	0.847	0.081 ± 0.008	0.221
NBFNet	0.252 ± 0.002	0.966	0.063 ± 0.037	0.294

Restricted to hard-node-seen queries and measured by the concentration ratio of §5.4, the non-message-passing baselines land at 1.11 (LASTFM), 1.08 (MATHOVERFLOW), 0.97 (WIKIPEDIA and REDDIT), 0.92 (EMAIL-EU), 0.77 (BITCOIN-OTC) and 0.66 (COLLEGEMSG), i.e. at or near parity and five of them below it, so whatever edge those baselines have is not specific to the flagged queries. TGB-REVIEW is the one public graph above that band at 1.40, and §5.4 shows that it does not qualify either: its baseline beats the message-passing models by 2–3× in every target-degree band, flagged or not. D1 (2.13) is the only case where the advantage is confined to the flagged set.

Two observations explain this pattern. First, a concentration near 1 is the default for a stronger model that improves everywhere, and the flagged set has no special claim on its advantage. Second, the condition that produces a concentrated advantage is narrow. It needs both (i) a covariate that predicts the target when no path exists (short-window popularity on a recommendation graph) and (ii) an OK set the message-passing models already solve, so the baseline has nowhere else to gain. D1 satisfies both (OK-set MRR 0.97); TGB-REVIEW satisfies (i) but not (ii), so its baseline gains uniformly rather than specifically. This makes class-specificity harder to demonstrate than we expected, and it is the reason we report it on one graph.

L THE SECOND-GRAPH SEARCH IS NARROW

The break of §5.4 rests on D1, which we cannot redistribute, so the obvious extension is to find a public graph with the same internal structure. Its requirements are testable before any training, from the diagnostic's own BFS, and we report the search because the failures are informative in a way a single success would not be.

Three conditions are checkable from the mask alone. (i) Discrimination: hard must not sit at 0 or 1. (ii) Spread: the flagged rows must not all sit at one distance, we use the share first reachable at exactly $L{+}1$, which must be below 0.80. (iii) Stratum size: the second-largest stratum needs enough rows to measure a difference; we require 300. On D1 all three hold ($\mathrm{hard} = 0.177$, share at $L{+}1$ equal to 0.482, strata of $2834/1644/1068$), and (iii) is not implied by (ii): BITCOIN-OTC passes (ii) at 0.822 and still fails, because its second and third strata hold 70 and 62 rows.

The two failure modes point in opposite directions. Six of our graphs fail (ii), between 82% and 98% of flagged rows sit at exactly $L{+}1$, so there is no gradient to measure inside the flagged set. Against that, a candidate chosen for having genuinely path-free covariates, the TGB flight network (Huang et al., 2023) whose nodes carry airport coordinates, fails (i) instead and decisively: with 16,344 airports and average degree 2252, hub-and-spoke connectivity puts almost every airport within two hops and $\mathrm{hard} = 0.001$, leaving two usable rows. Shrinking the history window twelvefold moves it only to 0.004, so this is the graph's density rather than an artefact of the split. A useful graph must therefore be sparse enough to flag queries and irregular enough that the flagged queries span multiple distances. The first requirement is predicted by whether $\overline{d}^2$, the squared average degree, is small against n, which predicted the flight failure before we built it.

The most informative attempt is the one selected to satisfy every criterion we could check in advance. The TGB token-transfer network has a single node type, high recurrence, and $\overline{d}^2$ around 3.7×10^3 against 6.4×10^5 nodes, so we expected (i) to hold, and it does: $\mathrm{hard} = 0.129$. It fails (ii) and (iii) anyway, 86.8% of flagged rows at $L{+}1$ and a second stratum of 20, because 64% of its flagged targets never appear in history, leaving 363 target-role-seen rows of which 315 sit at one distance. That is the useful negative result: across nine graphs, none of node count, average degree, recurrence, node-type count or feature availability predicted spread, and only building the mask revealed it. We

therefore stop rather than screen a tenth graph, and we do not have a cheap proxy to offer for the property that turns out to matter most.

One explanation suggests itself and does not survive contact with the data. If ball growth were geometric, almost all reachable targets would land within one hop of the mean distance, so whatever spilled past L would pile up at $L{+}1$ and condition (ii) would fail on any small-world graph. We measured it: over the graphs where both quantities are available, the concentration of reachable targets at their modal distance is uncorrelated with the spread surviving inside the flagged set (Spearman $+0.09$), and the counterexamples are clean: COLLEGEMSG concentrates only 0.39 of its reachable targets at one distance yet retains spread 0.02, while WIKIPEDIA concentrates 0.84 and retains 0.19. D1 is distinguished instead by the way its hop distribution decays just outside the flagged radius: its flagged rows fall off gradually ($2834/1644/759/309$ at $h{=}3,\dots,6$) where the failing graphs fall off a cliff ($1359 \to 28$ on COLLEGEMSG, $398 \to 26$ on MATHOVERFLOW).

We tried to turn that into the cheap proxy the previous paragraph says we lack, and report the attempt because it fails in a way worth knowing. The obvious statistic is circular: spread is $1 - n(L{+}1)/n_{\mathrm{hard}}$, so any decay ratio built from the same histogram shares its dominant bin, and the naive version correlates at $\rho = +0.96$ for that reason alone. A version sharing no bin with spread, the mass at $h \geq L{+}3$ relative to the $h{=}L{+}2$ bin, still correlates at $+0.89$ over the seven graphs with a usable histogram, so the relation is not purely definitional. But it does not survive as a predictor. Fitting it on half the graphs and predicting the other half over 200 random splits gives mean absolute error 0.122 against 0.153 for predicting the mean: a 21% reduction, on an error that is a quarter of spread's full range. With $n{=}7$, and with the six graphs other than D1 all falling in 0.02–0.18, the correlation rests on a single leverage point. We therefore keep the decay pattern as a description of the feature unique to D1, and we do not offer it as a screening rule.

M THE READOUT EXPERIMENT: SHELL CONTRAST AND ITS LIMITS

Motivation and the two arms. Among flagged queries on the low-signal graphs, 82–98% have the target at exactly $L{+}1$ (Table 8), so the path lies inside a depth-$(L{+}1)$ model's receptive field, yet counting paths between two endpoints is precisely what a readout over independent node embeddings cannot do. That literature therefore predicts a pairwise readout should recover those rows specifically. Both arms share encoder, depth, width, optimizer, seeds and queries, and differ only in the readout's input: $\mathrm{MLP}([\boldsymbol{h}_s, \boldsymbol{h}_c])$ against $\mathrm{MLP}([\boldsymbol{h}_s, \boldsymbol{h}_c, \text{path counts}])$, three seeds each. §5.4 reports the pairwise-readout arms on three public graphs and finds the gain smaller on the flagged shell than on the OK set. That comparison is forced: those graphs place 83–98% of flagged rows at exactly $L{+}1$, so their off-shell strata hold 28–148 rows and the contrast the mechanism actually predicts, shell against the rest of the flagged set, cannot be estimated. D1 is the one graph where it can, with 2834 shell rows against 3044 off-shell.

We run the same two arms there, differing only in whether the readout sees source-to-candidate path counts, at the depth and width of the paper's other D1 runs and with the graph-only feature interface, so no engineered covariate can recover the target by a path-free route. Confidence intervals are a query bootstrap over 4000 resamples: the estimand is a per-query mean, and with three seeds a seed interval is too wide to resolve effects of this size. Table 17 gives the result.

The ordering inside the flagged set matches the labeling-trick argument and the graded diagnostic's prediction. The shell gains $+0.041$; one hop further out the readout hurts (-0.076); further still the harm shrinks toward zero, including in the censored bin, as the path features become uninformative. The shell-minus-off-shell difference is $+0.091$ with a bootstrap interval excluding zero, and it is positive on each of the three seeds individually ($+0.099$, $+0.037$, $+0.135$), the per-seed spread is wide, so the query bootstrap rather than the seed interval carries the claim.

This result does not yield a design rule. The same readout is worth $+0.188$ on rows the diagnostic never flagged, four and a half times the shell gain, so a practitioner told to add a pairwise readout "where the diagnostic fires" would be spending effort in the wrong place. The diagnostic ranks flagged queries correctly and still does not identify where the remedy pays most, because the constraint it measures, absent path evidence, is not the same thing as headroom.

Table 17: Pairwise readout on D1, stratified by the training-free minimum reachable hop, three seeds, query bootstrap. Inside the flagged set the gain is ordered as the mechanism predicts; against the unflagged complement it is $4.5\times$ larger.

Stratum	n	Δ (arm B $-$ arm A)	95% CI
flagged, shell ($h{=}L{+}1$)	2834	$+0.0414$	$[+0.0308, +0.0525]$
flagged, $h{=}L{+}2$	1644	-0.0756	$[-0.0844, -0.0674]$
flagged, $h{\geq}L{+}3$	1068	-0.0217	$[-0.0302, -0.0134]$
flagged, censored	331	-0.0064	$[-0.0183, +0.0051]$
OK (not flagged)	27718	$+0.1881$	$[+0.1847, +0.1915]$
shell $-$ off-shell		$+0.0905$	$[+0.0784, +0.1033]$

N FLAG-GATED ROUTING DOES NOT PAY

If the diagnostic identifies queries where message passing has no evidence, the obvious application is to route: send flagged queries to a non-message-passing predictor and the rest to the GNN. We report this because it fails for a reason that follows directly from §5.4 and that a reader would otherwise have to rediscover.

Table 18 evaluates the oracle version of that router, per half, take whichever model is actually better, which is an upper bound on any real gating rule, against the better of the two models used alone. The mean gain is $+0.0017$ and only COLLEGEMSG exceeds $+0.005$. The cause is visible in Table 2: the non-message-passing baseline wins both halves on eight of nine graphs, so the router selects the same model twice and cannot beat it. Routing needs the two model families to be complementary (each better on one half) and a concentration ratio above 1 on its own does not supply that. COLLEGEMSG is the only graph with $\mathrm{conc} < 1$ and a message-passing model that wins the flagged half, and it is the only graph where routing gains anything at all.

This is a stronger statement than "the diagnostic is not a router". Under the standard reading of a hard/easy partition, flag the queries a class cannot handle, hand them to something else, the partition would have to identify queries where the other family is relatively stronger. In contrast, §5.4 measures whether the other family is absolutely stronger there, and on these graphs the answer is yes almost everywhere, flagged or not. The two criteria differ, and the routing failure separates them.

Table 18: Oracle flag-gated routing versus the best single model. "MP alone" and "non-MP alone" are hard/OK MRR recombined at each graph's hard fraction; "router" takes the better model on each half, an upper bound on any real gating rule. Computed from Table 2's inputs, no retraining.

Graph	hard	MP alone	non-MP alone	best single	router	gain
D1	0.177	0.839	0.899	0.899	0.899	$+0.000$
WIKIPEDIA	0.093	0.756	0.788	0.788	0.788	$+0.000$
LASTFM	0.360	0.182	0.257	0.257	0.257	$+0.000$
COLLEGEMSG	0.356	0.323	0.375	0.375	0.391	$+0.015$
MATHOVERFLOW	0.364	0.313	0.428	0.428	0.428	$+0.000$
BITCOIN-OTC	0.550	0.289	0.371	0.371	0.371	$+0.000$
EMAIL-EU	0.006	0.585	0.686	0.686	0.686	$+0.000$
TGB-REVIEW	0.561	0.323	0.469	0.469	0.469	$+0.000$
REDDIT	0.095	0.852	0.881	0.881	0.881	$+0.000$
mean						$+0.0017$

O WHICH CONFOUNDS COULD EXPLAIN THE GAP, AND WHICH ARE RULED OUT

Flagged queries may simply be hard for reasons unrelated to reachability: an inactive source, a sparse neighbourhood, a different connected component, a pair never seen before. This appendix

measures which of these covariates could account for the gap. §5.3 reweights target in-degree and §5.3 controls target novelty, but neither addresses the rest, and a reader is right to ask about them.

We proceed in two steps. First, which covariates are even capable of explaining the gap: if the flagged and unflagged sets carry the same distribution over a covariate, it cannot account for any difference between them, whatever the model. We report the overlap coefficient $\sum_b \min(p_b, q_b)$ on shared log-spaced bins, 1 indistinguishable and 0 disjoint (Table 19). Second, for the covariate that matters most, the matched gap itself.

One covariate is ruled out at the first step. Source and target lie in the same connected component at essentially the same rate on both sides (overlap ≥ 0.972 everywhere), so component structure is not the mechanism, unsurprising, since a target in another component is unreachable at every depth and such queries are flagged uniformly rather than selectively.

Source degree is not ruled out this way. It is imbalanced on all six graphs, D1 included, where flagged queries have median source degree 9 against 17 on the complement (overlap 0.774); on MATHOVERFLOW the imbalance is severe (20 against 276). Source activity behaves similarly, and target degree shows the imbalance that motivated §5.3 in the first place, its appearance here is a check that the measurement detects a confound we already knew about. So the reading has to be tested rather than dismissed: low-degree sources are both likelier to have distant targets and plausibly harder to predict for unrelated reasons, and matching on target in-degree leaves the source-degree distribution untouched.

The gap survives source-degree matching. We therefore recompute the gap reweighted onto the flagged set's source-degree distribution, using the same stratify-and-reweight estimator as §5.3 and the cached per-query reciprocal ranks, so no retraining is involved (Table 20). It survives everywhere it can be measured: all 25 rows with a defined estimate keep a negative gap whose source-clustered interval excludes zero. On D1 the matched gap is -0.646 for GCN and -0.720 for NBFNet at full coverage, against raw gaps of -0.663 and -0.712, source-degree matching moves the estimate by less than a hundredth. EMAIL-EU yields no usable stratum at 15 flagged queries and is reported as undefined rather than as a gap.

Clustered uncertainty. The bootstrap in §5.3 resamples queries, which treats queries sharing a source as independent when they are not. We therefore also resample whole sources with replacement. The clustered intervals are wider, by 0.98–$3.45\times$, median $1.60\times$, which confirms the concern is real and quantifies it. No interval crosses zero, on any graph or architecture, so the widening changes no conclusion.

The remaining uncertainty is narrower than before: source activity, pair-history frequency and candidate popularity are imbalanced and unmatched, and we have matched one covariate rather than audited all of them. The mechanism attribution on the feature-carrying graphs rests on the target-degree control, this source-degree control, the depth-matched sweep and the width control.

Table 19: Distributional overlap between the flagged and unflagged sets for each covariate, on shared log-spaced bins (1 indistinguishable, 0 disjoint). Only same comp is high everywhere, ruling out component structure; source degree and source activity are imbalanced on every graph and remain open confounds. EMAIL-EU has 15 flagged queries and its row is uninterpretable.

Graph	n_{flag}	src deg	src act	tgt deg	pair hist	cand pop	time	same comp
D1	5878	0.774	0.879	0.713	0.868	0.953	0.919	0.989
WIKIPEDIA	190	0.810	0.587	0.592	0.687	0.859	0.743	0.972
COLLEGEMSG	1384	0.707	0.762	0.730	0.769	0.946	0.867	1.000
MATHOVERFLOW	404	0.516	0.629	0.496	0.930	0.831	0.904	0.995
BITCOIN-OTC	808	0.634	0.803	0.620	1.000	0.873	0.883	1.000
EMAIL-EU	15	0.350	0.272	0.329	0.718	0.645	0.741	1.000

Table 20: The gap after matching on source degree, with source-clustered bootstrap intervals. raw: unmatched flagged-minus-unflagged MRR. src-matched: reweighted onto the flagged set's source-degree distribution. cov.: share of flagged queries in strata usable on both sides. width×: clustered interval width over the naive query-bootstrap width. EMAIL-EU has no usable stratum and is omitted.

Graph	Model	raw gap	src-matched	cov.	clustered CI	width×
D1	GCN	−0.663	−0.646	1.00	[−0.659, −0.634]	1.70
	APPNP	−0.665	−0.649	1.00	[−0.662, −0.637]	1.73
	SAGE	−0.662	−0.643	1.00	[−0.654, −0.629]	1.74
	NBFNet	−0.712	−0.720	1.00	[−0.733, −0.708]	1.56
	TGN	−0.478	−0.438	1.00	[−0.450, −0.425]	1.72
WIKIPEDIA	GCN	−0.559	−0.502	0.81	[−0.588, −0.429]	1.65
	APPNP	−0.263	−0.228	0.81	[−0.294, −0.179]	1.62
	SAGE	−0.510	−0.452	0.81	[−0.537, −0.377]	1.92
	NBFNet	−0.702	−0.631	0.81	[−0.718, −0.558]	1.43
	TGN	−0.667	−0.600	0.81	[−0.690, −0.533]	1.60
COLLEGEMSG	GCN	−0.074	−0.091	0.99	[−0.123, −0.059]	2.63
	APPNP	−0.061	−0.065	0.99	[−0.089, −0.048]	2.44
	SAGE	−0.090	−0.102	0.99	[−0.138, −0.073]	2.78
	NBFNet	−0.326	−0.376	0.99	[−0.445, −0.323]	3.45
	TGN	−0.093	−0.089	0.99	[−0.141, −0.040]	2.93
MATHOVERFLOW	GCN	−0.271	−0.318	0.77	[−0.345, −0.292]	1.08
	APPNP	−0.190	−0.172	0.77	[−0.197, −0.140]	1.12
	SAGE	−0.335	−0.360	0.77	[−0.383, −0.332]	1.09
	NBFNet	−0.322	−0.372	0.77	[−0.403, −0.337]	1.13
	TGN	−0.131	−0.184	0.77	[−0.204, −0.163]	0.98
BITCOIN-OTC	GCN	−0.098	−0.175	0.92	[−0.213, −0.142]	1.34
	APPNP	−0.133	−0.099	0.92	[−0.126, −0.066]	1.08
	SAGE	−0.079	−0.154	0.92	[−0.187, −0.124]	1.22
	NBFNet	−0.186	−0.275	0.92	[−0.312, −0.242]	1.09
	TGN	−0.164	−0.210	0.92	[−0.247, −0.173]	1.17

P SOURCES OF PERFORMANCE ON THE TWO HALVES: FEATURES, MEMORIZATION, AND MESSAGE PASSING

Proposition 1(i) says only that the target's representation receives nothing propagated from the source; it explicitly leaves the decoder free to use $\boldsymbol{h}_s$ and the endpoint features. The remaining issue is the source of accuracy on each half of the partition. We test it by ablating node representation while holding the encoder, depth, width, optimizer, seeds and queries fixed (Table 21). Three arms against a control: the engineered feature rows permuted across queries, which destroys the feature-target link while preserving every marginal distribution; the features zeroed; and the free node-embedding table zeroed and frozen.

The feature arms are defined only on D1, the one graph whose GNN receives engineered features; the public graphs run graph-only, where the feature array has zero width and shuffling it is a no-op. WIKIPEDIA therefore contributes the embedding arm, which is the memorization check on a second graph. All arms use the same budget as each other and none substitutes for the paper's own GCN row.

Features carry the unflagged half, not the flagged half. On D1, permuting the feature rows costs 0.4926 on unflagged queries against 0.0417 on flagged ones, a factor of $11.8\times$. Zeroing the features instead of shuffling them gives 0.4082 against 0.0125, a factor of $32.5\times$. The asymmetry is the point: the unflagged half is largely feature-driven, and the flagged half has almost no feature signal to lose. The diagnostic asserts this pattern, and the ablation measures it rather than relying on argument; it also explains the class-specificity break of §5.4, which needs features predictive of the target without a path rather than merely rich ones, though Appendix Q shows even that is not sufficient.

Memorization contributes little, and the two graphs disagree on its sign. The natural alternative reading of the small positive MRR on flagged queries is that the embedding table memorizes popular targets. If so, zeroing it should cost most there. The result does not support that reading, but it is not clean either. On WIKIPEDIA the flagged half improves by 0.0372 when the table is removed while the unflagged half loses 0.1367: there, the memorization channel is actively harmful where no propagated evidence exists. On D1 it goes the other way (the flagged half loses 0.0427 against 0.0289 on the unflagged half) so the table does supply a little of the flagged half's accuracy on that graph.

The common feature is the magnitude. The embedding effect on the flagged half is 0.0427 on D1 and 0.0372 on WIKIPEDIA, an order of magnitude below the 0.4926 that destroying the feature-target link costs on the unflagged half. So memorization does not account for the flagged half on either graph, but we cannot claim it is uniformly harmful there; one graph says it helps slightly and the other says it hurts. We report the split rather than the more quotable half of it. Read against §4.2, which shows that a free embedding table could still yield a perfect ranking on any pairwise-consistent query set, the message is that a trained table, one that must generalize to unseen future edges rather than memorize them, realizes almost none of that capacity on precisely the queries where the escape would have to operate.

Table 21: Feature / ID ablation, GCN, three seeds, arms differing only in the representation assigned to a node. Δ columns are the change from the control arm on each half of the partition. Destroying the feature-target link costs an order of magnitude more on the unflagged half; zeroing the embedding table has a small, dataset-dependent effect there (it helps on WIKIPEDIA and hurts on D1). Feature arms are defined only on D1, whose GNN receives engineered features.

Graph	Arm	flagged	unflagged	Δ flagged	Δ unflagged
D1	full (control)	0.1990	0.8606	—	—
	features shuffled	0.1573	0.3680	-0.0417	-0.4926
	features zeroed	0.1864	0.4524	-0.0125	-0.4082
	embedding table zeroed	0.1563	0.8317	-0.0427	-0.0289
WIKIPEDIA	full (control)	0.1153	0.6271	—	—
	embedding table zeroed	0.1525	0.4904	$+0.0372$	-0.1367

Q A PUBLIC BENCHMARK FOR THE CLASS-SPECIFICITY CONDITION

The break of §5.4 holds on D1, a graph we cannot redistribute, and appears on none of eight public graphs. Rather than continue searching, we test the condition that §5.4 identifies, not rich features, but features predictive of the target in the absence of a path, on a construction whose every input is public.

The history graph is real and unmodified: the same edges, cutoff, id space and query sources as our COLLEGEMSG results, so degree distribution, sparsity, components and temporal ordering are untouched. Only the query targets are synthesized, by a mixture with one knob α: with probability α the target maximizes a pairwise function of node latents, which is path-free by construction; otherwise it maximizes the number of length-$\leq L$ walks from the source, which is path-dependent. Neither family receives information the other lacks, both are given the same node latents, but their interfaces to that information differ, deliberately: the baseline gets the pairwise transformation computed for it as a feature, while the message-passing arm holds the latents as frozen node inputs and must recover the transformation from two independent node representations through its readout. We state this explicitly because "both see the same inputs" would overclaim: one side is handed the interaction and the other must learn it, and that gap is precisely the class difference the paper studies rather than an artefact of the setup. The message-passing arm is source-conditioned, since a model whose $\boldsymbol{h}_c$ ignores the source cannot represent reachability at all and would fail both rules for one reason. Oracle checks confirm the construction separates the mechanisms: walk counts score 0.973 on path-rule queries and 0.031 on feature-rule ones, the pairwise rule the reverse. No query pair appears in history, and the training split shares no pair with the evaluation split.

Path-free is not sufficient. With the pairwise rule taken to be the plain inner product, the concentration ratio falls as path-free signal is added: 1.15 at $\alpha = 0.25$, 0.89 at $\alpha = 0.50$, 0.84 at $\alpha = 0.75$, 1.02 at $\alpha = 1.00$. The reason is measurable rather than speculative. The message-passing arm holds the same latents, and its readout is an MLP on $[\boldsymbol{h}_s, \boldsymbol{h}_c]$, which approximates an inner product well; as α rises it has more feature-rule training queries and gets better at exactly the queries the flag selects. Its per-query agreement with an affinity oracle rises from 0.121 to 0.27 over the sweep while its agreement with a walk oracle falls to -0.03. So a feature can be perfectly path-free and still fail to produce class-specificity, because the class can learn it anyway. This is a condition §5.4 states too weakly, and we correct it here: the feature must be path-free and not recoverable by the class's own readout from the same inputs.

With that, the effect reproduces and exceeds D1. Replacing the rule by a high-frequency function of the same quantity, using no information the message-passing arm lacks, only one that a concatenation readout approximates badly, restores it (Table 27): concentration reaches 9.40 at $\alpha = 0.25$ and stays above it through $\alpha = 0.75$, against 2.144 on D1. The mechanism is legible in the per-rule split rather than inferred: message passing scores 0.055 on feature-rule queries and 0.424 on path-rule ones at $\alpha = 0.25$.

The sweep is not monotone, and the reason is the second condition. Concentration collapses to 0.58 at $\alpha = 1$. There, every target follows the pairwise rule, so the complement, queries whose target happens to be reachable anyway, carries no path signal either, and message passing is weak on both halves (0.151 flagged against 0.080 complement). A ratio of ratios cannot detect specificity when the class fails everywhere. That is precisely the second applicability condition of §5.4, the class must be able to solve the complement, emerging unbidden from a construction built to vary the first. At $\alpha = 0$ the first condition fails instead: the path rule cannot produce an unreachable target, the flagged set is empty, and concentration is undefined.

The isolation replicates on a second public history. To check that none of the above is specific to COLLEGEMSG, we repeat the construction verbatim on the BITCOIN-OTC history ($5{,}881$ nodes, $30{,}253$ edges; signed trust rather than social messaging), with the same protocol, seeds, and gates. The leak check is again zero, $\alpha{=}0$ again yields no flagged queries, and the oracles separate as sharply (walk counts 0.989 on path-rule queries against 0.039 on feature-rule ones; the pairwise rule 0.928, the reverse). Both regimes reproduce (Table 22): the high-frequency rule exceeds D1's 2.144 at every α in $[0.25, 0.75]$ and collapses at $\alpha{=}1$ for the same complement reason, while the smooth rule never rises above it. The per-rule split matches: message passing scores 0.063–0.083 on feature-rule queries against the baseline's 0.877–0.890.

Table 22: BITCOIN-OTC replication of the semi-synthetic sweep: concentration by α, three seeds. The high-frequency rule exceeds the D1 reference (2.144) throughout $[0.25, 0.75]$; the smooth rule never does; $\alpha{=}0$ has no flagged queries by construction.

pairwise rule	$\alpha{=}0.25$	$\alpha{=}0.50$	$\alpha{=}0.75$	$\alpha{=}1.00$
high-frequency	6.79	5.79	2.92	1.04
plain inner product	0.50	0.45	0.66	1.04

ROBUSTNESS OF THE PAIRWISE RULE: A FREQUENCY × CAPACITY SWEEP

The rule above is an oscillating function that a concatenation MLP approximates badly, so class-specificity could follow from that choice alone and vanish under a more expressive readout. We state the rule exactly and sweep both axes. The pairwise rule is

$$g(s, c) \;=\; \cos\big(f \cdot \langle \boldsymbol{u}_s, \boldsymbol{u}_c \rangle\big),$$

with $\boldsymbol{u}$ the unit-normalized node latents, drawn once per node from a standard normal in 8 dimensions and shared by both arms and both splits, and f the frequency; $f{=}0$ makes g constant, and f small makes it nearly affine in the inner product, recovering the smooth case. We sweep $f \in \{0, 2, 6, 12\}$ against message-passing width $d \in \{16, 64, 256\}$ (Table 23).

On frequency selection. $f{=}12$ was the first value we tried, chosen before any result was seen, and we report the full sweep rather than that cell alone. It is in fact the strongest cell here on both statistics, so we do not claim it was an arbitrary pick that happened to work; what the sweep establishes is the weaker and more useful statement that every $f > 0$ we tried yields a positive DiD ($+0.30$ to $+0.65$ over 9 cells), so the effect does not depend on tuning the frequency to a narrow band. The headline 9.40 of Table 27 comes from the α sweep at $d{=}32$, a different configuration from any cell here, and is not the maximum over this grid.

A denominator-free effect size. The concentration ratio divides by the baseline-to-model ratio on the complement, and that denominator is small and unstable here, MP complement moves 0.35, 0.10 and 0.45 across widths at one frequency, so the ratio inherits the instability and looks erratic in capacity when the underlying quantities are not. We therefore also report difference-in-differences,

$$\mathrm{DiD} \;=\; \big(\mathrm{nonMP}_{\text{flag}} - \mathrm{MP}_{\text{flag}}\big) \;-\; \big(\mathrm{nonMP}_{\text{compl}} - \mathrm{MP}_{\text{compl}}\big),$$

which measures whether the baseline's advantage is larger on the flagged half, without dividing. $\mathrm{DiD} > 0$ is the claim; the ratio is kept only for comparability with D1.

The effect does not decay with capacity. At $f{=}0$, DiD is negative at every width (-0.239, -0.258, -0.221): with no path-free signal the baseline is worse on the flagged half, a second null endpoint alongside $\alpha{=}0$. For $f > 0$, DiD stays in $[+0.30, +0.65]$ across a sixteen-fold increase in width, with no downward trend, at $f{=}6$ it goes $+0.462$ to $+0.425$ to $+0.646$ as d goes $16 \to 64 \to 256$. So the phenomenon is not a width artefact, and the ratio column is too noisy at one seed to support a trend in either direction, so we report DiD.

The sweep does not settle whether a readout of a different form, bilinear, Fourier-featured, or an explicit product, would recover g where a wider MLP does not. This concerns the pairwise families a class can represent, not its capacity, and we have not tested it. The claim we make is narrower: within the node-wise-representation-plus-MLP readout the paper studies, the effect is stable in width.

Correcting the $\alpha{=}1$ reading. We earlier attributed the collapse at $\alpha{=}1$ to message passing being weak on both halves, but its flagged MRR (0.151) in fact exceeds its complement MRR (0.080) there, which that account does not explain. DiD gives the right one: $\mathrm{DiD} = +0.0099$, essentially zero, because the baseline's advantage is 0.762 on the flagged half against 0.753 on the complement. At $\alpha{=}1$ the baseline is uniformly strong rather than specifically strong, which a concentration statistic is designed not to reward, and the flagged/complement inversion in the raw MRRs is a composition effect that DiD is insensitive to.

Table 23: Frequency $\times$ capacity sweep at $\alpha{=}0.5$, one seed. Left block: concentration ratio. Right block: difference-in-differences, which does not divide by the complement ratio and is therefore the primary effect size. $f{=}0$ makes the pairwise rule constant and DiD negative, a null endpoint. For $f > 0$ DiD shows no decay from $d{=}16$ to $d{=}256$, so the effect is not a readout-width artefact; the ratio column is visibly noisier because its denominator is small.

	concentration			DiD		
f	$d{=}16$	$d{=}64$	$d{=}256$	$d{=}16$	$d{=}64$	$d{=}256$
0	0.70	0.66	0.72	-0.239	-0.258	-0.221
2	6.64	2.11	7.76	$+0.541$	$+0.298$	$+0.629$
6	4.22	3.99	7.32	$+0.462$	$+0.425$	$+0.646$
12	8.67	2.32	6.69	$+0.649$	$+0.318$	$+0.614$

Capacity on natural data, and two questions the strata answer directly

The construction above varies decoder capacity, but the objection it answers, that the collapse is an artefact of a small encoder, is about the natural-data experiments, where $d{=}16$. Those runs exist at larger widths and the gap does not move (Table 24): on D1 the flagged-minus-OK gap goes -0.7141 to -0.7071 for NBFNet and -0.6674 to -0.6649 for GCN as d goes $16 \to 128$, and on WIKIPEDIA -0.7105 to -0.7003. An eight-fold width increase changes the gap by under a hundredth, while OK-set MRR is already near saturation on D1 (0.97), so there is little room for capacity to help there

and none of it reaches the flagged half. This is the natural-data counterpart of the semi-synthetic sweep, and it is the same conclusion: capacity raises accuracy where evidence exists and cannot manufacture it where it does not.

The WIKIPEDIA shell is empty. Appendix R reports no shell result for WIKIPEDIA, which reads like a missing measurement. It is not: the shell is empty (Table 25). The graph is bipartite and the diagnostic is reported at $L{=}3$, so a target first reachable at $L{+}1{=}4$ would sit at even distance from the source, where targets, being of the opposite node type, cannot be. Bipartite graphs at odd L therefore have no shell at all, and the shell/off-shell decomposition is only available on the non-bipartite graphs. All 190 of WIKIPEDIA's flagged queries are off-shell.

Pairwise consistency across benchmarks. Appendix E exhibits one query set violating pairwise consistency, which motivates measuring its frequency. It varies by two orders of magnitude across our graphs (Table 26): 0.02% (D1) at the low end and 24.72% (EMAIL-EU) at the high end. The distribution matters for how the proposition should be read. On D1, where the paper's class-specificity claim lives, only 3 conflicting pairs exist among 33673 queries, so the hypothesis is very nearly satisfied and the proposition applies as stated. On EMAIL-EU and COLLEGEMSG it is violated on a substantial share of queries, and there the cyclic structure of the candidate sets imposes its own ceiling below 1, one that has nothing to do with the history graph and is a property of how the benchmark samples negatives. So Proposition 2 should not be read as a statement about every sampled-query benchmark; it is a statement about pairwise-consistent ones, and whether a given benchmark qualifies is checkable in seconds by the count in that table.

Table 24: Flagged/OK gap against encoder width on natural data, seed 11. An eight-fold width increase moves the gap by under a hundredth on either graph, so the collapse is not an artefact of $d{=}16$.

Graph	Model	d	flagged	OK	gap	gap/OK
D1	NBFNet	16	0.2517	0.9658	−0.7141	−0.739
	NBFNet	64	0.2575	0.9667	−0.7092	−0.734
	NBFNet	128	0.2590	0.9660	−0.7071	−0.732
	GCN	16	0.1694	0.8368	−0.6674	−0.798
	GCN	64	0.1905	0.8592	−0.6687	−0.778
	GCN	128	0.1926	0.8575	−0.6649	−0.775
WIKIPEDIA	NBFNet	16	0.1158	0.8263	−0.7105	−0.860
	NBFNet	64	0.1212	0.8209	−0.6997	−0.852
	NBFNet	128	0.1208	0.8211	−0.7003	−0.853
	GCN	16	0.1093	0.6953	−0.5860	−0.843
	GCN	64	0.1309	0.6482	−0.5173	−0.798
	GCN	128	0.1083	0.6623	−0.5540	−0.837

Table 25: Stratum sizes for hard-target-seen queries. off-shell is the residual non-shell group and therefore includes collide-only rows and targets censored beyond the stored BFS cap. WIKIPEDIA's shell is empty because the graph is bipartite and the diagnostic is at $L{=}3$. EMAIL-EU is uninterpretable at 15 flagged queries.

Graph	L	flagged	shell	off-shell	OK
D1	2	5878	2834	3044	27578
WIKIPEDIA	3	190	0	190	5442
COLLEGEMSG	2	1384	1352	32	3861
MATHOVERFLOW	2	404	376	28	3809
BITCOIN-OTC	2	808	664	144	1630
EMAIL-EU	2	15	14	1	5965

This is a construction, not a discovered property of natural graphs. It does not show that real temporal graphs behave this way; D1 remains our only organic instance and eight public graphs suggest such graphs are uncommon. The result establishes that the mechanism is real, reproducible from public inputs, and governed by both stated conditions, and that the first condition needs the strengthening above.

Table 26: Frequency of violations of Proposition 2's pairwise-consistency hypothesis: pairs of same-source queries each holding the other's target as a negative. The rate spans two orders of magnitude and is near-zero on D1, where the class-specificity claim lives.

Graph	queries	sources	conflicting pairs	sources involved	% queries
D1	33673	8543	3	3	0.02%
WIKIPEDIA	6000	1243	135	38	3.12%
COLLEGEMSG	6000	419	684	63	13.78%
MATHOVERFLOW	6000	1178	1	1	0.03%
BITCOIN-OTC	3641	378	12	12	0.66%
EMAIL-EU	6000	547	1262	140	24.72%

Table 27: Semi-synthetic sweep on a real public history graph. α is the share of targets set by a path-free pairwise rule. conc. is the concentration ratio for the high-frequency rule, conc. (smooth) for the plain inner product. The smooth column falls with α because the message-passing readout learns the inner product; the hard column exceeds D1's 2.144 and then collapses at $\alpha = 1$, where the complement carries no path signal either. $\alpha = 0$ has no flagged queries by construction.

α	flagged	MP flag	non-MP flag	MP compl.	non-MP compl.	conc.	conc. (smooth)
0.00	0.000	—	—	0.7713	0.9620	undef.	undef.
0.25	0.174	0.0530	0.8530	0.3827	0.6561	9.40	1.15
0.50	0.331	0.0922	0.8850	0.2467	0.6186	3.83	0.89
0.75	0.501	0.0857	0.8930	0.1572	0.6392	2.56	0.84
1.00	0.611	0.1509	0.9133	0.0802	0.8327	0.58	1.02

R A REPRESENTATION-MATCHED PAIRWISE BASELINE, BY STRATUM

Two objections to the class-specificity evidence of §5.4 are fair and connected. Our non-message-passing baseline is a gradient-boosted tree over engineered covariates. On D1, the GNNs receive 26 core features while XGBoost adds 20 more, so the comparison is not feature-matched. And the comparison omits the subgraph-sketching family (SEAL, BUDDY, ELPH), which is the engineered response to precisely the limitation we quantify.

BUDDY-lite (Chamberlain et al., 2023) answers both at once because it omits those side-features. It scores a pair directly through the counts $|\{v : d(s, v)=i, d(c, v)=j\}|$, which a readout over independent node representations cannot express (Zhang et al., 2021), but it carries no degree, no popularity and no side-features. It is therefore outside the message-passing class in the way that matters here and matched to the GNNs on information: structure only.

We profile it against the strongest message-passing arm per graph on three strata: the shell, where the target is first reachable at exactly $L+1$ so a path exists just outside the radius; off-shell, the residual flagged group including later, censored, and collide-only rows; and the OK set (Table 28).

On two shells, pairwise structure recovers information absent from message passing. BUDDY-lite scores $0.09\times$ on D1, $1.80\times$ on COLLEGEMSG, $1.38\times$ on MATHOVERFLOW, $0.67\times$ on BITCOIN-OTC relative to the best message-passing arm, beating it on 2 of 4 graphs where the shell has enough rows to measure. These two cases match the labeling-trick prediction for our fixed-radius models; the other two do not.

Off-shell, the tested structure-only baseline also falls. BUDDY-lite collapses off-shell on all five graphs where it can be measured (D1, WIKIPEDIA, COLLEGEMSG, MATHOVERFLOW, BITCOIN-OTC), scoring as low as $0.16\times$ of the message-passing arm. The covariate-bearing baselines do not: adding popularity and recency lifts the off-shell ratio to at or above parity on 5 graphs. This is evidence about these fixed-radius arms: it does not establish that every structural method fails, especially because the residual group contains censored and collide-only rows.

Implications for the paper's claim. The empirical result splits. The flagged set is not homogeneous: pairwise structure helps on two measured shells, while the tested covariate-bearing arms are

stronger in the residual group. This explains both the concentration result of §5.4, which needs a baseline with such covariates to appear at all, and the fact that the pairwise readout of §5.4 helps on the shell but not beyond it, two findings the paper previously reported side by side without connecting them. It also means the diagnostic's graded form (§5.4) is doing more work than its binary form: which stratum a flagged query falls in determines which family of remedy applies.

Table 28: Non-message-passing arms by stratum, against the strongest message-passing arm per graph. shell: target first reached at $L{+}1$. off-shell: residual flagged group, including later, censored, and collide-only rows. Ratios are arm over message-passing MRR. BUDDY-lite wins the shell on 2 of 4 measurable graphs and falls in all five residual groups; covariate-bearing arms remain competitive there. These are fixed-radius empirical comparisons. Strata with fewer than 20 rows are omitted.

Graph	Non-MP arm	shell	off-shell	OK	shell/MP	off/MP
D1	BUDDY-lite (structure only)	0.039	0.022	0.831	0.086	0.325
	GBT, structure only	0.389	0.034	0.960	0.859	0.501
	GBT, + popularity/recency	0.447	0.063	0.963	0.987	0.923
	GBT, all covariates	0.445	0.116	0.965	0.984	1.707
	GBT, dataset features	0.646	0.457	0.975	1.428	6.718
WIKIPEDIA	BUDDY-lite (structure only)	—	0.102	0.837	—	0.862
	GBT, structure only	—	0.107	0.844	—	0.897
	GBT, + popularity/recency	—	0.123	0.854	—	1.037
	GBT, all covariates	—	0.130	0.856	—	1.094
COLLEGEMSG	BUDDY-lite (structure only)	0.097	0.019	0.429	1.801	0.371
	GBT, structure only	0.117	0.074	0.431	2.168	1.457
	GBT, + popularity/recency	0.172	0.051	0.485	3.183	1.007
	GBT, all covariates	0.188	0.069	0.466	3.479	1.353
MATHOVERFLOW	BUDDY-lite (structure only)	0.122	0.015	0.470	1.384	0.233
	GBT, structure only	0.098	0.091	0.513	1.113	1.412
	GBT, + popularity/recency	0.154	0.145	0.563	1.745	2.258
	GBT, all covariates	0.160	0.146	0.582	1.821	2.270
BITCOIN-OTC	BUDDY-lite (structure only)	0.151	0.028	0.355	0.669	0.159
	GBT, structure only	0.149	0.106	0.456	0.660	0.594
	GBT, + popularity/recency	0.249	0.234	0.561	1.107	1.310
	GBT, all covariates	0.221	0.186	0.540	0.984	1.044

S FORWARD APPLICABILITY TEST

The conditions of §5.4 were read off the same graphs Table 4 then scores, so that table summarizes ten outcomes rather than testing a rule. This appendix asks whether the conditions predict on graphs they were not derived from.

During the search for a second break (Appendix L) we screened seven graphs that are used nowhere else in the paper: COIN, COMMENT, FLIGHT and its one-month restriction FLIGHT-1MO are the TGB transaction, comment and flight networks (Huang et al., 2023); REDDITLINKS is the SNAP subreddit hyperlink network (Leskovec & Krevl, 2014); DATASET3 shares D1's 46-feature interface. We freeze condition 1 exactly as stated, $0.02 \leq \mathrm{hard} \leq 0.90$, and score it on all seven from the mask alone, with no training (Table 29). It splits them 4–3. The three failures are FLIGHT, FLIGHT-1MO, EMAIL-EU2, and all three fail for the density reason Appendix L documents: hub-and-spoke connectivity puts almost every candidate within two hops, so hard collapses toward 0 and reachability stops discriminating. That agreement is not new evidence, but it is the first time the condition is used to call a graph before rather than after the fact.

Condition 2 needs a trained model, and we score it on all four graphs that pass condition 1. COIN and COMMENT carry 19–24M edges over roughly 600k nodes, which we first judged intractable and which is not: the encoder allocates a node table of $n \times 16$ and a sparse adjacency, so COIN trains in 1.7 GB at 25 s per epoch. We train GCN on each (3 seeds), add NBFNet on REDDITLINKS, and evaluate four static arms on DATASET3, at the settings of Appendix H, and score the conditions' joint prediction. That prediction is fixed in advance and both branches are claims this paper already makes: with condition 2 also passing, the flagged/complement separation should be real and

survive degree matching; with condition 2 failing, the raw gap should be composition rather than mechanism.

Three graphs confirm and one refutes. On REDDITLINKS both conditions pass ($\mathrm{hard} = 0.1227$, best OK-set MRR 0.7803) and the prediction held: GCN scores 0.1672 on the flagged set against 0.5994 on the complement, a matched gap of -0.1923 (source-clustered CI $[-0.2294, -0.1577]$, coverage 0.94), and NBFNet 0.1387 against 0.7803, matched gap -0.5759 (CI $[-0.6164, -0.5226]$). COIN likewise passes both ($\mathrm{hard} = 0.1289$, OK 0.6479) and held: 0.2756 against 0.6479, matched gap -0.1109 (CI $[-0.1386, -0.0725]$, coverage 0.75). DATASET3, detailed below, also passes and holds.

COMMENT refutes it, and is the more informative case. Condition 2 fails there (best OK-set MRR $0.2421 < 0.30$), so the registered branch predicted the raw gap would be composition, i.e. that matching would leave nothing. Matching removes 90.2% of it, -0.1314 to -0.0129, by far the largest reduction of the three, so the prediction is right in direction. But the residual CI $[-0.0202, -0.0063]$ still excludes zero, and the criterion fixed in advance was that no arm remain significantly negative. By that criterion the prediction did not hold, and we do not relax it after the fact. The result is directionally correct but its stated form is too strict: a gap can be 90% composition and still leave a detectable mechanism.

One pattern across the three original trained graphs is worth recording, since it was not predicted. The share of the gap that matching removes tracks condition 2 inversely: 55.5% at OK-set MRR 0.7803, 70.2% at 0.6479, 90.2% at 0.2421. The weaker the class on the complement, the more of its flagged deficit is degree composition, which is the mechanism §5.3 describes, arriving here from three graphs the conditions had never seen.

Condition 3, tested forward twice. COIN and DATASET3 both admit a non-message-passing arm, so the third condition can be scored on graphs it never saw. Both failed as predicted. On COIN, XGBoost is far stronger overall (0.8247 against GCN's 0.5783) but uniformly so: its edge is $1.42\times$ on the complement against $1.16\times$ on the flagged set, giving $\mathrm{conc} = 0.817$, and its ratio exceeds 1 in all four target-degree bands with enough rows to measure. On DATASET3, which passes conditions 1–2 and has verified spread in its flagged set, the full 46-dimensional feature set gives $\mathrm{conc} = 1.14$: a complement ratio of $1.61\times$ absorbs the baseline's advantage uniformly, because the class is not yet saturated on the complement (OK MRR 0.54 against D1's 0.97). That is the second applicability condition doing its work: class-specificity needs the class to already solve the complement, and DATASET3 shows that passing conditions 1–2 is not sufficient.

On DATASET3 the collapse is not a degree artefact either: standardizing the complement to the flagged rows' target-degree distribution leaves a matched gap of -0.2722 for GCN (95% CI $[-0.296, -0.250]$; OK MRR 0.533 raw, 0.429 matched), with -0.256 for APPNP and -0.274 for SAGE, and the within-stratum contrast holds in every target-degree band (e.g. 0.207 against 0.661 in the most popular band). The non-message-passing arm is affected more than the class itself there (-0.570 for XGBoost), which is the same non-concentration that condition 3 records.

Three limits bound the interpretation. Condition 2 rests on four graphs and condition 3 on two, so this is not a prospective study and does not show the conditions transfer generally. The non-message-passing arm is single-seed on COIN, and on DATASET3 only its full 46-dimensional arm is: the three nested feature sets each run at three seeds against the same message-passing reference and put the seed-to-seed spread of conc below 0.008 (1.013 ± 0.005 for the 30-dimensional rich set, 0.927 ± 0.008 for pop, 0.879 ± 0.004 for struct, every individual run far below the 1.3 threshold), so the single-seed verdict there is not a sampling artefact. And COIN's coverage is 0.75. The surviving claim is that the conditions are falsifiable: condition 2 was refuted once in four trained outcomes, and condition 3 failed twice for reasons the framework predicts.

T AUDITING THE DIAGNOSTIC ITSELF, NOT ITS TRANSCRIPTION

Every number in this paper is generated into the text by script, and four gates check that the generated tables and the prose agree with the stored result files: 15 consistency checks, 8 table-provenance checks, an untraceable-number sweep over the 579 numerals in the body and appendices, and a meta-gate that injects wrong values into each table and requires the gates to fail (a green gate is only evidence if it has been shown capable of going red, and this caught one gate that passed while

Table 29: The applicability conditions applied forward to seven graphs that support no claim elsewhere in this paper. Condition 1 is frozen at $0.02 \leq \mathrm{hard} \leq 0.90$ and computed from the mask alone, with no training; its three failures are density cases where almost every candidate sits within two hops. Condition 2 is scored on the four that pass, on the arm with the best OK-set MRR (named in the gap column), with the matched gap and its source-clustered interval as in §5.3 where available. The prediction registered in advance held on three, was refuted on COMMENT (condition 2 fails, matching removes 90.2% of the gap but not all), and condition 3 failed on both graphs tested, for reasons the framework predicts.

Graph	hard	n_{flag}	C1	OK MRR	C2	matched gap	prediction
COMMENT	0.3540	2832	pass	0.2421	**fail**	−0.0129 (gcn)	**refuted**
COIN	0.1289	1031	pass	0.6479	pass	−0.1109 (gcn)	held
REDDITLINKS	0.1227	982	pass	0.7803	pass	−0.5759 (nbfnet)	held
DATASET3	0.0976	781	pass	0.5408	pass	−0.2722 (gcn)	held
EMAIL-EU2	0.0057	46	fail	–	–	–	–
FLIGHT-1MO	0.0043	34	fail	–	–	–	–
FLIGHT	0.0010	8	fail	–	–	–	–

checking nothing). None of these tests inspects the diagnostic implementation: a bug in the mask builder would leave all of them green, because they verify that the paper reports the computed code output, not that the code computed Definition 1.

We therefore reimplemented the diagnostic from the definition as printed, without reference to the mask builder, and deliberately unlike it: breadth-first balls as an explicit frontier loop rather than sparse-matrix reachability, and WL as exact refinement on sorted multisets of neighbour colours rather than the reference's Monte-Carlo fingerprint (two summed random 60-bit hashes per colour). The exact version is the stronger check, since hash collisions would coarsen the reference's colouring without coarsening ours. The independent implementation reproduces hard, and both of its terms separately, on all six graphs to floating-point equality.

The audit was informative before it agreed, and the disagreement is worth reporting because it locates an ambiguity in how we state the definition. A first version kept one adjacency entry per temporal edge, so a node pair interacting k times contributed k entries to its neighbour multiset. That version matched the unreachability term exactly on all six graphs, breadth-first reach is insensitive to edge multiplicity, and disagreed on the collision term on all six, always with fewer collisions (for instance 0 against 24 on WIKIPEDIA, whose repeated user–item interactions make it the most multi-edge-heavy graph we use). Multiplicity is extra information for WL: it refines the colouring, and a finer colouring collides less. Deduplicating to the simple graph recovers the reference's collision count exactly on every graph. Since 1-WL is defined on the simple graph this is the correct reading of "the (undirected) history graph", but an earlier version of Definition 1 did not say so, and a reader reimplementing it from that text could reasonably have made the other choice; the definition now names the deduplicated simple graph explicitly. The consequence for our results is confined to the collision term, which is below 0.4% of queries on nine of ten graphs; the receptive-field term that carries every claim in §5.2 is unaffected either way.

U SCOPE OF THE RESULTS

The break of §5.4 rests on D1, which is unredistributable and which we could not reproduce on any public graph; the construction of App. Q is the closest public substitute. TGN does not satisfy Proposition 1(i)'s locality premise and is only an empirical stress test; sequence-based temporal predictors such as CAW, GraphMixer and DyGFormer are not trained at all, although their one-hop sequence encoders and two-hop pairwise features place flagged targets beyond their structural evidence as well. The class-specificity comparison on D1 is not feature-matched: the tree receives 20 covariates the encoders do not, four of them transductive, and App. R shows the break depends on them; whether an encoder given the same covariates through its readout would recover the flagged queries is untested. The audit is topological rather than time-respecting: it uses the history graph's L-hop structure and not the order of edges within it. The 100-candidate evaluation protocol is held fixed throughout and not varied. Neither depth (App. I) nor width (App. Q) closes the flagged-query

gap, and on low-signal graphs the raw flagged/complement gap is degree composition rather than propagated evidence (App. K); we therefore report the graded form as the primary object there. Finally, under Proposition 2's pairwise-consistency and model-class conditions, the history graph and Definition 1 do not upper-bound achievable accuracy, so every consequence drawn here is empirical rather than a numerical ceiling on the class's attainable accuracy.

V Extended related work

This appendix expands §2.

Expressiveness, stated at the resolution we need. The 1-WL bound says nodes with identical L-round colours receive identical embeddings (Xu et al., 2019; Morris et al., 2019). The labeling-trick framework (Zhang & Chen, 2018; Zhang et al., 2021; Wang et al., 2025) shows that a decoder reading two independently computed node embeddings through an inner product cannot express pairwise quantities such as path counts between the endpoints, and that endpoint-conditioned labeling exceeds 1-WL. Loukas (2020b;a) lower-bound the distinctions available to depth- and width-restricted networks, which we use in App. E. Note the difference in the quantities these results and our diagnostic quantify: expressiveness fixes the representational ceiling of a model class once the input is rich enough, whereas the audit asks whether a given evaluation set ever supplies that richness.

The temporal receptive field is a different quantity. Karmim et al. (2024) study, under the same name, how many past snapshots a model should see. That quantity is a window hyper-parameter of the temporal encoder; ours is hop-reachability in the history graph, and the two are independent: a model with a long temporal window can still fail on a query whose target lies outside its L-hop ball. The subgraph-sketching line (BUDDY/ELPH (Chamberlain et al., 2023), Neural Common Neighbor (Wang et al., 2024)) is, in effect, an engineered response to the limitation the diagnostic quantifies.

Post-hoc structural attribution. Jiang & Deng (2025) resample connectivity-preserving subgraphs of FB15k-237, retrain, correlate structural statistics with accuracy, and use LIME to attribute the effect to how models weight entity versus relation embeddings. That analysis is post-hoc and per-dataset; the audit here needs no training, speaks about a model class instead of a fitted model, and rests on a specific mechanism instead of correlations. We also differ in our claim: under its stated conditions, §4.2 shows that the history-derived hard fraction cannot bound the class supremum.