GRAPHSIEVE: GRAPH PURIFICATION VIA ATTACK-AGNOSTIC CALIBRATION ACROSS PERTURBATION SEVERITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Structural poisoning attacks degrade node classification in graph neural networks by editing edges before training. Purification defenses respond by editing the observed graph, and their test accuracy depends on both the editing rule and the procedure that selects its hyperparameters. We study that selection when it must be made without access to the evaluation attack instances. We introduce GraphSieve, a purification defense for homophilous graphs that prunes low-similarity edges and adds high-similarity absent edges. Its attack-agnostic calibration selects configurations by mean validation accuracy over a trusted clean graph and label-blind random perturbations at prespecified severities. The same procedure calibrates GraphSieve and the purification baselines, separating calibration from graph editing. A prespecified evaluation on three citation graphs identifies gains in attacked-graph test accuracy over the undefended GCN, RGCN, and GCN-SVD, with no statistically significant loss to RGCN or GCN-SVD. An attack that adapts to the published editing criterion leaves GraphSieve ahead of the undefended GCN in 20 of 36 cells, with no loss. Against equally calibrated GCN-Jaccard, which applies the pruning rule alone, the evaluation finds both gains and losses. In exploratory ablations, calibration also lifts the baselines, raising GCN-Jaccard's mean attacked-graph accuracy over clean-only tuning by +0.44 and +4.18 percentage points on the fresh and reproduction profiles, and the benefit extends to GCN-SVD and GraphSieve. Single-severity comparisons show that a selection rule must be reported with its severity distribution. A follow-up replication, run after the prespecified evaluation with new attack graphs and model seeds, reproduces the augmentation gains in the Cora deletion-Metattack conditions, where calibration disables pruning. The gains are +1.21 and +0.93 percentage points over calibrated GCN-Jaccard at deletion budgets of 15 and 25 percent, respectively. These gains identify a replicated use case for augmentation rather than a general benefit from edge addition. Recovery diagnostics show that most added edges had never been deleted. These results motivate a common calibration protocol for purification defenses and distinguish predictive augmentation from reconstruction of the original topology.

1 INTRODUCTION

Graph neural networks (GNNs) classify nodes by aggregating features along the edges of a graph (Kipf & Welling, 2017), and they underpin recommendation, route planning, antibiotic discovery, and fraud detection (Ying et al., 2018; Derrow-Pinion et al., 2021; Stokes et al., 2020; Dou et al., 2020). This reliance on structure is also an exposure. The model propagates whatever the edges carry, including edits made before training. Manipulation of relational structure is documented (Boshmaf et al., 2011) and has motivated defenses in web and social systems (Gyöngyi et al., 2004; Hooi et al., 2016). Against a GNN, a small edit budget suffices, targeted at one node (Zügner et al., 2018) or untargeted at overall accuracy (Zügner & Günnemann, 2019; Xu et al., 2019); we study the untargeted case.

Purification defenses modify the observed graph before model training. GCN-Jaccard removes edges between feature-dissimilar nodes (Wu et al., 2019), whereas GCN-SVD replaces the adjacency matrix with a low-rank approximation from its singular value decomposition (Entezari et al., 2020). Their performance depends on both the editing rule and its configuration. A comparison against a

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.

default threshold or rank can therefore confound the benefit of a new editing rule with the benefit of hyperparameter selection. Re-evaluations under tuned configurations and adaptive attacks make this distinction consequential (Mujkanovic et al., 2022).

We study configuration selection before the evaluation attack is observed. A trusted clean graph and validation labels are available for calibration; the resulting configuration is then fixed for purification. We ask whether random perturbations at prespecified severities can select useful configurations without access to the evaluation attack instances.

The editing problem also depends on attack composition. Pruning can remove suspicious insertions but cannot replace missing connections. Selective edge addition provides a complementary intervention, although an edge that improves classification need not reconstruct a deleted connection. We introduce GraphSieve to investigate these distinctions on homophilous citation graphs with informative features. Our contributions are as follows.

- **A calibration protocol for graph purification.** Severity-ladder calibration selects defense hyperparameters by aggregating validation accuracy over random perturbations at severities fixed in advance. Perturbations are generated without labels, and configurations are frozen before attack evaluation. The same procedure calibrates GraphSieve and the purification baselines.
- **Graph purification through pruning and addition.** GraphSieve combines similarity-based edge pruning with selective edge addition on homophilous graphs. Together the two operations reach both insertions and deletions, and calibration decides which of them is active for a given graph.
- **An empirical separation of calibration, purification, and repair.** Prespecified tests, exploratory ablations, and a focused replication separate calibration effects from editing effects and locate where addition helps. Recovery diagnostics then separate predictive gains from structural repair and identify what limits it.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 presents the GraphSieve framework, Section 4 states the prespecified experimental protocol, Section 5 reports the results, and Section 6 discusses them. Finally, Section 7 concludes the paper.

2 Related Work

This section reviews the attacks that motivate purification, the defenses that edit the graph or harden the model, their evaluation standards, and the homophily assumption behind similarity-based editing.

Structural attacks. Poisoning attacks modify the graph before training, targeting individual nodes (Zügner et al., 2018) or overall test accuracy (Zügner & Günnemann, 2019; Xu et al., 2019). Metattack selects edge flips through a meta-gradient objective. Its realized edits need not balance insertions and deletions; as Table A9 shows, the released matrices evaluated here are predominantly insertion attacks. Li et al. (2023) examine the concentration of adversarial edits around training nodes. DICE attacks (Disconnect Internally, Connect Externally) delete within-class edges and insert cross-class edges (Waniek et al., 2018; Zügner & Günnemann, 2019; Jin et al., 2021). The protocol therefore varies the insertion-deletion composition to separate the regimes in which pruning and addition can help.

Purification defenses. GCN-Jaccard (Wu et al., 2019) and GCN-SVD (Entezari et al., 2020) edit the graph once before training, whereas Pro-GNN jointly learns graph structure and model parameters (Jin et al., 2020), coupling the defense to the training procedure. GraphSieve instead keeps a preprocessing interface. It edits the observed graph once and then trains a standard graph convolutional network (GCN) unchanged, and its addition stage extends similarity pruning to absent candidate edges.

Robust architectures. Robust GCN (RGCN) attenuates neighbors whose Gaussian hidden states have high variance (Zhu et al., 2019), GNNGuard adjusts neighbor contributions using feature similarity (Zhang & Zitnik, 2020), and robust aggregation reduces sensitivity to contaminated neighborhoods (Geisler et al., 2020). RGCN provides an architectural comparator. The calibration analysis is scoped to preprocessing defenses, which edit the graph once and expose a hyperparameter grid the ladder can select over; joint structure learning (Jin et al., 2020) falls outside that scope.

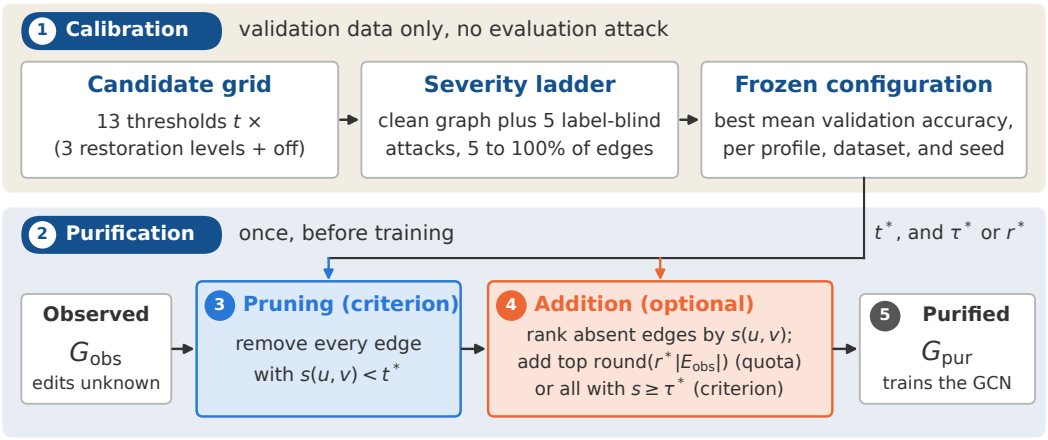


Figure 1: The GraphSieve pipeline. **Calibration (top)** scores candidate configurations by mean validation accuracy over a trusted clean graph and label-blind perturbations at prespecified severities, and freezes the best one per profile, dataset, and replicate. **Purification (bottom)** prunes edges below the frozen threshold t, optionally adds high-similarity absent edges under a quota or a criterion (restoration in the registered protocol), and trains a standard GCN on G_{pur}.

Evaluation standards and configuration choice. DeepRobust provides common implementations (Li et al., 2021), and surveys organize the threat models (Jin et al., 2021). Adaptive re-evaluation exposes weaknesses in several defenses and motivates tuning for both clean accuracy and adversarial robustness (Mujkanovic et al., 2022; Li et al., 2024). That work varies the attacker's knowledge, whereas this study fixes what the defender may use when choosing a configuration. Perturbation semantics also affect interpretation (Gosch et al., 2023). Ngo et al. (2026) re-evaluate attacks and defenses under a protocol that selects every model on clean data, before any attack exists; that policy is one of the controls evaluated here.

G-RNA is a close precedent for selection using synthetic perturbations: it scores architectures by validation accuracy and by prediction stability under perturbations from an attack proxy at a fixed budget (Xie et al., 2023). This paper selects defense hyperparameters rather than architectures, using aggregate validation accuracy across several prespecified severities. The clean-only and single-severity controls (Section 5.1) test that choice.

The homophily regime. Feature-similarity purification assumes that informative connections join feature-similar nodes, and the value of graph structure varies with the task, features, and model (Zhu et al., 2020; Chien et al., 2021; Ma et al., 2022; Platonov et al., 2023). Evaluation is therefore limited to homophilous citation graphs with informative node features, and so are the conclusions.

3 The GraphSieve Framework

GraphSieve selects a configuration on a trusted clean graph and its validation labels and then purifies the observed graph once before model training, as depicted in Figure 1. Section 3.1 defines the two editing stages and their control laws, and Section 3.2 the calibration that fixes their hyperparameters.

3.1 Two-Stage Similarity Editing

Let $G_{\text{obs}} = (V, E_{\text{obs}})$ denote the observed graph, where V is its set of nodes and E_{obs} its set of observed edges, which may carry an attacker's insertions and deletions. Let $X \in \mathbb{R}^{|V| \times F}$ hold the nonnegative node features, where F is the feature dimension, so node $u \in V$ has feature vector x_u with entries $x_{u,i}$. For nodes $u, v \in V$, the similarity score $s(u, v)$ is the Jaccard coefficient of the feature supports S_u and S_v, the index sets of positive entries, for binary features, and the cosine similarity of the feature vectors otherwise:

$$s(u,v) = \frac{|S_u \cap S_v|}{|S_u \cup S_v|}, \quad S_u = \{i : x_{u,i} > 0\} \qquad \text{or} \qquad s(u,v) = \frac{x_u^\top x_v}{\|x_u\| \, \|x_v\|}. \tag{1}$$

Both scores lie in $[0, 1]$ and are defined as zero when a denominator vanishes. Neither uses labels at edit time.

Stage 1: pruning as a criterion. The pruning stage removes every observed edge whose score is strictly below a frozen threshold t:

$$E_{\text{pruned}} = \{\, (u, v) \in E_{\text{obs}} : s(u, v) < t \,\}. \tag{2}$$

The threshold fixes the selection rule, while the observed score distribution determines the number of removed edges. We call this a *criterion-based* control law. Since scores are nonnegative, $t = 0$ disables pruning. A safeguard listed in Table A4 retains the single highest-scoring observed edge if every observed edge falls below t.

Stage 2: restoration under a quota or a criterion. The addition stage inserts absent edges between nodes with high feature similarity; the registered protocol calls this stage *restoration*, and the notation below keeps that name. Each node contributes up to 30 of its highest-scoring non-neighbors with positive similarity, and up to 30 of its two-hop non-neighbors with the highest Adamic–Adar weight (Adamic & Adar, 2003), a shared-neighbor score that discounts high-degree neighbors. Because both shortlists are built on the observed graph before pruning, a pruned edge is never a restoration candidate. The shortlist width bounds the candidate pool for tractability, and it was fixed before any evaluation rather than chosen by calibration, with the pool's coverage reported in Appendix D. The union of these shortlists, taken as unordered pairs, forms the candidate set $\mathcal{C}$, ranked by descending feature similarity and, among equal scores, by ascending node index.

Under a *quota*, a frozen rate r sets $R = \text{round}(r \cdot |E_{\text{obs}}|)$, and E_{restored} is the first R elements of $\mathcal{C}$. Under a *criterion*, a frozen threshold τ sets $E_{\text{restored}} = \{\, (u, v) \in \mathcal{C} : s(u, v) \geq \tau \,\}$, truncated to 1.00 times $|E_{\text{obs}}|$. A standard GCN is trained on $(E_{\text{obs}} \setminus E_{\text{pruned}}) \cup E_{\text{restored}}$.

The *criterion arm*, denoted GraphSieve unless qualified, draws candidates from the full union pool defined above, whereas the *quota arm* draws them from the similarity shortlists alone and binarizes features, so it scores PubMed by Jaccard rather than cosine (Section 4). Both use the same pruning rule, backbone, and calibration procedure, although each searches its own grid, so their comparison contrasts two complete variants, not an isolated change in the budget rule.

Response to structural damage. At a fixed threshold, pruning tracks the number of observed edges of low feature similarity: insertions between dissimilar nodes satisfy Equation 2, so the edit count follows the attack without retuning, whereas deletions leave surviving scores unchanged (Section 5.3). A quota tracks graph size instead. Under generalized DICE with budget k and deletion share δ, the observed edge count is $|E_{\text{clean}}| + k - 2d$ for $d = \text{round}(\delta k)$ deletions, so R cannot increase with d, and the quota admits fewer candidates as more connections are deleted. Criterion-based addition avoids that dependence, except through its cap. Neither rule estimates how many edges were deleted; both rank candidates by similarity (Appendix D).

3.2 Severity-Ladder Calibration

In the evaluated setting, the configuration is fixed before G_{obs} is available, so GraphSieve calibrates on a trusted clean copy of the graph and its validation labels alone. Each arm searches a grid of 52 configurations, crossing 13 unevenly spaced pruning thresholds from 0 to 0.50 with 3 active restoration levels plus restoration off. The pruning ladder includes GCN-Jaccard's DeepRobust default of 0.01, and the restoration levels are similarity thresholds $\tau \in \{0.50, 0.35, 0.25\}$ for the criterion arm and rates $r \in \{0.05, 0.10, 0.15\}$ for the quota arm. Both grids were registered before any evaluation, as Table A4 records.

Selection maximizes mean validation accuracy over 6 rungs, the clean graph and 5 label-blind perturbations of the random composition family (Section 4) at deletion share $\delta = 0.50$ and budgets of 0.05, 0.15, 0.25, 0.50, 1.00 of the clean edge count. When configurations tie, selection takes the one that makes the fewest edits, breaking the tie first on whether restoration is active, then on its level, and last on the pruning threshold. Clean-only and single-severity selection serve as controls for this rule, isolating the effect of scoring across several severities (Appendix C).

Selection uses model and perturbation seeds disjoint from those of evaluation (Appendix B). Every calibrated method receives the same trusted clean graph and validation labels, and purification then

uses only the frozen configuration, features, and observed graph. *Attack-agnostic* is therefore a claim about what selection may use, no evaluation attack instance and no knowledge of its mechanism, rather than a guarantee of robustness. Sharing one trusted graph makes the comparison of selection procedures controlled rather than realistic, since a defender who held one could train on it directly, and calibration from the observed graph alone is outside this study. The registered evaluation repeats calibration per profile, dataset, and replicate, so its intervals include calibration variation.

4 Experimental Protocol

The protocol was registered before any evaluation ran. Section 4.1 fixes the datasets, attacks, and methods, Section 4.2 the paired statistics, and Section 4.3 the hypotheses and their decision rules.

4.1 Datasets, Attacks, and Methods

We evaluate Cora, CiteSeer, and PubMed under the two preprocessing profiles summarized in Table A3, with edge label agreement h ranging from 0.736 to 0.810. The *fresh* profile uses standard public splits (Yang et al., 2016); the *reproduction* profile uses Pro-GNN's component-restricted graphs and splits, on which its released Metattack matrices were generated (Jin et al., 2020).

A *cell* is one dataset under one attack condition, with 10 replicates, each fixing an attack seed and a model seed that every method shares. Cells form analysis *populations* by dataset group and profile, analyzed separately and never pooled, with 130 attacked cells in citation-fresh and 18 in citation-reproduction, plus one clean cell per profile and dataset.

The citation-fresh cells come from four attack families, which vary attack composition and attacker knowledge. (1) *Generalized DICE* (Waniek et al., 2018) fixes a budget at 5, 15, 25 percent of the clean edge count and sweeps the *deletion share* over 0.00, 0.25, 0.50, 0.75, 1.00 (45 cells), where that share is the fraction of the budget spent on deletions rather than insertions. (2) *Random composition* is its label-blind control, deleting observed edges and inserting absent pairs uniformly at random (45 cells), and is also the calibration family, with disjoint attack seeds. (3) *Adaptive deletion DICE* is defense-aware, deleting the same-class edges scoring lowest under GraphSieve's similarity and inserting by the DICE rule (36 cells), reading the similarity signal and shortlist width but not the frozen thresholds or the trained model. (4) *Deletion Metattack* restricts the meta-gradient objective of Zügner & Günnemann (2019) to deletions, at 15 and 25 percent on Cora and CiteSeer (4 cells).

The other population, citation-reproduction, holds the *released Metattack* matrices at rates 0.15, 0.25, a *standard DICE* attack matched to their edit counts, and a *legacy DICE* regime at 0.50, 1.00 times the clean edge count (6 cells each). The 6 released matrices and the 9 pure-deletion adaptive cells hold one graph across seeds; every other attack is regenerated per seed.

Every cell evaluates the same 9 methods, namely the undefended GCN (Kipf & Welling, 2017); the robust architecture RGCN (Zhu et al., 2019) at its reference settings; the purification baselines GCN-SVD (Entezari et al., 2020) and GCN-Jaccard (Wu et al., 2019); a *structure-free comparator*, a GCN trained with no edges, which Mujkanovic et al. (2022) argue for; the two GraphSieve arms; and pruning-only and restoration-only ablations matched to the criterion arm's frozen configuration. All but RGCN share one backbone. GCN-SVD, GCN-Jaccard, and both arms are calibrated on their own registered grids under the ladder (4 ranks, 13 thresholds, and 52 candidates per arm). GCN-Jaccard applies the Stage 1 rule, so it is the pruning stage under independent calibration.

4.2 Statistical Analysis

Each contrast is a paired mean difference in test accuracy over 10 replicates, reported in percentage points (pp) with a 95 percent Student-t interval (df $= 9$). A win or loss requires the interval to exclude zero. NS denotes no significant difference, not equivalence. Comparisons that are identical by construction fall in the NS counts, and the tables that separate them report an I column. All attacked cells enter the analysis; damage to the undefended GCN is significant in 123 of 130 citation-fresh cells and all 18 citation-reproduction cells, and Appendix A reports the comparison restricted to them. Means across cells are descriptive summaries, and no cell-wise interval is adjusted for multiplicity: the registered rule addresses multiplicity only at the population level (Section 4.3).

4.3 Prespecified Hypotheses

The hypotheses, contrasts, populations, and decision rule were committed before the first evaluation run. Appendix A gives their complete clauses. **E1** predicts no accumulation of significant losses to calibrated GCN-Jaccard; **E2** predicts no accumulation of significant gains over pruning alone. **E3** predicts that the adaptive attack's added damage over generalized DICE depends on whether calibration disables pruning. **E4** places the structure-free comparator outside the top three in at least 80 percent of citation-fresh cells. **E5** bounds candidate coverage on PubMed and the accuracy benefit of exact deleted-edge restoration.

E1 and E2 are refuted in a population only when its count of significant wrong-way cells reaches the one-sided binomial 95 percent bound at a 2.5 percent per-cell rate, 7 cells on citation-fresh and 3 on citation-reproduction. The bound treats cells as independent, which they are not, so Appendix A adds a dependence-aware sensitivity analysis. Throughout, "held" means not refuted. E3 to E5 are decided by their literal clauses, and everything else is exploratory. The public test splits were examined while the framework was developed, so the registration governs hypotheses, contrasts, and decision rule rather than the splits. We added the comparison with GCN-Jaccard at its default threshold after the study.

5 Results

The two analysis populations remain separate throughout; Appendix B reports every registered cell.

5.1 Calibration beyond Default Settings

Exploratory ablations isolate configuration choice with fixed grids, splits, seeds, attack graphs, and backbone. Each follow-up fixed its design and decision rule after the registered study but before any follow-up outcome was computed (Appendix C). Table 1 shows that across GCN-Jaccard, GCN-SVD, and GraphSieve, mean attacked-graph accuracy is higher under the ladder than under clean-only tuning in both profiles, and Figure 2 shows the cell-wise spread. GraphSieve's mean differences are +0.49 and +3.92 pp, with the largest dataset-level benefits on PubMed. GCN-Jaccard also improves over its DeepRobust default, with two exceptions visible in Table A12: its mean difference is negative on fresh-profile CiteSeer against both alternatives, and on reproduction-profile Cora against the default.

Against GCN-SVD's default rank, the ladder records 19 significant gains and no losses, all on fresh-profile PubMed, because 88 of the 130 fresh cells are identical by construction. Where no reproduction cell is identical, the ladder records 10 gains and no losses in 18 cells. Gains over clean-only tuning are smaller and never significant on citation-reproduction. On reproduction-profile CiteSeer, the mean attacked-graph difference from the default is +5.63 pp against a clean-accuracy difference of -1.64 [-2.76, -0.53] pp, a clean-versus-attacked tradeoff.

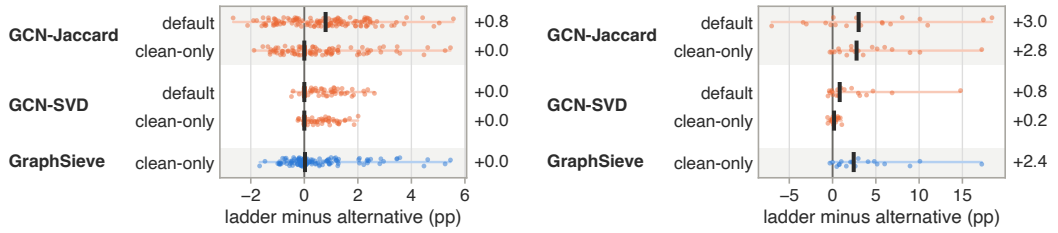
Single-severity controls show that the calibration distribution matters, and not always in the ladder's favor. Against clean-plus-100-percent tuning, the ladder trails on citation-reproduction: over the 18 cells, the mean difference is -1.39 pp with 4 significant losses for GCN-Jaccard and -1.44 pp with 3 for GraphSieve.

5.2 Comparative Performance

The criterion arm significantly improves on GCN-SVD in 117 of 130 citation-fresh and 14 of 18 citation-reproduction cells, and on RGCN in 94 and 17, as Table 2 reports; neither GraphSieve arm loses significantly to either baseline in either population. The criterion arm improves on the undefended GCN in 52 of 130 citation-fresh cells, with 6 significant losses, and in 13 of 18 citation-reproduction cells, with none. Against equally calibrated GCN-Jaccard, significant differences are infrequent: 4 gains and 6 losses on citation-fresh, and 1 gain and 0 losses on citation-reproduction. The two build the same graph in 42 fresh and 6 reproduction cells, so the fresh contrast rests on 88 cells (Section 5.6). The largest absolute cell-wise mean difference from the calibrated GCN-Jaccard is 0.99 pp on citation-fresh and 1.09 pp on citation-reproduction.

Table 1: Exploratory calibration effects across the evaluated purification methods. Each contrast is severity-ladder selection minus the named alternative, with W/NS/L over attacked cells and the mean paired difference in percentage points. Positive differences favor the ladder. NS includes identical selections; means are descriptive, and cell-wise intervals are unadjusted. Dataset-level results and clean-accuracy costs appear in Appendix C.

		Citation-fresh		Citation-reproduction	
Method	**Alternative**	**W/NS/L**	**Mean (pp)**	**W/NS/L**	**Mean (pp)**
GCN-Jaccard	library default	69/39/22	+0.77	11/5/2	+4.26
	clean-only tuning	36/82/12	+0.44	6/12/0	+4.18
GCN-SVD	library default	19/111/0	+0.31	10/8/0	+2.30
	clean-only tuning	10/120/0	+0.23	0/18/0	+0.21
GraphSieve	clean-only tuning	35/91/4	+0.49	7/11/0	+3.92



(a) Citation-fresh

(b) Citation-reproduction

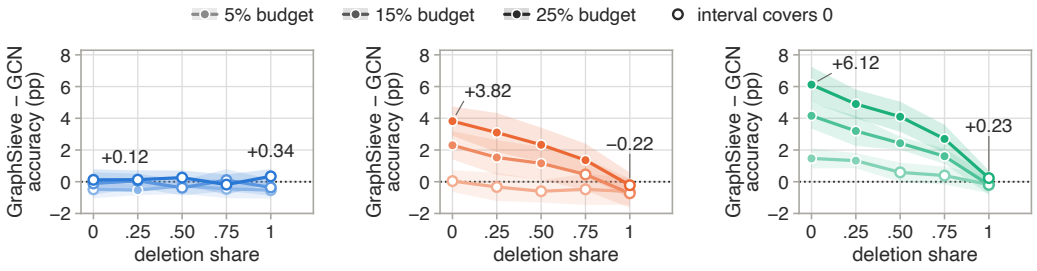
Figure 2: Exploratory gain from severity-ladder calibration over a cheaper policy, per attacked cell, with the defense held fixed. Each dot is one cell's mean paired difference over 10 seeds, positive where the ladder wins. The black tick marks the median, printed at the right, and the line spans the range. The spread is descriptive, not a confidence interval, because cells share datasets, splits, and seeds. The panels use different horizontal scales, and single-severity controls appear in Appendix C.

Table 2: Registered paired comparisons, each arm of GraphSieve minus the column's method over the population's 130 citation-fresh and 18 citation-reproduction attacked cells. Each entry reads wins/NS/losses and sums to the cell count, where a win or loss requires the paired interval to exclude zero and NS is never equivalence. Feat.-only is the structure-free comparator and Prune-only the pruning-only ablation. NS includes the 42 and 6 PubMed cells in which the criterion arm builds the same graph as GCN-Jaccard and its pruning-only ablation.

Population	**Arm**	**GCN**	**RGCN**	**GCN-SVD**	**GCN-Jaccard**	**Feat.-only**	**Prune-only**
citation-fresh	quota	52/68/10	102/28/0	113/17/0	6/118/6	125/5/0	6/113/11
	criterion	52/72/6	94/36/0	117/13/0	4/120/6	126/4/0	3/118/9
citation-reproduction	quota	12/6/0	16/2/0	14/4/0	1/13/4	14/0/4	0/14/4
	criterion	13/5/0	17/1/0	14/4/0	1/17/0	13/1/4	1/17/0

The ladder independently selects the same pruning threshold for GCN-Jaccard and the criterion arm in 27 of 30 fresh and 25 of 30 reproduction calibrations, so in most replicates the contrast with GCN-Jaccard isolates the additional editing stage. On the fresh profile, the ladder also freezes pruning off on Cora in all 10 replicates and declines addition throughout PubMed. The criterion arm therefore runs addition alone on Cora, pruning alone on PubMed, and both stages only on CiteSeer (Appendix C). Every significant gain over GCN-Jaccard occurs at a deletion share of at least 0.50, as do 4 of the 6 losses, all on citation-fresh, so composition alone does not predict a benefit. Clean-graph costs also remain: Table A5 shows the criterion arm losing significantly to the undefended GCN on two of the three datasets in each profile, with the difference reaching -0.56 pp on the fresh profile and -2.43 pp on the reproduction profile.

Severe attacks. In the severe cells, the larger rate of legacy DICE and released Metattack, the criterion arm improves on the undefended GCN under legacy DICE by +22.70 [+21.11, +24.29] pp on CiteSeer and +18.78 [+18.02, +19.55] pp on PubMed. Under released Metattack, the corresponding gains are +11.64 [+9.78, +13.49] and +11.58 [+11.39, +11.76] pp, as shown in Figure A1, while



(a) Cora (b) CiteSeer (c) PubMed

Figure 3: GraphSieve's advantage as the attack budget shifts from insertions to deletions (generalized DICE, citation-fresh). Each panel plots the criterion arm minus the undefended GCN (pp) against the deletion share, one line per budget (5, 15, 25 percent of the clean edge count) with paired 95 percent interval bands; filled markers exclude zero.

Cora's differences from the undefended GCN are not significant. The structure-free comparator significantly outperforms every graph-based method in 3 of the six severe cells, with the criterion arm at −12.77 [−15.86, −9.68] pp against it in the worst of them. Overall, the comparator beats the criterion arm in 4 of the 18 reproduction cells. In those cells, no graph-based defense beats it.

5.3 Attack Composition and Pruning

At the largest budget, the criterion arm's mean advantage over the undefended GCN across the fresh-profile datasets declines from +3.35 pp under pure insertion to +0.12 pp under pure deletion, as Figure 3 illustrates. Across dataset-by-budget cells, the comparison changes from 5W/4NS/0L to 0W/9NS/0L. Calibration disables pruning on Cora, where no comparison shows a significant gain.

This pattern is consistent with the criterion's response to low-similarity insertions. CiteSeer's pruned fraction rises from 11.8 percent on the clean graph to 21.8 percent under pure insertion at the largest budget, compared with 12.4 percent under pure deletion, and PubMed follows the same pattern (Appendix D). Figure A4 shows that in every matched dataset-by-budget comparison, insertions cause more damage to the undefended GCN than deletions, by factors of 1.48 to 5.07. The criterion arm's strongest gains therefore fall in the more damaging composition regime.

5.4 Addition under Deletion

The Cora deletion-Metattack conditions isolate addition: calibration disables pruning for both GCN-Jaccard and the criterion arm in every fresh-profile replicate. Against GCN-Jaccard, addition improves accuracy by +0.89 [+0.33, +1.45] and +0.74 [+0.19, +1.29] pp at the 15 and 25 percent budgets. A follow-up with a prespecified protocol, new attack graphs, and new model seeds reproduces both gains, at +1.21 [+0.77, +1.65] and +0.93 [+0.26, +1.60] pp respectively, as Table A16 reports.

These replicated gains identify a specific use case for addition, not a population-wide restoration benefit. In the registered comparison with its pruning-only ablation, the criterion arm records 3W/118NS/9L on citation-fresh and 1W/17NS/0L on citation-reproduction. E2 is not refuted: positive counts remain below its registered rejection thresholds, and the negative direction, which E2's clause does not cover, exceeds the same bound (Appendix A). Calibration declines addition throughout PubMed, so the two coincide in the PubMed cells of Table 2. The Cora cells were selected after the original outcomes were observed; the follow-up remains exploratory and does not alter E2.

Recovery diagnostics show what addition adds. In the registered oracle experiment, exact restoration of deleted edges improves accuracy by up to +0.98 [+0.61, +1.35] pp, a point estimate below E5's registered 1 pp bound, although the intervals reach +1.40 pp (Appendix A). On Cora and CiteSeer, the candidate pool contains 52 to 82 percent of deleted edges, yet similarity ranking with the true deletion count recovers only 0.5 to 4.2 percent. Ranking therefore limits reconstruction even when candidates are available; PubMed also has limited candidate coverage. In the deployed stage, Table A19 shows that 97 percent of added edges on citation-fresh and 91 percent on citation-reproduction had never been deleted, which supports the distinction between predictive augmentation and topological repair.

5.5 Sensitivity to the Adaptive Attack

On Cora, where pruning is inactive, the adaptive family lowers the criterion arm's accuracy significantly more than generalized DICE does in 10 of 12 cells at matched budget and share. The mean gap, generalized-DICE accuracy minus adaptive accuracy, reaches +2.47 pp at the largest budget. As Figure A5 depicts, where pruning is active the gap is significant in 0 of the 12 CiteSeer cells and 2 of the 12 PubMed cells, which contradicts E3's literal prediction on PubMed. Across this family's 36 cells, the criterion arm is 20W/16NS/0L against the undefended GCN, with every gain on CiteSeer and PubMed and 0 on Cora, where pruning is inactive. Against calibrated GCN-Jaccard, it is 0W/36NS/0L; the criterion arm builds GCN-Jaccard's graph in every PubMed cell.

5.6 Registered Outcomes

E1 and E2 are not refuted under the registered rule. E1's citation-fresh loss count nevertheless reaches both dependence-aware bounds, so E1's holding does not establish parity with GCN-Jaccard (Appendix A). E4 holds with room to spare: the structure-free comparator, predicted outside the top three in at least 80 percent of citation-fresh cells, reaches the top three in none of the 130. E3 fails as written on PubMed. E5's accuracy clause holds on the point estimate. Its candidate-reach ceiling fails for 4 of 6 PubMed families: we had set the prediction from feature-pool measurements, whereas the registered endpoint concerns the union pool.

6 Discussion

Configuration choice changes what a purification benchmark measures. The exploratory ablations show that the ladder raises mean attacked-graph accuracy over clean-only tuning for GCN-Jaccard and GCN-SVD as well as GraphSieve, in both profiles, so tuning only the new method does not isolate editing. Under common calibration, the differences from GCN-Jaccard are small, yet E1's citation-fresh loss count reaches the dependence-aware bounds (Section 5.6), so parity is not established.

Two results show why the ladder's severity distribution must be reported: its shared loss to strong single-severity tuning on citation-reproduction for GraphSieve and GCN-Jaccard, and GCN-SVD's clean-accuracy cost on CiteSeer. Weighting milder rungs can favor configurations that underperform when the evaluation attack rewards stronger pruning. Three reporting practices follow. Report clean accuracy and each selection's fit count beside attacked-graph accuracy, and flag cells in which calibrated methods build the same graph, where no difference is possible. Treat cell-wise counts as descriptive, because cells sharing calibration draws and seeds are not independent.

Limitations. The conclusions are bounded by the evaluation setting and by dependence among cells. Calibration assumes a trusted clean graph and validation labels that a defender could train on directly. The replication holds the split and calibrated configurations fixed, and its Cora cells were chosen after the original outcomes. The adaptive family knows the published criterion but not the frozen thresholds or the trained model (Mujkanovic et al., 2022), and injection attacks that preserve the similarity signal need separate study (Chen et al., 2022). The evidence concerns three homophilous citation graphs small enough for CPU training and the evaluated baselines, without establishing performance on heterophilous or larger graphs, or against architectural defenses beyond RGCN.

7 Conclusion

GraphSieve gains significantly over GCN-SVD in 117 of 130 citation-fresh and 14 of 18 citation-reproduction cells, and over RGCN in 94 and 17, with no significant loss to either baseline. In exploratory ablations, calibration raises mean attacked-graph accuracy over clean-only tuning for each purification method, in both profiles, although its advantage over single-severity selection depends on the evaluation regime. A focused replication reproduces augmentation gains in Cora deletion settings chosen after the original outcomes, despite limited recovery of original edges. These results motivate a common calibration protocol and distinguish predictive augmentation from reconstruction of topology.

Ethics statement

This work evaluates defenses against structural poisoning of graph neural networks on three public citation benchmarks (Cora, CiteSeer, PubMed), whose nodes are scientific publications with text-derived features. The attacks studied follow established academic threat models, some as restrictions or reparameterizations of published attacks; the only new attack, an adaptive deletion rule for DICE, targets the similarity signal of the defense under test and is evaluated only on these benchmarks. The study reports both benefits and adverse effects of the proposed components, and it reports the registered predictions that were not supported.

Reproducibility statement

Section 4 specifies the statistics, seed pairing, and information-access controls. Appendix A records the prespecified hypotheses and decision rule; Appendix B reports every registered per-cell result. Appendices C and D describe the exploratory calibration ablations and replication. An anonymized repository (Appendix E) contains the study package, the configuration, the scripts that rerun every reported experiment, and a Docker environment pinning the CPU software stack.

References

Lada A. Adamic and Eytan Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, 2003. doi: 10.1016/S0378-8733(03)00009-1.

Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. The socialbot network: When bots socialize for fame and money. In *Proceedings of the Annual Computer Security Applications Conference*, pp. 93–102, 2011. doi: 10.1145/2076732.2076746.

Yongqiang Chen, Han Yang, Yonggang Zhang, Kaili Ma, Tongliang Liu, Bo Han, and James Cheng. Understanding and improving graph injection attack by promoting unnoticeability. In *International Conference on Learning Representations*, 2022.

Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized PageRank graph neural network. In *International Conference on Learning Representations*, 2021.

Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, Peter W. Battaglia, Vishal Gupta, Ang Li, Zhongwen Xu, Alvaro Sanchez-Gonzalez, Yujia Li, and Petar Veličković. ETA prediction with graph neural networks in Google Maps. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 3767–3776, 2021. doi: 10.1145/3459637.3481916.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S. Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 315–324, 2020. doi: 10.1145/3340531.3411903.

Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, and Evangelos E. Papalexakis. All you need is low (rank): Defending against adversarial attacks on graphs. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pp. 169–177, 2020. doi: 10.1145/3336191.3371789.

Simon Geisler, Daniel Zügner, and Stephan Günnemann. Reliable graph neural networks via robust aggregation. In *Advances in Neural Information Processing Systems*, 2020.

Lukas Gosch, Daniel Sturm, Simon Geisler, and Stephan Günnemann. Revisiting robustness in graph machine learning. In *International Conference on Learning Representations*, 2023.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with TrustRank. In *Proceedings of the International Conference on Very Large Data Bases*, pp. 576–587, 2004. doi: 10.1016/B978-012088469-8/50052-8.

Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. FRAUDAR: Bounding graph fraud in the face of camouflage. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 895–904, 2016. doi: 10.1145/2939672.2939747.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 66–74, 2020. doi: 10.1145/3394486.3403049.

Wei Jin, Yaxin Li, Han Xu, Yiqi Wang, Shuiwang Ji, Charu Aggarwal, and Jiliang Tang. Adversarial attacks and defenses on graphs: A review, a tool and empirical studies. *ACM SIGKDD Explorations Newsletter*, 22(2):19–34, 2021. doi: 10.1145/3447556.3447566.

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

Kuan Li, Yang Liu, Xiang Ao, and Qing He. Revisiting graph adversarial attack and defense from a data distribution perspective. In *International Conference on Learning Representations*, 2023.

Kuan Li, Yiwen Chen, Yang Liu, Jin Wang, Qing He, Minhao Cheng, and Xiang Ao. Boosting the adversarial robustness of graph neural networks: An OOD perspective. In *International Conference on Learning Representations*, 2024.

Yaxin Li, Wei Jin, Han Xu, and Jiliang Tang. DeepRobust: A platform for adversarial attacks and defenses. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):16078–16080, 2021. doi: 10.1609/aaai.v35i18.18017.

Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? In *International Conference on Learning Representations*, 2022.

Felix Mujkanovic, Simon Geisler, Stephan Günnemann, and Aleksandar Bojchevski. Are defenses for graph neural networks robust? In *Advances in Neural Information Processing Systems*, pp. 8954–8968, 2022.

Tran Gia Bao Ngo, Zulfikar Alom, Federico Errica, Murat Kantarcioglu, and Cuneyt Gurcan Akcora. Adversarial graph neural network benchmarks: Towards practical and fair evaluation, 2026. arXiv:2605.05534.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations*, 2023.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020. doi: 10.1016/j.cell.2020.01.021.

Marcin Waniek, Tomasz P. Michalak, Michael J. Wooldridge, and Talal Rahwan. Hiding individuals and communities in a social network. *Nature Human Behaviour*, 2:139–147, 2018. doi: 10.1038/s41562-017-0290-3.

Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. Adversarial examples for graph data: Deep insights into attack and defense. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4816–4823, 2019. doi: 10.24963/ijcai.2019/669.

Beini Xie, Heng Chang, Ziwei Zhang, Xin Wang, Daixin Wang, Zhiqiang Zhang, Rex Ying, and Wenwu Zhu. Adversarially robust neural architecture search for graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Kaidi Xu, Hongge Chen, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Mingyi Hong, and Xue Lin. Topology attack and defense for graph neural networks: An optimization perspective. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3961–3967, 2019. doi: 10.24963/ijcai.2019/550.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48, 2016.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 974–983, 2018. doi: 10.1145/3219819.3219890.

Xiang Zhang and Marinka Zitnik. GNNGuard: Defending graph neural networks against adversarial attacks. In *Advances in Neural Information Processing Systems*, 2020.

Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1399–1407, 2019. doi: 10.1145/3292500.3330851.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, 2020.

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

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2847–2856, 2018. doi: 10.1145/3219819.3220078.

A Registered Hypotheses and Decision Rule

We use the paired statistic defined in Section 4: the mean test-accuracy difference over 10 replicates, with a 95 percent Student-t interval at 9 degrees of freedom ($t = 2.2622$). Populations are analyzed separately. A win or loss requires the interval to exclude zero. Attacks significantly reduce the undefended GCN's accuracy in 123 of 130 fresh cells and all 18 reproduction cells. The main analysis retains all cells. Restricted to harmed fresh cells, the criterion arm records 3W/114NS/6L against GCN-Jaccard, 88W/35NS/0L against RGCN, 110W/13NS/0L against GCN-SVD, and 52W/65NS/6L against the undefended GCN.

Decision rule. E1 concerns significant losses to GCN-Jaccard; E2 concerns significant gains over pruning alone. Under a zero-effect null and a calibrated two-sided interval, a result in either specified direction occurs with probability 0.025. The registered rule therefore rejects a population-level prediction when the count in its specified direction reaches the smallest k satisfying $P(X \geq k) \leq 0.05$ for $X \sim \mathrm{Binomial}(\mathrm{cells}, 0.025)$. This gives thresholds of 7 for citation-fresh and 3 for citation-reproduction. It changes neither individual intervals nor cell-wise verdicts. Its independence assumption motivates the sensitivity analysis below. E3–E5 retain the literal refutation clauses in the registration. Findings outside these hypotheses are exploratory.

Sensitivity to dependence (exploratory). The binomial bound treats the cells of a population as independent trials. They are not. Every cell of a dataset shares the 10 model seeds and frozen calibrations, and in 42 citation-fresh and 6 citation-reproduction cells, all on PubMed, the criterion arm coincides by construction with both GCN-Jaccard and its pruning-only ablation, so no wrong-way result can occur there. Over the remaining 88 and 12 cells the same binomial bound is 6 and 2, so E1's 6 citation-fresh losses reach the citation-fresh bound there as well, while E1 on citation-reproduction and E2 in both populations stay below theirs. A sign-flip null keeps the dependence among cells that share a seed, assuming that each seed's paired differences are symmetric about zero under the null and that seeds are exchangeable. It flips the sign of each seed's paired difference in every cell at once, recomputes every interval, and enumerates all 2^{10} patterns, either jointly over the population or independently per dataset. Under it, the refutation bound for E1 on citation-fresh is 6 (joint) and 6 (per dataset), and 6 or more losses have probability 0.050 and 0.048. Under this analysis, E1 on citation-fresh reaches the bound, although it held under the registered rule. E2 holds under it, with 3 positives against a bound of 6 on citation-fresh (probability 0.306) and 1 against 2 on citation-reproduction (0.186). E1 on citation-reproduction, with no loss, holds under both the registered rule and the sign-flip null. The analysis is post hoc and changes no registered verdict.

- **E1. No accumulated loss to GCN-Jaccard, per population.** Contrast: GraphSieve (criterion arm) minus the calibrated GCN-Jaccard, paired accuracy, per attacked cell. Predicted-against direction: significant losses. Outcome: 6 losses of 130 on citation-fresh against a threshold of 7, and 0 of 18 on citation-reproduction against 3; held in both under the registered rule (Table A1; see the sensitivity analysis above). The record is 4W/120NS/6L and 1W/17NS/0L. The quota arm, which is not the registered contrast, recorded 6 losses on citation-fresh and 4 on citation-reproduction; the latter count reaches the reproduction threshold, so only the criterion arm holds E1 there.
- **E2. Restoration's marginal effect does not accumulate significant positives.** Contrast: GraphSieve minus its pruning-only ablation, per attacked cell, per population and per family. Predicted-against direction: significant positives. Outcome: the criterion arm was significantly positive in 3 citation-fresh cells and 1 of the 18 reproduction cells, and significantly negative in 9 and 0; held in both populations (by family in Table A2). Its clause counts only significant positives. The 9 significant negatives among the 88 citation-fresh cells in which the two can differ exceed the same binomial bound of 6, the mirror image of the clause; the direction was not registered, the bound assumes independence, and this reading is exploratory. The quota arm, reported alongside, was positive in 6 and 0 cells and negative in 11 and 4, also below both thresholds.
- **E3. The adaptive family's defended advantage depends on the frozen pruning threshold.** Contrast: generalized-DICE accuracy minus adaptive-family accuracy on the defended model at matched budget and deletion share, per cell, conditional on the threshold calibration froze. Prediction: positive and growing with budget where $t = 0$ is frozen, slightly negative where $t > 0$; refuted by a significant defended advantage under $t > 0$ or by its absence under $t = 0$. Outcome: Cora, where $t = 0$ was frozen in 10 of 10 replicates, showed the advantage in 10 of 12 cells, rising

from +0.66 pp at the 5 percent budget to +2.47 pp at 25 percent. The 24 cells with active pruning were significant in 2, both on PubMed at the 25 percent budget (Table A22). The literal clause counts those 2 cells against the prediction, and we report it as not held as written; the registration set a multiplicity rule for E1 and E2 only.

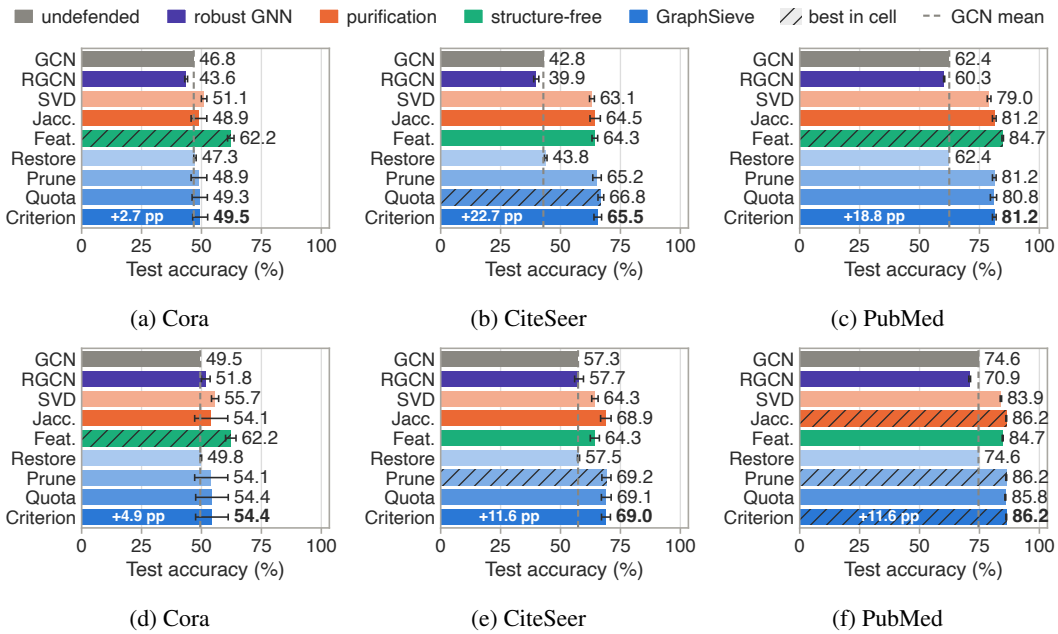
- **E4. The structure-free comparator ranks outside the top three on citation-fresh.** Contrast: the rank of the feature-only GCN among the 9 methods by mean test accuracy, per cell. Prediction: outside the top three in at least 80 percent of citation-fresh cells. Outcome: mean rank 8.68, top three in 0 of 130 cells; held. Citation-reproduction is not part of the hypothesis and is reported separately: mean rank 5.67, top three in 4 of 18 cells (22 percent), and first in 4 (Table A9).
- **E5. Reach has a ceiling below the damage.** Contrast: the fraction of attacker-deleted edges present in the restoration candidate pool, per dataset and family, and the accuracy gained by restoring every deleted edge exactly: the undefended GCN's accuracy on the clean graph plus the attack's insertions, minus its accuracy on the poisoned graph, in the oracle mechanism diagnostic (generalized DICE at the 5 percent budget, every deletion share, citation-fresh). Prediction: reach on PubMed stays below 20 percent under every registered family, and the perfect-restoration bound stays under 1 pp on citation-fresh; refuted by either. Outcome: the ceiling failed as written for 4 of 6 PubMed families, because the registered quantity is the union pool the stage ranks while the 20 percent figure had been set from feature-pool measurements that the study reproduces (Table A18, Appendix D). The bound held on the point estimate: at most +0.98 pp [+0.61, +1.35], on Cora at deletion shares 0.50 and 1.00, although the intervals extend to +1.40 pp. The diagnostic covers the 5 percent budget only. Similarity-ranked restoration at the true deletion count, in the same diagnostic, gains at most +0.29 pp [+0.14, +0.44] over count-matched pruning; this is a different quantity from the perfect-restoration bound.

Table A1: Outcome of the two population-level hypotheses under the registered multiplicity-aware rule. E1: GraphSieve minus GCN-Jaccard, refuted when significant losses reach the threshold. E2: GraphSieve minus its pruning-only ablation, refuted when significant positives reach it. Both contrasts are the criterion arm. Thresholds are the smallest k with $P(X \geq k) \leq 0.05$ for $X \sim \mathrm{Binomial}(\mathrm{cells}, 0.025)$.

			E1		E2	
Population	Cells	Refute at	Losses	Outcome	Positives	Outcome
citation-fresh	130	7	6	held	3	held
citation-reproduction	18	3	0	held	1	held

Table A2: E2 by population and attack family, the registered reporting unit: the criterion arm minus its pruning-only ablation. "Identical" counts cells in which calibration declined restoration, so the two build the same graph on every seed; the no-significant-difference column excludes them. The columns are the registered endpoint's own vocabulary for the four categories written W/NS/I/L elsewhere: Positive is a significant gain for the arm and Negative a significant loss, each by the paired 95 percent interval.

Population	Family	Cells	Positive	No sig. diff.	Identical	Negative
citation-fresh	generalized DICE	45	0	28	15	2
	random composition	45	1	23	15	6
	adaptive deletion DICE	36	0	23	12	1
	deletion Metattack	4	2	2	0	0
citation-reproduction	released Metattack	6	0	4	2	0
	standard DICE	6	0	4	2	0
	legacy DICE	6	1	3	2	0



(a) Cora (b) CiteSeer (c) PubMed

(d) Cora (e) CiteSeer (f) PubMed

Figure A1: All 9 methods in the six severe citation-reproduction cells; **top row:** legacy DICE, **bottom row:** released Metattack, each at its larger rate (100 and 25 percent of the clean edge count). Bars are mean accuracy; whiskers are the paired 95 percent interval of each method minus the undefended GCN, drawn around the method's mean, so a whisker clearing the dashed undefended mean marks a significant difference. Hatching marks the best mean, and the criterion arm's bar is labeled with its gain over the undefended GCN in pp. The structure-free comparator significantly outperforms every graph-based method in 3 of the six cells (Section 5.2). Jacc. is GCN-Jaccard, Feat. the structure-free comparator, and Restore and Prune the ablations.

B Settings and Per-cell Results

Table A4 specifies the fixed settings and candidate grids; Table A5 reports clean-graph accuracy differences. Tables A6–A9 give all 9 methods in the 148 attacked cells. The 133 cells with varying attacks use 1,330 graphs. A further 15 cells hold one graph fixed across model seeds: 9 deterministic adaptive-family cells at deletion share 1.00 and 6 released Metattack cells. The study therefore evaluates 1,345 attack graphs in total. Intervals for fixed-graph cells quantify model-seed variation conditional on that graph. Bold entries identify the highest replicate mean, including ties; significance is assessed through paired intervals.

Table A3: Datasets under the two profiles; h is the fraction of edges whose endpoints share a label. The reproduction profile keeps the largest connected component and the splits of Pro-GNN.

Profile	Dataset	Nodes	Edges	Feat.	Cls.	Train/val/test	h
Fresh	Cora	2,708	5,278	1,433	7	140/500/1000	0.810
	CiteSeer	3,327	4,552	3,703	6	120/500/1000	0.736
	PubMed	19,717	44,324	500	3	60/500/1000	0.802
Reproduction	Cora	2,485	5,069	1,433	7	247/249/1988	0.804
	CiteSeer	2,110	3,668	3,703	6	210/211/1688	0.736
	PubMed	19,717	44,324	500	3	1971/1972/15774	0.802

Table A4: Fixed settings and registered candidate grids, as recorded in the study specification of every artifact. Every method except RGCN, which runs at its authors' reference settings, shares the upper block; the lower block gives each method's own grid and its size. Grids are searched on validation data alone, and the selected configuration is then frozen for every evaluation cell, so a grid's size is a validation budget rather than a test-time choice. Pruning thresholds apply to the similarity of Equation 1, positive-support Jaccard on Cora and CiteSeer and cosine on PubMed, except that the quota arm binarizes features on every graph. GCN-SVD clips the negative entries of its rank-restricted reconstruction when a self-looped degree is not positive, which happened in 20 of 1,480 attacked fits and in no clean fit. The criterion arm's restoration levels are similarity thresholds, and the quota arm's are rates of the observed edge count. The bullet before each method gives its family as in Figure A1: undefended, robust architecture, purification, structure-free, and GraphSieve with its ablations. Neither guard was ever active: the pruning safeguard in none of the 6,160 attacked and clean fits of the two GraphSieve arms and their two ablations, and the restoration cap in none of the 1,480 attacked criterion-arm fits.

Setting or method	Value or registered grid	Candidates
Shared across every method except RGCN		
Backbone	GCN, 16 hidden units, dropout 0.5	
Optimizer	learning rate 0.01, weight decay 0.0005, 200 epochs, evaluating the weights of the last epoch to improve validation accuracy or validation loss	
Replicates per cell	10, paired by model seed 42–51	
Calibration seeds	model seeds 10042–10051, attack seeds 10042–10091, disjoint from evaluation	
Severity ladder	clean rung plus 0.05, 0.15, 0.25, 0.50, 1.00 of the clean edge count, random composition family at deletion share 0.50	
Attack budgets	5, 15, 25 percent at deletion shares 0.00, 0.25, 0.50, 0.75, 1.00 (adaptive family: 4 shares, no pure insertion); released and standard DICE at rates 0.15, 0.25; legacy DICE at 0.50, 1.00	
Statistics	paired difference, 95 percent Student-t, df 9, $t = 2.2622$	
Device	cpu	
Registered candidate grid, searched on validation data only		
• GCN	none	fixed
• RGCN	reference values: dropout 0.6, learning rate 0.01, $\beta = 0.0005$, $\gamma = 1.0$	fixed
• GCN-SVD	rank $\{5, 10, 15, 50\}$	4
• Feat.-only	none, the graph is discarded	fixed
• GCN-Jaccard	threshold t on the 13-point pruning ladder below	13
• GraphSieve, criterion arm	the same ladder crossed with restoration off or a similarity threshold in $\{0.50, 0.35, 0.25\}$	52
• GraphSieve, quota arm	the same ladder crossed with restoration off or a rate in $\{0.05, 0.10, 0.15\}$ of the observed edge count	52
• Prune-only, restore-only	parent-matched ablations of the criterion arm: its frozen t with restoration off, and its frozen restoration level with $t = 0$	fixed
Pruning ladder: $t \in \{0, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5\}$		
Restoration shortlist	per node, the 30 highest-similarity non-neighbors and, for the criterion arm, the 30 two-hop non-neighbors of highest Adamic–Adar weight; symmetrically unioned	

Table A5: Cost on the clean graph for the two GraphSieve arms, GCN-Jaccard and the structure-free comparator: each minus the undefended GCN in percentage points, beside the undefended GCN's own accuracy. The GraphSieve and GCN-Jaccard columns are paired over 10 seeds with the 95 percent interval, and an interval entirely below zero indicates a significant clean-accuracy loss; the Feat.-only column gives the mean difference alone. The two GraphSieve arms are tinted.

			Difference from the undefended GCN (pp)			
Profile	Dataset	GCN (%)	Criterion arm	Quota arm	GCN-Jaccard	Feat.-only
Fresh	Cora	80.42	−0.52 [−0.86, −0.18]	−0.43 [−0.79, −0.07]	+0.00 [+0.00, +0.00]	−26.73
	CiteSeer	68.30	−0.59 [−1.32, +0.14]	−0.96 [−1.79, −0.13]	−0.92 [−1.89, +0.05]	−16.89
	PubMed	78.72	−0.56 [−1.09, −0.03]	−0.19 [−0.57, +0.19]	−0.56 [−1.09, −0.03]	−5.99
Reproduction	Cora	83.48	−0.61 [−1.67, +0.45]	−0.66 [−1.72, +0.39]	−0.66 [−1.67, +0.34]	−21.25
	CiteSeer	72.04	−2.43 [−3.47, −1.39]	−2.17 [−3.22, −1.13]	−2.04 [−2.88, −1.20]	−7.73
	PubMed	86.85	−0.37 [−0.52, −0.22]	−0.49 [−0.70, −0.28]	−0.37 [−0.52, −0.22]	−2.16

Table A6: Citation-fresh population, Cora: mean test accuracy (percent) of the 9 methods in every attacked cell, over 10 seed-paired replicates. Columns are the undefended GCN, RGCN, GCN-SVD (SVD), the structure-free comparator (Feat.), GCN-Jaccard (Jacc.), the restoration-only and pruning-only ablations (Restore, Prune), and the quota and criterion arms of GraphSieve (Quota, Criterion). Budget is the edit budget as a percent of the clean edge count and Del. share the registered fraction of it spent on deletions. The four GraphSieve columns are tinted. Bold marks the row's highest mean, ties all bold.

				Baselines					GraphSieve			
Dataset	Attack	Budget (%)	Del. share	GCN	RGCN	SVD	Feat.	Jacc.	Restore	Prune	Quota	Criterion
Cora	Gen. DICE	5	0.00	**78.89**	77.58	70.21	53.69	**78.89**	78.41	**78.89**	78.53	78.41
		5	0.25	**79.07**	77.90	70.54	53.69	**79.07**	78.54	**79.07**	78.62	78.54
		5	0.50	78.83	77.88	71.04	53.69	78.83	78.57	78.83	**79.05**	78.57
		5	0.75	**79.29**	78.19	70.98	53.69	**79.29**	78.84	**79.29**	78.88	78.84
		5	1.00	**79.44**	78.52	70.74	53.69	**79.44**	78.91	**79.44**	78.99	78.91
		15	0.00	75.07	72.90	66.42	53.69	75.07	74.96	75.07	**75.55**	74.96
		15	0.25	75.20	73.83	66.57	53.69	75.20	**75.26**	75.20	75.09	**75.26**
		15	0.50	**75.90**	74.29	66.49	53.69	**75.90**	75.52	**75.90**	75.83	75.52
		15	0.75	76.55	75.57	68.25	53.69	76.55	**76.65**	76.55	76.56	**76.65**
		15	1.00	**77.64**	76.40	68.56	53.69	**77.64**	77.28	**77.64**	**77.64**	77.28
		25	0.00	72.44	69.06	62.67	53.69	72.44	72.56	72.44	**73.14**	72.56
		25	0.25	72.42	69.54	63.04	53.69	72.42	72.55	72.42	**72.83**	72.55
		25	0.50	72.39	70.08	62.88	53.69	72.39	**72.65**	72.39	72.48	**72.65**
		25	0.75	73.55	71.94	64.43	53.69	73.55	73.37	73.55	**73.78**	73.37
		25	1.00	75.04	74.25	66.42	53.69	75.04	75.38	75.04	**75.48**	75.38
	Random comp.	5	0.00	**79.15**	78.06	71.72	53.69	**79.15**	78.88	**79.15**	78.81	78.88
		5	0.25	**79.35**	78.64	71.54	53.69	**79.35**	78.79	**79.35**	78.91	78.79
		5	0.50	**79.48**	78.55	71.64	53.69	**79.48**	79.02	**79.48**	79.38	79.02
		5	0.75	**79.80**	78.63	72.18	53.69	**79.80**	79.42	**79.80**	79.63	79.42
		5	1.00	**79.71**	78.48	72.35	53.69	**79.71**	79.30	**79.71**	79.47	79.30
		15	0.00	77.14	74.68	69.14	53.69	77.14	76.93	77.14	**77.37**	76.93
		15	0.25	**77.70**	75.31	69.26	53.69	**77.70**	77.50	**77.70**	77.51	77.50
		15	0.50	77.70	76.17	69.05	53.69	77.70	77.67	77.70	**77.74**	77.67
		15	0.75	**78.38**	76.85	69.82	53.69	**78.38**	77.75	**78.38**	77.88	77.75
		15	1.00	**78.83**	77.79	70.75	53.69	**78.83**	78.18	**78.83**	78.29	78.18
		25	0.00	75.65	71.93	66.52	53.69	75.65	75.91	75.65	**76.39**	75.91
		25	0.25	75.93	73.03	66.63	53.69	75.93	**76.22**	75.93	76.19	**76.22**
		25	0.50	**75.97**	73.46	67.68	53.69	**75.97**	75.93	**75.97**	75.74	75.93
		25	0.75	**76.66**	74.22	67.72	53.69	**76.66**	76.09	**76.66**	76.40	76.09
		25	1.00	**77.57**	76.01	68.63	53.69	**77.57**	77.13	**77.57**	77.52	77.13
	Adaptive DICE	5	0.25	**77.65**	76.65	70.97	53.69	**77.65**	77.49	**77.65**	77.60	77.49
		5	0.50	78.07	77.61	70.16	53.69	78.07	78.05	78.07	**78.14**	78.05
		5	0.75	**78.54**	77.94	70.70	53.69	**78.54**	78.21	**78.54**	78.25	78.21
		5	1.00	**78.73**	78.23	70.96	53.69	**78.73**	78.46	**78.73**	78.46	78.46
		15	0.25	74.44	73.04	66.89	53.69	74.44	74.39	74.44	**74.62**	74.39
		15	0.50	74.83	73.48	67.42	53.69	74.83	**74.89**	74.83	74.83	**74.89**
		15	0.75	74.91	73.68	67.21	53.69	74.91	74.60	74.91	**74.94**	74.60
		15	1.00	**75.18**	74.40	65.93	53.69	**75.18**	74.94	**75.18**	75.17	74.94
		25	0.25	71.23	68.97	63.72	53.69	71.23	71.19	71.23	**71.37**	71.19
		25	0.50	**70.42**	68.33	62.28	53.69	**70.42**	70.33	**70.42**	70.25	70.33
		25	0.75	70.88	69.09	62.72	53.69	70.88	70.88	70.88	**71.24**	70.88
		25	1.00	**72.13**	70.65	62.09	53.69	**72.13**	71.68	**72.13**	**72.13**	71.68
	Del. Metattack	15	1.00	74.60	73.60	64.50	53.69	74.60	**75.49**	74.60	75.22	**75.49**
		25	1.00	72.76	71.44	63.30	53.69	72.76	**73.50**	72.76	72.95	**73.50**

Table A7: Citation-fresh population, CiteSeer: mean test accuracy of the 9 methods in every attacked cell. Columns and conventions as in Table A6.

				Baselines					GraphSieve			
Dataset	Attack	Budget (%)	Del. share	GCN	RGCN	SVD	Feat.	Jacc.	Restore	Prune	Quota	Criterion
CiteSeer	Gen. DICE	5	0.00	66.76	58.15	58.36	51.41	66.54	66.31	**66.82**	66.28	66.80
		5	0.25	**66.87**	58.83	58.37	51.41	66.39	66.20	66.59	66.29	66.54
		5	0.50	**67.19**	58.86	59.27	51.41	66.67	66.91	67.18	66.77	66.59
		5	0.75	**67.44**	59.75	59.40	51.41	66.64	66.83	66.90	66.88	66.96
		5	1.00	**67.74**	59.78	60.40	51.41	66.68	67.49	66.99	66.79	67.12
		15	0.00	63.16	53.89	54.98	51.41	65.28	63.22	65.36	65.39	**65.46**
		15	0.25	63.99	54.54	56.12	51.41	65.30	63.90	65.47	**65.54**	65.52
		15	0.50	64.18	55.37	56.83	51.41	65.42	64.15	**65.79**	65.49	65.34
		15	0.75	65.32	56.92	57.03	51.41	65.36	65.14	65.53	65.57	**65.78**
		15	1.00	**66.76**	57.76	58.75	51.41	65.59	66.54	65.86	66.13	66.03
		25	0.00	60.11	50.53	53.24	51.41	**64.47**	60.29	64.12	64.40	63.93
		25	0.25	60.60	51.37	53.54	51.41	63.88	61.00	63.90	**63.94**	63.70
		25	0.50	61.61	52.93	54.65	51.41	64.08	62.09	64.10	**64.27**	63.94
		25	0.75	62.67	53.68	55.86	51.41	63.74	63.19	63.69	**64.22**	64.03
		25	1.00	64.92	55.43	58.15	51.41	63.95	**65.26**	64.33	64.22	64.70
	Random comp.	5	0.00	**67.33**	59.72	59.81	51.41	66.78	67.14	67.11	66.55	66.97
		5	0.25	**67.58**	59.92	59.54	51.41	67.11	67.46	67.41	66.97	66.98
		5	0.50	**67.79**	60.76	60.24	51.41	67.41	67.44	67.78	66.99	67.56
		5	0.75	**68.05**	60.20	60.67	51.41	67.20	68.02	67.80	66.92	67.41
		5	1.00	67.86	60.40	60.68	51.41	66.99	**67.88**	67.42	66.91	67.71
		15	0.00	64.99	56.30	56.76	51.41	66.64	65.03	**66.65**	66.05	66.38
		15	0.25	65.18	57.02	57.13	51.41	66.48	65.42	66.55	66.32	**66.63**
		15	0.50	66.19	57.46	58.07	51.41	66.03	65.99	66.44	**66.48**	66.06
		15	0.75	67.00	59.00	58.69	51.41	65.90	**67.11**	66.33	66.52	66.50
		15	1.00	**67.59**	59.59	59.75	51.41	66.25	67.56	66.62	66.48	67.05
		25	0.00	63.83	54.25	55.25	51.41	**66.15**	63.95	65.83	65.74	65.95
		25	0.25	64.42	54.61	55.38	51.41	66.02	64.35	66.09	66.07	**66.15**
		25	0.50	63.88	55.61	56.88	51.41	65.50	64.25	**66.06**	65.91	65.83
		25	0.75	65.44	56.84	58.46	51.41	65.64	65.47	65.72	65.68	**65.92**
		25	1.00	67.04	58.01	59.21	51.41	65.20	**67.07**	65.58	66.08	66.05
	Adaptive DICE	5	0.25	66.90	58.67	59.68	51.41	67.41	66.54	**67.80**	67.15	67.39
		5	0.50	67.24	59.09	58.50	51.41	67.03	67.27	**67.44**	67.02	67.20
		5	0.75	**67.55**	58.99	59.77	51.41	66.91	67.19	67.30	66.88	67.26
		5	1.00	67.33	59.00	58.04	51.41	67.32	67.22	**67.72**	67.01	67.46
		15	0.25	63.72	55.85	56.19	51.41	66.02	63.96	**66.33**	65.94	65.98
		15	0.50	63.20	54.99	55.38	51.41	65.82	63.41	**66.28**	65.77	65.73
		15	0.75	64.43	56.00	57.03	51.41	66.28	64.42	**66.34**	66.20	66.13
		15	1.00	64.27	56.81	57.71	51.41	65.81	64.55	65.66	**65.93**	65.61
		25	0.25	60.69	51.74	52.62	51.41	65.07	60.78	**65.09**	64.71	64.91
		25	0.50	60.34	51.71	53.79	51.41	64.33	60.68	64.05	**64.46**	64.23
		25	0.75	60.97	52.36	55.06	51.41	64.09	60.97	63.69	63.84	**64.10**
		25	1.00	62.40	54.13	56.72	51.41	64.83	62.54	64.41	**64.93**	64.67
	Del. Metattack	15	1.00	67.38	56.92	55.79	51.41	66.03	**68.27**	66.33	66.67	66.79
		25	1.00	66.58	55.93	56.18	51.41	65.05	**67.64**	65.48	65.20	66.04

Table A8: Citation-fresh population, PubMed: mean test accuracy of the 9 methods in every attacked cell. Columns and conventions as in Table A6. Because calibration froze restoration off on PubMed in every replicate of the criterion arm (Appendix C), the Criterion column coincides with Prune and the Restore column with GCN.

				Baselines					GraphSieve			
Dataset	Attack	Budget (%)	Del. share	GCN	RGCN	SVD	Feat.	Jacc.	Restore	Prune	Quota	Criterion
PubMed	Gen. DICE	5	0.00	75.63	75.33	74.15	72.73	77.10	75.63	77.10	**77.23**	77.10
		5	0.25	76.12	75.68	74.70	72.73	**77.45**	76.12	**77.45**	77.37	**77.45**
		5	0.50	76.75	76.28	75.31	72.73	77.34	76.75	77.34	**77.76**	77.34
		5	0.75	77.28	76.73	75.69	72.73	77.67	77.28	77.67	**77.87**	77.67
		5	1.00	78.11	77.68	75.87	72.73	77.93	78.11	77.93	**78.19**	77.93
		15	0.00	70.91	70.56	72.01	72.73	**75.07**	70.91	**75.07**	74.94	**75.07**
		15	0.25	71.96	71.56	73.63	72.73	75.16	71.96	75.16	**75.27**	75.16
		15	0.50	73.32	72.85	74.49	72.73	**75.75**	73.32	**75.75**	75.69	**75.75**
		15	0.75	74.81	74.27	74.42	72.73	76.42	74.81	76.42	**76.58**	76.42
		15	1.00	**77.10**	76.27	75.30	72.73	76.90	**77.10**	76.90	76.99	76.90
		25	0.00	67.04	66.97	70.61	72.73	73.16	67.04	73.16	**73.46**	73.16
		25	0.25	68.66	68.33	72.20	72.73	73.56	68.66	73.56	**73.75**	73.56
		25	0.50	70.17	69.58	73.00	72.73	74.27	70.17	74.27	**74.30**	74.27
		25	0.75	72.21	71.79	74.15	72.73	**74.91**	72.21	**74.91**	74.74	**74.91**
		25	1.00	75.43	74.74	74.63	72.73	**75.66**	75.43	**75.66**	75.61	**75.66**
	Random comp.	5	0.00	77.21	76.78	74.90	72.73	77.66	77.21	77.66	**77.82**	77.66
		5	0.25	77.49	77.05	75.03	72.73	77.74	77.49	77.74	**77.80**	77.74
		5	0.50	77.51	77.04	75.20	72.73	**77.89**	77.51	**77.89**	77.78	**77.89**
		5	0.75	**78.27**	77.64	75.30	72.73	77.91	**78.27**	77.91	78.11	77.91
		5	1.00	**78.50**	77.89	75.86	72.73	78.11	**78.50**	78.11	78.23	78.11
		15	0.00	75.41	74.99	73.67	72.73	76.82	75.41	76.82	**76.89**	76.82
		15	0.25	75.54	75.23	74.05	72.73	76.48	75.54	76.48	**77.00**	76.48
		15	0.50	76.00	75.61	74.98	72.73	76.92	76.00	76.92	**77.09**	76.92
		15	0.75	76.39	75.94	76.11	72.73	76.82	76.39	76.82	**77.03**	76.82
		15	1.00	**77.51**	76.82	76.17	72.73	77.09	**77.51**	77.09	77.06	77.09
		25	0.00	73.56	73.92	73.38	72.73	**76.09**	73.56	**76.09**	76.05	**76.09**
		25	0.25	73.64	73.67	74.02	72.73	**76.08**	73.64	**76.08**	75.86	**76.08**
		25	0.50	74.02	74.06	74.79	72.73	75.65	74.02	75.65	**75.76**	75.65
		25	0.75	75.24	75.09	75.64	72.73	**76.43**	75.24	**76.43**	76.09	**76.43**
		25	1.00	**76.96**	76.45	76.05	72.73	76.59	**76.96**	76.59	76.70	76.59
	Adaptive DICE	5	0.25	75.77	75.36	74.23	72.73	77.15	75.77	77.15	**77.26**	77.15
		5	0.50	76.26	75.77	75.09	72.73	**77.55**	76.26	**77.55**	77.54	**77.55**
		5	0.75	76.64	75.89	76.14	72.73	**77.80**	76.64	**77.80**	77.67	**77.80**
		5	1.00	77.31	76.54	75.81	72.73	77.85	77.31	77.85	**77.90**	77.85
		15	0.25	71.37	70.55	74.27	72.73	**75.34**	71.37	**75.34**	75.16	**75.34**
		15	0.50	72.17	71.67	74.38	72.73	**75.80**	72.17	**75.80**	75.16	**75.80**
		15	0.75	74.05	73.57	74.27	72.73	**76.50**	74.05	**76.50**	75.59	**76.50**
		15	1.00	75.74	74.96	75.01	72.73	**76.68**	75.74	**76.68**	75.87	**76.68**
		25	0.25	67.48	67.37	72.37	72.73	**73.69**	67.48	**73.69**	72.61	**73.69**
		25	0.50	68.90	68.67	72.77	72.73	**74.22**	68.90	**74.22**	73.48	**74.22**
		25	0.75	69.68	69.60	72.92	72.73	**73.44**	69.68	**73.44**	72.86	**73.44**
		25	1.00	72.79	73.38	72.41	72.73	**73.56**	72.79	**73.56**	72.98	**73.56**

Table A9: Citation-reproduction population, all 18 attacked cells: mean test accuracy of the 9 methods on the component-restricted graphs and splits. Budget is the perturbation rate of the released Metattack matrices and of the size-matched standard DICE attack, and the edge budget of the legacy DICE regime; Del. share is measured on the released matrices and registered for the DICE families. The released Metattack cells hold one attack graph across the 10 seeds. Columns and conventions as in Table A6.

				Baselines					GraphSieve			
Dataset	**Attack**	**Budget (%)**	**Del. share**	**GCN**	**RGCN**	**SVD**	**Feat.**	**Jacc.**	**Restore**	**Prune**	**Quota**	**Criterion**
Cora	Rel. Metattack	15	0.01	65.31	63.40	**69.12**	62.23	67.47	65.44	67.47	67.73	67.60
		25	0.04	49.49	51.80	55.67	**62.23**	54.06	49.77	54.06	54.38	54.35
	Std. DICE	15	0.50	79.75	78.71	73.08	62.23	79.44	**79.76**	79.44	79.36	79.45
		25	0.50	76.40	75.29	70.11	62.23	76.62	76.68	76.62	76.83	**76.90**
	Legacy DICE	50	0.50	67.93	65.19	62.35	62.23	68.62	68.32	68.62	68.99	**69.01**
		100	0.50	46.78	43.61	51.11	**62.23**	48.95	47.28	48.95	49.28	49.46
CiteSeer	Rel. Metattack	15	0.02	64.61	64.21	68.89	64.32	69.28	64.47	**69.43**	69.15	69.12
		25	0.02	57.33	57.74	64.35	64.32	68.91	57.52	**69.16**	69.09	68.96
	Std. DICE	15	0.50	68.64	67.93	66.49	64.32	69.24	68.96	69.38	**69.48**	69.09
		25	0.50	66.05	64.92	65.16	64.32	69.03	66.07	68.86	**69.41**	68.90
	Legacy DICE	50	0.50	59.30	56.48	63.09	64.32	67.35	59.34	67.54	**68.28**	67.65
		100	0.50	42.84	39.88	63.12	64.32	64.45	43.83	65.23	**66.83**	65.55
PubMed	Rel. Metattack	15	0.00	78.40	76.14	84.11	84.69	**86.30**	78.40	**86.30**	86.05	**86.30**
		25	0.00	74.59	70.92	83.94	84.69	**86.16**	74.59	**86.16**	85.79	**86.16**
	Std. DICE	15	0.50	82.25	81.30	83.16	84.69	**85.76**	82.25	**85.76**	85.59	**85.76**
		25	0.50	79.16	78.37	82.49	84.69	**85.27**	79.16	**85.27**	85.05	**85.27**
	Legacy DICE	50	0.50	72.44	71.45	80.75	**84.69**	84.03	72.44	84.03	83.67	84.03
		100	0.50	62.40	60.27	79.02	**84.69**	81.18	62.40	81.18	80.84	81.18

C CALIBRATION

This appendix records what calibration selected and what it cost on clean graphs, then reports the four exploratory calibration-policy ablations that Section 5.1 summarizes. For each profile, dataset, and replicate, calibration scores the method's registered candidate grid using validation labels and seed streams disjoint from evaluation. The ladder comprises the clean graph and 5 random perturbations at budgets of 0.05, 0.15, 0.25, 0.50, 1.00 of the clean edge count. Selection follows Section 3.2, and the configuration is reused for every evaluation cell of that replicate. Table A10 reports the modal configuration and its frequency, listing all modes when tied.

Selected configurations. Independent calibration frequently agrees on the pruning threshold. GCN-Jaccard's 13-candidate search selects the same threshold as the criterion arm in 27 of 30 fresh calibrations and 25 of 30 reproduction calibrations. The default threshold of 0.01 is selected in 0 of 60 calibrations. The mean validation gain over that candidate lies between 0.6 and 4.0 pp across profile-dataset pairs. These are selection-set quantities; test-set effects are assessed below.

On Cora, pruning is disabled in 10 of 10 fresh-profile replicates and 8 reproduction-profile replicates of the criterion arm. Fresh-profile Cora therefore isolates addition, with pruning alone reproducing the undefended graph. Restoration is active in 33 of 60 criterion-arm calibrations but disabled throughout PubMed. By comparison, the quota arm enables it in 47 calibrations, including 9 reproduction-profile PubMed replicates.

Edits on clean graphs. The criterion arm removes 0.0, 11.8, and 9.4 percent of observed edges on fresh-profile Cora, CiteSeer, and PubMed, respectively; the reproduction-profile values are 4.0, 60.3, and 33.9 percent. Addition amounts to 14.4, 7.5, and 0.0 percent of observed edges on the fresh profile, and 2.5, 5.3, and 0.0 percent on the reproduction profile. Values are averaged over replicates; Table A5 reports the associated accuracy differences.

Table A10: Modal frozen configuration per profile, dataset, and arm over 10 replicate calibrations: the pruning threshold t and the restoration level (a similarity threshold for the criterion arm, a rate of the observed edge count for the quota arm, or off). Calibration selects one configuration per model seed, so the last column reports how concentrated the selection was; a tie contributes one row per tied configuration. $t = 0$ is a pruning no-op.

			Frozen configuration		
Profile	Dataset	Arm	Prune t	Restore	Replicates
Fresh	Cora	criterion	0.000	0.25	9 of 10
	Cora	quota	0.000	0.15	8 of 10
	CiteSeer	criterion	0.020	0.25	4 of 10
	CiteSeer	quota	0.020	0.10	2 of 10 (tie)
	CiteSeer	quota	0.020	0.15	2 of 10 (tie)
	CiteSeer	quota	0.040	0.15	2 of 10 (tie)
	CiteSeer	quota	0.050	0.05	2 of 10 (tie)
	PubMed	criterion	0.100	off	5 of 10
	PubMed	quota	0.100	off	4 of 10
Reproduction	Cora	criterion	0.000	0.35	3 of 10 (tie)
	Cora	criterion	0.000	off	3 of 10 (tie)
	Cora	quota	0.000	off	5 of 10
	CiteSeer	criterion	0.100	0.35	3 of 10
	CiteSeer	quota	0.100	0.15	5 of 10
	PubMed	criterion	0.150	off	5 of 10 (tie)
	PubMed	criterion	0.200	off	5 of 10 (tie)
	PubMed	quota	0.150	0.15	7 of 10

Clean-only and light calibration (exploratory). This follow-up distinguishes the effect of the ladder from tuning alone on GCN-Jaccard's registered grid. Its design was fixed before computing its outcomes, after the registered study. Both alternatives use the same 13-point grid, calibration seeds, validation labels, and tie rule: clean-only selection scores the clean rung; light selection averages the clean and 5 percent rungs. Alternative thresholds are evaluated with the original splits, seeds,

attack graphs, and backbone. A pilot reproduced recorded ladder accuracies on one cell per family and dataset.

Against clean-only tuning, the ladder records 36W/82NS/12L on citation-fresh and 6W/12NS/0L on citation-reproduction, with mean differences of +0.44 and +4.18 pp (Table A11). Fresh-profile Cora contributes 44 identical cells because both policies select the same threshold. Fresh-profile PubMed contributes 28 significant gains and no losses. All 12 fresh-profile losses occur on CiteSeer, including 10 deletion-dominated conditions; the largest loss is -1.88 [-3.32, -0.44] pp. On clean graphs, the maximum mean cost is 1.10 pp, with a significant loss only on reproduction-profile PubMed (-0.36 [-0.52, -0.20] pp). Relative to light calibration, the mean differences are +0.30 and +3.33 pp, and the cell-wise records are 34W/80NS/16L on citation-fresh, whose NS count includes 44 identical cells on fresh-profile Cora, and 6W/12NS/0L on citation-reproduction.

Table A11: Exploratory calibration-policy ablation for GCN-Jaccard: the registered ladder selection minus a selection from the clean rung alone (clean-only) or from the clean and 5 percent rungs, per attacked cell by population and family, and on the clean graphs of each profile. W/NS/I/L counts ladder wins, cells with no significant difference, cells identical by construction (the two policies select the same threshold on every seed), and ladder losses; a win or loss requires the paired 95 percent interval to exclude zero. Mean is the mean of the per-cell paired differences.

			Ladder − clean-only		Ladder − clean + 5%	
Population	**Family**	**Cells**	**W/NS/I/L**	**Mean (pp)**	**W/NS/I/L**	**Mean (pp)**
citation-fresh	generalized DICE	45	11/14/15/5	+0.54	11/13/15/6	+0.42
	random composition	45	7/18/15/5	+0.02	6/18/15/6	−0.14
	adaptive deletion DICE	36	18/5/12/1	+0.97	17/5/12/2	+0.80
	deletion Metattack	4	0/1/2/1	−0.60	0/0/2/2	−0.69
	all attacked cells	130	36/38/44/12	+0.44	34/36/44/16	+0.30
	clean graph	3	0/2/1/0	−0.51	0/1/1/1	−0.49
citation-reproduction	released Metattack	6	2/4/0/0	+4.48	2/4/0/0	+3.61
	standard DICE	6	2/4/0/0	+1.46	2/4/0/0	+1.30
	legacy DICE	6	2/4/0/0	+6.61	2/4/0/0	+5.09
	all attacked cells	18	6/12/0/0	+4.18	6/12/0/0	+3.33
	clean graph	3	0/2/0/1	−0.61	0/2/0/1	−0.29

Table A12: Exploratory calibration effects for GCN-Jaccard by dataset: ladder selection minus the DeepRobust default or clean-only selection, per attacked cell. Conventions as in Table 1; "identical" marks a dataset whose two policies select the same threshold on every seed.

Population	**Ladder minus**	**Cora**	**CiteSeer**	**PubMed**	**All cells**
citation-fresh	DeepRobust default	33/5/6 (+1.10)	7/21/16 (−0.33)	29/13/0 (+1.57)	69/39/22 (+0.77)
	clean-only tuning	identical	8/24/12 (−0.16)	28/14/0 (+1.54)	36/82/12 (+0.44)
citation-reproduction	DeepRobust default	0/4/2 (−2.31)	5/1/0 (+5.89)	6/0/0 (+9.20)	11/5/2 (+4.26)
	clean-only tuning	0/6/0 (+1.59)	0/6/0 (+2.56)	6/0/0 (+8.40)	6/12/0 (+4.18)

Stronger single-severity policies (exploratory). A further prespecified follow-up asks whether the clean rung plus one stronger rung selects as well as the full ladder. The alternatives average validation accuracy over the clean rung and either the 25 or 100 percent rung, keeping the grid, seeds, labels, and fitting protocol fixed. These severities were selected before their test outcomes were computed, after the clean-only and light-policy results were known. Against the middle-severity policy, the ladder records 23W/98NS/9L on citation-fresh and 3W/15NS/0L on citation-reproduction. Mean differences are +0.25 and +1.15 pp, respectively (Table A13). Fresh-profile Cora is identical by construction.

Against the strongest policy, the direction depends on the profile. On citation-fresh, the ladder records 31W/95NS/4L and a mean difference of +0.33 pp. The gain concentrates on Cora (+1.06 pp), where the strongest policy leaves pruning off in only 4 of 10 replicates while the ladder disables it in every replicate. On citation-reproduction, the record is 1W/13NS/4L, with a mean difference of -1.39 pp. Cora contributes a difference of -3.85 pp and 4 significant losses. There, the ladder disables pruning in 8 replicates, compared with 2 for the strongest policy. The result illustrates how the weighting of milder calibration conditions can favor configurations that underperform when evaluation attacks reward stronger pruning.

Table A13: Exploratory single-severity calibration policies for GCN-Jaccard: the registered ladder selection minus a selection scored on the clean rung plus one attacked rung, per attacked cell by population and dataset, with the mean paired difference in pp. Conventions as in Table 1; a positive difference favors the ladder.

Population	Ladder minus	Cora	CiteSeer	PubMed	All cells
citation-fresh	clean plus the 25 percent rung	identical	4/31/9 (−0.17)	19/23/0 (+0.96)	23/98/9 (+0.25)
	clean plus the 100 percent rung	31/9/4 (+1.06)	0/44/0 (+0.03)	0/42/0 (−0.12)	31/95/4 (+0.33)
citation-reproduction	clean plus the 25 percent rung	0/6/0 (+1.59)	0/6/0 (+0.33)	3/3/0 (+1.55)	3/15/0 (+1.15)
	clean plus the 100 percent rung	1/1/4 (−3.85)	0/6/0 (−0.21)	0/6/0 (−0.12)	1/13/4 (−1.39)

GCN-SVD calibration policies (exploratory). The follow-up uses the registered 4-rank grid and compares the ladder with clean-only, light, middle, and strong calibration, plus DeepRobust's fixed default rank of 50 (Li et al., 2021). Table A14 reports the default, clean-only, middle, and strong comparisons. Against the default, the ladder records 19 significant gains among 130 fresh cells and 10 among 18 reproduction cells, with no significant losses. Mean differences are +0.31 and +2.30 pp. The largest dataset-level gain occurs on reproduction-profile CiteSeer (+5.63 pp). The absence of significant losses does not imply dominance: 8 attacked cells have negative point estimates.

Mean gains over clean-only tuning are +0.23 and +0.21 pp; those over middle-severity tuning are +0.00 and +0.08 pp. No significant losses occur against clean-only, light, or middle calibration. Against strong calibration, the ladder records 2 losses and 3 gains on citation-reproduction. Reproduction-profile CiteSeer shows the clearest clean-versus-attacked tradeoff: its clean-accuracy difference from the default is −1.64 [−2.76, −0.53] pp, the only significant clean-graph difference against that comparator.

Table A14: Exploratory calibration policies for GCN-SVD: the registered ladder selection minus four alternatives on the same rank grid, per attacked cell by population and dataset, with the mean paired difference in pp. Conventions as in Table 1; a positive difference favors the ladder.

Population	Ladder minus	Cora	CiteSeer	PubMed	All cells
citation-fresh	DeepRobust default rank 50	identical	identical	19/23/0 (+0.97)	19/111/0 (+0.31)
	clean-only tuning	identical	identical	10/32/0 (+0.73)	10/120/0 (+0.23)
	clean plus the 5 percent rung	identical	identical	0/42/0 (+0.33)	0/130/0 (+0.11)
	clean plus the 25 percent rung	identical	identical	0/42/0 (+0.00)	0/130/0 (+0.00)
	clean plus the 100 percent rung	0/44/0 (+1.53)	0/44/0 (+0.07)	identical	0/130/0 (+0.54)
citation-reproduction	DeepRobust default rank 50	0/6/0 (−0.11)	5/1/0 (+5.63)	5/1/0 (+1.36)	10/8/0 (+2.30)
	clean-only tuning	0/6/0 (−0.02)	0/6/0 (+0.36)	0/6/0 (+0.30)	0/18/0 (+0.21)
	clean plus the 5 percent rung	identical	0/6/0 (+0.31)	0/6/0 (+0.30)	0/18/0 (+0.20)
	clean plus the 25 percent rung	0/6/0 (−0.11)	0/6/0 (+0.36)	0/6/0 (−0.01)	0/18/0 (+0.08)
	clean plus the 100 percent rung	3/1/2 (+0.36)	identical	0/6/0 (−0.07)	3/13/2 (+0.10)

GraphSieve calibration policies (exploratory). A separate follow-up applies the same clean-only, light, middle, and strong selection policies to the criterion arm's registered grid, holding calibration traces, splits, evaluation graphs, seeds, and backbone fixed. Its design was frozen after the earlier calibration findings were known and before computing these outcomes. No library-default comparator is defined for GraphSieve. Table A15 reports all policies by population and dataset.

Relative to clean-only tuning, the ladder records 35W/91NS/4L on citation-fresh and 7W/11NS/0L on citation-reproduction, with mean differences of +0.49 and +3.92 pp. PubMed contributes the largest dataset-level gains (+1.54 and +8.40 pp); fresh-profile Cora's mean difference is −0.06 pp.

These results locate the calibration benefit within GraphSieve itself, rather than attributing it solely to baseline configuration.

The strongest-severity policy exposes the same limitation seen for GCN-Jaccard. The ladder's mean difference is +0.15 pp on citation-fresh but −1.44 pp on citation-reproduction, with 3 significant losses on reproduction-profile Cora. Clean-graph effects also vary: relative to clean-only tuning, mean clean accuracy decreases in every profile-dataset pair, with a significant loss on reproduction-profile PubMed. The attacked-graph gains therefore do not imply a cost-free calibration policy. All means are descriptive, the paired cell-wise intervals are unadjusted, and cells share calibration and seed dependence.

Table A15: Exploratory calibration policies for GraphSieve's criterion arm: ladder selection minus each alternative on the same configuration grid, per attacked cell by population and dataset. W/NS/L and mean differences follow Table 1; positive differences favor the ladder.

Population	Ladder minus	Cora	CiteSeer	PubMed	All cells
citation-fresh	clean-only tuning	0/44/0 (−0.06)	7/33/4 (+0.03)	28/14/0 (+1.54)	35/91/4 (+0.49)
	clean plus the 5 percent rung	0/44/0 (−0.05)	6/34/4 (−0.10)	27/15/0 (+1.24)	33/93/4 (+0.35)
	clean plus the 25 percent rung	0/44/0 (−0.07)	1/43/0 (+0.04)	22/20/0 (+0.97)	23/107/0 (+0.30)
	clean plus the 100 percent rung	0/44/0 (+0.27)	3/41/0 (+0.29)	0/42/0 (−0.12)	3/127/0 (+0.15)
citation-reproduction	clean-only tuning	1/5/0 (+1.81)	0/6/0 (+1.55)	6/0/0 (+8.40)	7/11/0 (+3.92)
	clean plus the 5 percent rung	2/4/0 (+1.85)	0/6/0 (+1.16)	6/0/0 (+7.15)	8/10/0 (+3.39)
	clean plus the 25 percent rung	0/6/0 (+1.78)	0/6/0 (+0.48)	3/3/0 (+1.55)	3/15/0 (+1.27)
	clean plus the 100 percent rung	2/1/3 (−3.64)	0/6/0 (−0.56)	0/6/0 (−0.12)	2/13/3 (−1.44)

D RESTORATION AND REACH DIAGNOSTICS

This appendix examines what the editing stages change and where those changes help or fail. It reports the two deployed variants, candidate coverage, realized edits, recovery of deleted edges, the Cora replication, and sensitivity to adaptive deletion DICE (Section 4), called the adaptive family below.

The two control laws. The quota and criterion arms share the pruning rule and the calibration procedure but differ in more than the budget rule. The quota arm binarizes features, so it scores PubMed by Jaccard rather than cosine; it ranks the feature pool alone; and it is calibrated separately, freezing a different pruning threshold from the criterion arm's in 16 of 60 replicates. The pruning-only ablation copies the criterion arm's threshold. Per attacked cell, they showed no significant difference in 117 of 130 citation-fresh cells, with the quota arm ahead in 9 and behind in 4, and in 14 of 18 reproduction cells, with the quota arm behind in 4 and ahead in none (Table A17). Against the pruning-only ablation, the two arms are likewise close: 3W/118NS/9L for the criterion arm and 6W/113NS/11L for the quota arm on citation-fresh, 1W/17NS/0L and 0W/14NS/4L on citation-reproduction. The comparison describes two deployed variants and does not isolate the budget rule.

Reach and realized edits. Table A18 and Figure A2 give the percent of attacker-deleted edges present in the feature pool (the similarity shortlists alone, which is all the quota arm ranks) and in the union pool the criterion arm ranks, which adds the structural candidates, per dataset and family, across every cell and every attack realization. The union pool holds 1.07 to 2.05 times as many of the deleted edges as the feature pool over the 16 dataset-family pairs with at least 1 percent feature reach. Two cases are excluded: the adaptive family, whose feature reach is near zero by construction because its deletion rule removes low-similarity edges and the shortlist ranks by similarity, and the released PubMed Metattack matrix, whose few deletions fall in neither pool.

On PubMed, the feature pool reaches 14.4 to 16.8 percent of the deleted edges under generalized DICE, and the union pool reaches 28.5 to 35.6 percent; this is the comparison behind the E5 outcome of Appendix A.

Figure A3 plots the realized prune fraction of the criterion arm under generalized DICE on the fresh profile. Pruning removes nothing on Cora, where $t = 0$ is frozen. On CiteSeer it removes 11.8 percent of the clean graph's edges, 14.1, 18.3, and 21.8 percent of the observed edges under pure insertion at the 5, 15, 25 percent budgets, and 11.9, 12.1, and 12.4 percent under pure deletion. On PubMed, the corresponding values are 9.4, then 12.0, 16.4, and 20.2 under insertion and 9.5, 9.5, and 9.7 under deletion.

The prune fraction rises with the insertion budget and stays near its clean-graph level under deletion, while the absolute count falls with the observed edge count, which is the response Section 3.1 attributes to the criterion. The criterion arm's restored fraction moves the other way. On Cora it is 13.7, 12.5, and 11.5 percent under pure insertion and 15.4, 17.8, and 20.7 percent under pure deletion at the three budgets, and on CiteSeer 7.1, 6.5, and 6.0 against 8.2, 10.0, and 12.1. Criterion-based addition therefore increases with deletion severity in these conditions, but this response does not establish a general accuracy benefit: E2, which predicts no accumulated gain from restoration, remains unrefuted.

What the deployed stage restores. Table A19 follows the criterion arm's restoration from reach to recovery over 722 citation-fresh and 84 citation-reproduction replicates of 76 and 12 Cora and CiteSeer cells (PubMed is excluded because calibration declined restoration there). On citation-fresh the pool holds 51 percent of the deleted edges on average, but the stage recovers 3.7 percent, at most 14.9 percent in any replicate, while adding 2.4 edges per deleted edge, 97 percent of them edges the attacker never deleted. On citation-reproduction the pool holds 59 percent, the stage recovers 3.2 percent, at most 10.0 percent, and 91 percent of the 5.5 edges added per deleted edge were never deleted. Under the adaptive family it recovers none: that attack deletes the lowest-similarity edges, which a similarity ranking places last. Recovery measures reconstruction of the original topology; an added edge the attacker never deleted can still help classification, and E2 measures that net effect on accuracy directly.

Where the criterion arm and GCN-Jaccard separated. The two methods separated in 5 cells in the criterion arm's favor (4 of 130 on citation-fresh, 1 of 18 on citation-reproduction) and 6 against it, all on citation-fresh (Tables A20 and A21). All 5 wins sit at deletion share 0.50 or above, with margins from +0.51 to +0.89 pp, and so do 4 of the 6 losses. The concentration of gains in deletion-dominated settings is consistent with a role for addition but does not make deletion share sufficient to predict improvement. All 6 losses are on Cora, 5 of them under the random-composition family, and none is larger than 0.65 pp. On Cora the criterion arm runs restoration alone (Appendix C), so each loss is a cell in which restoring edges cost accuracy. The losses are E1's registered count on citation-fresh, below its threshold of 7 (citation-reproduction recorded 0 against a threshold of 3). No registered hypothesis covers the wins, which are exploratory.

Replication of the Cora deletion-Metattack cells (exploratory). The original favorable cells were selected after their results were observed. A follow-up fixed its design and decision rule before generating 20 new attack graphs and evaluating new model seeds, starting at 20042 and disjoint from the original evaluation and calibration streams. A cell replicates when its paired interval excludes zero in the positive direction. Both conditions meet this criterion (Table A16). GCN-Jaccard again matches the undefended GCN on every replicate, so the contrast isolates addition. The follow-up tests repeatability over attack and model randomness conditional on the original dataset, public split, and frozen configurations; it does not resample calibration or extend the finding to other datasets.

Table A16: Original and follow-up Cora deletion-Metattack results. Values are GraphSieve minus calibrated GCN-Jaccard in percentage points, with paired 95 percent Student-t intervals over 10 replicates. The follow-up uses new attack graphs and model seeds, holding the split and configurations fixed. Pruning is disabled in both methods. Positive pairs count follow-up replicates in which GraphSieve is more accurate. Both conditions satisfy the prespecified follow-up criterion.

Edge budget	Original gain	Follow-up gain	Positive pairs
15 percent	+0.89 [+0.33, +1.45]	+1.21 [+0.77, +1.65]	10 of 10
25 percent	+0.74 [+0.19, +1.29]	+0.93 [+0.26, +1.60]	8 of 10

The adaptive family by dataset. Table A22 breaks the E3 contrast down by dataset. The defended advantage of the adaptive family concentrates on Cora, where pruning is frozen off, and appears on PubMed only at the largest budget. On CiteSeer, the contrast is never significant in either direction, and no cell of any dataset is significantly negative. Under this family, the criterion arm beat the undefended GCN in 20 of 36 cells and showed no significant losses (Cora 0W/12NS/0L, CiteSeer 8W/4NS/0L, PubMed 12W/0NS/0L). It showed no significant loss to the calibrated GCN-Jaccard and was 20W/6NS/10L against GCN-Jaccard at its default threshold (Appendix D.1).

Table A17: The quota and criterion variants compared: the quota arm against the criterion arm, per attacked cell, by population. A win requires the paired 95 percent interval to exclude zero.

Population	Cells	Quota wins	No sig. diff.	Criterion wins
citation-fresh	130	9	117	4
citation-reproduction	18	0	14	4

D.1 GCN-Jaccard at the Default Threshold

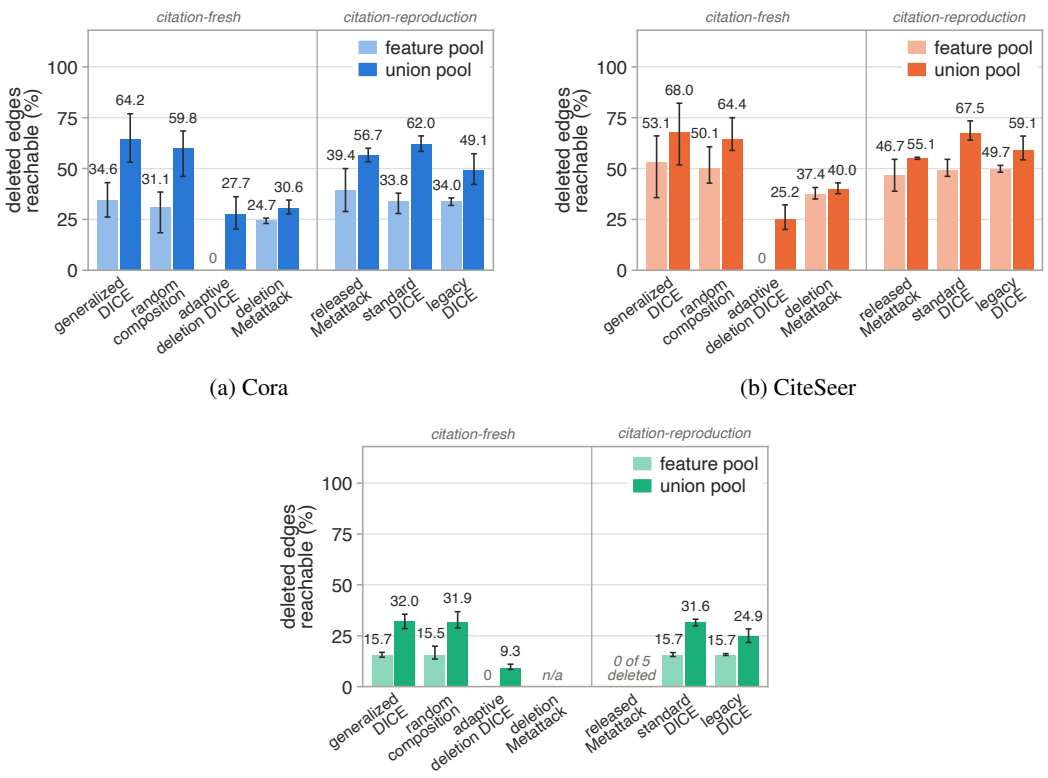
This subsection is exploratory and is not part of the registered study. GCN-Jaccard's reference implementation, DeepRobust (Li et al., 2021), sets a default threshold of 0.01. Wu et al. (2019) remove the edges of zero similarity, and on the clean fresh-profile Cora and CiteSeer graphs no edge scores strictly between 0 and 0.01, so there the default removes exactly those edges (572 and 117); PubMed, scored by cosine, has no counterpart. The ladder never selected the default (Appendix C). To measure the corresponding test-set effect, we evaluated an arm fixed at the default threshold after the registered study on copies of the six study artifacts (one per profile and dataset) with identical seeds, splits, attack graphs, and backbone. Each of the 148 paired contrasts below therefore rests on the evaluations reported elsewhere in the paper, and only the added arm is new. Its candidate grid has size 1, so its calibration is a no-op: all 60 replicate calibrations select 0.01. Table A23 gives four

Table A18: Restoration reach: percent of attacker-deleted edges present in the feature pool (the similarity shortlists alone, all the quota arm ranks) and in the union pool the criterion arm ranks, which adds structural candidates, minimum and maximum across all cells of the family and 10 attack realizations each. On PubMed the released Metattack matrices are almost purely insertions, and deletion Metattack is not run there because it does not scale to PubMed. The registered E5 ceiling of 20 percent on PubMed applies to the union pool.

			Feature pool (%)		Union pool (%)	
Population	Dataset	Family	min	max	min	max
citation-fresh	Cora	generalized DICE	26.2	43.1	53.1	76.9
	Cora	random composition	18.5	38.5	46.2	68.5
	Cora	adaptive deletion DICE	0.0	0.3	20.3	36.2
	Cora	deletion Metattack	23.0	25.6	27.7	34.5
	CiteSeer	generalized DICE	35.7	66.1	51.8	82.1
	CiteSeer	random composition	42.9	60.7	58.9	75.0
	CiteSeer	adaptive deletion DICE	0.0	0.2	20.1	32.1
	CiteSeer	deletion Metattack	35.0	40.7	37.7	43.0
	PubMed	generalized DICE	14.4	16.8	28.5	35.6
	PubMed	random composition	13.5	19.9	28.8	36.8
	PubMed	adaptive deletion DICE	0.0	0.0	8.4	11.0
citation-reproduction	Cora	released Metattack	28.9	50.0	53.3	60.0
	Cora	standard DICE	27.9	37.9	58.5	66.1
	Cora	legacy DICE	31.9	35.6	42.2	57.3
	CiteSeer	released Metattack	38.9	54.5	54.5	55.6
	CiteSeer	standard DICE	46.2	54.5	64.0	73.5
	CiteSeer	legacy DICE	48.2	51.6	54.3	66.0
	PubMed	released Metattack	0 of 5 deleted edges (the matrix inserts 11,076)			
	PubMed	standard DICE	14.8	16.7	29.8	33.1
	PubMed	legacy DICE	15.3	16.2	21.7	28.4

Table A19: Exploratory: what the deployed criterion arm restores, on every Cora and CiteSeer cell with deletions and every replicate in which calibration turned restoration on. The edited graphs were rebuilt from the frozen configurations without training, and each matches the fingerprint recorded in the study. Reach is the percent of attacker-deleted edges in the candidate pool; recovered is the percent of those edges the stage restored; precision is the percent of restored edges that were attacker-deleted; and the last column gives the restored edges per deleted edge. All are means over replicates.

Population	Dataset	Family	Reach (%)	Recovered (%)	Precision (%)	Added/deleted
citation-fresh	Cora	generalized DICE	64.2	5.1	3.3	3.14
	Cora	random composition	59.8	4.2	2.7	3.13
	Cora	adaptive deletion DICE	27.7	0.0	0.0	3.09
	Cora	deletion Metattack	30.6	2.6	4.0	0.79
	CiteSeer	generalized DICE	68.0	7.5	6.9	1.86
	CiteSeer	random composition	64.3	5.7	5.2	1.84
	CiteSeer	adaptive deletion DICE	25.2	0.0	0.0	1.79
	CiteSeer	deletion Metattack	40.0	3.0	6.8	0.47
citation-reproduction	Cora	released Metattack	56.7	6.6	0.7	15.61
	Cora	standard DICE	60.8	2.2	5.0	0.56
	Cora	legacy DICE	48.9	2.1	16.0	0.17
	CiteSeer	released Metattack	55.1	0.0	0.0	15.87
	CiteSeer	standard DICE	67.8	4.3	6.6	0.67
	CiteSeer	legacy DICE	59.2	4.5	21.3	0.22

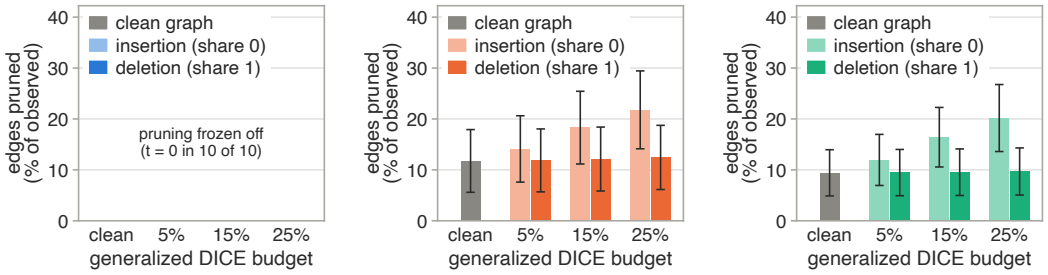


(a) Cora

(b) CiteSeer

(c) PubMed

Figure A2: Reach of the feature pool (light bars) and the union pool (dark bars), as in Table A18: one panel per dataset, the attack families of each population along the horizontal axis, and the percent of attacker-deleted edges each pool holds on the vertical axis. Bars are the mean over every replicate of every deleting cell in the family, whiskers span the replicate minimum and maximum, and a 0 marks a bar whose mean reach rounds to zero. On PubMed the released Metattack matrices are almost purely insertions, so their slot gives counts rather than bars, and deletion Metattack, marked n/a, is not run there because it does not scale to PubMed.

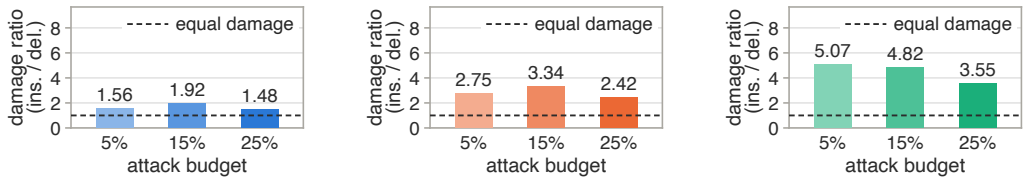


(a) Cora

(b) CiteSeer

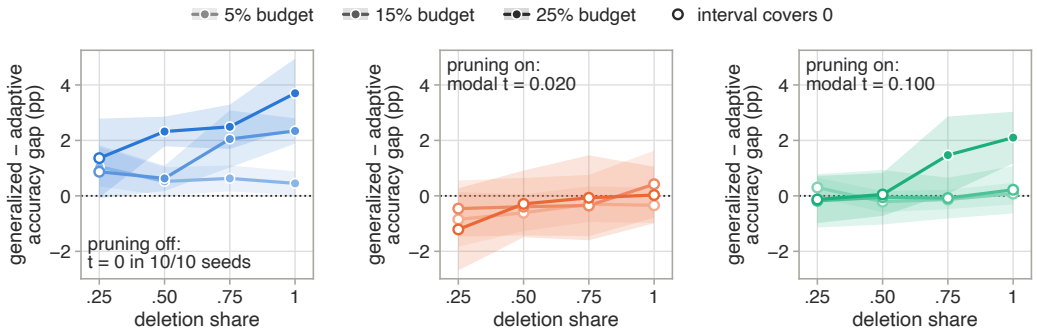
(c) PubMed

Figure A3: Realized prune fraction of the criterion arm, pruned edges as a percent of observed edges, on the clean graph and under generalized DICE at each budget for pure insertion (deletion share 0.00) and pure deletion (share 1.00), mean over 10 seeds with the 95 percent Student-t interval; fresh profile, the only one with a composition sweep. Cora is empty because calibration froze $t = 0$ there in every replicate.



(a) Cora (b) CiteSeer (c) PubMed

Figure A4: Insertions damage the undefended GCN more than deletions at every budget (citation-fresh, generalized DICE). Each bar is the accuracy lost to a budget spent entirely on insertions divided by the accuracy lost to the same budget spent entirely on deletions; the dashed line marks equal damage.



(a) Cora (b) CiteSeer (c) PubMed

Figure A5: The adaptive deletion family, which reads GraphSieve's own similarity signal, against the criterion arm (E3). Each panel plots generalized-DICE accuracy minus adaptive-family accuracy on the defended model, in pp, at matched budget and deletion share, so positive values mean the adaptive attack did more damage; one line per budget, paired 95 percent interval bands, filled markers where the interval excludes zero. Where calibration left pruning on (CiteSeer, PubMed), the gap stays near zero in all but 2 PubMed cells at the largest budget; it opens on Cora, where calibration froze the pruning no-op.

Table A20: The 5 attacked cells in which the criterion arm significantly beat GCN-Jaccard: paired margin in percentage points with its 95 percent interval. Every one sits at deletion share 0.50 or above.

					Paired difference (pp)	
Profile	Dataset	Family	Budget (%)	Del. share	Margin	95% CI
Fresh	Cora	deletion Metattack	15	1.00	+0.89	[+0.33, +1.45]
Fresh	Cora	deletion Metattack	25	1.00	+0.74	[+0.19, +1.29]
Fresh	CiteSeer	generalized DICE	25	1.00	+0.75	[+0.03, +1.47]
Fresh	CiteSeer	random composition	5	1.00	+0.72	[+0.05, +1.39]
Reproduction	Cora	legacy DICE	100	0.50	+0.51	[+0.01, +1.00]

Table A21: The 6 attacked cells in which the criterion arm significantly lost to GCN-Jaccard, all on Cora, where the arm's frozen configuration prunes nothing and restores alone. The count is below the registered refutation threshold of 7.

					Paired difference (pp)	
Profile	Dataset	Family	Budget (%)	Del. share	Margin	95% CI
Fresh	Cora	generalized DICE	5	0.25	−0.53	[−0.86, −0.20]
Fresh	Cora	random composition	5	0.25	−0.56	[−1.01, −0.11]
Fresh	Cora	random composition	5	0.50	−0.46	[−0.88, −0.04]
Fresh	Cora	random composition	5	1.00	−0.41	[−0.69, −0.13]
Fresh	Cora	random composition	15	0.75	−0.63	[−1.13, −0.13]
Fresh	Cora	random composition	15	1.00	−0.65	[−1.23, −0.07]

Table A22: The adaptive deletion family against the defended model (E3): generalized-DICE accuracy minus adaptive-family accuracy for the criterion arm at matched budget and deletion share, so a positive value means the adaptive attack did more damage. "Pruning off" counts replicates whose calibration froze $t = 0$. The range spans the 12 per-dataset cell means.

				Significant cells		
Population	Dataset	Pruning off	Cells	Positive	Negative	Range (pp)
citation-fresh	Cora	10/10	12	10	0	$+0.45$ to $+3.70$
	CiteSeer	0/10	12	0	0	-1.21 to $+0.42$
	PubMed	0/10	12	2	0	-0.21 to $+2.10$

contrasts per population; the calibrated baseline remains the primary comparator, and no registered hypothesis covers the default-threshold comparison.

The calibrated GCN-Jaccard beats the default one in 69 of 130 citation-fresh cells and loses 22, with a mean paired difference of +0.77 pp; on citation-reproduction, it wins 11 of 18 cells and loses 2, with a mean of +4.26 pp. The default arm has mixed differences from the undefended GCN on citation-fresh (43W/53NS/34L, mean +0.05 pp) and is ahead of it on citation-reproduction (16W/1NS/1L, mean +2.32 pp).

The criterion arm's three largest gains over the default arm are +5.57 [+3.95, +7.19] pp (PubMed, adaptive deletion DICE, 25 percent, deletion share 0.25), +5.04 [+3.76, +6.32] pp (PubMed, generalized DICE, 25 percent, deletion share 0.00), and +4.84 [+3.16, +6.52] pp (PubMed, adaptive deletion DICE, 25 percent, deletion share 0.50) on citation-fresh, and +18.51 [+17.09, +19.93] pp (CiteSeer, legacy DICE, 100 percent), +18.42 [+17.63, +19.21] pp (PubMed, legacy DICE, 100 percent), and +10.99 [+10.56, +11.41] pp (PubMed, legacy DICE, 50 percent) on citation-reproduction.

The criterion arm loses 17 citation-fresh and 1 citation-reproduction cells to the default arm, 18 in all: 9 on Cora, 9 on CiteSeer, and none on PubMed. Of these, 13 fall at budgets of at most 15 percent, and all but one lie within 2.75 pp.

The one loss beyond 5 pp is Cora under released Metattack at 25 percent, -6.75 [-12.86, -0.64] pp. There the ladder froze the pruning no-op in 8 of 10 replicates of the criterion arm and in 8 of 10 of the calibrated GCN-Jaccard (Table A10), while the default threshold prunes; the calibrated GCN-Jaccard's own difference from the default arm in the same cell is -7.04 [-13.26, -0.81] pp.

Table A23: Exploratory: GCN-Jaccard at the DeepRobust default threshold ($t = 0.01$) against the registered arms, per population. Not part of the registered study: the arm was computed after it on copies of the six study artifacts (one per profile and dataset) with identical seeds, splits, attack graphs, and backbone, and its calibration is a no-op (a candidate grid of size 1). Each row is a paired contrast over the population's cells: wins/no significant difference/losses by the 95 percent interval, the mean of the per-cell paired differences, and the largest per-cell difference with its interval. "Default" is the exploratory arm; "calibrated" is the registered GCN-Jaccard of Table 2.

				Paired difference (pp)		
Population	Contrast	Cells	W/NS/L	Mean	Max	95% CI of max
citation-fresh	Criterion arm − default	130	69/44/17	$+0.78$	$+5.57$	[$+3.95$, $+7.19$]
	Quota arm − default	130	65/46/19	$+0.76$	$+5.34$	[$+4.63$, $+6.05$]
	Calibrated − default	130	69/39/22	$+0.77$	$+5.57$	[$+3.95$, $+7.19$]
	Default − undefended GCN	130	43/53/34	$+0.05$	$+2.69$	[$+1.75$, $+3.63$]
citation-reproduction	Criterion arm − default	18	11/6/1	$+4.41$	$+18.51$	[$+17.09$, $+19.93$]
	Quota arm − default	18	11/6/1	$+4.46$	$+19.80$	[$+18.58$, $+21.02$]
	Calibrated − default	18	11/5/2	$+4.26$	$+18.42$	[$+17.63$, $+19.21$]
	Default − undefended GCN	18	16/1/1	$+2.32$	$+11.61$	[$+10.20$, $+13.03$]

E Code and Environment Availability

An anonymized repository accompanies the paper at `https://anonymous.4open.science/r/GraphSieve`. It contains the study package that produced every reported result, the study configuration, the analysis scripts for the exploratory calibration ablations, the recovery diagnostics and the replication, the registered configuration for the comparison with GCN-Jaccard at its default threshold, the pinned deletion-Metattack matrices, and a Docker environment pinning the CPU software stack. A runbook covers dataset download and the full study run. Calibration is the largest source of model fits: for each profile, dataset, and replicate, a method's full grid is scored on all 6 rungs, so the criterion arm alone fits 52 candidates on 6 rungs per replicate before any evaluation cell is trained.