CLASS PAIR DISTANCE CALIBRATION FOR ROBUST GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Distance based graph defenses can suppress useful messages when edges between different classes have large representation distances. CLASP divides each edge distance by the median for its predicted class pair, then applies a robust penalty to the relative distance. Each pair thus has its own reference scale. If retained edges form a majority of a perturbed bucket and their distance drift is bounded, we bound the new median using their clean order statistics and the drift bound. We evaluate CLASP in a two layer GCN and in HARA, which combines ego, one hop, and two hop channels. With full training labels and five shared attack edits, matched GCN controls show mean accuracy gains of 7.7 percentage points on Cora and 4.0 on Actor, and losses of 1.0 point on Roman-Empire and Amazon-Ratings. The system comparison uses unequal tuning budgets, with HARA's search expanded after earlier test results and final parameters selected on validation data. Across seven datasets and four attack families, HARA exceeds the strongest recorded fixed baseline in 24 of 48 budget scenarios, mostly under shared and local attacks.

1 INTRODUCTION

Graph classifiers need to use informative neighbor messages while limiting the influence of adversarial edges. Distance based aggregation reduces the weight of messages from distant node representations. Yet a useful edge between different classes can also span a large distance. When typical distances differ across class pairs, a shared cutoff can suppress these messages along with adversarial ones. The question is how to judge an edge's distance relative to the relation between its endpoints.

We introduce CLASP, Class Pair Length Adjusted Scale Penalty, to give each predicted class pair its own distance reference. A frozen preliminary classifier predicts node labels. CLASP groups edges by the ordered pair of predicted endpoint classes, then computes a distance median within each group. Dividing by this median places each edge on the scale of its class pair. The robust penalty then acts on relative distance, so a large raw distance need not receive a small weight if it is typical of its pair (Figure 1). Within each pair, the penalty still suppresses edges with large normalized distances. For a fixed bucket with an inactive numerical floor, a common positive rescaling leaves all normalized distances unchanged.

We implement CLASP in two hosts for node classification. GCN+CLASP uses two layers of projection and calibrated aggregation. HARA combines an ego channel with calibrated one hop and two hop channels, allowing each node to mix its own features with information from two neighborhoods. Within each host, a matched control uses one global median at the same cutoff. This comparison measures the effect of grouping distances by predicted class pair, with the architecture and training held fixed.

Graph edits change which edges enter a bucket and can change the distances of edges that remain. If retained edges form more than half of the perturbed bucket and their distance drift is bounded, its lower median lies between two retained clean order statistics expanded by that bound. The interval depends on the spread of these order statistics. We also account for untouched edges reassigned after preliminary predictions change. The result concerns the calibration median at a specified aggregation call, and the experiments measure retained membership and distance drift at those calls.

The experiments assess node classification under structural attacks on seven datasets. Matched controls test the grouping rule at fixed settings, while a complete grid of 48 budget scenarios

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.

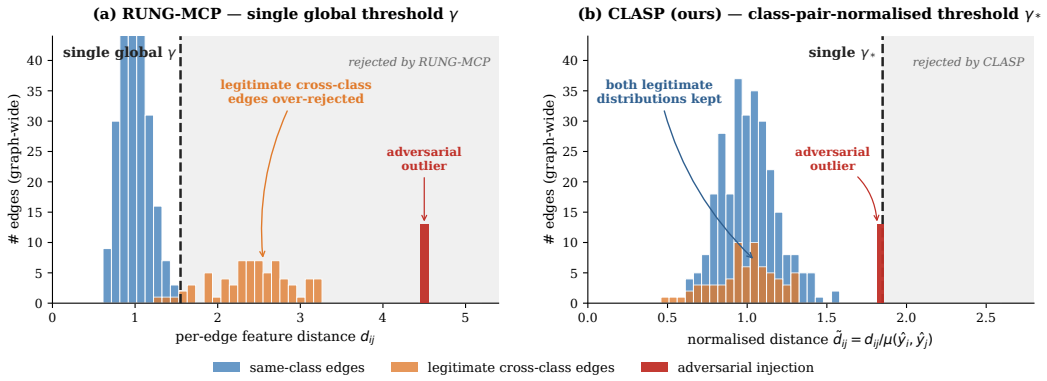


Figure 1: Class pair distance calibration. The mechanism illustration compares raw distances in (a) with distances normalized by predicted class pair medians in (b). A single raw distance cutoff suppresses many useful edges between classes. Calibration places the typical distances of both groups on a common relative scale. Histograms, outliers, and cutoffs are schematic.

compares selected HARA and GCN+CLASP configurations with recorded fixed baselines. HARA has higher mean accuracy in 24 scenarios, including 14 of 21 shared attacks and five of nine local attacks. The search budgets differ across methods, and earlier test outcomes guided an additional HARA search before final validation selection. Studies of training label budgets and bucket retention examine how preliminary predictions and graph edits affect calibration.

2 Related Work

Structural attacks provide several ways to perturb graph learning. Nettack and related methods optimize local graph edits, while Metattack differentiates through surrogate training to construct poisoning attacks (Zügner et al., 2018; Dai et al., 2018; Zügner & Günnemann, 2019). PR-BCD uses randomized blocks to optimize attacks on larger graphs (Geisler et al., 2021). Attacks against the evaluated defense complement transfer attacks by testing its own response to perturbations (Mujkanovic et al., 2022). EvA searches discrete perturbations with an evolutionary algorithm (Akhondzadeh et al., 2026). Our evaluation distinguishes shared transfer attacks, evasion against each fitted model, and transferred training poisoning.

Graph purification and robust aggregation reduce the influence of suspicious edges or messages (Wu et al., 2019; Jin et al., 2020; Geisler et al., 2020; Zhang & Zitnik, 2020; Hou et al., 2024). RUNG applies robust penalties to representation distances. CLASP adds a reference scale for each predicted class pair before applying an MCP penalty. This calibration step addresses differences in distance distributions across relations.

Architectures for heterophily separate channels or adapt propagation (Zhu et al., 2020; Lim et al., 2021; Chien et al., 2021; Bo et al., 2021; Platonov et al., 2023). Separating ego and neighbor embeddings also supports structural robustness (Zhu et al., 2022). HARA uses this separation to combine CLASP with multiple graph channels. Selective blocking handles uncertainty in edge classifications by choosing where to suppress propagation (Choi et al., 2025). CLASP uses predicted classes to estimate a distance scale, then weights messages according to their normalized distances.

3 Method

3.1 Graph and preliminary predictions

We study transductive node classification with access to the graph structure and node features. Let $G = (V, E)$ be an undirected graph with features X and C node classes. We represent each undirected edge by directed adjacency entries. Training labels define the fitting objective, and validation labels

select checkpoints. We first fit a preliminary GCN g_ϕ with one layer (Kipf & Welling, 2017). We select its checkpoint by validation loss, then freeze its weights before downstream training. For an evaluation graph G', the predicted labels are

$$\hat{\mathbf{y}}(G') = \arg\max g_{\phi^\star}(X, G'). \tag{1}$$

These predictions adapt to the evaluation graph with $\phi^\star$ held fixed. Both classifiers are fitted on the poisoned graph in the main poisoning experiment. They share the reduced training mask in the label budget studies. CLASP always constructs its buckets from predictions, without access to true test labels.

3.2 Class pair distance calibration

For directed edge (j, i), let $d_{ji} = \|h_j - h_i\|_2$. At each call, bucket B_{ab} contains active edges with ordered predicted labels $(\hat{y}_j, \hat{y}_i) = (a, b)$. CLASP computes

$$\mu_{ab} = \mathrm{med}_\downarrow\{d_{ji} : (j, i) \in B_{ab}\}, \tag{2}$$

$$\tilde{d}_{ji} = \frac{d_{ji}}{\max(\mu_{\hat{y}_j \hat{y}_i}, \epsilon)}, \qquad r_{ji} = \left[\frac{1}{2\max(\tilde{d}_{ji}, \epsilon)} - \frac{1}{2\gamma_\star}\right]_0^{10^3}. \tag{3}$$

Here h denotes the current node representations, $\epsilon = 10^{-9}$, and $\gamma_\star > 0$ is the selected cutoff. The lower median is the $\lceil n/2 \rceil$th order statistic of n values. An empty bucket uses the median of all active edges, or one if none is active. During weighted attack search, entries with $a_{ji} > 0.5$ define the median. Original self loops can contribute to this statistic, but their message factors are zero. Given adjacency weights a_{ji}, the aggregate is

$$\mathcal{A}_i(H; A, r) = \frac{\sum_j a_{ji} r_{ji} h_j}{\max(\sum_j a_{ji} r_{ji}, \epsilon)}. \tag{4}$$

Calibration gives the penalty a relative scale within each bucket. For a fixed nonempty bucket, multiplying every distance by $s > 0$ preserves d/μ if the median floor is inactive before and after scaling. Thus a common change in scale leaves the calibrated penalty unchanged. Changes in the shape of the distance distribution remain relevant.

3.3 Hosts and matched controls

We use GCN+CLASP with two layers of projection and aggregation, followed by a classifier. Its matched control, GCN+Global, places all edges in one calibration bucket. Both models share parameter initialization, training, preliminary prediction computation, cutoff, and robust penalty. Their comparison isolates grouping at a shared cutoff. The comparison with GCN measures the complete aggregation change.

HARA retains a node's own features in an ego channel and gathers neighbor information through two graph channels in each layer (Figure 2). The graph channels aggregate one hop and two hop messages, with T refinements and anchor coefficient α. A learned convex mixer uses predicted local homophily and log degree. HARA+Global keeps this architecture and merges only the calibration buckets. The matched HARA models use two layers, $T = 2$, $\alpha = 0.15$, and $\gamma_\star = 1.25$. The fixed GCN comparison uses the same cutoff.

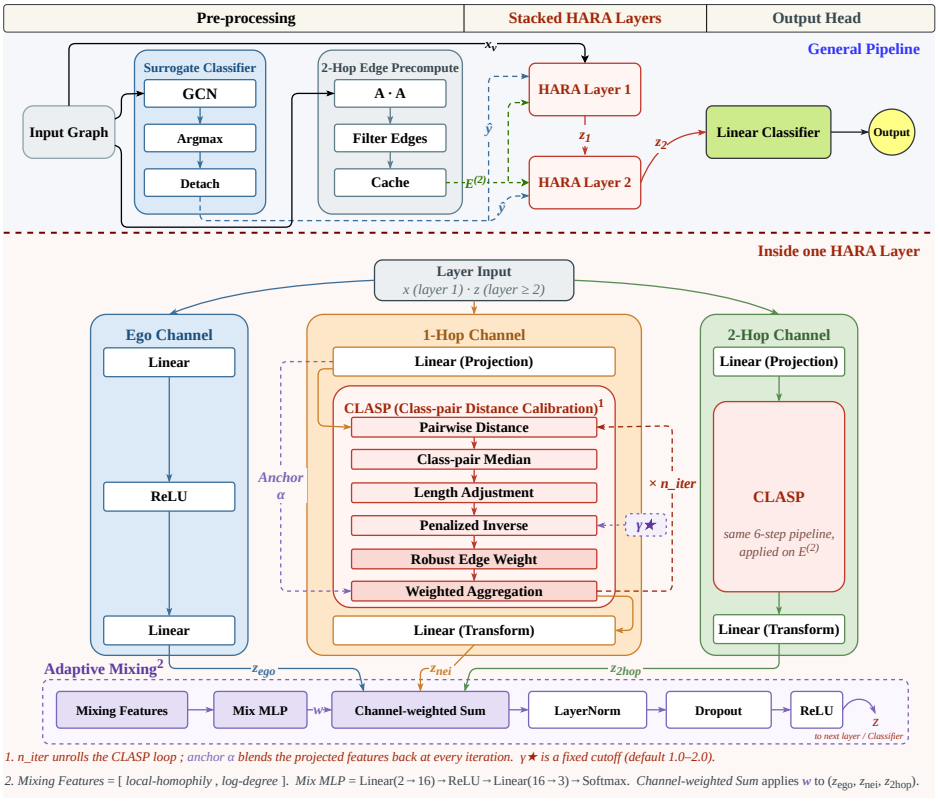


Figure 2: HARA host architecture, illustrated with the fixed setting configuration ($T = 2$, $\alpha = 0.15$, and $\gamma_\star = 1.25$). Selected settings are listed separately in Table 3. The preliminary GCN is fitted and frozen before downstream training. Its predictions and the two hop graph are recomputed on each evaluation graph. CLASP is used in both graph channels, while GCN+CLASP removes the surrounding channel architecture.

Additional HARA variants retain the channels and mixer. HARA-Mean uses weighted means, and HARA-Median uses soft medoid aggregation. Both use one refinement without anchoring. Their comparison with HARA therefore covers aggregation, refinement, and anchoring together. HARA-Mean also uses preliminary predictions in its mixer. Appendix C gives the architecture equations and configurations.

All graph computations share the same edge weights, including the two hop graph formed from weighted A^2. During attack ranking, the relaxation retains the original discrete two hop sample. It adds a bounded set of candidate paths with zero initial value, preserving the binary forward state. CLASP computes robust factors without autograd. Attack gradients pass through the weighted message path with these factors and hard labels held fixed. Exact discrete trials then recompute labels, factors, and the two hop graph. This procedure uses gradients to propose edits and complete forward evaluations to select them.

4 Median Changes After Graph Edits

After a graph edit, how far can the median used by CLASP move? Insertions and deletions change the edges in a bucket. Recomputed predictions can also move untouched edges to other buckets, and later representations can change the distances of retained edges. We bound the perturbed median using the clean distances of retained edges and a bound on their drift. CLASP uses the lower median, the $\lceil n/2 \rceil$th smallest of n values.

Proposition 1 (Median bound with changing bucket membership)**.** *Let B and B' be the clean and perturbed buckets of directed edges for one ordered predicted class pair. Write $I = B \cap B'$, $r = |I|$, $n = |B'|$, and $a = n - r$. Assume $n > 0$, $r > n/2$, and $|d'_e - d_e| \le \delta$ for every $e \in I$. Sort the clean distances on I as $z_{(1)} \le \cdots \le z_{(r)}$ and set $k = \lceil n/2 \rceil$. The perturbed lower median satisfies*

$$\max\{0, z_{(k-a)} - \delta\} \;\le\; \mu' \;\le\; z_{(k)} + \delta. \tag{5}$$

The lower endpoint admits fewer than k smaller perturbed values, even if all a new members lie below it. The upper endpoint includes at least k retained values. These two counts establish the bound. Appendix A gives the proof and a bound on normalized distances.

For a one hop call, let $m = |B|$. Let R count surviving directed edges with a changed ordered pair of predicted classes. After K undirected edge edits, $|B \triangle B'| \le 2K + R$. Thus $m > 2K + R$ suffices for a retained majority. With fixed labels, this reduces to $m > 2K$. The first layer's first one hop call uses fixed projected features, so $\delta = 0$ under structural evasion with frozen weights. Later calls require their own distance drift bounds. Changes in preliminary labels can reassign many untouched edges, making R relevant beyond the edit budget K. We measure these quantities at named one hop calls and relate R to preliminary logit changes in Appendix A. The interval can remain wide when its retained order statistics are far apart. A prediction certificate would also require bounds on the resulting message weights and their propagation through the classifier.

5 Experimental Protocol

We evaluate Cora, CiteSeer, Chameleon-F, Squirrel-F, Actor, Roman-Empire, and Amazon-Ratings with fixed dataset variants and splits. HARA and GCN+CLASP each use one parameter vector per dataset across all attacks and budgets, with seeds $0, 1, 2$. Validation data select the parameters. The initial search evaluates 48 configurations per host and dataset. A subsequent search evaluates 64 additional HARA configurations on each of Cora, CiteSeer, Actor, and Amazon-Ratings. It targets six combinations of dataset and attack identified from earlier test results. Parameters are sealed before the corresponding new tests. Appendix D.1 describes the search spaces, validation attacks, and checkpoint selection.

The baseline comparison uses previously evaluated fixed configurations. When several cohorts are available for a method, we use the complete cohort with the highest mean at that budget. Each cohort contains three seeds. HARA and GCN+CLASP receive additional tuning, so the resulting accuracy differences compare configurations obtained with unequal search budgets. The matched controls in Section 6.2 and Appendix E hold the architecture and training fixed to isolate the grouping rule. N/A denotes unevaluated table settings.

We first assess transfer behavior with shared targeted attacks. Each dataset has 100 test targets, selected independently of the models and stratified by class and degree. For each target and budget $k \in \{1, 3, 5\}$, all methods use the same frozen Nettack graph with structural edits only. The selected configurations reuse these graphs. The appendix also varies training labels through percentage budgets and five labels per class. Both classifiers are retrained for each mask, with the original validation labels and the attacker's full label surrogate retained.

We next attack each fitted model separately on CiteSeer, Squirrel-F, and Actor, using the same 100 targets. Candidate deletions lie in the target's current two hop neighborhood. Additions connect the target to nonneighbors, including up to 512 sampled across the graph. The combined candidate set is capped at 5,000. Gradients rank the edits, and exact discrete trials evaluate the top 50. Each trial recomputes all labels and graph operations. The attack accepts a decrease above 10^{-8} in the margin of the true class. It stops after misclassification, failure to improve, or five flips. Prefixes supply budgets one and three. Five variants on 20 validation targets each from Cora and Actor measure sensitivity to support, top-k, candidate seed, and loss. The primary configuration is fixed independently of these diagnostics.

For global evasion, PR-BCD attacks each fitted model on Cora, CiteSeer, Squirrel-F, and Actor. Its white box objective uses all test labels, with budgets of $1\%, 3\%, 5\%$ of editable undirected edges. At each budget, we report the feasible whole graph with the lowest test accuracy among that and smaller budgets, including the clean graph.

Table 1: Accuracy of the selected configurations, in percent. Clean accuracy uses all test nodes; shared Nettack accuracy uses the same 100 targets per dataset at five edits. Entries are mean and sample standard deviation over three seeds.

	Clean, all test nodes		Shared, five edits	
Dataset	HARA	GCN+CLASP	HARA	GCN+CLASP
Cora	76.4 ± 0.2	74.4 ± 1.8	52.7 ± 1.5	32.0 ± 1.0
CiteSeer	59.6 ± 1.5	61.8 ± 1.7	46.0 ± 2.0	37.3 ± 1.5
Chameleon-F	39.3 ± 3.6	38.4 ± 2.3	38.0 ± 1.7	35.3 ± 0.6
Squirrel-F	39.9 ± 0.9	35.3 ± 0.4	31.3 ± 4.2	36.7 ± 2.1
Actor	38.4 ± 0.7	29.1 ± 0.4	36.0 ± 1.0	22.0 ± 1.0
Roman-Empire	76.0 ± 0.5	55.4 ± 0.4	46.3 ± 3.5	22.3 ± 0.6
Amazon-Ratings	48.0 ± 0.1	46.9 ± 0.7	33.3 ± 5.1	26.0 ± 3.5

For training poisoning, exact Metattack constructs graphs on Cora and Chameleon-F at the same percentages, using a GCN surrogate. All defenses share these graphs. Each complete pipeline is retrained and evaluated on the poisoned graph with the original full training mask. This includes its preliminary classifier when present. An appendix control poisons only preliminary fitting, with clean downstream training and evaluation. True test labels are withheld from poisoning optimization. Original self loops are fixed and excluded from percentage budgets.

We report attacked accuracy over all evaluation nodes. Conditional ASR measures newly incorrect predictions among each model's initially correct nodes. It is undefined for an empty denominator. Shared transfer graphs can also correct earlier errors, so their attacked accuracy can exceed clean accuracy. For shared attacks, normalized RAUC integrates ASR over budgets $(0, 1, 3, 5)$ by the trapezoidal rule and divides the area by five. ASR is zero at budget zero. Summaries use means and sample standard deviations over three fitted replicates. Label budget replicates combine mask and training variation, while global evasion and poisoning vary both attack and model seeds. Targets and buckets are dependent observations within each fit. The appendix reports denominators, seed results, and contributing replicate counts for ASR summaries.

Figure 3 contains all 48 budget scenarios, and the absolute accuracy table uses the same shared attack budget of five edits for every dataset. GCN+CLASP is reported separately from the fixed baseline criterion. Appendix D also reports selected budget tables. Their retrospective rule chooses the budget with HARA's largest observed margin for each dataset and attack. The full grid count uses all budgets.

6 Results

6.1 Performance across the complete attack grid

Figure 3 compares configurations obtained with the unequal searches described in Section 5. HARA exceeds the strongest recorded fixed baseline in 24 of 48 scenarios, and GCN+CLASP does so in 9. HARA's gains are concentrated in shared and local attacks, with higher mean accuracy in 14 of 21 shared scenarios and five of nine local scenarios. It also exceeds the reference in four of 12 global evasion scenarios and one of six poisoning scenarios. These counts compare means across three seeds, with ties excluded.

On CiteSeer, HARA's margin over the fixed reference increases from -17.0 to $+16.3$ percentage points as shared edits increase from one to five. Under local attacks, its margins on Actor remain positive across all three budgets. Under global evasion, HARA exceeds the reference on Actor at every budget, while GCN+CLASP does so on Squirrel-F. Both methods remain below the reference on Cora under global evasion and transferred poisoning.

Table 1 reports absolute accuracy at a common shared attack budget, alongside clean accuracy on the full test set. At five edits, HARA reaches 46.0% on CiteSeer and 36.0% on Actor. GCN+CLASP reaches 36.7% on Squirrel-F, exceeding HARA's 31.3%. The clean and attacked columns use different evaluation populations, so their difference is not a paired measure of attack damage. All budget specific means and sample standard deviations appear in Appendix D.

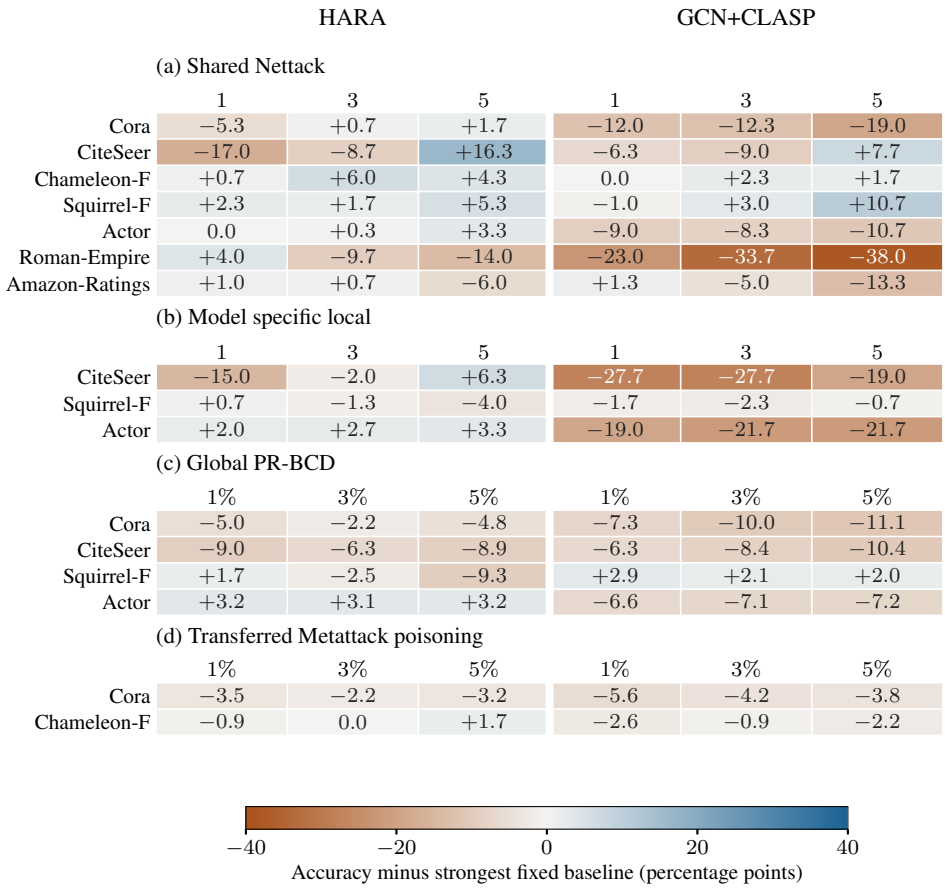


Figure 3: Complete attack grid for the selected configurations. Each cell shows mean attacked accuracy minus the strongest recorded fixed baseline at the same budget, in percentage points. HARA and GCN+CLASP use separate column groups with a shared color scale centered at zero. Panels (a) and (b) use 100 targets and budgets of 1, 3, and 5 edge edits. Panels (c) and (d) use all test nodes and budgets of 1%, 3%, and 5% of editable edges. All means use three seeds. Parameters are selected on validation data; tuning budgets differ across methods. Differences are descriptive. Appendix D gives absolute accuracies, sample standard deviations, and baseline identities for every cell.

6.2 Calibration within matched hosts

At five shared Nettack edits with full training labels, Table 2 compares class pair and global calibration on the same 100 targets per dataset, using three fits with the architecture, cutoff, initialization, and training held fixed. In the GCN host, class pair grouping raises mean attacked accuracy by 7.7 percentage points on Cora and 4.0 on Actor. It lowers the mean by 1.0 point on both Roman-Empire and Amazon-Ratings. In HARA, grouping raises the mean by 1.0 point on Amazon-Ratings and lowers it on the other three datasets. These differences measure grouping at the specified settings. The HARA results in Figure 3 also depend on the channel architecture, refinement, and parameter search.

The fixed configuration studies in Appendix E measure how reduced training masks affect classification and whether attacked buckets satisfy the retained majority condition. Appendix G reports bucket support, preliminary prediction accuracy, and comparisons with mean and soft medoid aggregation. These studies use the fixed recipes specified in Appendix B.

Table 2: Matched calibration controls with full training labels and five shared Nettack edits. Accuracy uses the same 100 targets per dataset. Global uses one distance median; Pair uses predicted class pairs. Δ is Pair minus Global in percentage points, paired by model seed. Entries are mean and sample standard deviation over three fits. Both hosts use the fixed settings in Section 3.3.

Dataset	GCN Global	GCN Pair	Δ	HARA Global	HARA Pair	Δ
Cora	37.3 ± 3.8	45.0 ± 0.0	7.7 ± 3.8	50.3 ± 2.1	49.0 ± 1.7	-1.3 ± 0.6
Actor	19.0 ± 1.7	23.0 ± 4.0	4.0 ± 5.6	31.3 ± 3.2	31.0 ± 4.6	-0.3 ± 2.5
Roman-Empire	13.7 ± 1.2	12.7 ± 1.5	-1.0 ± 1.0	51.7 ± 3.8	50.3 ± 6.1	-1.3 ± 2.5
Amazon-Ratings	39.3 ± 3.2	38.3 ± 2.3	-1.0 ± 4.6	36.0 ± 1.0	37.0 ± 3.0	1.0 ± 2.6

7 Scope and Open Questions

The experiments cover transductive node classification with fixed features and class sets. In the label budget experiments, only the fitting labels are reduced; validation labels and the attacker's full label surrogate remain available. Performance under feature attacks, inductive deployment, changing classes, or combined label scarcity and poisoning remains untested.

The local and global evasion results give accuracy under the feasible perturbations found by the specified searches. The gradients hold hard assignments and robust factors fixed, with exact discrete evaluations used to select the final edits. Poisoning uses transfer from a GCN surrogate; optimization through the defended training pipeline is unevaluated. Evolutionary search, including EvA (Akhondzadeh et al., 2026), could test sensitivity to a different search procedure.

The median bound assumes a retained majority and bounded distance drift at a specified call. The analysis does not derive a bound on preliminary label reassignment from the edit budget alone. Drift at later calls remains uncontrolled without further assumptions. Relating these quantities to a final prediction requires bounds through the full message passing computation.

8 Conclusion

CLASP judges an edge's distance relative to its predicted class pair, allowing robust aggregation to use a separate scale for each relation. With a retained majority and bounded distance drift, the perturbed median is bounded by retained clean order statistics and the drift bound. At five shared edits, matched GCN experiments give higher mean attacked accuracy with class pair grouping on Cora and Actor, and with a global scale on Roman-Empire and Amazon-Ratings. Under unequal tuning budgets, selected HARA configurations exceed the recorded fixed baselines in 24 of 48 scenarios. Most of these gains occur under shared and local attacks. The remaining question is which properties of the predicted class pairs explain when separate scales improve node classification.

USE OF AI ASSISTANCE

AI tools assisted with wording and proofreading of the manuscript.

REPRODUCIBILITY STATEMENT

Sections 3 and 5 specify the method and attack procedures. Appendix B gives the detailed settings. Appendix A provides the assumptions and proofs for the median bounds. The accompanying code includes frozen configurations, dataset cache identities, result validators, and generators for the reported tables and figures. Complete numerical comparisons, label counts, and ASR denominators appear in Appendices D and F; the shared attack study is documented separately in Appendices H through Q.

REFERENCES

Mohammad Sadegh Akhondzadeh, Soroush H. Zargarbashi, Jimin Cao, and Aleksandar Bojchevski. Eva: Evolutionary attacks on graphs. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 64694–64723, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/69ffb2b7055546c4c2d559025286675b-Paper-Conference.pdf.

Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3950–3957, 2021.

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

Yoonhyuk Choi, Taewook Ko, Jiho Choi, and Chong-Kwon Kim. Selective blocking for message-passing neural networks on heterophilic graphs. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 737–751. PMLR, 2025. URL https://proceedings.mlr.press/v286/choi25a.html.

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

Simon Geisler, Daniel Zügner, and Stephan Günnemann. Reliable graph neural networks via robust aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 13272–13284, 2020.

Simon Geisler, Tobias Schmidt, Hakan Şirin, Daniel Zügner, Aleksandar Bojchevski, and Stephan Günnemann. Robustness of graph neural networks at scale. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7637–7649. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/3ea2db50e62ceefceaf70a9d9a56a6f4-Paper.pdf.

Zhichao Hou, Ruiqi Feng, Tyler Derr, and Xiaorui Liu. Robust graph neural networks via unbiased aggregation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

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

Runlin Lei, Zhen Wang, Yaliang Li, Bolin Ding, and Zhewei Wei. EvenNet: Ignoring odd-hop neighbors improves robustness of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser-Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 20887–20902, 2021.

Felix Mujkanovic, Simon Geisler, Stephan Günnemann, and Aleksandar Bojchevski. Are defenses for graph neural networks robust? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8954–8968. Curran Associates, Inc., 2022. doi: 10.52202/068431-0651. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3ac904a31f9141444009777abef2ed8e-Paper-Conference.pdf.

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 (ICLR)*, 2023.

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 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4816–4823, 2019.

Xiang Zhang and Marinka Zitnik. GNNGuard: Defending graph neural networks against adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9263–9275, 2020.

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 (NeurIPS)*, volume 33, pp. 7793–7804, 2020.

Jiong Zhu, Junchen Jin, Donald Loveland, Michael T. Schaub, and Danai Koutra. How does heterophily impact the robustness of graph neural networks? theoretical connections and practical implications. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. doi: 10.1145/3534678.3539418. URL https://arxiv.org/abs/2106.07767.

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

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

A Estimator Bounds and Proofs

In Proposition 1, the retained majority $r > n/2$ implies $a < k \le r$, so both order statistics in Eq. (5) exist. At most $k - a - 1$ retained clean distances are strictly below $z_{(k-a)}$. With drift at most δ, at most the same number of retained perturbed distances can fall below $z_{(k-a)} - \delta$. Adding all a new members still leaves fewer than k values below this threshold. The perturbed kth order statistic is therefore at least the threshold. Nonnegative distances supply the additional lower bound zero.

For the upper bound, at least k retained clean distances are at most $z_{(k)}$. Their perturbed distances are at most $z_{(k)} + \delta$, so the perturbed kth order statistic is at most this value. The argument allows arbitrary distances for newly assigned edges. It also allows arbitrary deletions and reassignments from the clean bucket.

For a nonempty perturbed bucket with a retained majority, the interval also bounds displacement from the clean lower median μ over B:

$$|\mu' - \mu| \le \max\{|z_{(k-a)} - \mu|,\ |z_{(k)} - \mu|\} + \delta. \tag{6}$$

Its width depends on the retained rank interval and the drift. A large gap between retained distances can yield a wide interval despite a retained majority.

For a one hop bucket, each undirected insertion or deletion changes at most two directed members. With fixed labels, K edge edits therefore give $|B \triangle B'| \le 2K$. Let $m = |B|$, $D = |B \setminus B'|$, and $a = |B' \setminus B|$. Then $r = m - D$ and $n = m - D + a$, so $r > n/2$ is equivalent to $m > D + a$. Consequently, $m > 2K$ ensures a retained majority. For an ordered pair of different classes, one undirected edit changes at most one member. This gives the sharper sufficient condition $m > K$. The argument allows different clean and perturbed bucket sizes, with no replacement partner required for a deleted edge.

One original edge edit can change many two hop paths, so the one hop count does not bound changes in two hop membership. A change in a node's predicted label can also reassign all of its incident edges, including untouched edges. The one hop diagnostics count these reassignments separately from direct edits.

To account for recomputed labels, let R count surviving directed one hop edges with different ordered predicted pairs on the two graphs. Every member of $B \triangle B'$ is either an inserted or deleted edge, or a reassigned survivor. A surviving edge cannot both enter and leave the same bucket. Hence $D + a \le 2K + R$, and $m > 2K + R$ suffices for $r > n/2$. This count permits simultaneous deletions and label changes. At the first layer's first one hop call, either host aggregates a fixed projection of X. Frozen weights and unchanged features therefore give $\delta = 0$ for retained distances. Subsequent calls use representations from earlier graph operations and require a separate drift bound.

Predicted labels remain fixed when the preliminary GCN's logit changes are small relative to its clean winning margins. Let S and S' be its normalized adjacency operators on the two graphs, including its self loops. The frozen linear layer produces logits $Z = SXW + \mathbf{1}b^\top$ and $Z' = S'XW + \mathbf{1}b^\top$. For $C \ge 2$, let m_i be node i's clean winning logit margin over the largest competing logit. Define

$$b_i^{\mathrm{drift}} = \|[(S' - S)XW]_{i,:}\|_\infty, \qquad U = \{i : m_i \le 2b_i^{\mathrm{drift}}\}.$$

If $i \notin U$, each pairwise logit margin can decrease by at most $2b_i^{\mathrm{drift}} < m_i$, so the predicted class remains unchanged. Every reassigned surviving edge therefore has an endpoint in U. Let D'_U count directed entries of the perturbed one hop graph incident to U. Then $R \le D'_U$, and $m > 2K + D'_U$ suffices for a retained majority. The logit drift and reassignment bounds apply to a specified perturbation with frozen preliminary weights. Uniform label stability over an attack set requires a uniform bound on b_i^{drift}. Retraining under poisoning requires a separate analysis of parameter changes.

To bound changes in normalized distances on retained edges, we also need a positive lower bound on the calibration scales. For clean and perturbed distances d, d', put $s = \max(\mu, \epsilon)$ and $s' = \max(\mu', \epsilon)$. If $s, s' \ge \eta > 0$,

$$\left|\frac{d'}{s'} - \frac{d}{s}\right| \le \frac{|d' - d|}{\eta} + \frac{d\,|\mu' - \mu|}{\eta^2}. \tag{7}$$

Add and subtract d/s', then apply the triangle inequality with $|s' - s| \le |\mu' - \mu|$ and $ss' \ge \eta^2$. The numerical floor permits $\eta = \epsilon$, although this choice can give a loose bound. A classification margin

bound would additionally need to control all normalized message weights and their propagation through subsequent layers.

If bucket size and predicted labels remain fixed, replacing fewer than half the distances keeps the median within the range of the unreplaced distances. The replacement count in the following proposition measures changed observations. Proposition 1 separately accounts for graph edits through membership changes and distance drift.

Proposition 2 (Replacement stability of a fixed median)**.** *Fix predicted labels and one nonempty ordered bucket containing m nonnegative distances. Replace $q < m/2$ distances by arbitrary nonnegative values. Every sample median of the contaminated bucket lies between the minimum and maximum of the unreplaced distances.*

Proof. More than half of the observations remain unchanged. A median below their minimum would have fewer than half the observations at or below it. A median above their maximum would have fewer than half at or above it. Both contradict the median property. The argument includes both central order statistics when m is even, and every value between them. □

The fixed bucket bound permits dependent edge distances. Its interval width is the range of the unreplaced distances, so a retained majority alone does not imply a small median displacement.

B Stress Study Settings and Interpretation

B.1 Label budgets and model selection

The shared attack benchmark and the stress study use independent model cohorts and separate result summaries. Within the stress study, neutral attacks with full training labels and evasion against individual models reuse the same checkpoints on overlapping datasets. Each preliminary model is fitted within its host's training run. Matched GCN+Global and GCN+CLASP fits share preliminary initialization and training settings.

For each original training class, we randomly order its nodes with mask seed $1000 + s$, where $s \in \{0, 1, 2\}$ is the model seed. At fraction f, we retain the first $\max(1, \lfloor f n_c \rfloor)$ nodes of class c. Masks are nested across fractions and shared by all methods. At $f = 1$, the original mask is reused and its subset seed has no effect. The study of training label percentages contains 288 cells across methods, datasets, fractions, and seeds. It includes 86,400 target and budget records. The three replicates jointly measure training and subset variation.

The study with five labels per class limits fitting labels while retaining the original validation sets for model selection. It uses exactly 25 training labels on Actor, 90 on Roman-Empire, and 25 on Amazon-Ratings, with validation set sizes of 2,432, 7,251, and 7,837, respectively. Validation labels select checkpoints for both preliminary and downstream models. The study contains 54 fitted cells and 16,200 target budget records. The transferred Nettack graphs were constructed with the original full label surrogate and are identical across defender label budgets.

At 10% on Cora and Actor, the crossed mask diagnostic fits GCN+CLASP and HARA for all nine combinations of three mask and three model seeds, giving 36 cells. Twelve coincide with the main study of label percentages, and 24 are additional fits. The endpoints are clean accuracy and preliminary classifier diagnostics. The crossed diagnostic adds no attack replicates.

B.2 Matched HARA calibration control

HARA+Global retains the HARA architecture, with two refinements and anchor coefficient 0.15. Both models share parameter initialization, the frozen preliminary classifier procedure, and the mixer. Only the calibration groups in its one hop and two hop aggregators are merged. The control contains 24 shared evasion cells on Cora, Actor, Roman-Empire, and Amazon-Ratings, using full and 10% training labels. Another 18 cells evaluate complete pipeline poisoning on Cora and Chameleon-F. Masks, attacks, and model seeds match the corresponding HARA fits.

We specified the control after the initial six method results and fixed its full scope before fitting the additional models. HARA+Global is absent from the model specific or attack grids with five labels per class.

B.3 Bounded model specific search

For target t with true label y_t, the primary objective minimizes

$$L_t(A) = z_{t,y_t}(A) - \max_{c \neq y_t} z_{t,c}(A).$$

Attack success is measured only for initially correct targets. Each attack follows this procedure:

1. Construct the target's current two hop neighborhood. Candidate deletions are existing edges in this neighborhood, including edges away from the target. Additions connect the target to local nonneighbors and up to 512 nonneighbors sampled across the graph. A stable hash of the candidate seed and canonical endpoints selects the additional nodes.
2. Cap the combined support at 5,000 candidates using the same deterministic hash ordering. When both edit types are available, retain at least one of each. Represent additions with zero weight and current edges with unit weight.
3. Differentiate the implemented weighted forward computation. The preliminary classes and CLASP robust factors are held fixed during this derivative. The distance penalty is computed under `no_grad`. The derivative supplies a proposal score through the remaining message, degree, mixer, and two hop computations.
4. Rank an addition by $-\partial L_t/\partial a_{uv}$ and a deletion by $+\partial L_t/\partial a_{uv}$. Evaluate the top 50 finite candidates as separate discrete graph edits. Each trial recomputes all preliminary predictions, class pair medians, robust factors, and graph channels.
5. Accept the trial with the largest exact margin decrease if that decrease exceeds 10^{-8}. Stop at misclassification, five edits, or failure to improve. Evaluate the accepted path prefixes at budgets one, three, and five, retaining the shorter path if the attack stopped early.

The primary candidate seed is zero. The study contains 54 cells across methods, datasets, and seeds, with 16,200 target and budget records.

The strength diagnostic compares five search variants for GCN+CLASP and HARA. It uses model seed zero and 20 fixed validation targets each on Cora and Actor. The first three variants use local support with top 20, expanded support with top 20, and expanded support with top 50. The last two modify the third variant, using candidate seed one or cross entropy loss. The cross entropy variant maximizes this loss for both gradient proposals and exact discrete acceptance. The other four variants minimize the logit margin of the true class. The 400 diagnostic records at budget five measure sensitivity to these search choices. The primary configuration is fixed independently of these outcomes.

A separate HARA study uses local support with top 20, producing 2,700 records in separately labeled tables. The local support variant also samples global additions when no local addition is available. Expanded support variants include global additions regardless of local availability.

The evaluated RUNG-MCP implementation also detaches the distances used to calculate its IRLS weights. Gradient ranking follows the implemented gradient path, and discrete trials reevaluate the model.

On Actor, the full label shared and model specific local evaluations use the same fitted checkpoints and targets. A per target attack portfolio selects between these two evaluated graphs and no edit separately for each target. An initially incorrect target remains incorrect under the no edit choice. Each target can use a different graph, so the portfolio measures complementary errors under separate targeted perturbations.

HARA caps its discrete two hop support at 200,000 directed entries with sample seed zero. Sampling from an enlarged support containing zero weight candidate paths can displace active paths, even before an edit is made. For attack ranking, we threshold current one hop weights at 0.5. We then retain exactly the original selected two hop entries of this binary graph. We add up to 200,000 paths

absent from the full binary two hop support and take all path values from weighted A^2. At binary weights, these additional paths have zero value, while the retained sample matches the discrete model. At binary weights and in exact arithmetic, the extension therefore preserves the model's forward state. Candidate derivatives use this extension, while accepted discrete graphs use the original model. PR-BCD uses the same extension. Sampling thresholds and hard label assignments still make the overall search nonsmooth.

B.4 Global PR-BCD evasion

We use PyG 2.7.0's `PRBCDAttack` with 125 optimization epochs, 100 resampling epochs, block size $\max(4096, 4K)$, probability margin loss, and attack seed $3000 + s$. It attacks all test nodes of each fitted model, using their true labels only in the evasion objective. The perturbation budget is $K = \lfloor p|E_{\text{editable}}| \rfloor$ for $p \in \{0.01, 0.03, 0.05\}$, where E_{editable} counts off diagonal undirected edges. Original self loops are preserved. Four datasets, six methods, three model seeds and three rates give 216 attacks in 72 model cells.

We retain the clean graph if a sampled attack increases test accuracy. For a result with at most K edits, we also compare feasible graphs from smaller budgets and select the single graph with lowest overall test accuracy. The selected result always corresponds to one complete graph. Raw and selected outcomes, perturbations, prediction vectors and optimization logs are retained. We check edit feasibility and finite final discrete logits for the reported perturbations. These checks provide no bound on search optimality.

B.5 Transferred training poisoning

DeepRobust 0.2.11 exact Metattack uses a linear GCN surrogate with hidden width 16, dropout 0.5, no hidden ReLU and no bias. Surrogate training uses 200 epochs with validation selection. The meta gradient inner loop uses 100 epochs. We set $\lambda = 0$, disable the likelihood ratio constraint, and use attack seed $2000 + s$. The implementation's default singleton mask suppresses deletion scores for edges incident to degree one nodes, preventing node isolation. Training labels fit the surrogate and validation labels select its checkpoint. The attack uses surrogate pseudo labels outside the training set. True test labels are replaced before being passed to the attack implementation and do not enter its optimization.

The 18 poisoned graphs span two datasets, three seeds and three edge fractions. Metattack operates on the off diagonal graph. Original loops are restored before defended training. Each graph is shared by all defenses. The main condition fits and evaluates the complete pipeline on the poisoned graph, including any preliminary model. This produces 108 fits. A second condition poisons only preliminary fitting for GCN+CLASP and HARA. Downstream fitting and final inference use the clean graph, producing 36 fits. Matched clean fits use the same model seeds. The main condition measures graph corruption during both training and evaluation. Both conditions use transfer from the GCN surrogate, with defended fitting applied after the attack graphs are fixed.

B.6 Aggregation diagnostics

For all 1,800 budget five adaptive records of GCN+CLASP and HARA, we capture the first one hop aggregation call at each of the two layers. For each ordered class pair, we match clean and perturbed edge identities. We record retained membership, maximum distance drift on retained edges, medians, and the rank interval in Proposition 1. We also count inserted directed edges and surviving edges reassigned by changed preliminary predictions, along with deleted directed edges. The per bucket symmetric difference is checked against the total edit and reassignment count. Summaries include initially correct targets with at least one accepted edit. All other target records remain available with flags. Empty buckets and failures of retained majority are counted explicitly. Numerical interval checks test the bound's implementation at the recorded calls. Appendix A proves the bound under its stated conditions.

B.7 Compute and provenance

The shared attack study used an RTX 5090 Laptop GPU with approximately 24 GB of memory. The six method stress study used an RTX 5090 with 32 GB. Concurrent jobs affect elapsed times, so these records describe resource use under the execution schedule. HARA-Median remains in the original nine method comparison but is absent from the six method factorial study. One original Amazon-Ratings training seed took 14,229 seconds. The stress study includes the simpler GCN and HARA-Mean controls, RUNG-MCP, and the matched global calibration control.

Dataset caches, training masks, configurations, implementation sources, checkpoints, attack graphs and result cells are hashed. Reported summaries include every configured cell in each study. Tables and figures are generated from the validated machine readable records. The shared attack and stress study references are distributed separately.

C Host Architecture

HARA combines an ego representation with one hop and two hop graph representations, as shown in Figure 2 and described in Section 3.3. For edge weights a_{ji}, the mixer inputs are weighted degree and the fraction of incoming weight from nodes with the same predicted label:

$$\delta_i = \sum_j a_{ji}, \qquad \widehat{h}_i = \frac{\sum_j a_{ji} \mathbb{1}[\hat{y}_j = \hat{y}_i]}{\max(\delta_i, \epsilon)}. \tag{8}$$

These statistics include the graph's fixed self loops. To construct the two hop channel, we remove self loops and active one hop entries from weighted A^2. We clip the remaining values to $[0, 1]$ and apply the fixed cap in Appendix B.3.

For layer input $\mathbf{H}^{(\ell)}$, the ego state is $\mathbf{z}_i^{(0)} = \mathrm{MLP}_\ell(\mathbf{h}_i^{(\ell)})$. The two graph channels use projected anchors $\mathbf{p}_i^{(q)} = W_\ell^{(q)} \mathbf{h}_i^{(\ell)}$, $q \in \{1, 2\}$. Each graph channel starts at its anchor and performs T refinements,

$$\mathbf{u}_i^{(q,t+1)} = (1 - \alpha)\, \mathcal{A}_i^{(q)}(\mathbf{U}^{(q,t)}; A^{(q)}, r^{(q)}) + \alpha \mathbf{p}_i^{(q)}, \qquad 0 \le \alpha \le 1, \tag{9}$$

followed by a learned linear transform. For $q = 1$, $A^{(1)} = A$. For $q = 2$, $A^{(2)}$ is the weighted two hop graph. The CLASP cutoff $\gamma_\star$ is shared by both channels. The fixed setting controls use $T = 2$, $\alpha = 0.15$, and $\gamma_\star = 1.25$. Table 3 lists the selected configurations. The mixer uses predicted homophily and weighted degree to assign convex weights to the three channels:

$$\boldsymbol{\pi}_i = \mathrm{softmax}\Big(\mathrm{MLP}_{\mathrm{mix}}[\widehat{h}_i, \log(1 + \delta_i)]\Big), \qquad \mathbf{h}_i^{(\ell+1)} = \mathrm{ReLU}\left(\mathrm{LN}\left[\sum_{q=0}^{2} \pi_{iq} \mathbf{z}_i^{(q)}\right]\right). \tag{10}$$

Dropout is applied after normalization. The two layer GCN+CLASP host uses the same preliminary classifier semantics and omits Eqs. (9) through (10). GCN+CLASP evaluates the complete aggregation module without HARA's additional channels, refinements, and mixer.

D Validation Selected Configurations and Complete Budget Results

D.1 Search and selection protocol

We select one configuration per host and dataset, then hold it fixed across test attack families and budgets. The first search evaluates two hosts on seven datasets. Each combination uses 32 initial configurations and 16 refinements guided by validation, for 672 configurations in total. Screening uses seed zero and ranks mean validation accuracy on the clean graph and six fixed transfer graphs. PR-BCD generates these graphs against GCN and RUNG-MCP using validation objectives. Three candidates per host and dataset are confirmed with seeds zero, one, and two. Confirmation also evaluates attacks against each fitted model. Local attacks cover three datasets, and global attacks cover four. Local validation uses 20 fixed targets.

Downstream training selects checkpoints by validation loss, with at most 300 epochs and patience 60. Preliminary training uses at most 200 epochs and patience 50. This stage selects one parameter vector for each host and dataset before evaluating the full test grid.

The first search's test results determine where we allocate additional HARA search. Six dataset and attack pairs had no strictly winning budget: global evasion and poisoning on Cora, local and global evasion on CiteSeer, local evasion on Actor, and shared attacks on Amazon-Ratings. The second search evaluates 64 new HARA configurations on each of these four datasets. Candidate selection within this search uses validation data. Candidates include previous parameters retrained under the expanded schedule, parents selected on validation data, structured variations, and deterministically sampled settings. The fitted model selected in the first search remains a separate incumbent. New fits use at most 800 epochs with patience 100. Preliminary training uses at most 300 epochs with patience 75. Both stages select checkpoints by validation loss, and fitting labels outside the training and validation sets are masked.

The expanded search uses learning rates $\{0.0003, 0.001, 0.003, 0.01, 0.03\}$ and weight decay $\{0, 0.00005, 0.0005, 0.005, 0.02\}$. Hidden widths are $\{32, 64, 128\}$, and dropout values are $\{0, 0.2, 0.5, 0.7\}$. Cutoffs range from 0.25 to 50, with one, two, four, or eight refinements. Anchor coefficients range from zero to 0.95, and models use one to three layers. The frozen catalogue specifies 64 combinations for each dataset, sampled from this space.

Screening reuses the six first stage validation transfer graphs. We shortlist eight new candidates by cycling through four rankings and taking the next unused candidate from each. The first three rank clean accuracy, mean accuracy across seven conditions, and worst transfer accuracy. The fourth ranks mean accuracy among configurations with anchor coefficient at least 0.8. The incumbent is added as a ninth candidate. All nine receive seed zero attack validation. Three new candidates and the incumbent then receive confirmation with the other two seeds, so final comparisons use three seed means for all four candidates.

The shared family uses an AdaptiveNettack transfer proxy directed at a GCN on 20 fixed validation targets. Final tests retain the original frozen Nettack graphs. Model specific local validation uses the same 20 targets. Global validation attacks all validation nodes with the same PR-BCD settings as test evasion. All three budgets of each registered family are included. Cora additionally refits each candidate on the three fixed poisoned graphs per seed and measures validation accuracy. Validation uses fixed RUNG-MCP on Cora, CiteSeer, and Amazon-Ratings, and fixed HARA-Mean on Actor. The strongest baseline in the test comparison can vary with the budget.

For each attack family, the validation score is the largest budget specific difference of three seed mean accuracy from its reference. We rank candidates lexicographically. The first criterion is the number of targeted families with a positive score, followed by the minimum score among targeted families. Next come the positive count and mean score across all evaluated families. Clean validation accuracy and candidate identifier resolve the remaining ties, in this order. The seed zero shortlist uses the same ordering with that single seed. Validation comparisons use integer correct counts and exact fractions to preserve ties and the ordering of subsequent criteria.

Parameters for all four datasets are sealed before the final test evaluation. Every original test budget is retained. One parameter vector per dataset is used across attack families and budgets. The other three datasets and all GCN+CLASP configurations retain first stage selections.

The final comparison contains 48 scenarios across datasets, attacks, and budgets: 21 shared, nine local, 12 global, and six poisoning. It adds 90 HARA evaluations across scenarios and seeds on the four targeted datasets. The remaining HARA evaluations and all GCN+CLASP evaluations come from the first search. We count overall wins on this full grid. The selected budget display maximizes the observed HARA margin over the strongest fixed baseline independently for each of the 16 dataset and attack pairs. This use of test outcomes is retrospective.

D.2 Fixed references, uncertainty, and numerical execution

The shared attack study, fixed stress study, and matched controls retain separate cohorts. The stress study training recipe uses hidden width 64, learning rate 0.01, weight decay 5×10^{-4}, at most 200 epochs, and validation loss early stopping with patience 50. All retained baseline cohorts have three seeds. When a method has multiple eligible cohorts, the table takes the complete cohort with the highest mean at the given budget. Cell provenance records the selected cohort, keeping its three seeds together. HARA receives the additional search described above, while these baseline configurations remain fixed. The comparison therefore uses unequal tuning budgets.

Shared and model specific local accuracies use 100 fixed test targets per dataset. Global evasion, poisoning, and the clean table use all test nodes. Reported standard deviations summarize three fitted replicates. Global attacks and poisoning also vary the attack seed. Three seed means and standard deviations support descriptive comparisons; we do not test statistical significance. GCN+CLASP appears in every applicable table, as a separate proposed method outside the fixed baseline set. Bold values identify the largest unrounded mean, including GCN+CLASP.

The targeted runs use segmented lower median evaluation and, on Actor and Amazon-Ratings, CUDA sparse two hop evaluation with the original sampling rule. Cora and CiteSeer retain CPU two hop construction. Numerical checks compare the operators and fitted model forward outputs. Full attack trajectories can vary across backends, because floating point differences can change graph selection near a tied proposal. Each dataset's backend is fixed throughout the targeted run, and the source and checkpoint hashes identify the executed calculation.

Table 3: Selected parameters, held fixed across attack types and budgets within each dataset. T is the refinement count and α the anchor coefficient. GCN+CLASP has two layers. N/A marks inapplicable refinement parameters.

Dataset	Host	lr	decay	width	drop	$\gamma_\star$	T	α	layers
Cora	HARA	0.01	0.0005	64	0.5	1.25	2	0.15	2
Cora	GCN+CLASP	0.01	0.0005	32	0	2	N/A	N/A	2
CiteSeer	HARA	0.003	5e-05	64	0.5	50	1	0.8	1
CiteSeer	GCN+CLASP	0.01	0.0005	32	0.5	2	N/A	N/A	2
Chameleon-F	HARA	0.003	5e-05	128	0	0.75	2	0	3
Chameleon-F	GCN+CLASP	0.003	0.005	64	0	0.75	N/A	N/A	2
Squirrel-F	HARA	0.01	0	64	0.5	10	4	0	2
Squirrel-F	GCN+CLASP	0.001	0.005	64	0.2	0.5	N/A	N/A	2
Actor	HARA	0.003	0.0005	64	0.5	20	1	0.95	2
Actor	GCN+CLASP	0.001	0.005	32	0.2	0.75	N/A	N/A	2
Roman-Empire	HARA	0.001	0	64	0.2	1.5	1	0.05	3
Roman-Empire	GCN+CLASP	0.01	0.0005	128	0.5	10	N/A	N/A	2
Amazon-Ratings	HARA	0.0003	0.02	128	0.7	5	2	0.05	2
Amazon-Ratings	GCN+CLASP	0.01	0.0005	128	0.5	5	N/A	N/A	2

Table 4: Clean accuracy of the selected configurations on all test nodes; mean and sample SD over three seeds, in percent.

Dataset	HARA	GCN+CLASP
Cora	$\mathbf{76.4 \pm 0.2}$	74.4 ± 1.8
CiteSeer	59.6 ± 1.5	$\mathbf{61.8 \pm 1.7}$
Chameleon-F	$\mathbf{39.3 \pm 3.6}$	38.4 ± 2.3
Squirrel-F	$\mathbf{39.9 \pm 0.9}$	35.3 ± 0.4
Actor	$\mathbf{38.4 \pm 0.7}$	29.1 ± 0.4
Roman-Empire	$\mathbf{76.0 \pm 0.5}$	55.4 ± 0.4
Amazon-Ratings	$\mathbf{48.0 \pm 0.1}$	46.9 ± 0.7

Table 5: The six targeted pairs at their retrospectively chosen display budgets. Accuracies are three seed means in percent. The gap is HARA minus the strongest recorded fixed baseline, in percentage points. Ties do not count as wins.

Dataset	Attack	Budget	HARA	Baseline	Baseline acc.	Gap
Amazon-Ratings	Shared Nettack	1 edit	42.33	SoftMedian	41.33	+1.00
CiteSeer	Local evasion	5 edits	33.00	RUNG-MCP	26.67	+6.33
Actor	Local evasion	5 edits	28.67	HARA-Mean	25.33	+3.33
Cora	Global PR-BCD	3%	73.60	RUNG-MCP	75.80	-2.20
CiteSeer	Global PR-BCD	3%	58.40	RUNG-MCP	64.67	-6.27
Cora	Metattack	3%	73.77	RUNG-MCP	75.97	-2.20

D.3 RETROSPECTIVELY SELECTED BUDGETS

Tables 6 through 9 present one budget for each dataset and attack. The budget maximizes HARA's observed mean margin over the strongest fixed baseline, with ties resolved toward the smaller budget. HARA exceeds this reference in 13 of 16 displayed comparisons. The displayed budgets are selected using test outcomes and favor HARA within each dataset and attack pair. Figure 3 in the main text and the numerical tables below report all 48 scenarios.

Table 6: Shared Nettack: attacked accuracy on the same 100 fixed test targets. Each dataset uses its retrospectively selected budget. Values are percentages, mean and sample SD over three seeds. Bold denotes the largest unrounded mean, including GCN+CLASP.

Method	Cora 5 edits	CiteSeer 5 edits	Chameleon-F 3 edits	Squirrel-F 5 edits
GCN	1.3 ± 0.6	3.7 ± 0.6	3.3 ± 2.3	3.0 ± 1.7
RUNG-L1	28.0 ± 2.6	15.0 ± 0.0	26.3 ± 1.5	19.7 ± 0.6
RUNG-MCP	51.0 ± 3.0	29.7 ± 2.1	27.0 ± 1.7	21.3 ± 1.5
SoftMedian	11.0 ± 1.0	6.3 ± 0.6	31.3 ± 0.6	25.3 ± 2.3
EvenNet	5.7 ± 0.6	7.3 ± 0.6	10.7 ± 1.5	5.3 ± 0.6
HARA-Mean	14.3 ± 3.2	17.3 ± 1.5	32.0 ± 6.9	15.7 ± 1.2
HARA-Median	11.3 ± 0.6	20.7 ± 6.4	35.7 ± 6.7	26.0 ± 7.0
GCN+Global	37.3 ± 3.8	N/A	N/A	N/A
HARA+Global	50.3 ± 2.1	N/A	N/A	N/A
GCN+CLASP	32.0 ± 1.0	37.3 ± 1.5	38.0 ± 1.7	**36.7 ± 2.1**
HARA	**52.7 ± 1.5**	**46.0 ± 2.0**	**41.7 ± 0.6**	31.3 ± 4.2

Method	Actor 5 edits	Roman-Empire 1 edit	Amazon-Ratings 1 edit
GCN	1.0 ± 0.0	27.7 ± 1.2	24.7 ± 3.1
RUNG-L1	6.7 ± 0.6	60.7 ± 0.6	39.3 ± 2.5
RUNG-MCP	19.7 ± 2.3	63.7 ± 0.6	41.0 ± 1.0
SoftMedian	23.0 ± 1.0	39.7 ± 1.2	41.3 ± 0.6
EvenNet	28.7 ± 4.9	66.3 ± 0.6	27.0 ± 1.0
HARA-Mean	32.7 ± 1.5	68.7 ± 2.1	29.0 ± 4.6
HARA-Median	32.7 ± 5.1	61.0 ± 3.5	37.0 ± 4.0
GCN+Global	19.0 ± 1.7	28.3 ± 5.0	39.3 ± 4.2
HARA+Global	31.3 ± 3.2	64.7 ± 3.5	39.7 ± 3.1
GCN+CLASP	22.0 ± 1.0	45.7 ± 1.2	**42.7 ± 0.6**
HARA	**36.0 ± 1.0**	**72.7 ± 2.1**	42.3 ± 5.5

Table 7: Model specific local evasion: attacked accuracy on the same 100 fixed test targets. Each dataset uses its retrospectively selected budget. Values are percentages, mean and sample SD over three seeds. Bold denotes the largest unrounded mean, including GCN+CLASP.

Method	CiteSeer 5 edits	Squirrel-F 1 edit	Actor 5 edits
GCN	3.7 ± 1.5	8.0 ± 0.0	2.0 ± 0.0
RUNG-MCP	26.7 ± 1.2	27.0 ± 0.0	4.0 ± 1.0
HARA-Mean	2.3 ± 1.5	14.3 ± 5.1	25.3 ± 1.5
GCN+Global	14.0 ± 1.7	24.7 ± 4.7	1.7 ± 0.6
GCN+CLASP	7.7 ± 1.5	25.3 ± 2.1	3.7 ± 1.2
HARA	**33.0 ± 3.6**	**27.7 ± 1.5**	**28.7 ± 4.0**

Table 8: Global PR-BCD evasion: attacked accuracy on all test nodes. Each dataset uses its retrospectively selected budget. Values are percentages, mean and sample SD over three seeds. Bold denotes the largest unrounded mean, including GCN+CLASP.

Method	Cora 3%	CiteSeer 3%	Squirrel-F 1%	Actor 1%
GCN	67.2 ± 0.4	56.3 ± 1.4	20.5 ± 2.0	26.5 ± 0.3
RUNG-MCP	$\mathbf{75.8 \pm 1.0}$	$\mathbf{64.7 \pm 0.1}$	30.8 ± 0.2	30.0 ± 0.6
HARA-Mean	61.7 ± 1.0	55.1 ± 3.0	16.2 ± 2.6	34.9 ± 1.3
GCN+Global	63.8 ± 2.8	52.4 ± 2.3	29.4 ± 2.0	24.6 ± 1.1
GCN+CLASP	65.8 ± 2.6	56.3 ± 2.5	$\mathbf{33.7 \pm 0.4}$	28.4 ± 0.2
HARA	73.6 ± 0.8	58.4 ± 2.0	32.5 ± 2.5	$\mathbf{38.1 \pm 1.0}$

Table 9: Transferred Metattack poisoning: attacked accuracy on all test nodes. Each dataset uses its retrospectively selected budget. Values are percentages, mean and sample SD over three seeds. Bold denotes the largest unrounded mean, including GCN+CLASP.

Method	Cora 3%	Chameleon-F 5%
GCN	70.8 ± 1.6	32.8 ± 4.2
RUNG-MCP	$\mathbf{76.0 \pm 1.3}$	30.4 ± 1.7
HARA-Mean	73.3 ± 1.0	38.0 ± 4.8
GCN+Global	64.2 ± 1.2	33.5 ± 0.6
HARA+Global	74.2 ± 1.7	38.0 ± 3.7
GCN+CLASP	71.7 ± 0.8	35.8 ± 1.0
HARA	73.8 ± 0.9	$\mathbf{39.7 \pm 1.1}$

D.4 ALL EVALUATED BUDGETS

Table 10: Cora: Shared Nettack at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	43.7 ± 2.1	9.7 ± 1.2	1.3 ± 0.6
RUNG-L1	71.7 ± 1.2	45.0 ± 0.0	28.0 ± 2.6
RUNG-MCP	**77.3 ± 4.7**	64.0 ± 4.0	51.0 ± 3.0
SoftMedian	61.0 ± 3.6	29.0 ± 2.0	11.0 ± 1.0
EvenNet	55.3 ± 1.5	22.7 ± 2.5	5.7 ± 0.6
HARA-Mean	57.0 ± 3.6	32.3 ± 6.1	14.3 ± 3.2
HARA-Median	59.7 ± 1.2	32.0 ± 0.0	11.3 ± 0.6
GCN+Global	64.7 ± 2.5	54.3 ± 4.7	37.3 ± 3.8
HARA+Global	72.3 ± 2.5	63.0 ± 2.6	50.3 ± 2.1
GCN+CLASP	65.3 ± 0.6	51.7 ± 2.1	32.0 ± 1.0
HARA	72.0 ± 1.0	**64.7 ± 2.5**	**52.7 ± 1.5**

Table 11: CiteSeer: Shared Nettack at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	29.0 ± 2.0	5.0 ± 0.0	3.7 ± 0.6
RUNG-L1	67.3 ± 2.5	32.3 ± 1.5	15.0 ± 0.0
RUNG-MCP	**70.7 ± 0.6**	**57.3 ± 1.2**	29.7 ± 2.1
SoftMedian	49.3 ± 1.2	12.7 ± 0.6	6.3 ± 0.6
EvenNet	40.7 ± 1.5	14.3 ± 1.2	7.3 ± 0.6
HARA-Mean	49.7 ± 1.2	27.3 ± 4.5	17.3 ± 1.5
HARA-Median	52.7 ± 3.8	30.3 ± 9.3	20.7 ± 6.4
GCN+CLASP	64.3 ± 0.6	48.3 ± 1.5	37.3 ± 1.5
HARA	53.7 ± 0.6	48.7 ± 2.1	**46.0 ± 2.0**

Table 12: Chameleon-F: Shared Nettack at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	7.7 ± 2.1	3.3 ± 2.3	2.3 ± 0.6
RUNG-L1	28.7 ± 1.2	26.3 ± 1.5	25.0 ± 1.7
RUNG-MCP	28.0 ± 1.0	27.0 ± 1.7	24.7 ± 2.1
SoftMedian	36.7 ± 0.6	31.3 ± 0.6	27.3 ± 4.0
EvenNet	21.3 ± 2.5	10.7 ± 1.5	7.3 ± 1.5
HARA-Mean	38.0 ± 5.3	32.0 ± 6.9	27.3 ± 9.0
HARA-Median	37.0 ± 7.0	35.7 ± 6.7	33.7 ± 5.0
GCN+CLASP	38.0 ± 4.4	38.0 ± 1.7	35.3 ± 0.6
HARA	**38.7 ± 3.2**	**41.7 ± 0.6**	**38.0 ± 1.7**

Table 13: Squirrel-F: Shared Nettack at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	9.0 ± 1.0	5.7 ± 0.6	3.0 ± 1.7
RUNG-L1	28.3 ± 1.5	22.7 ± 0.6	19.7 ± 0.6
RUNG-MCP	28.7 ± 2.1	24.7 ± 1.5	21.3 ± 1.5
SoftMedian	29.0 ± 1.7	28.0 ± 0.0	25.3 ± 2.3
EvenNet	10.7 ± 2.5	6.7 ± 1.2	5.3 ± 0.6
HARA-Mean	34.0 ± 5.6	24.0 ± 6.6	15.7 ± 1.2
HARA-Median	34.3 ± 0.6	32.7 ± 3.8	26.0 ± 7.0
GCN+CLASP	33.3 ± 1.2	$\mathbf{35.7 \pm 2.1}$	$\mathbf{36.7 \pm 2.1}$
HARA	$\mathbf{36.7 \pm 2.1}$	34.3 ± 2.3	31.3 ± 4.2

Table 14: Actor: Shared Nettack at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	7.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
RUNG-L1	18.7 ± 1.5	9.0 ± 1.7	6.7 ± 0.6
RUNG-MCP	27.7 ± 2.1	24.7 ± 3.2	19.7 ± 2.3
SoftMedian	25.0 ± 1.7	24.3 ± 1.5	23.0 ± 1.0
EvenNet	31.3 ± 2.1	27.7 ± 4.0	28.7 ± 4.9
HARA-Mean	33.7 ± 2.1	32.7 ± 1.5	32.7 ± 1.5
HARA-Median	33.0 ± 3.6	33.7 ± 4.0	32.7 ± 5.1
GCN+Global	24.3 ± 3.2	22.0 ± 1.0	19.0 ± 1.7
HARA+Global	$\mathbf{34.0 \pm 4.6}$	33.3 ± 2.5	31.3 ± 3.2
GCN+CLASP	25.0 ± 1.0	25.3 ± 1.5	22.0 ± 1.0
HARA	$\mathbf{34.0 \pm 2.6}$	$\mathbf{34.0 \pm 2.6}$	$\mathbf{36.0 \pm 1.0}$

Table 15: Roman-Empire: Shared Nettack at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	27.7 ± 1.2	4.3 ± 1.5	0.7 ± 0.6
RUNG-L1	60.7 ± 0.6	53.0 ± 1.7	45.0 ± 1.0
RUNG-MCP	63.7 ± 0.6	$\mathbf{65.0 \pm 0.0}$	$\mathbf{60.3 \pm 0.6}$
SoftMedian	39.7 ± 1.2	33.3 ± 2.3	16.3 ± 1.2
EvenNet	66.3 ± 0.6	60.0 ± 1.0	51.7 ± 0.6
HARA-Mean	68.7 ± 2.1	54.3 ± 2.1	48.0 ± 3.5
HARA-Median	61.0 ± 3.5	52.7 ± 1.2	53.7 ± 1.2
GCN+Global	28.3 ± 5.0	20.7 ± 2.1	13.7 ± 1.2
HARA+Global	64.7 ± 3.5	55.0 ± 2.6	51.7 ± 3.8
GCN+CLASP	45.7 ± 1.2	31.3 ± 2.1	22.3 ± 0.6
HARA	$\mathbf{72.7 \pm 2.1}$	55.3 ± 1.2	46.3 ± 3.5

Table 16: Amazon-Ratings: Shared Nettack at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	24.7 ± 3.1	11.3 ± 1.5	9.3 ± 0.6
RUNG-L1	39.3 ± 2.5	37.0 ± 2.0	32.7 ± 0.6
RUNG-MCP	41.0 ± 1.0	40.0 ± 1.0	35.0 ± 1.0
SoftMedian	41.3 ± 0.6	36.0 ± 1.0	30.0 ± 2.0
EvenNet	27.0 ± 1.0	16.7 ± 0.6	16.0 ± 1.0
HARA-Mean	29.0 ± 4.6	19.7 ± 2.5	19.7 ± 2.5
HARA-Median	37.0 ± 4.0	28.7 ± 2.1	25.7 ± 1.5
GCN+Global	39.3 ± 4.2	39.7 ± 3.1	**39.3 ± 3.2**
HARA+Global	39.7 ± 3.1	37.3 ± 2.3	36.0 ± 1.0
GCN+CLASP	**42.7 ± 0.6**	35.0 ± 2.6	26.0 ± 3.5
HARA	42.3 ± 5.5	**40.7 ± 3.8**	33.3 ± 5.1

Table 17: CiteSeer: Model specific local evasion at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	26.3 ± 3.8	5.7 ± 2.1	3.7 ± 1.5
RUNG-MCP	**63.3 ± 2.3**	**39.0 ± 1.0**	26.7 ± 1.2
HARA-Mean	12.7 ± 2.1	4.3 ± 0.6	2.3 ± 1.5
GCN+Global	32.3 ± 2.3	17.3 ± 3.5	14.0 ± 1.7
GCN+CLASP	35.7 ± 1.5	11.3 ± 1.2	7.7 ± 1.5
HARA	48.3 ± 3.2	37.0 ± 2.6	**33.0 ± 3.6**

Table 18: Squirrel-F: Model specific local evasion at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	8.0 ± 0.0	5.0 ± 0.0	5.0 ± 0.0
RUNG-MCP	27.0 ± 0.0	**24.0 ± 1.0**	**21.7 ± 1.2**
HARA-Mean	14.3 ± 5.1	5.3 ± 2.3	3.0 ± 0.0
GCN+Global	24.7 ± 4.7	21.3 ± 7.6	20.7 ± 8.3
GCN+CLASP	25.3 ± 2.1	21.7 ± 2.1	21.0 ± 2.0
HARA	**27.7 ± 1.5**	22.7 ± 2.9	17.7 ± 6.0

Table 19: Actor: Model specific local evasion at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1 edit	3 edits	5 edits
GCN	6.3 ± 0.6	2.0 ± 0.0	2.0 ± 0.0
RUNG-MCP	9.7 ± 1.5	5.3 ± 0.6	4.0 ± 1.0
HARA-Mean	28.7 ± 0.6	26.3 ± 0.6	25.3 ± 1.5
GCN+Global	7.3 ± 1.2	3.0 ± 1.0	1.7 ± 0.6
GCN+CLASP	9.7 ± 0.6	4.7 ± 0.6	3.7 ± 1.2
HARA	**30.7 ± 2.3**	**29.0 ± 3.5**	**28.7 ± 4.0**

Table 20: Cora: Global PR-BCD evasion at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1%	3%	5%
GCN	75.3 ± 0.3	67.2 ± 0.4	60.8 ± 0.2
RUNG-MCP	**79.4 ± 1.3**	**75.8 ± 1.0**	**73.1 ± 1.2**
HARA-Mean	76.0 ± 1.0	61.7 ± 1.0	53.2 ± 0.8
GCN+Global	68.1 ± 2.1	63.8 ± 2.8	61.0 ± 2.1
GCN+CLASP	72.1 ± 1.1	65.8 ± 2.6	62.0 ± 2.1
HARA	74.4 ± 0.7	73.6 ± 0.8	68.3 ± 1.2

Table 21: CiteSeer: Global PR-BCD evasion at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1%	3%	5%
GCN	63.6 ± 0.7	56.3 ± 1.4	49.8 ± 1.7
RUNG-MCP	**67.8 ± 0.3**	**64.7 ± 0.1**	**62.6 ± 0.4**
HARA-Mean	57.6 ± 2.2	55.1 ± 3.0	45.6 ± 1.6
GCN+Global	53.3 ± 2.2	52.4 ± 2.3	49.6 ± 2.3
GCN+CLASP	61.5 ± 1.5	56.3 ± 2.5	52.2 ± 1.8
HARA	58.8 ± 1.4	58.4 ± 2.0	53.7 ± 1.7

Table 22: Squirrel-F: Global PR-BCD evasion at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1%	3%	5%
GCN	20.5 ± 2.0	14.5 ± 3.6	12.1 ± 3.2
RUNG-MCP	30.8 ± 0.2	30.6 ± 0.2	30.3 ± 0.1
HARA-Mean	16.2 ± 2.6	9.0 ± 0.4	6.0 ± 1.0
GCN+Global	29.4 ± 2.0	27.5 ± 3.6	26.6 ± 4.3
GCN+CLASP	**33.7 ± 0.4**	**32.7 ± 1.1**	**32.3 ± 1.3**
HARA	32.5 ± 2.5	28.1 ± 5.1	21.0 ± 2.1

Table 23: Actor: Global PR-BCD evasion at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1%	3%	5%
GCN	26.5 ± 0.3	20.9 ± 0.4	17.3 ± 0.6
RUNG-MCP	30.0 ± 0.6	28.6 ± 0.6	27.6 ± 0.5
HARA-Mean	34.9 ± 1.3	34.9 ± 1.2	34.5 ± 1.1
GCN+Global	24.6 ± 1.1	21.8 ± 1.4	20.4 ± 1.6
GCN+CLASP	28.4 ± 0.2	27.8 ± 0.2	27.3 ± 0.4
HARA	**38.1 ± 1.0**	**38.0 ± 0.8**	**37.6 ± 0.8**

Table 24: Cora: Transferred Metattack poisoning at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1%	3%	5%
GCN	77.4 ± 0.7	70.8 ± 1.6	68.1 ± 0.9
RUNG-MCP	**78.9 ± 0.6**	**76.0 ± 1.3**	**74.5 ± 1.4**
HARA-Mean	76.6 ± 0.6	73.3 ± 1.0	70.2 ± 3.2
GCN+Global	67.8 ± 2.7	64.2 ± 1.2	64.2 ± 1.9
HARA+Global	76.2 ± 0.4	74.2 ± 1.7	69.9 ± 3.6
GCN+CLASP	73.3 ± 1.2	71.7 ± 0.8	70.7 ± 1.3
HARA	75.4 ± 1.3	73.8 ± 0.9	71.2 ± 0.5

Table 25: Chameleon-F: Transferred Metattack poisoning at every evaluated budget. Mean and sample SD over three seeds; percentages.

Method	1%	3%	5%
GCN	34.1 ± 0.0	35.2 ± 1.5	32.8 ± 4.2
RUNG-MCP	31.8 ± 2.4	30.5 ± 1.4	30.4 ± 1.7
HARA-Mean	$\mathbf{39.7 \pm 0.6}$	37.1 ± 2.1	38.0 ± 4.8
GCN+Global	32.4 ± 1.1	31.8 ± 1.7	33.5 ± 0.6
HARA+Global	38.2 ± 3.1	$\mathbf{38.5 \pm 2.4}$	38.0 ± 3.7
GCN+CLASP	37.1 ± 4.2	37.6 ± 0.3	35.8 ± 1.0
HARA	38.7 ± 3.3	$\mathbf{38.5 \pm 2.6}$	$\mathbf{39.7 \pm 1.1}$

E CONTROLLED CALIBRATION AND STRESS COMPARISONS

The following experiments use the fixed training recipe in Appendix B. Matched controls compare calibration rules within each host. Further studies vary the training labels and attack setting, then measure retained bucket membership and distance drift in the attacked models.

E.1 CLASS PAIR CALIBRATION WITHIN EACH HOST

The matched comparison holds the penalty cutoff, initialization, and training fixed, while changing only the calibration groups (Table 2). With full training labels and five shared Nettack edits, GCN+CLASP reaches 45.0% attacked accuracy on Cora, compared with 37.3% for GCN+Global. On Actor, the corresponding accuracies are 23.0% and 19.0%. Roman-Empire and Amazon-Ratings have lower mean attacked accuracy with class pair calibration. Clean accuracies and the reduced label comparisons appear in Appendix F.

For HARA, the comparison also holds both anchored refinements fixed. Class pair grouping gives higher mean attacked accuracy on 1 of 4 datasets. On Amazon-Ratings, HARA reaches 37.0%, compared with 36.0% for HARA+Global. Table 2 reports the paired differences and their variation over three seeds. These descriptive differences concern the fixed host and training recipe used in this study.

The complete aggregation change also affects clean accuracy. On Cora, GCN+CLASP's full test clean accuracy is 68.8%, compared with 80.5% for GCN. This comparison measures the clean accuracy cost of replacing GCN aggregation, including both calibration and the robust penalty.

E.2 REDUCED TRAINING LABELS

The label study reduces training labels under the fixed training recipe, while retaining validation labels and the attacker's full label surrogate. On Cora, reducing training labels from 100% to 10% changes HARA's attacked accuracy at budget five from 49.0% to 40.0%. GCN+CLASP changes from 45.0% to 23.0% (Figure 4). On Actor, HARA reaches 22.7% at 10%, below HARA-Mean's 29.7%. GCN+CLASP and GCN+Global reach 19.0% and 18.7%, respectively. HARA-Mean also differs in refinement count and anchoring, so its comparison with HARA concerns the configured systems rather than calibration alone.

At 10% on Roman-Empire, GCN+CLASP reaches 14.3%, HARA reaches 53.7%, and RUNG-MCP reaches 58.7%. On Amazon-Ratings, HARA's mean changes from 37.0% at full training labels to 36.0% at 10%. The study with five labels per class further reduces the absolute training budget on these graphs.

The five labels per class study retains the original validation sets. The absolute training budgets on Actor, Roman-Empire, and Amazon-Ratings are 25, 90, and 25 labels. Their validation sets contain 2,432, 7,251, and 7,837 labels, respectively. Under these conditions, HARA reaches 19.3% attacked accuracy on Actor and GCN+CLASP reaches 21.0%. On Roman-Empire, the corresponding accuracies are 22.7% and 8.7%, compared with 28.0% for RUNG-MCP. On Amazon-Ratings, HARA reaches 36.3% and GCN+CLASP reaches 27.0%. Table 30 reports all three datasets.

Preliminary accuracy, bucket occupancy, and downstream outcomes accompany the reduced label results in Appendix F. Table 52 reports clean accuracy for the nine combinations of mask and model seed, separating variation due to the training labels from variation due to model fitting. Attack replicates are reported separately.

E.3 MODEL SPECIFIC STRUCTURAL EVASION

The local searches use the same candidate construction and evaluation budget for all six methods. Each model's own loss ranks its proposals. On CiteSeer, the budget five attack leaves HARA at 18.7% accuracy and RUNG-MCP at 26.7% (Table 26). GCN+CLASP reaches 11.7%, below the matched global scale control's 14.0%. On Squirrel-F, GCN+CLASP reaches 18.0%, compared with 20.7% for GCN+Global. HARA reaches 13.7% and RUNG-MCP reaches 21.7%. On Actor, HARA reaches 26.7%, HARA-Mean reaches 25.3%, and RUNG-MCP reaches 4.0%.

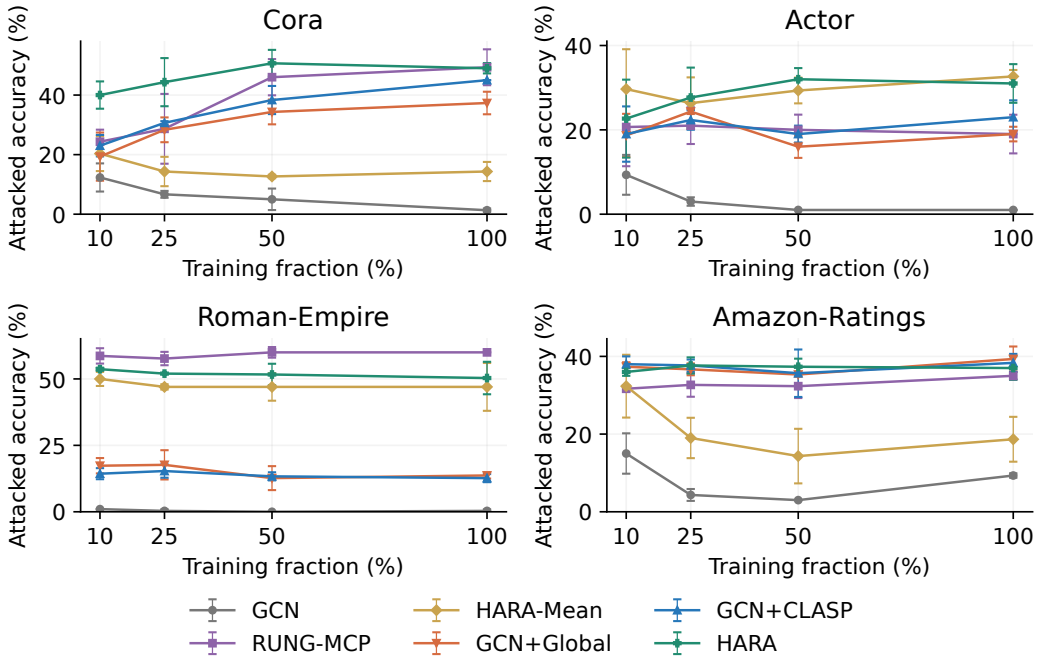


Figure 4: Accuracy on the same 100 targets after five shared Nettack edge flips, as training labels are reduced under the fixed training recipe. Points and error bars show mean and sample standard deviation over three joint mask/training replicates. Validation labels and the attacker's full label surrogate remain available. Panels have separate accuracy scales.

Table 26: Model specific local evasion with at most five undirected edits per target. Attacked accuracy and conditional ASR are percentages, mean and sample standard deviation over three fitted models, using the same 100 targets per dataset. ASR denominators are model specific and appear in Table 68.

Method	CiteSeer Acc.	ASR	Squirrel-F Acc.	ASR	Actor Acc.	ASR
GCN	3.7 ± 1.5	95.0 ± 2.2	5.0 ± 0.0	84.5 ± 1.0	2.0 ± 0.0	93.7 ± 0.4
RUNG-MCP	26.7 ± 1.2	62.4 ± 1.7	21.7 ± 1.2	25.1 ± 6.4	4.0 ± 1.0	87.1 ± 3.8
HARA-Mean	2.3 ± 1.5	96.3 ± 2.6	3.0 ± 0.0	91.5 ± 0.9	25.3 ± 1.5	28.3 ± 4.7
GCN+Global	14.0 ± 1.7	76.3 ± 2.2	20.7 ± 8.3	34.8 ± 25.7	1.7 ± 0.6	92.9 ± 2.8
GCN+CLASP	11.7 ± 2.1	80.2 ± 2.3	18.0 ± 2.6	44.6 ± 10.9	2.3 ± 1.2	91.3 ± 3.8
HARA	18.7 ± 4.2	72.9 ± 5.4	13.7 ± 6.1	59.4 ± 20.1	26.7 ± 4.7	26.2 ± 9.1

The global PR-BCD comparison attacks all test nodes on one graph. At the 5% edge budget on Cora, HARA reaches 68.1%, compared with 73.1% for RUNG-MCP. GCN+CLASP and GCN+Global reach 61.5% and 61.0% (Table 27). RUNG-MCP also has higher mean accuracy than HARA on CiteSeer and Squirrel-F under this global attack. On Actor, HARA reaches 34.6%, HARA-Mean reaches 34.5%, and RUNG-MCP reaches 27.6%. Full budget curves appear in Figure 5.

Attack strength depends on the candidates and objective used in the search. Table 51 varies candidate support, attack objective, top-k, and candidate seed on fixed validation targets. On Cora's single diagnostic HARA fit, expanded support with top 50 breaks 10 of 14 initially correct targets. Local support with top 20 breaks 7, and the cross entropy objective breaks 11.

On Actor, Table 71 compares shared and model specific local attacks on the same full label checkpoints. The accompanying portfolio selects among the shared, model specific, and clean graphs for each target. These selections combine errors from separate targeted attacks, so they do not give simultaneous errors on a single graph.

Table 27: Global PR-BCD evasion (E) and transferred Metattack poisoning (P), each at 5% of editable undirected edges. Accuracy uses all test nodes and is reported as mean and sample standard deviation across three joint attack/training replicates. E keeps model weights fixed and selects one feasible graph from that and smaller budgets, including the clean graph. P retrains the complete pipeline and evaluates on the poisoned graph. The two column types represent different threat models.

Method	Cora (E)	CiteSeer (E)	Squirrel-F (E)	Actor (E)	Cora (P)	Chameleon-F (P)
GCN	60.8 ± 0.2	49.8 ± 1.7	12.1 ± 3.2	17.3 ± 0.6	68.1 ± 0.9	32.8 ± 4.2
RUNG-MCP	73.1 ± 1.2	62.6 ± 0.4	30.3 ± 0.1	27.6 ± 0.5	74.5 ± 1.4	30.4 ± 1.7
HARA-Mean	53.2 ± 0.8	45.6 ± 1.6	6.0 ± 1.0	34.5 ± 1.1	70.2 ± 3.2	38.0 ± 4.8
GCN+Global	61.0 ± 2.1	49.6 ± 2.3	26.6 ± 4.3	20.4 ± 1.6	64.2 ± 1.9	33.5 ± 0.6
GCN+CLASP	61.5 ± 2.1	48.0 ± 1.1	24.8 ± 1.5	20.3 ± 1.8	64.5 ± 2.0	32.8 ± 2.3
HARA	68.1 ± 1.9	60.3 ± 2.8	21.3 ± 6.6	34.6 ± 0.8	71.2 ± 0.5	34.8 ± 0.3

E.4 Transferred training poisoning

The poisoning study transfers structural edits from a GCN surrogate, then retrains the complete pipeline and evaluates it on the poisoned graph. At the 5% budget on Cora, HARA's test accuracy changes from 76.6% to 71.2%. RUNG-MCP reaches 74.5%. GCN+CLASP and the matched global scale control reach 64.5% and 64.2%, respectively (Table 27). On Chameleon-F, HARA changes from 36.5% to 34.8%. HARA-Mean reaches 38.0%. GCN+CLASP reaches 32.8%, compared with 33.5% for its global scale control. GCN's relative performance with class pair and global calibration changes between these graphs.

The comparison between HARA and HARA-Mean changes refinement and anchoring along with aggregation. To isolate class pair grouping, Table 59 compares HARA with HARA+Global. At the same 5% rate, HARA+Global reaches 69.9% on Cora and 38.0% on Chameleon-F. HARA's mean difference from this matched control also changes sign between datasets.

With only preliminary fitting poisoned, HARA reaches 77.6% on Cora and 35.9% on Chameleon-F. Its preliminary accuracies in that condition are 57.4% and 36.7%. Compared with preliminary fitting alone, poisoning the complete pipeline also changes downstream fitting and the evaluation graph. The accuracy gap combines these effects with the resulting preliminary predictions.

Several Chameleon-F baselines retain similar accuracy under transfer from the GCN surrogate. Their response to attacks optimized through defended training remains unmeasured. Tables 49 and 50 report all rates and paired changes from clean training.

E.5 Bucket retention and classification outcomes

A target can be misclassified even when retained edges form a majority in every perturbed bucket. Among initially correct targets with at least one accepted edit, 751 attacked observations retain a majority in every nonempty perturbed bucket at the second layer's first one hop call. Of these, 496 are misclassified after the attack (Table 28). The counts pool dependent target observations across the three fitted models of each method.

Retained majority supplies the membership premise of Proposition 1. Together with a bound on retained distance drift, the retained distance distribution determines the median interval. Neither the membership condition nor the median interval alone bounds classification error.

Table 72 counts observations satisfying the sufficient graph edit condition $m > 2K + R$. Exact retained majority can also hold outside this criterion. Table 53 reports empty buckets, membership, median shifts, and interval widths for the first one hop call at each layer. Later refinements and the two hop channel are outside these measurements. Table 54 reports label changes, reassigned surviving edges, and retained distance drift. The observed distance drift is zero at the first layer's first call and nonzero at the second layer's first call.

Table 28: Retention and prediction outcomes at the second layer's first one hop call, under model specific budget five attacks. "All retain majority" counts edited, initially correct targets with a retained majority in every nonempty bucket. "Failures" counts misclassifications within this subset. Counts pool dependent target observations across three fitted models.

Dataset	Method	Edited targets	All retain majority	Failures among these
CiteSeer	GCN+CLASP	176	176	141
CiteSeer	HARA	206	206	150
Squirrel-F	GCN+CLASP	89	89	44
Squirrel-F	HARA	103	103	62
Actor	GCN+CLASP	75	75	71
Actor	HARA	102	102	28

F Complete Stress Study Results

This section reports every configured condition in the fixed training stress study. Local attacked accuracy uses the fixed target manifest; global evasion and poisoning use every test node. Reported standard deviations summarize three fitted replicates, with joint mask or attack variation as specified in Appendix B. Undefined ASR values retain their contributing replicate counts.

Table 29: Original label splits and editable graph size. Unassigned nodes have features and graph structure available but belong to none of the three label masks. Reduced label experiments change only the training mask. Percentage attack budgets count off diagonal undirected edges; original self loops stay fixed.

Dataset	Train	Val.	Test	Unassigned	Editable edges	Loops
Cora	140	500	1000	1068	5278	0
CiteSeer	120	500	1000	1707	4552	0
Chameleon-F	427	284	179	0	8854	50
Squirrel-F	1067	711	445	0	46998	140
Actor	3648	2432	1520	0	26659	93
Roman-Empire	10877	7251	4534	0	32927	0
Amazon-Ratings	11756	7837	4899	0	93050	0

Table 30: Exactly five training labels per class: 25 labels on Actor and Amazon-Ratings, 90 on Roman-Empire. Original validation labels remain available (2,432; 7,251; 7,837 respectively). Clean and attacked accuracy use the same fixed 100 targets, with five neutral Nettack edge flips for the latter. Values are percentages, mean and sample standard deviation across three joint mask/training replicates.

Method	Actor clean	Attacked	Roman clean	Attacked	Amazon clean	Attacked
GCN	18.0 ± 6.6	17.0 ± 6.0	21.0 ± 2.6	7.7 ± 4.0	21.3 ± 5.7	30.0 ± 4.6
RUNG-MCP	20.7 ± 3.5	20.7 ± 3.5	32.7 ± 3.8	28.0 ± 5.6	29.3 ± 6.7	30.0 ± 5.3
HARA-Mean	21.0 ± 0.0	23.0 ± 1.0	26.3 ± 2.5	16.0 ± 7.9	32.7 ± 4.5	33.0 ± 7.8
GCN+Global	21.3 ± 3.1	20.3 ± 5.0	11.0 ± 2.6	9.7 ± 5.7	29.7 ± 4.0	30.3 ± 5.5
GCN+CLASP	21.0 ± 3.0	21.0 ± 4.0	10.7 ± 3.8	8.7 ± 4.5	29.3 ± 5.1	27.0 ± 1.0
HARA	19.3 ± 3.2	19.3 ± 2.5	25.3 ± 3.8	22.7 ± 3.2	30.3 ± 2.5	36.3 ± 2.1

Table 31: Cora: reduced training labels with the original validation and test masks. Clean accuracy uses all test nodes; attacked accuracy and ASR use the fixed 100 targets at five edge flips. Values are mean and sample standard deviation over three paired mask/training replicates.

Train	Method	Labels	Clean test	Clean target	Attacked	ASR	RAUC
100%	GCN	140	80.5 ± 0.8	75.0 ± 2.6	1.3 ± 0.6	98.2 ± 0.7	0.670 ± 0.008
100%	RUNG-MCP	140	81.9 ± 1.4	80.0 ± 4.6	49.3 ± 6.0	38.9 ± 6.1	0.169 ± 0.029
100%	HARA-Mean	140	78.2 ± 0.9	70.0 ± 2.0	14.3 ± 3.2	79.4 ± 5.1	0.429 ± 0.068
100%	GCN+Global	140	68.9 ± 2.6	65.0 ± 2.6	37.3 ± 3.8	42.7 ± 3.6	0.158 ± 0.017
100%	GCN+CLASP	140	68.8 ± 2.8	62.3 ± 3.8	45.0 ± 0.0	32.4 ± 4.7	0.118 ± 0.021
100%	HARA	140	76.6 ± 1.8	73.0 ± 2.0	49.0 ± 1.7	36.5 ± 2.4	0.148 ± 0.007
50%	GCN	70	78.9 ± 2.9	77.0 ± 3.6	5.0 ± 3.6	93.6 ± 4.3	0.634 ± 0.035
50%	RUNG-MCP	70	77.6 ± 2.0	75.0 ± 2.6	46.0 ± 6.1	38.9 ± 9.8	0.158 ± 0.040
50%	HARA-Mean	70	75.2 ± 2.0	70.0 ± 3.5	12.7 ± 0.6	81.8 ± 1.7	0.425 ± 0.030
50%	GCN+Global	70	63.0 ± 0.9	59.7 ± 4.6	34.3 ± 4.2	43.9 ± 7.7	0.174 ± 0.052
50%	GCN+CLASP	70	65.4 ± 3.4	62.7 ± 2.5	38.3 ± 4.7	41.5 ± 6.7	0.174 ± 0.039
50%	HARA	70	73.4 ± 3.7	71.7 ± 4.0	50.7 ± 4.5	31.8 ± 4.6	0.145 ± 0.008
25%	GCN	35	71.5 ± 2.3	67.3 ± 2.9	6.7 ± 1.2	90.0 ± 2.2	0.591 ± 0.016
25%	RUNG-MCP	35	71.7 ± 4.4	67.3 ± 3.2	28.7 ± 11.7	57.9 ± 15.1	0.260 ± 0.065
25%	HARA-Mean	35	72.1 ± 4.4	66.3 ± 3.2	14.3 ± 4.9	78.6 ± 6.2	0.439 ± 0.026
25%	GCN+Global	35	51.7 ± 8.8	51.7 ± 7.6	28.3 ± 4.2	46.4 ± 1.8	0.208 ± 0.043
25%	GCN+CLASP	35	54.1 ± 5.2	52.3 ± 2.5	30.7 ± 0.6	46.5 ± 0.7	0.228 ± 0.030
25%	HARA	35	70.7 ± 3.8	65.7 ± 6.7	44.3 ± 8.1	37.4 ± 5.1	0.163 ± 0.021
10%	GCN	14	54.7 ± 8.4	52.0 ± 7.9	12.3 ± 4.7	75.9 ± 11.2	0.467 ± 0.091
10%	RUNG-MCP	14	59.3 ± 3.6	57.3 ± 5.7	24.3 ± 4.0	57.0 ± 10.8	0.265 ± 0.087
10%	HARA-Mean	14	59.1 ± 8.4	57.7 ± 10.7	20.3 ± 5.9	67.4 ± 10.3	0.372 ± 0.089
10%	GCN+Global	14	29.2 ± 16.2	26.7 ± 17.8	19.3 ± 8.1	41.7 ± 12.5	0.176 ± 0.074
10%	GCN+CLASP	14	33.6 ± 8.8	31.3 ± 13.5	23.0 ± 3.6	42.1 ± 13.1	0.194 ± 0.068
10%	HARA	14	57.5 ± 2.2	56.7 ± 4.0	40.0 ± 4.6	35.3 ± 2.7	0.181 ± 0.020

Table 32: Cora: reduced training labels at the two smaller shared Nettack budgets. Attacked accuracy uses all 100 targets; ASR uses each fitted model's initially correct targets. Values are percentages, mean and sample standard deviation across three joint mask/training replicates.

Train	Method	Acc. (1)	ASR (1)	Acc. (3)	ASR (3)
100%	GCN	43.7 ± 2.1	41.8 ± 1.3	9.7 ± 1.2	87.1 ± 1.1
100%	RUNG-MCP	77.3 ± 4.7	3.3 ± 0.8	64.0 ± 4.0	20.4 ± 4.0
100%	HARA-Mean	57.0 ± 3.6	18.5 ± 6.3	32.3 ± 6.1	53.7 ± 9.8
100%	GCN+Global	64.7 ± 2.5	1.5 ± 1.5	54.3 ± 4.7	17.0 ± 3.7
100%	GCN+CLASP	63.0 ± 1.0	1.5 ± 1.5	58.7 ± 3.2	12.2 ± 1.8
100%	HARA	73.7 ± 1.2	3.6 ± 2.1	64.0 ± 2.6	16.0 ± 3.2
50%	GCN	44.7 ± 5.5	42.0 ± 6.9	15.3 ± 2.9	80.1 ± 3.2
50%	RUNG-MCP	73.0 ± 1.7	2.6 ± 1.2	61.3 ± 1.5	18.1 ± 4.4
50%	HARA-Mean	57.3 ± 4.7	18.0 ± 6.9	34.0 ± 2.0	51.9 ± 2.3
50%	GCN+Global	58.7 ± 5.5	3.4 ± 1.9	49.0 ± 4.0	19.0 ± 10.5
50%	GCN+CLASP	61.0 ± 2.6	4.8 ± 4.2	52.3 ± 4.9	19.2 ± 4.7
50%	HARA	68.0 ± 2.6	6.0 ± 2.6	61.7 ± 3.5	15.9 ± 1.6
25%	GCN	41.3 ± 0.6	39.0 ± 3.0	18.0 ± 4.4	73.3 ± 5.8
25%	RUNG-MCP	64.7 ± 4.7	5.0 ± 1.9	46.3 ± 7.5	32.3 ± 7.9
25%	HARA-Mean	51.0 ± 4.0	23.7 ± 2.6	31.3 ± 2.5	52.8 ± 1.7
25%	GCN+Global	48.0 ± 8.7	8.6 ± 8.2	41.7 ± 8.6	22.4 ± 5.0
25%	GCN+CLASP	48.7 ± 4.7	10.9 ± 7.3	41.0 ± 5.3	25.6 ± 3.3
25%	HARA	62.0 ± 6.9	7.5 ± 2.3	57.0 ± 6.6	16.4 ± 2.5
10%	GCN	37.7 ± 8.4	28.9 ± 11.6	22.0 ± 4.6	57.1 ± 10.8
10%	RUNG-MCP	53.3 ± 2.3	8.4 ± 5.9	39.3 ± 6.7	31.6 ± 12.6
10%	HARA-Mean	48.0 ± 9.0	19.3 ± 9.2	34.3 ± 5.8	44.9 ± 10.5
10%	GCN+Global	27.3 ± 15.3	6.5 ± 6.5	24.7 ± 13.2	18.2 ± 8.0
10%	GCN+CLASP	31.7 ± 11.6	6.9 ± 5.5	27.0 ± 7.2	22.3 ± 9.7
10%	HARA	53.0 ± 7.8	10.8 ± 4.3	49.0 ± 6.2	19.5 ± 2.2

Table 33: Actor: reduced training labels with the original validation and test masks. Clean accuracy uses all test nodes; attacked accuracy and ASR use the fixed 100 targets at five edge flips. Values are mean and sample standard deviation over three paired mask/training replicates.

Train	Method	Labels	Clean test	Clean target	Attacked	ASR	RAUC
100%	GCN	3648	30.6 ± 0.1	32.0 ± 2.0	1.0 ± 0.0	96.9 ± 0.2	0.815 ± 0.005
100%	RUNG-MCP	3648	31.8 ± 0.4	31.3 ± 1.5	19.0 ± 4.6	50.7 ± 12.5	0.285 ± 0.074
100%	HARA-Mean	3648	35.5 ± 1.5	35.3 ± 0.6	32.7 ± 1.5	11.3 ± 5.7	0.106 ± 0.046
100%	GCN+Global	3648	26.2 ± 0.8	24.0 ± 4.6	19.0 ± 1.7	34.5 ± 15.5	0.174 ± 0.064
100%	GCN+CLASP	3648	26.4 ± 0.9	26.0 ± 2.6	23.0 ± 4.0	26.4 ± 12.4	0.173 ± 0.039
100%	HARA	3648	35.2 ± 0.6	36.0 ± 2.0	31.0 ± 4.6	18.9 ± 9.6	0.141 ± 0.060
50%	GCN	1824	29.2 ± 0.4	29.0 ± 3.5	1.0 ± 0.0	96.5 ± 0.4	0.798 ± 0.027
50%	RUNG-MCP	1824	31.5 ± 1.5	28.3 ± 1.5	20.0 ± 3.6	47.6 ± 14.8	0.254 ± 0.061
50%	HARA-Mean	1824	33.2 ± 0.3	31.7 ± 3.1	29.3 ± 3.1	13.4 ± 5.6	0.109 ± 0.042
50%	GCN+Global	1824	26.8 ± 1.0	22.7 ± 3.8	16.0 ± 2.6	38.8 ± 5.9	0.210 ± 0.021
50%	GCN+CLASP	1824	26.5 ± 0.8	26.3 ± 1.2	19.0 ± 2.0	36.5 ± 8.2	0.168 ± 0.020
50%	HARA	1824	34.3 ± 0.4	31.0 ± 2.6	32.0 ± 2.6	7.5 ± 3.7	0.066 ± 0.035
25%	GCN	910	28.2 ± 1.0	29.7 ± 7.4	3.0 ± 1.0	89.1 ± 5.0	0.709 ± 0.077
25%	RUNG-MCP	910	30.4 ± 0.4	28.3 ± 3.1	21.0 ± 4.4	38.7 ± 2.7	0.201 ± 0.036
25%	HARA-Mean	910	31.4 ± 1.6	27.7 ± 5.1	26.3 ± 6.1	28.1 ± 16.1	0.182 ± 0.122
25%	GCN+Global	910	25.1 ± 1.4	26.7 ± 3.2	24.3 ± 0.6	29.8 ± 5.2	0.183 ± 0.031
25%	GCN+CLASP	910	25.1 ± 1.8	25.7 ± 0.6	22.3 ± 2.3	27.1 ± 13.6	0.152 ± 0.081
25%	HARA	910	31.8 ± 0.9	30.0 ± 1.7	27.7 ± 7.1	29.9 ± 26.7	0.229 ± 0.200
10%	GCN	363	27.9 ± 1.2	25.7 ± 3.2	9.3 ± 4.7	69.7 ± 10.9	0.503 ± 0.128
10%	RUNG-MCP	363	27.0 ± 2.2	25.3 ± 2.1	20.7 ± 9.3	24.0 ± 29.2	0.187 ± 0.215
10%	HARA-Mean	363	30.0 ± 1.5	27.3 ± 5.7	29.7 ± 9.5	31.3 ± 5.9	0.270 ± 0.093
10%	GCN+Global	363	25.9 ± 0.9	29.7 ± 3.1	18.7 ± 5.1	42.6 ± 19.3	0.212 ± 0.046
10%	GCN+CLASP	363	25.8 ± 0.6	26.3 ± 1.5	19.0 ± 6.6	35.8 ± 16.5	0.178 ± 0.128
10%	HARA	363	29.5 ± 2.3	29.7 ± 7.5	22.7 ± 9.2	39.2 ± 18.2	0.267 ± 0.119

Table 34: Actor: reduced training labels at the two smaller shared Nettack budgets. Attacked accuracy uses all 100 targets; ASR uses each fitted model's initially correct targets. Values are percentages, mean and sample standard deviation across three joint mask/training replicates.

Train	Method	Acc. (1)	ASR (1)	Acc. (3)	ASR (3)
100%	GCN	7.0 ± 0.0	78.1 ± 1.4	1.0 ± 0.0	96.9 ± 0.2
100%	RUNG-MCP	26.0 ± 1.0	19.0 ± 5.0	24.7 ± 3.2	31.6 ± 9.8
100%	HARA-Mean	32.7 ± 0.6	11.3 ± 5.7	32.3 ± 1.2	12.3 ± 4.4
100%	GCN+Global	24.3 ± 3.2	6.9 ± 1.8	22.0 ± 1.0	21.1 ± 8.7
100%	GCN+CLASP	27.0 ± 1.0	8.9 ± 1.7	23.0 ± 1.7	23.3 ± 4.7
100%	HARA	31.7 ± 3.1	13.1 ± 4.9	32.3 ± 2.1	16.0 ± 6.7
50%	GCN	6.3 ± 1.2	77.7 ± 6.2	2.0 ± 1.0	93.0 ± 3.8
50%	RUNG-MCP	25.7 ± 2.1	11.9 ± 5.6	23.0 ± 4.6	30.7 ± 9.7
50%	HARA-Mean	29.7 ± 2.5	12.3 ± 5.1	31.0 ± 3.6	11.4 ± 3.9
50%	GCN+Global	22.0 ± 2.6	7.0 ± 3.6	17.7 ± 2.3	27.8 ± 3.5
50%	GCN+CLASP	27.0 ± 1.0	6.3 ± 4.2	24.7 ± 1.5	19.0 ± 3.8
50%	HARA	29.3 ± 1.5	7.3 ± 4.2	31.3 ± 1.5	7.3 ± 4.2
25%	GCN	10.0 ± 2.0	64.3 ± 12.8	4.3 ± 1.5	84.4 ± 7.2
25%	RUNG-MCP	26.3 ± 3.1	10.4 ± 2.5	25.0 ± 2.6	23.1 ± 6.2
25%	HARA-Mean	25.7 ± 5.1	15.3 ± 6.8	27.3 ± 7.4	20.1 ± 18.2
25%	GCN+Global	26.0 ± 4.0	11.4 ± 1.5	24.3 ± 1.2	22.2 ± 5.7
25%	GCN+CLASP	25.3 ± 1.5	10.3 ± 7.9	23.0 ± 1.7	16.8 ± 8.8
25%	HARA	28.7 ± 4.0	19.4 ± 14.9	27.7 ± 6.8	27.6 ± 25.7
10%	GCN	15.0 ± 5.3	46.4 ± 11.8	14.0 ± 7.9	56.0 ± 18.4
10%	RUNG-MCP	21.7 ± 5.5	18.2 ± 19.7	21.0 ± 7.2	21.1 ± 24.4
10%	HARA-Mean	25.0 ± 8.9	29.3 ± 9.8	27.0 ± 7.5	29.8 ± 13.7
10%	GCN+Global	28.0 ± 3.0	12.3 ± 4.7	25.0 ± 2.0	22.4 ± 1.8
10%	GCN+CLASP	26.0 ± 4.0	6.3 ± 8.1	22.7 ± 7.1	21.8 ± 18.3
10%	HARA	23.7 ± 5.9	23.2 ± 11.7	23.7 ± 8.4	29.7 ± 14.6

Table 35: Roman-Empire: reduced training labels with the original validation and test masks. Clean accuracy uses all test nodes; attacked accuracy and ASR use the fixed 100 targets at five edge flips. Values are mean and sample standard deviation over three paired mask/training replicates.

Train	Method	Labels	Clean test	Clean target	Attacked	ASR	RAUC
100%	GCN	10877	46.4 ± 0.2	48.7 ± 1.2	0.3 ± 0.6	99.3 ± 1.2	0.742 ± 0.025
100%	RUNG-MCP	10877	57.7 ± 0.4	63.0 ± 1.0	60.0 ± 0.0	7.4 ± 1.7	0.018 ± 0.010
100%	HARA-Mean	10877	77.0 ± 0.4	77.7 ± 2.1	47.0 ± 9.0	43.8 ± 12.1	0.274 ± 0.040
100%	GCN+Global	10877	26.1 ± 0.2	28.3 ± 3.8	13.7 ± 1.2	54.9 ± 5.8	0.286 ± 0.026
100%	GCN+CLASP	10877	27.4 ± 0.2	25.7 ± 1.2	12.7 ± 1.5	67.6 ± 4.1	0.393 ± 0.021
100%	HARA	10877	70.2 ± 0.3	70.7 ± 1.5	50.3 ± 6.1	31.5 ± 10.2	0.185 ± 0.054
50%	GCN	5435	45.3 ± 0.3	46.7 ± 2.1	0.0 ± 0.0	100.0 ± 0.0	0.743 ± 0.026
50%	RUNG-MCP	5435	57.3 ± 0.2	61.7 ± 1.5	60.0 ± 2.0	5.9 ± 1.8	0.014 ± 0.004
50%	HARA-Mean	5435	73.6 ± 0.5	73.3 ± 1.5	47.0 ± 5.2	39.6 ± 5.5	0.267 ± 0.037
50%	GCN+Global	5435	23.4 ± 0.3	29.3 ± 4.0	12.7 ± 4.5	62.7 ± 13.1	0.335 ± 0.030
50%	GCN+CLASP	5435	25.6 ± 1.2	27.7 ± 1.2	13.3 ± 1.5	66.1 ± 5.5	0.372 ± 0.042
50%	HARA	5435	67.0 ± 0.4	68.7 ± 3.1	51.7 ± 4.0	28.1 ± 5.3	0.166 ± 0.031
25%	GCN	2714	42.3 ± 0.4	44.0 ± 1.7	0.3 ± 0.6	99.2 ± 1.4	0.728 ± 0.035
25%	RUNG-MCP	2714	56.7 ± 0.7	59.7 ± 0.6	57.7 ± 2.5	6.7 ± 4.4	0.022 ± 0.016
25%	HARA-Mean	2714	69.2 ± 1.0	73.0 ± 3.0	47.0 ± 1.0	38.3 ± 2.5	0.272 ± 0.023
25%	GCN+Global	2714	22.5 ± 0.4	24.3 ± 2.3	17.7 ± 5.5	50.0 ± 12.1	0.240 ± 0.070
25%	GCN+CLASP	2714	22.4 ± 1.0	25.0 ± 4.4	15.3 ± 2.5	60.4 ± 3.5	0.313 ± 0.020
25%	HARA	2714	62.1 ± 1.7	62.7 ± 5.7	52.0 ± 0.0	19.3 ± 5.9	0.113 ± 0.052
10%	GCN	1079	36.9 ± 0.3	35.0 ± 2.6	1.0 ± 0.0	97.1 ± 0.2	0.744 ± 0.043
10%	RUNG-MCP	1079	55.9 ± 0.2	58.3 ± 3.5	58.7 ± 2.9	3.9 ± 2.4	0.024 ± 0.010
10%	HARA-Mean	1079	62.3 ± 1.5	63.7 ± 2.1	50.0 ± 2.6	26.8 ± 3.5	0.180 ± 0.042
10%	GCN+Global	1079	21.0 ± 0.4	23.3 ± 1.5	17.3 ± 2.9	47.0 ± 5.7	0.214 ± 0.025
10%	GCN+CLASP	1079	20.6 ± 1.1	22.7 ± 0.6	14.3 ± 2.1	59.0 ± 8.3	0.352 ± 0.068
10%	HARA	1079	56.5 ± 0.4	60.0 ± 2.0	53.7 ± 0.6	12.7 ± 3.6	0.087 ± 0.011

Table 36: Roman-Empire: reduced training labels at the two smaller shared Nettack budgets. Attacked accuracy uses all 100 targets; ASR uses each fitted model's initially correct targets. Values are percentages, mean and sample standard deviation across three joint mask/training replicates.

Train	Method	Acc. (1)	ASR (1)	Acc. (3)	ASR (3)
100%	GCN	26.7 ± 3.2	56.1 ± 5.2	4.3 ± 1.5	93.8 ± 2.2
100%	RUNG-MCP	63.7 ± 0.6	0.5 ± 0.9	65.0 ± 0.0	0.5 ± 0.9
100%	HARA-Mean	68.7 ± 2.1	14.2 ± 1.6	53.0 ± 2.6	36.1 ± 3.5
100%	GCN+Global	28.3 ± 5.0	8.9 ± 7.0	20.7 ± 2.1	37.4 ± 8.0
100%	GCN+CLASP	28.0 ± 1.7	18.3 ± 3.0	20.7 ± 1.2	50.7 ± 2.2
100%	HARA	66.0 ± 2.6	8.0 ± 4.2	55.0 ± 3.0	24.4 ± 6.6
50%	GCN	26.7 ± 1.2	55.2 ± 5.4	4.3 ± 1.5	94.4 ± 3.1
50%	RUNG-MCP	62.0 ± 1.0	0.0 ± 0.0	64.7 ± 1.5	0.6 ± 1.0
50%	HARA-Mean	62.3 ± 3.8	17.8 ± 3.9	51.3 ± 4.0	33.7 ± 4.5
50%	GCN+Global	27.7 ± 2.5	12.7 ± 3.0	19.0 ± 1.0	42.8 ± 4.9
50%	GCN+CLASP	26.7 ± 1.5	10.9 ± 3.7	17.7 ± 3.2	51.8 ± 7.4
50%	HARA	65.7 ± 2.1	6.3 ± 1.5	55.7 ± 3.5	22.8 ± 5.7
25%	GCN	22.0 ± 0.0	55.3 ± 2.3	5.0 ± 2.6	91.0 ± 7.6
25%	RUNG-MCP	60.7 ± 1.2	0.6 ± 1.0	61.0 ± 1.7	1.7 ± 1.7
25%	HARA-Mean	62.7 ± 0.6	17.8 ± 1.7	49.7 ± 0.6	35.5 ± 3.5
25%	GCN+Global	26.7 ± 3.2	5.6 ± 2.7	21.0 ± 3.0	30.8 ± 10.9
25%	GCN+CLASP	28.7 ± 3.8	9.1 ± 8.7	21.0 ± 1.0	41.1 ± 2.1
25%	HARA	61.7 ± 2.9	4.1 ± 2.9	55.0 ± 1.0	15.5 ± 8.0
10%	GCN	15.3 ± 6.7	65.1 ± 14.9	6.0 ± 1.0	88.5 ± 0.9
10%	RUNG-MCP	57.7 ± 2.1	1.6 ± 1.6	59.0 ± 3.6	2.9 ± 1.0
10%	HARA-Mean	58.3 ± 5.9	12.7 ± 4.5	52.7 ± 4.5	22.1 ± 6.0
10%	GCN+Global	23.7 ± 2.5	5.8 ± 2.8	20.0 ± 3.0	25.7 ± 7.5
10%	GCN+CLASP	22.7 ± 0.6	19.2 ± 5.3	18.0 ± 4.4	44.1 ± 13.1
10%	HARA	58.3 ± 4.0	5.6 ± 2.7	55.0 ± 0.0	11.1 ± 2.4

Table 37: Amazon-Ratings: reduced training labels with the original validation and test masks. Clean accuracy uses all test nodes; attacked accuracy and ASR use the fixed 100 targets at five edge flips. Values are mean and sample standard deviation over three paired mask/training replicates.

Train	Method	Labels	Clean test	Clean target	Attacked	ASR	RAUC
100%	GCN	11756	43.0 ± 0.2	42.3 ± 0.6	9.3 ± 0.6	77.9 ± 1.5	0.576 ± 0.019
100%	RUNG-MCP	11756	41.9 ± 0.3	41.7 ± 0.6	35.0 ± 1.0	16.8 ± 2.2	0.058 ± 0.010
100%	HARA-Mean	11756	42.9 ± 0.6	39.3 ± 2.5	18.7 ± 5.8	62.0 ± 8.8	0.455 ± 0.042
100%	GCN+Global	11756	43.6 ± 0.3	39.7 ± 4.0	39.3 ± 3.2	3.3 ± 4.0	0.018 ± 0.017
100%	GCN+CLASP	11756	43.1 ± 0.5	39.3 ± 3.5	38.3 ± 2.3	7.6 ± 2.4	0.027 ± 0.011
100%	HARA	11756	39.5 ± 2.1	38.0 ± 2.6	37.0 ± 3.0	3.5 ± 3.0	0.023 ± 0.021
50%	GCN	5876	42.4 ± 0.1	41.3 ± 1.5	3.0 ± 0.0	92.7 ± 0.3	0.675 ± 0.019
50%	RUNG-MCP	5876	41.5 ± 0.5	39.0 ± 0.0	32.3 ± 3.1	17.1 ± 7.8	0.059 ± 0.025
50%	HARA-Mean	5876	42.0 ± 0.3	38.3 ± 1.2	14.3 ± 7.0	66.0 ± 16.4	0.473 ± 0.117
50%	GCN+Global	5876	41.2 ± 1.0	35.3 ± 2.3	35.3 ± 0.6	6.6 ± 1.2	0.038 ± 0.010
50%	GCN+CLASP	5876	42.0 ± 0.5	38.3 ± 4.5	35.7 ± 6.1	8.2 ± 5.6	0.051 ± 0.024
50%	HARA	5876	40.7 ± 0.5	39.3 ± 2.5	37.3 ± 2.1	11.4 ± 10.8	0.089 ± 0.077
25%	GCN	2937	41.3 ± 0.3	41.0 ± 2.6	4.3 ± 1.5	89.4 ± 3.7	0.630 ± 0.059
25%	RUNG-MCP	2937	41.4 ± 0.1	41.7 ± 1.5	32.7 ± 3.1	21.6 ± 6.3	0.097 ± 0.033
25%	HARA-Mean	2937	41.4 ± 0.4	40.0 ± 1.0	19.0 ± 5.2	59.0 ± 14.7	0.401 ± 0.069
25%	GCN+Global	2937	39.3 ± 1.5	38.3 ± 3.2	36.7 ± 1.5	7.5 ± 6.0	0.039 ± 0.028
25%	GCN+CLASP	2937	39.9 ± 1.0	37.3 ± 0.6	37.7 ± 1.5	3.6 ± 1.5	0.023 ± 0.008
25%	HARA	2937	39.4 ± 2.1	37.0 ± 3.0	37.7 ± 2.1	13.4 ± 11.8	0.099 ± 0.088
10%	GCN	1172	39.3 ± 0.5	36.0 ± 4.6	15.0 ± 5.2	59.6 ± 8.9	0.408 ± 0.021
10%	RUNG-MCP	1172	39.4 ± 0.7	35.3 ± 1.5	31.7 ± 0.6	12.2 ± 3.7	0.056 ± 0.008
10%	HARA-Mean	1172	38.3 ± 2.1	38.0 ± 1.7	32.3 ± 8.1	15.8 ± 27.4	0.113 ± 0.196
10%	GCN+Global	1172	38.4 ± 0.7	36.7 ± 1.5	37.3 ± 1.2	1.8 ± 3.1	0.010 ± 0.017
10%	GCN+CLASP	1172	39.2 ± 0.0	38.7 ± 3.1	38.0 ± 2.0	3.3 ± 2.9	0.010 ± 0.010
10%	HARA	1172	38.6 ± 1.4	38.0 ± 1.0	36.0 ± 1.0	8.6 ± 9.0	0.074 ± 0.074

Table 38: Amazon-Ratings: reduced training labels at the two smaller shared Nettack budgets. Attacked accuracy uses all 100 targets; ASR uses each fitted model's initially correct targets. Values are percentages, mean and sample standard deviation across three joint mask/training replicates.

Train	Method	Acc. (1)	ASR (1)	Acc. (3)	ASR (3)
100%	GCN	24.3 ± 1.2	42.5 ± 2.4	11.3 ± 1.5	73.3 ± 3.2
100%	RUNG-MCP	41.0 ± 1.0	1.6 ± 1.4	40.0 ± 1.0	4.8 ± 2.4
100%	HARA-Mean	27.7 ± 3.2	35.5 ± 5.7	19.7 ± 2.5	56.0 ± 3.4
100%	GCN+Global	39.3 ± 4.2	1.6 ± 1.4	39.7 ± 3.1	1.6 ± 1.4
100%	GCN+CLASP	40.3 ± 2.1	0.8 ± 1.3	40.7 ± 2.5	2.5 ± 2.3
100%	HARA	38.0 ± 3.6	1.7 ± 1.5	38.0 ± 3.6	2.7 ± 2.8
50%	GCN	21.7 ± 1.2	48.4 ± 3.9	5.7 ± 2.1	86.2 ± 5.5
50%	RUNG-MCP	38.0 ± 1.0	2.6 ± 2.6	37.3 ± 0.6	4.3 ± 1.5
50%	HARA-Mean	29.0 ± 1.0	33.8 ± 8.7	17.0 ± 5.6	59.9 ± 14.8
50%	GCN+Global	35.3 ± 2.3	2.0 ± 1.7	35.7 ± 2.1	4.8 ± 1.9
50%	GCN+CLASP	38.7 ± 3.8	1.8 ± 3.0	36.0 ± 5.6	7.2 ± 3.9
50%	HARA	42.3 ± 3.5	3.3 ± 2.9	38.3 ± 1.2	14.0 ± 12.1
25%	GCN	22.7 ± 3.5	44.2 ± 12.2	8.3 ± 2.5	79.6 ± 6.1
25%	RUNG-MCP	39.7 ± 2.1	4.8 ± 2.4	37.7 ± 3.5	9.7 ± 5.2
25%	HARA-Mean	34.0 ± 2.0	26.7 ± 1.6	22.3 ± 2.9	50.7 ± 9.4
25%	GCN+Global	37.3 ± 2.5	2.5 ± 2.4	38.0 ± 2.0	4.2 ± 2.5
25%	GCN+CLASP	37.3 ± 1.5	1.8 ± 3.1	37.3 ± 1.5	2.7 ± 0.0
25%	HARA	38.7 ± 1.5	9.8 ± 8.9	38.0 ± 3.6	10.7 ± 9.4
10%	GCN	25.0 ± 1.7	29.9 ± 9.1	18.3 ± 2.5	49.9 ± 1.5
10%	RUNG-MCP	34.7 ± 2.1	1.9 ± 1.7	33.0 ± 1.0	6.6 ± 1.3
10%	HARA-Mean	35.7 ± 2.3	8.3 ± 14.4	33.3 ± 6.4	14.2 ± 24.5
10%	GCN+Global	37.0 ± 1.0	0.9 ± 1.6	37.7 ± 1.5	0.9 ± 1.6
10%	GCN+CLASP	39.0 ± 2.6	0.0 ± 0.0	39.3 ± 1.5	0.8 ± 1.4
10%	HARA	37.3 ± 1.5	5.2 ± 5.1	35.0 ± 2.0	10.3 ± 10.3

Table 39: Actor: all shared Nettack budgets with exactly five training labels per class. Original validation labels remain available. Clean/attacked accuracy use the fixed 100 targets; ASR uses the actual initially correct denominator. Values are percentages, mean and sample standard deviation across three joint mask/training replicates.

Method	Budget	Clean target	Attacked	ASR
GCN	1	18.0 ± 6.6	17.3 ± 6.5	10.2 ± 14.1
GCN	3	18.0 ± 6.6	18.0 ± 6.1	11.5 ± 13.4
GCN	5	18.0 ± 6.6	17.0 ± 6.0	18.5 ± 16.1
RUNG-MCP	1	20.7 ± 3.5	20.7 ± 3.5	0.0 ± 0.0
RUNG-MCP	3	20.7 ± 3.5	20.7 ± 3.5	0.0 ± 0.0
RUNG-MCP	5	20.7 ± 3.5	20.7 ± 3.5	0.0 ± 0.0
HARA-Mean	1	21.0 ± 0.0	21.7 ± 0.6	15.9 ± 7.3
HARA-Mean	3	21.0 ± 0.0	20.0 ± 1.7	31.7 ± 7.3
HARA-Mean	5	21.0 ± 0.0	23.0 ± 1.0	33.3 ± 9.5
GCN+Global	1	21.3 ± 3.1	20.7 ± 1.5	5.7 ± 6.3
GCN+Global	3	21.3 ± 3.1	22.3 ± 2.3	4.3 ± 4.2
GCN+Global	5	21.3 ± 3.1	20.3 ± 5.0	13.1 ± 8.9
GCN+CLASP	1	21.0 ± 3.0	22.0 ± 3.5	3.0 ± 2.6
GCN+CLASP	3	21.0 ± 3.0	23.7 ± 4.0	3.0 ± 2.6
GCN+CLASP	5	21.0 ± 3.0	21.0 ± 4.0	11.3 ± 6.0
HARA	1	19.3 ± 3.2	22.7 ± 4.2	25.8 ± 16.2
HARA	3	19.3 ± 3.2	19.3 ± 1.2	42.8 ± 19.8
HARA	5	19.3 ± 3.2	19.3 ± 2.5	46.1 ± 20.4

Table 40: Roman-Empire: all shared Nettack budgets with exactly five training labels per class. Original validation labels remain available. Clean/attacked accuracy use the fixed 100 targets; ASR uses the actual initially correct denominator. Values are percentages, mean and sample standard deviation across three joint mask/training replicates.

Method	Budget	Clean target	Attacked	ASR
GCN	1	21.0 ± 2.6	14.0 ± 1.7	39.5 ± 2.5
GCN	3	21.0 ± 2.6	11.0 ± 3.5	58.0 ± 9.8
GCN	5	21.0 ± 2.6	7.7 ± 4.0	76.1 ± 8.3
RUNG-MCP	1	32.7 ± 3.8	32.3 ± 3.2	1.8 ± 3.1
RUNG-MCP	3	32.7 ± 3.8	30.0 ± 4.6	10.1 ± 9.7
RUNG-MCP	5	32.7 ± 3.8	28.0 ± 5.6	18.4 ± 12.9
HARA-Mean	1	26.3 ± 2.5	19.7 ± 3.5	34.0 ± 9.6
HARA-Mean	3	26.3 ± 2.5	19.3 ± 5.0	44.5 ± 8.3
HARA-Mean	5	26.3 ± 2.5	16.0 ± 7.9	54.6 ± 9.4
GCN+Global	1	11.0 ± 2.6	11.3 ± 3.1	8.5 ± 7.5
GCN+Global	3	11.0 ± 2.6	10.3 ± 4.2	30.3 ± 9.0
GCN+Global	5	11.0 ± 2.6	9.7 ± 5.7	57.8 ± 21.1
GCN+CLASP	1	10.7 ± 3.8	10.7 ± 3.1	17.8 ± 16.8
GCN+CLASP	3	10.7 ± 3.8	10.0 ± 5.0	29.5 ± 16.0
GCN+CLASP	5	10.7 ± 3.8	8.7 ± 4.5	41.1 ± 8.4
HARA	1	25.3 ± 3.8	28.3 ± 2.3	10.9 ± 9.7
HARA	3	25.3 ± 3.8	24.3 ± 4.5	27.3 ± 12.1
HARA	5	25.3 ± 3.8	22.7 ± 3.2	32.9 ± 13.4

Table 41: Amazon-Ratings: all shared Nettack budgets with exactly five training labels per class. Original validation labels remain available. Clean/attacked accuracy use the fixed 100 targets; ASR uses the actual initially correct denominator. Values are percentages, mean and sample standard deviation across three joint mask/training replicates.

Method	Budget	Clean target	Attacked	ASR
GCN	1	21.3 ± 5.7	22.0 ± 5.3	7.5 ± 3.7
GCN	3	21.3 ± 5.7	26.0 ± 3.0	24.3 ± 13.7
GCN	5	21.3 ± 5.7	30.0 ± 4.6	26.0 ± 6.3
RUNG-MCP	1	29.3 ± 6.7	29.7 ± 6.4	0.9 ± 1.6
RUNG-MCP	3	29.3 ± 6.7	29.3 ± 5.8	0.9 ± 1.6
RUNG-MCP	5	29.3 ± 6.7	30.0 ± 5.3	1.8 ± 3.1
HARA-Mean	1	32.7 ± 4.5	35.0 ± 5.3	20.2 ± 7.5
HARA-Mean	3	32.7 ± 4.5	32.3 ± 6.4	26.1 ± 15.9
HARA-Mean	5	32.7 ± 4.5	33.0 ± 7.8	25.6 ± 17.0
GCN+Global	1	29.7 ± 4.0	30.3 ± 4.0	0.0 ± 0.0
GCN+Global	3	29.7 ± 4.0	31.0 ± 5.3	1.3 ± 2.2
GCN+Global	5	29.7 ± 4.0	30.3 ± 5.5	6.0 ± 5.4
GCN+CLASP	1	29.3 ± 5.1	29.3 ± 4.5	3.1 ± 2.9
GCN+CLASP	3	29.3 ± 5.1	28.7 ± 2.1	7.1 ± 7.1
GCN+CLASP	5	29.3 ± 5.1	27.0 ± 1.0	10.2 ± 10.0
HARA	1	30.3 ± 2.5	34.7 ± 1.5	18.4 ± 6.8
HARA	3	30.3 ± 2.5	35.3 ± 3.1	12.2 ± 2.8
HARA	5	30.3 ± 2.5	36.3 ± 2.1	15.4 ± 6.9

Table 42: CiteSeer: model specific local structural evasion, 100 fixed targets. Values are percentages, mean and sample standard deviation across three trained models.

Method	Budget	Clean	Attacked	ASR
GCN	1	73.7 ± 2.5	26.3 ± 3.8	64.3 ± 4.1
GCN	3	73.7 ± 2.5	5.7 ± 2.1	92.3 ± 2.8
GCN	5	73.7 ± 2.5	3.7 ± 1.5	95.0 ± 2.2
RUNG-MCP	1	71.0 ± 1.0	63.3 ± 2.3	10.8 ± 2.2
RUNG-MCP	3	71.0 ± 1.0	39.0 ± 1.0	45.1 ± 1.2
RUNG-MCP	5	71.0 ± 1.0	26.7 ± 1.2	62.4 ± 1.7
HARA-Mean	1	63.7 ± 3.1	12.7 ± 2.1	80.2 ± 2.3
HARA-Mean	3	63.7 ± 3.1	4.3 ± 0.6	93.2 ± 1.2
HARA-Mean	5	63.7 ± 3.1	2.3 ± 1.5	96.3 ± 2.6
GCN+Global	1	59.0 ± 2.0	32.3 ± 2.3	45.2 ± 2.4
GCN+Global	3	59.0 ± 2.0	17.3 ± 3.5	70.7 ± 5.4
GCN+Global	5	59.0 ± 2.0	14.0 ± 1.7	76.3 ± 2.2
GCN+CLASP	1	58.7 ± 3.8	31.3 ± 1.5	46.5 ± 1.7
GCN+CLASP	3	58.7 ± 3.8	16.3 ± 4.0	72.3 ± 5.8
GCN+CLASP	5	58.7 ± 3.8	11.7 ± 2.1	80.2 ± 2.3
HARA	1	68.7 ± 3.8	43.0 ± 5.3	37.5 ± 5.8
HARA	3	68.7 ± 3.8	25.3 ± 3.2	63.2 ± 2.6
HARA	5	68.7 ± 3.8	18.7 ± 4.2	72.9 ± 5.4

Table 43: Squirrel-F: model specific local structural evasion, 100 fixed targets. Values are percentages, mean and sample standard deviation across three trained models.

Method	Budget	Clean	Attacked	ASR
GCN	1	32.3 ± 2.1	8.0 ± 0.0	75.2 ± 1.6
GCN	3	32.3 ± 2.1	5.0 ± 0.0	84.5 ± 1.0
GCN	5	32.3 ± 2.1	5.0 ± 0.0	84.5 ± 1.0
RUNG-MCP	1	29.0 ± 1.0	27.0 ± 0.0	6.8 ± 3.2
RUNG-MCP	3	29.0 ± 1.0	24.0 ± 1.0	17.1 ± 5.6
RUNG-MCP	5	29.0 ± 1.0	21.7 ± 1.2	25.1 ± 6.4
HARA-Mean	1	35.7 ± 4.0	14.3 ± 5.1	59.1 ± 16.1
HARA-Mean	3	35.7 ± 4.0	5.3 ± 2.3	84.9 ± 6.8
HARA-Mean	5	35.7 ± 4.0	3.0 ± 0.0	91.5 ± 0.9
GCN+Global	1	31.7 ± 0.6	24.7 ± 4.7	22.2 ± 14.0
GCN+Global	3	31.7 ± 0.6	21.3 ± 7.6	32.7 ± 23.3
GCN+Global	5	31.7 ± 0.6	20.7 ± 8.3	34.8 ± 25.7
GCN+CLASP	1	32.7 ± 1.5	23.3 ± 1.5	28.5 ± 4.2
GCN+CLASP	3	32.7 ± 1.5	19.0 ± 1.7	41.6 ± 8.1
GCN+CLASP	5	32.7 ± 1.5	18.0 ± 2.6	44.6 ± 10.9
HARA	1	34.3 ± 2.1	23.3 ± 4.2	31.9 ± 12.2
HARA	3	34.3 ± 2.1	16.0 ± 6.1	52.7 ± 19.7
HARA	5	34.3 ± 2.1	13.7 ± 6.1	59.4 ± 20.1

Table 44: Actor: model specific local structural evasion, 100 fixed targets. Values are percentages, mean and sample standard deviation across three trained models.

Method	Budget	Clean	Attacked	ASR
GCN	1	32.0 ± 2.0	6.3 ± 0.6	80.1 ± 3.0
GCN	3	32.0 ± 2.0	2.0 ± 0.0	93.7 ± 0.4
GCN	5	32.0 ± 2.0	2.0 ± 0.0	93.7 ± 0.4
RUNG-MCP	1	31.3 ± 1.5	9.7 ± 1.5	68.9 ± 6.3
RUNG-MCP	3	31.3 ± 1.5	5.3 ± 0.6	82.9 ± 2.1
RUNG-MCP	5	31.3 ± 1.5	4.0 ± 1.0	87.1 ± 3.8
HARA-Mean	1	35.3 ± 0.6	28.7 ± 0.6	18.9 ± 1.5
HARA-Mean	3	35.3 ± 0.6	26.3 ± 0.6	25.4 ± 2.5
HARA-Mean	5	35.3 ± 0.6	25.3 ± 1.5	28.3 ± 4.7
GCN+Global	1	24.0 ± 4.6	7.3 ± 1.2	69.2 ± 3.7
GCN+Global	3	24.0 ± 4.6	3.0 ± 1.0	87.5 ± 3.3
GCN+Global	5	24.0 ± 4.6	1.7 ± 0.6	92.9 ± 2.8
GCN+CLASP	1	26.0 ± 2.6	10.3 ± 3.2	60.4 ± 10.2
GCN+CLASP	3	26.0 ± 2.6	3.0 ± 1.0	88.6 ± 3.1
GCN+CLASP	5	26.0 ± 2.6	2.3 ± 1.2	91.3 ± 3.8
HARA	1	36.0 ± 2.0	30.0 ± 3.6	16.9 ± 5.5
HARA	3	36.0 ± 2.0	27.0 ± 5.3	25.4 ± 10.6
HARA	5	36.0 ± 2.0	26.7 ± 4.7	26.2 ± 9.1

Table 45: Cora: global PR-BCD evasion on all test nodes. At each budget, one feasible graph with the lowest test accuracy is selected from that and smaller budgets, including the clean graph. Values are percentages, mean and sample standard deviation across three trained models.

Method	Budget	Clean	Attacked	ASR
GCN	1%	80.5 ± 0.8	75.3 ± 0.3	6.5 ± 0.5
GCN	3%	80.5 ± 0.8	67.2 ± 0.4	16.7 ± 0.6
GCN	5%	80.5 ± 0.8	60.8 ± 0.2	24.6 ± 0.7
RUNG-MCP	1%	81.9 ± 1.4	79.4 ± 1.3	3.1 ± 0.1
RUNG-MCP	3%	81.9 ± 1.4	75.8 ± 1.0	7.4 ± 0.7
RUNG-MCP	5%	81.9 ± 1.4	73.1 ± 1.2	10.8 ± 0.5
HARA-Mean	1%	78.2 ± 0.9	76.0 ± 1.0	2.9 ± 0.4
HARA-Mean	3%	78.2 ± 0.9	61.7 ± 1.0	21.2 ± 0.4
HARA-Mean	5%	78.2 ± 0.9	53.2 ± 0.8	32.0 ± 0.3
GCN+Global	1%	68.9 ± 2.6	68.1 ± 2.1	1.4 ± 0.8
GCN+Global	3%	68.9 ± 2.6	63.8 ± 2.8	8.5 ± 0.5
GCN+Global	5%	68.9 ± 2.6	61.0 ± 2.1	12.4 ± 0.4
GCN+CLASP	1%	68.8 ± 2.8	67.9 ± 2.9	2.1 ± 0.2
GCN+CLASP	3%	68.8 ± 2.8	63.6 ± 2.1	9.6 ± 0.5
GCN+CLASP	5%	68.8 ± 2.8	61.5 ± 2.1	12.6 ± 1.6
HARA	1%	76.6 ± 1.8	74.7 ± 1.2	3.4 ± 0.8
HARA	3%	76.6 ± 1.8	73.6 ± 1.6	5.4 ± 0.1
HARA	5%	76.6 ± 1.8	68.1 ± 1.9	12.8 ± 0.7

Table 46: CiteSeer: global PR-BCD evasion on all test nodes. At each budget, one feasible graph with the lowest test accuracy is selected from that and smaller budgets, including the clean graph. Values are percentages, mean and sample standard deviation across three trained models.

Method	Budget	Clean	Attacked	ASR
GCN	1%	68.7 ± 0.8	63.6 ± 0.7	7.4 ± 0.9
GCN	3%	68.7 ± 0.8	56.3 ± 1.4	18.1 ± 2.1
GCN	5%	68.7 ± 0.8	49.8 ± 1.7	27.6 ± 2.4
RUNG-MCP	1%	70.0 ± 0.4	67.8 ± 0.3	3.1 ± 0.1
RUNG-MCP	3%	70.0 ± 0.4	64.7 ± 0.1	7.7 ± 0.5
RUNG-MCP	5%	70.0 ± 0.4	62.6 ± 0.4	10.7 ± 1.1
HARA-Mean	1%	60.2 ± 2.2	57.6 ± 2.2	4.4 ± 0.4
HARA-Mean	3%	60.2 ± 2.2	55.1 ± 3.0	9.3 ± 1.6
HARA-Mean	5%	60.2 ± 2.2	45.6 ± 1.6	24.9 ± 0.5
GCN+Global	1%	53.5 ± 2.1	53.3 ± 2.2	1.0 ± 0.1
GCN+Global	3%	53.5 ± 2.1	52.4 ± 2.3	4.3 ± 0.2
GCN+Global	5%	53.5 ± 2.1	49.6 ± 2.3	9.6 ± 1.6
GCN+CLASP	1%	52.7 ± 1.3	52.0 ± 1.1	2.0 ± 0.7
GCN+CLASP	3%	52.7 ± 1.3	51.1 ± 1.6	5.8 ± 0.7
GCN+CLASP	5%	52.7 ± 1.3	48.0 ± 1.1	13.0 ± 0.9
HARA	1%	62.6 ± 3.8	61.3 ± 3.5	2.4 ± 0.1
HARA	3%	62.6 ± 3.8	60.9 ± 3.7	4.8 ± 0.7
HARA	5%	62.6 ± 3.8	60.3 ± 2.8	5.8 ± 0.5

Table 47: Squirrel-F: global PR-BCD evasion on all test nodes. At each budget, one feasible graph with the lowest test accuracy is selected from that and smaller budgets, including the clean graph. Values are percentages, mean and sample standard deviation across three trained models.

Method	Budget	Clean	Attacked	ASR
GCN	1%	33.9 ± 1.4	20.5 ± 2.0	39.7 ± 4.0
GCN	3%	33.9 ± 1.4	14.5 ± 3.6	57.5 ± 9.3
GCN	5%	33.9 ± 1.4	12.1 ± 3.2	64.3 ± 8.6
RUNG-MCP	1%	32.5 ± 0.3	30.8 ± 0.2	5.8 ± 0.4
RUNG-MCP	3%	32.5 ± 0.3	30.6 ± 0.2	6.7 ± 1.0
RUNG-MCP	5%	32.5 ± 0.3	30.3 ± 0.1	7.6 ± 0.6
HARA-Mean	1%	36.8 ± 2.2	16.2 ± 2.6	58.8 ± 9.1
HARA-Mean	3%	36.8 ± 2.2	9.0 ± 0.4	77.2 ± 0.6
HARA-Mean	5%	36.8 ± 2.2	6.0 ± 1.0	85.2 ± 3.2
GCN+Global	1%	31.7 ± 1.0	29.4 ± 2.0	11.1 ± 8.8
GCN+Global	3%	31.7 ± 1.0	27.5 ± 3.6	17.6 ± 13.5
GCN+Global	5%	31.7 ± 1.0	26.6 ± 4.3	20.3 ± 15.2
GCN+CLASP	1%	33.6 ± 0.7	30.0 ± 0.8	15.2 ± 0.4
GCN+CLASP	3%	33.6 ± 0.7	27.4 ± 2.4	22.8 ± 4.5
GCN+CLASP	5%	33.6 ± 0.7	24.8 ± 1.5	31.1 ± 2.2
HARA	1%	39.1 ± 0.2	31.9 ± 2.7	20.9 ± 7.1
HARA	3%	39.1 ± 0.2	26.4 ± 6.2	35.9 ± 13.6
HARA	5%	39.1 ± 0.2	21.3 ± 6.6	47.6 ± 16.4

Table 48: Actor: global PR-BCD evasion on all test nodes. At each budget, one feasible graph with the lowest test accuracy is selected from that and smaller budgets, including the clean graph. Values are percentages, mean and sample standard deviation across three trained models.

Method	Budget	Clean	Attacked	ASR
GCN	1%	30.6 ± 0.1	26.5 ± 0.3	14.6 ± 0.4
GCN	3%	30.6 ± 0.1	20.9 ± 0.4	33.4 ± 1.4
GCN	5%	30.6 ± 0.1	17.3 ± 0.6	44.5 ± 2.0
RUNG-MCP	1%	31.8 ± 0.4	30.0 ± 0.6	6.3 ± 1.3
RUNG-MCP	3%	31.8 ± 0.4	28.6 ± 0.6	13.1 ± 0.8
RUNG-MCP	5%	31.8 ± 0.4	27.6 ± 0.5	17.5 ± 2.1
HARA-Mean	1%	35.5 ± 1.5	34.9 ± 1.3	3.3 ± 2.9
HARA-Mean	3%	35.5 ± 1.5	34.9 ± 1.2	3.5 ± 3.0
HARA-Mean	5%	35.5 ± 1.5	34.5 ± 1.1	5.1 ± 1.4
GCN+Global	1%	26.2 ± 0.8	24.6 ± 1.1	8.5 ± 2.5
GCN+Global	3%	26.2 ± 0.8	21.8 ± 1.4	22.2 ± 4.0
GCN+Global	5%	26.2 ± 0.8	20.4 ± 1.6	29.0 ± 4.6
GCN+CLASP	1%	26.4 ± 0.9	24.1 ± 0.8	12.0 ± 1.2
GCN+CLASP	3%	26.4 ± 0.9	21.4 ± 1.4	24.3 ± 1.8
GCN+CLASP	5%	26.4 ± 0.9	20.3 ± 1.8	30.1 ± 3.0
HARA	1%	35.2 ± 0.6	34.9 ± 0.9	1.5 ± 1.4
HARA	3%	35.2 ± 0.6	34.6 ± 0.8	3.2 ± 1.1
HARA	5%	35.2 ± 0.6	34.6 ± 0.8	3.2 ± 1.1

Table 49: Cora: transferred Metattack structure poisoning. Both denotes training the complete pipeline (including the preliminary model when present) and testing on the poisoned graph. Preliminary denotes fitting only the preliminary classifier on the poisoned graph, with clean downstream training and evaluation. The final column is the paired accuracy change in percentage points relative to clean training; negative values indicate a loss.

Method	Training	Budget	Clean	Poisoned	Change
GCN	Both	1%	80.5 ± 0.8	77.4 ± 0.7	-3.1 ± 1.3
GCN	Both	3%	80.5 ± 0.8	70.8 ± 1.6	-9.7 ± 2.2
GCN	Both	5%	80.5 ± 0.8	68.1 ± 0.9	-12.4 ± 1.5
RUNG-MCP	Both	1%	81.9 ± 1.4	78.9 ± 0.6	-2.9 ± 0.8
RUNG-MCP	Both	3%	81.9 ± 1.4	76.0 ± 1.3	-5.9 ± 0.2
RUNG-MCP	Both	5%	81.9 ± 1.4	74.5 ± 1.4	-7.4 ± 0.5
HARA-Mean	Both	1%	78.2 ± 0.9	76.6 ± 0.6	-1.7 ± 0.5
HARA-Mean	Both	3%	78.2 ± 0.9	73.3 ± 1.0	-4.9 ± 2.0
HARA-Mean	Both	5%	78.2 ± 0.9	70.2 ± 3.2	-8.1 ± 3.5
GCN+Global	Both	1%	68.9 ± 2.6	67.8 ± 2.7	-1.1 ± 2.2
GCN+Global	Both	3%	68.9 ± 2.6	64.2 ± 1.2	-4.7 ± 3.4
GCN+Global	Both	5%	68.9 ± 2.6	64.2 ± 1.9	-4.7 ± 2.9
GCN+CLASP	Both	1%	68.8 ± 2.8	68.3 ± 4.6	-0.5 ± 2.0
GCN+CLASP	Both	3%	68.8 ± 2.8	64.2 ± 0.9	-4.6 ± 2.8
GCN+CLASP	Both	5%	68.8 ± 2.8	64.5 ± 2.0	-4.4 ± 3.8
HARA	Both	1%	76.6 ± 1.8	75.3 ± 0.0	-1.3 ± 1.8
HARA	Both	3%	76.6 ± 1.8	73.8 ± 0.9	-2.8 ± 2.7
HARA	Both	5%	76.6 ± 1.8	71.2 ± 0.5	-5.4 ± 1.6
GCN+CLASP	Preliminary	1%	68.8 ± 2.8	69.0 ± 4.1	0.2 ± 1.5
GCN+CLASP	Preliminary	3%	68.8 ± 2.8	69.4 ± 3.3	0.5 ± 0.5
GCN+CLASP	Preliminary	5%	68.8 ± 2.8	68.5 ± 2.9	-0.4 ± 2.4
HARA	Preliminary	1%	76.6 ± 1.8	76.8 ± 1.7	0.2 ± 1.5
HARA	Preliminary	3%	76.6 ± 1.8	77.1 ± 1.8	0.5 ± 1.5
HARA	Preliminary	5%	76.6 ± 1.8	77.6 ± 1.6	1.0 ± 1.7

Table 50: Chameleon-F: transferred Metattack structure poisoning. Both denotes training the complete pipeline (including the preliminary model when present) and testing on the poisoned graph. Preliminary denotes fitting only the preliminary classifier on the poisoned graph, with clean downstream training and evaluation. The final column is the paired accuracy change in percentage points relative to clean training; negative values indicate a loss.

Method	Training	Budget	Clean	Poisoned	Change
GCN	Both	1%	34.8 ± 0.9	34.1 ± 0.0	-0.7 ± 0.9
GCN	Both	3%	34.8 ± 0.9	35.2 ± 1.5	0.4 ± 1.2
GCN	Both	5%	34.8 ± 0.9	32.8 ± 4.2	-2.0 ± 3.4
RUNG-MCP	Both	1%	30.5 ± 2.3	31.8 ± 2.4	1.3 ± 1.3
RUNG-MCP	Both	3%	30.5 ± 2.3	30.5 ± 1.4	0.0 ± 1.0
RUNG-MCP	Both	5%	30.5 ± 2.3	30.4 ± 1.7	-0.2 ± 0.9
HARA-Mean	Both	1%	38.0 ± 2.8	39.7 ± 0.6	1.7 ± 3.4
HARA-Mean	Both	3%	38.0 ± 2.8	37.1 ± 2.1	-0.9 ± 4.8
HARA-Mean	Both	5%	38.0 ± 2.8	38.0 ± 4.8	0.0 ± 7.4
GCN+Global	Both	1%	32.2 ± 2.3	32.4 ± 1.1	0.2 ± 1.2
GCN+Global	Both	3%	32.2 ± 2.3	31.8 ± 1.7	-0.4 ± 0.6
GCN+Global	Both	5%	32.2 ± 2.3	33.5 ± 0.6	1.3 ± 2.8
GCN+CLASP	Both	1%	36.9 ± 2.2	37.1 ± 4.0	0.2 ± 2.3
GCN+CLASP	Both	3%	36.9 ± 2.2	35.2 ± 3.9	-1.7 ± 5.4
GCN+CLASP	Both	5%	36.9 ± 2.2	32.8 ± 2.3	-4.1 ± 4.0
HARA	Both	1%	36.5 ± 2.3	35.6 ± 1.8	-0.9 ± 2.0
HARA	Both	3%	36.5 ± 2.3	35.6 ± 1.2	-0.9 ± 3.1
HARA	Both	5%	36.5 ± 2.3	34.8 ± 0.3	-1.7 ± 2.6
GCN+CLASP	Preliminary	1%	36.9 ± 2.2	35.9 ± 0.9	-0.9 ± 2.1
GCN+CLASP	Preliminary	3%	36.9 ± 2.2	34.1 ± 3.4	-2.8 ± 4.9
GCN+CLASP	Preliminary	5%	36.9 ± 2.2	33.7 ± 3.8	-3.2 ± 5.7
HARA	Preliminary	1%	36.5 ± 2.3	35.9 ± 2.0	-0.6 ± 1.5
HARA	Preliminary	3%	36.5 ± 2.3	36.3 ± 1.5	-0.2 ± 2.6
HARA	Preliminary	5%	36.5 ± 2.3	35.9 ± 2.8	-0.6 ± 1.1

Table 51: Candidate search settings on 20 fixed validation targets per dataset, model seed zero, budget five. Local20 uses local support, with global additions only if no local addition is available; Global variants always include sampled nonlocal additions. The CE variant uses cross entropy for both gradient proposals and exact discrete acceptance; other variants use the logit margin. Counts retain the actual initially correct denominator. These diagnostics do not establish search optimality.

Dataset	Method	Variant	Clean correct	Successes	Attacked acc.
Actor	GCN+CLASP	global20	6	4	10.0
Actor	GCN+CLASP	global50	6	4	10.0
Actor	GCN+CLASP	global50_ce	6	3	15.0
Actor	GCN+CLASP	global50_seed1	6	4	10.0
Actor	GCN+CLASP	local20	6	4	10.0
Actor	HARA	global20	6	1	25.0
Actor	HARA	global50	6	1	25.0
Actor	HARA	global50_ce	6	1	25.0
Actor	HARA	global50_seed1	6	1	25.0
Actor	HARA	local20	6	1	25.0
Cora	GCN+CLASP	global20	15	12	15.0
Cora	GCN+CLASP	global50	15	13	10.0
Cora	GCN+CLASP	global50_ce	15	13	10.0
Cora	GCN+CLASP	global50_seed1	15	13	10.0
Cora	GCN+CLASP	local20	15	12	15.0
Cora	HARA	global20	14	10	20.0
Cora	HARA	global50	14	10	20.0
Cora	HARA	global50_ce	14	11	15.0
Cora	HARA	global50_seed1	14	10	20.0
Cora	HARA	local20	14	7	35.0

Table 52: Clean test accuracy with 10% of original training labels, crossing three label subset seeds with three model seeds. Each cell is one fitted model; rows share a label subset and columns share a model seed. These are clean accuracy diagnostics, not additional attack replicates.

Dataset	Method	Mask seed	Model 0	Model 1	Model 2
Cora	GCN+CLASP	1000	36.6	42.5	39.2
Cora	GCN+CLASP	1001	25.7	23.6	31.4
Cora	GCN+CLASP	1002	45.6	37.9	40.5
Cora	HARA	1000	57.0	57.5	64.0
Cora	HARA	1001	62.7	55.5	58.5
Cora	HARA	1002	66.6	42.5	59.9
Actor	GCN+CLASP	1000	25.9	26.7	27.0
Actor	GCN+CLASP	1001	26.3	26.3	23.0
Actor	GCN+CLASP	1002	26.4	25.3	25.1
Actor	HARA	1000	26.9	29.5	27.0
Actor	HARA	1001	29.3	30.2	29.7
Actor	HARA	1002	28.5	30.1	31.2

Table 53: Bucket stability at the first one hop aggregation call of each layer. Only initially correct targets with at least one accepted edit enter these descriptive counts. Empty and nonempty count perturbed ordered bucket observations; majority counts those satisfying the retained member condition. The last two columns are mean absolute median shift and mean bound width, in representation distance units. Bucket observations are dependent and are not independent training replicates.

Dataset	Method	Seed	Layer	Targets	Empty	Nonempty	Majority	Shift	Width
Actor	GCN+CLASP	0	0	27	0	675	675	5.426e-05	8.453e-05
Actor	GCN+CLASP	0	1	27	0	675	675	0.0001207	0.09435
Actor	GCN+CLASP	1	0	22	0	550	550	0.0001011	0.0002386
Actor	GCN+CLASP	1	1	22	0	550	550	0.0001496	0.03318
Actor	GCN+CLASP	2	0	26	0	650	650	0.0004436	0.0008423
Actor	GCN+CLASP	2	1	26	0	650	650	0.0003108	0.08353
Actor	HARA	0	0	33	0	825	825	6.109e-05	0.0001279
Actor	HARA	0	1	33	0	825	825	0.08095	1.951
Actor	HARA	1	0	33	0	825	825	6.486e-05	0.0006908
Actor	HARA	1	1	33	0	825	825	0.02228	0.6584
Actor	HARA	2	0	36	0	900	900	0.0001406	0.0007942
Actor	HARA	2	1	36	0	900	900	0.07812	2.936
CiteSeer	GCN+CLASP	0	0	56	0	2016	2016	0.004267	0.009283
CiteSeer	GCN+CLASP	0	1	56	0	2016	2016	0.006692	0.1897
CiteSeer	GCN+CLASP	1	0	63	0	2268	2268	0.001342	0.002506
CiteSeer	GCN+CLASP	1	1	63	0	2268	2268	0.001647	0.08321
CiteSeer	GCN+CLASP	2	0	57	0	2052	2052	0.002358	0.003145
CiteSeer	GCN+CLASP	2	1	57	0	2052	2052	0.003429	0.1433
CiteSeer	HARA	0	0	73	0	2628	2628	0.004103	0.008028
CiteSeer	HARA	0	1	73	0	2628	2628	0.01333	0.5579
CiteSeer	HARA	1	0	67	0	2412	2412	0.002329	0.004727
CiteSeer	HARA	1	1	67	0	2412	2412	0.008194	0.3342
CiteSeer	HARA	2	0	66	0	2376	2376	0.003695	0.007616
CiteSeer	HARA	2	1	66	0	2376	2376	0.01091	0.6071
Squirrel-F	GCN+CLASP	0	0	31	0	775	775	0.0008419	0.0009947
Squirrel-F	GCN+CLASP	0	1	31	0	775	775	0.0008593	0.1465
Squirrel-F	GCN+CLASP	1	0	28	0	700	700	0.0006466	0.0007945
Squirrel-F	GCN+CLASP	1	1	28	0	700	700	0.0008983	0.1138
Squirrel-F	GCN+CLASP	2	0	30	0	750	750	0.0005933	0.00104
Squirrel-F	GCN+CLASP	2	1	30	0	750	750	0.0008965	0.2202
Squirrel-F	HARA	0	0	36	0	900	900	0.0005315	0.0011
Squirrel-F	HARA	0	1	36	0	900	900	0.01172	4.513
Squirrel-F	HARA	1	0	35	0	875	875	0.001708	0.004119
Squirrel-F	HARA	1	1	35	0	875	875	0.01064	6.003
Squirrel-F	HARA	2	0	32	0	800	800	0.0005477	0.00112
Squirrel-F	HARA	2	1	32	0	800	800	0.05065	9.858

Table 54: Label reassignment and retained distance drift at the first one hop call in layers 0 and 1. Targets are initially correct and have accepted edits. Mean labels counts nodes with a changed preliminary prediction per evaluated attack graph; mean R counts surviving directed edges with a changed ordered predicted pair. Means pool those target graphs across three fitted models. Max drift is the largest retained edge distance change across their buckets and target graphs, in representation distance units. These are observed values for the evaluated graphs, not uniform bounds over the attack set; target observations are dependent.

Dataset	Method	Layer	Targets	Mean labels	Mean R	Max drift
CiteSeer	GCN+CLASP	0	176	0.35	1.22	0
CiteSeer	GCN+CLASP	1	176	0.35	1.22	5.37
CiteSeer	HARA	0	206	0.55	2.23	0
CiteSeer	HARA	1	206	0.55	2.23	8.1
Squirrel-F	GCN+CLASP	0	89	0.48	58.18	0
Squirrel-F	GCN+CLASP	1	89	0.48	58.18	3.89
Squirrel-F	HARA	0	103	0.40	17.92	0
Squirrel-F	HARA	1	103	0.40	17.92	17.3
Actor	GCN+CLASP	0	75	0.60	12.53	0
Actor	GCN+CLASP	1	75	0.60	12.53	1.9
Actor	HARA	0	102	1.01	12.84	0
Actor	HARA	1	102	1.01	12.84	3.27

F.1 MATCHED CALIBRATION COMPARISONS

Table 55: GCN+CLASP minus GCN+Global under shared Nettack with reduced training labels. Differences are percentage points, paired by model and mask/attack seed, and summarized by mean and sample standard deviation over three replicates. Clean accuracy uses all test nodes. Attacked accuracy uses 100 targets for local attacks and all test nodes for global evasion.

Dataset	Train	Budget	Clean difference	Attacked difference
Actor	10%	5	-0.1 ± 0.3	0.3 ± 9.9
Actor	25%	5	0.0 ± 0.5	-2.0 ± 2.6
Actor	50%	5	-0.3 ± 0.5	3.0 ± 4.6
Actor	100%	5	0.2 ± 1.5	4.0 ± 5.6
Amazon-Ratings	10%	5	0.7 ± 0.7	0.7 ± 2.3
Amazon-Ratings	25%	5	0.6 ± 0.9	1.0 ± 0.0
Amazon-Ratings	50%	5	0.8 ± 1.4	0.3 ± 6.7
Amazon-Ratings	100%	5	-0.5 ± 0.6	-1.0 ± 4.6
Cora	10%	5	4.4 ± 15.8	3.7 ± 9.3
Cora	25%	5	2.4 ± 4.1	2.3 ± 3.8
Cora	50%	5	2.4 ± 2.7	4.0 ± 8.9
Cora	100%	5	-0.1 ± 0.2	7.7 ± 3.8
Roman-Empire	10%	5	-0.4 ± 1.4	-3.0 ± 4.0
Roman-Empire	25%	5	-0.1 ± 1.0	-2.3 ± 4.9
Roman-Empire	50%	5	2.2 ± 1.5	0.7 ± 6.0
Roman-Empire	100%	5	1.3 ± 0.1	-1.0 ± 1.0

Table 56: GCN+CLASP minus GCN+Global under model specific local evasion. Differences are percentage points, paired by model and mask/attack seed, and summarized by mean and sample standard deviation over three replicates. Clean accuracy uses all test nodes. Attacked accuracy uses 100 targets for local attacks and all test nodes for global evasion.

Dataset	Train	Budget	Clean difference	Attacked difference
Actor	100%	1	0.2 ± 1.5	3.0 ± 2.6
Actor	100%	3	0.2 ± 1.5	0.0 ± 1.0
Actor	100%	5	0.2 ± 1.5	0.7 ± 0.6
CiteSeer	100%	1	-0.8 ± 2.1	-1.0 ± 1.0
CiteSeer	100%	3	-0.8 ± 2.1	-1.0 ± 1.0
CiteSeer	100%	5	-0.8 ± 2.1	-2.3 ± 0.6
Squirrel-F	100%	1	1.9 ± 0.7	-1.3 ± 4.9
Squirrel-F	100%	3	1.9 ± 0.7	-2.3 ± 5.9
Squirrel-F	100%	5	1.9 ± 0.7	-2.7 ± 5.7

Table 57: GCN+CLASP minus GCN+Global under global PR-BCD evasion. Differences are percentage points, paired by model and mask/attack seed, and summarized by mean and sample standard deviation over three replicates. Clean accuracy uses all test nodes. Attacked accuracy uses 100 targets for local attacks and all test nodes for global evasion.

Dataset	Train	Budget	Clean difference	Attacked difference
Actor	100%	1%	0.2 ± 1.5	-0.5 ± 1.7
Actor	100%	3%	0.2 ± 1.5	-0.4 ± 2.8
Actor	100%	5%	0.2 ± 1.5	-0.1 ± 3.2
CiteSeer	100%	1%	-0.8 ± 2.1	-1.4 ± 1.7
CiteSeer	100%	3%	-0.8 ± 2.1	-1.4 ± 2.3
CiteSeer	100%	5%	-0.8 ± 2.1	-1.6 ± 1.3
Cora	100%	1%	-0.1 ± 0.2	-0.2 ± 0.8
Cora	100%	3%	-0.1 ± 0.2	-0.2 ± 0.7
Cora	100%	5%	-0.1 ± 0.2	0.5 ± 1.2
Squirrel-F	100%	1%	1.9 ± 0.7	0.6 ± 1.9
Squirrel-F	100%	3%	1.9 ± 0.7	-0.1 ± 2.4
Squirrel-F	100%	5%	1.9 ± 0.7	-1.8 ± 2.9

Table 58: Matched HARA+Global versus HARA under shared Nettack evasion on 100 targets, with 100% or 10% of training labels. Both models use two anchored refinements with coefficient 0.15 and the same channel/mixer architecture. Only the calibration groups differ. Clean/attacked accuracy are percentages; difference is HARA minus Global in percentage points, paired by model and mask/attack seed. Values are mean and sample standard deviation over three replicates.

Dataset	Train	Budget	Global clean	Global attacked	HARA clean	HARA attacked	Difference
Actor	10%	1	28.0 ± 6.2	23.0 ± 7.0	29.7 ± 7.5	23.7 ± 5.9	0.7 ± 1.2
Actor	10%	3	28.0 ± 6.2	24.0 ± 6.9	29.7 ± 7.5	23.7 ± 8.4	-0.3 ± 1.5
Actor	10%	5	28.0 ± 6.2	23.0 ± 9.5	29.7 ± 7.5	22.7 ± 9.2	-0.3 ± 0.6
Actor	100%	1	36.0 ± 1.0	34.0 ± 4.6	36.0 ± 2.0	31.7 ± 3.1	-2.3 ± 2.9
Actor	100%	3	36.0 ± 1.0	33.3 ± 2.5	36.0 ± 2.0	32.3 ± 2.1	-1.0 ± 2.6
Actor	100%	5	36.0 ± 1.0	31.3 ± 3.2	36.0 ± 2.0	31.0 ± 4.6	-0.3 ± 2.5
Amazon-Ratings	10%	1	38.7 ± 2.9	37.3 ± 1.5	38.0 ± 1.0	37.3 ± 1.5	0.0 ± 3.0
Amazon-Ratings	10%	3	38.7 ± 2.9	36.7 ± 0.6	38.0 ± 1.0	35.0 ± 2.0	-1.7 ± 2.1
Amazon-Ratings	10%	5	38.7 ± 2.9	36.0 ± 1.7	38.0 ± 1.0	36.0 ± 1.0	0.0 ± 2.0
Amazon-Ratings	100%	1	41.7 ± 1.2	39.7 ± 3.1	38.0 ± 2.6	38.0 ± 3.6	-1.7 ± 0.6
Amazon-Ratings	100%	3	41.7 ± 1.2	37.3 ± 2.3	38.0 ± 2.6	38.0 ± 3.6	0.7 ± 1.5
Amazon-Ratings	100%	5	41.7 ± 1.2	36.0 ± 1.0	38.0 ± 2.6	37.0 ± 3.0	1.0 ± 2.6
Cora	10%	1	59.7 ± 5.9	56.3 ± 8.1	56.7 ± 4.0	53.0 ± 7.8	-3.3 ± 14.5
Cora	10%	3	59.7 ± 5.9	51.3 ± 5.7	56.7 ± 4.0	49.0 ± 6.2	-2.3 ± 10.4
Cora	10%	5	59.7 ± 5.9	45.7 ± 6.8	56.7 ± 4.0	40.0 ± 4.6	-5.7 ± 11.4
Cora	100%	1	74.7 ± 2.1	72.3 ± 2.5	73.0 ± 2.0	73.7 ± 1.2	1.3 ± 1.5
Cora	100%	3	74.7 ± 2.1	63.0 ± 2.6	73.0 ± 2.0	64.0 ± 2.6	1.0 ± 5.0
Cora	100%	5	74.7 ± 2.1	50.3 ± 2.1	73.0 ± 2.0	49.0 ± 1.7	-1.3 ± 0.6
Roman-Empire	10%	1	59.7 ± 4.6	57.3 ± 5.0	60.0 ± 2.0	58.3 ± 4.0	1.0 ± 1.0
Roman-Empire	10%	3	59.7 ± 4.6	53.7 ± 3.8	60.0 ± 2.0	55.0 ± 0.0	1.3 ± 3.8
Roman-Empire	10%	5	59.7 ± 4.6	52.3 ± 4.2	60.0 ± 2.0	53.7 ± 0.6	1.3 ± 4.7
Roman-Empire	100%	1	67.0 ± 1.0	64.7 ± 3.5	70.7 ± 1.5	66.0 ± 2.6	1.3 ± 3.8
Roman-Empire	100%	3	67.0 ± 1.0	55.0 ± 2.6	70.7 ± 1.5	55.0 ± 3.0	0.0 ± 1.0
Roman-Empire	100%	5	67.0 ± 1.0	51.7 ± 3.8	70.7 ± 1.5	50.3 ± 6.1	-1.3 ± 2.5

Table 59: Matched HARA+Global versus HARA under transferred Metattack poisoning with both classifiers refitted and evaluated on the poisoned graph; accuracy uses all test nodes. Both models use two anchored refinements with coefficient 0.15 and the same channel/mixer architecture. Only the calibration groups differ. Clean/attacked accuracy are percentages; difference is HARA minus Global in percentage points, paired by model and mask/attack seed. Values are mean and sample standard deviation over three replicates.

Dataset	Train	Budget	Global clean	Global attacked	HARA clean	HARA attacked	Difference
Chameleon-F	N/A	1%	37.1 ± 0.9	38.2 ± 3.1	36.5 ± 2.3	35.6 ± 1.8	-2.6 ± 4.8
Chameleon-F	N/A	3%	37.1 ± 0.9	38.5 ± 2.4	36.5 ± 2.3	35.6 ± 1.2	-3.0 ± 3.1
Chameleon-F	N/A	5%	37.1 ± 0.9	38.0 ± 3.7	36.5 ± 2.3	34.8 ± 0.3	-3.2 ± 3.9
Cora	N/A	1%	78.5 ± 1.3	76.2 ± 0.4	76.6 ± 1.8	75.3 ± 0.0	-0.9 ± 0.4
Cora	N/A	3%	78.5 ± 1.3	74.2 ± 1.7	76.6 ± 1.8	73.8 ± 0.9	-0.4 ± 1.2
Cora	N/A	5%	78.5 ± 1.3	69.9 ± 3.6	76.6 ± 1.8	71.2 ± 0.5	1.4 ± 3.6

F.2 PRELIMINARY PREDICTIONS AND BUCKET OCCUPANCY

Table 60: Cora: preliminary model test accuracy, number of predicted classes, empty ordered buckets, and downstream clean test accuracy under reduced training labels. Accuracy columns are percentages; class and bucket columns are counts. Values are mean and sample standard deviation across three joint mask/training replicates.

Method	Train	Prelim. acc.	Classes	Empty	Downstream acc.
GCN+CLASP	100%	75.1 ± 0.1	7.0 ± 0.0	0.0 ± 0.0	68.8 ± 2.8
GCN+CLASP	50%	68.9 ± 1.5	7.0 ± 0.0	0.0 ± 0.0	65.4 ± 3.4
GCN+CLASP	25%	60.2 ± 5.5	7.0 ± 0.0	0.0 ± 0.0	54.1 ± 5.2
GCN+CLASP	10%	46.0 ± 6.7	7.0 ± 0.0	1.3 ± 1.2	33.6 ± 8.8
HARA	100%	75.1 ± 0.3	7.0 ± 0.0	0.0 ± 0.0	76.6 ± 1.8
HARA	50%	68.8 ± 2.0	7.0 ± 0.0	0.0 ± 0.0	73.4 ± 3.7
HARA	25%	60.3 ± 5.4	7.0 ± 0.0	0.0 ± 0.0	70.7 ± 3.8
HARA	10%	45.9 ± 7.0	7.0 ± 0.0	0.7 ± 1.2	57.5 ± 2.2

Table 61: Actor: preliminary model test accuracy, number of predicted classes, empty ordered buckets, and downstream clean test accuracy under reduced training labels. Accuracy columns are percentages; class and bucket columns are counts. Values are mean and sample standard deviation across three joint mask/training replicates.

Method	Train	Prelim. acc.	Classes	Empty	Downstream acc.
GCN+CLASP	100%	29.1 ± 0.1	5.0 ± 0.0	0.0 ± 0.0	26.4 ± 0.9
GCN+CLASP	50%	28.1 ± 1.0	5.0 ± 0.0	4.0 ± 3.6	26.5 ± 0.8
GCN+CLASP	25%	26.5 ± 0.3	4.7 ± 0.6	5.7 ± 3.1	25.1 ± 1.8
GCN+CLASP	10%	26.8 ± 0.9	5.0 ± 0.0	5.7 ± 2.3	25.8 ± 0.6
HARA	100%	29.0 ± 0.1	5.0 ± 0.0	0.0 ± 0.0	35.2 ± 0.6
HARA	50%	28.1 ± 0.9	4.7 ± 0.6	5.3 ± 4.7	34.3 ± 0.4
HARA	25%	27.3 ± 1.2	4.3 ± 0.6	7.7 ± 2.3	31.8 ± 0.9
HARA	10%	26.6 ± 0.8	4.7 ± 0.6	7.0 ± 2.0	29.5 ± 2.3

Table 62: Roman-Empire: preliminary model test accuracy, number of predicted classes, empty ordered buckets, and downstream clean test accuracy under reduced training labels. Accuracy columns are percentages; class and bucket columns are counts. Values are mean and sample standard deviation across three joint mask/training replicates.

Method	Train	Prelim. acc.	Classes	Empty	Downstream acc.
GCN+CLASP	100%	36.6 ± 0.1	15.0 ± 0.0	139.7 ± 1.2	27.4 ± 0.2
GCN+CLASP	50%	36.2 ± 0.4	15.0 ± 0.0	141.7 ± 3.1	25.6 ± 1.2
GCN+CLASP	25%	35.0 ± 0.7	15.7 ± 0.6	130.7 ± 16.6	22.4 ± 1.0
GCN+CLASP	10%	34.1 ± 0.4	17.0 ± 0.0	120.0 ± 11.5	20.6 ± 1.1
HARA	100%	36.5 ± 0.1	15.0 ± 0.0	139.0 ± 0.0	70.2 ± 0.3
HARA	50%	36.1 ± 0.5	15.0 ± 0.0	140.3 ± 4.2	67.0 ± 0.4
HARA	25%	35.1 ± 0.7	15.7 ± 0.6	131.3 ± 16.5	62.1 ± 1.7
HARA	10%	34.0 ± 0.5	17.0 ± 0.0	118.0 ± 11.5	56.5 ± 0.4

Table 63: Amazon-Ratings: preliminary model test accuracy, number of predicted classes, empty ordered buckets, and downstream clean test accuracy under reduced training labels. Accuracy columns are percentages; class and bucket columns are counts. Values are mean and sample standard deviation across three joint mask/training replicates.

Method	Train	Prelim. acc.	Classes	Empty	Downstream acc.
GCN+CLASP	100%	39.4 ± 0.1	3.0 ± 0.0	16.0 ± 0.0	43.1 ± 0.5
GCN+CLASP	50%	39.4 ± 0.2	3.0 ± 0.0	16.0 ± 0.0	42.0 ± 0.5
GCN+CLASP	25%	39.3 ± 0.8	3.0 ± 0.0	16.0 ± 0.0	39.9 ± 1.0
GCN+CLASP	10%	38.7 ± 1.0	3.3 ± 0.6	13.7 ± 4.0	39.2 ± 0.0
HARA	100%	39.5 ± 0.1	3.0 ± 0.0	16.0 ± 0.0	39.5 ± 2.1
HARA	50%	39.4 ± 0.2	3.0 ± 0.0	16.0 ± 0.0	40.7 ± 0.5
HARA	25%	39.3 ± 0.7	3.0 ± 0.0	16.0 ± 0.0	39.4 ± 2.1
HARA	10%	38.8 ± 0.9	3.3 ± 0.6	13.7 ± 4.0	38.6 ± 1.4

Table 64: Cora: preliminary model accuracy and empty ordered buckets after transferred poisoning, with downstream accuracy. Both evaluates on the poisoned graph; Preliminary evaluates on the clean graph after only preliminary fitting was poisoned. Accuracy columns are percentages.

Method	Training	Budget	Prelim. acc.	Empty	Downstream acc.
GCN+CLASP	Both	1%	72.0 ± 1.3	0.0 ± 0.0	68.3 ± 4.6
GCN+CLASP	Both	3%	62.7 ± 2.9	0.0 ± 0.0	64.2 ± 0.9
GCN+CLASP	Both	5%	58.5 ± 1.3	0.0 ± 0.0	64.5 ± 2.0
GCN+CLASP	Preliminary	1%	72.1 ± 1.4	0.0 ± 0.0	69.0 ± 4.1
GCN+CLASP	Preliminary	3%	62.8 ± 2.7	0.0 ± 0.0	69.4 ± 3.3
GCN+CLASP	Preliminary	5%	58.4 ± 1.0	0.0 ± 0.0	68.5 ± 2.9
HARA	Both	1%	71.7 ± 1.6	0.0 ± 0.0	75.3 ± 0.0
HARA	Both	3%	62.5 ± 2.6	0.0 ± 0.0	73.8 ± 0.9
HARA	Both	5%	57.4 ± 1.3	0.0 ± 0.0	71.2 ± 0.5
HARA	Preliminary	1%	71.8 ± 1.7	0.0 ± 0.0	76.8 ± 1.7
HARA	Preliminary	3%	62.5 ± 2.6	0.0 ± 0.0	77.1 ± 1.8
HARA	Preliminary	5%	57.4 ± 1.0	0.0 ± 0.0	77.6 ± 1.6

Table 65: Chameleon-F: preliminary model accuracy and empty ordered buckets after transferred poisoning, with downstream accuracy. Both evaluates on the poisoned graph; Preliminary evaluates on the clean graph after only preliminary fitting was poisoned. Accuracy columns are percentages.

Method	Training	Budget	Prelim. acc.	Empty	Downstream acc.
GCN+CLASP	Both	1%	35.6 ± 1.3	0.0 ± 0.0	37.1 ± 4.0
GCN+CLASP	Both	3%	38.4 ± 2.5	0.0 ± 0.0	35.2 ± 3.9
GCN+CLASP	Both	5%	35.2 ± 2.4	0.0 ± 0.0	32.8 ± 2.3
GCN+CLASP	Preliminary	1%	35.9 ± 0.9	0.0 ± 0.0	35.9 ± 0.9
GCN+CLASP	Preliminary	3%	38.7 ± 1.8	0.0 ± 0.0	34.1 ± 3.4
GCN+CLASP	Preliminary	5%	37.4 ± 3.9	0.0 ± 0.0	33.7 ± 3.8
HARA	Both	1%	35.6 ± 0.9	0.0 ± 0.0	35.6 ± 1.8
HARA	Both	3%	37.2 ± 3.4	0.3 ± 0.6	35.6 ± 1.2
HARA	Both	5%	34.1 ± 2.6	0.7 ± 1.2	34.8 ± 0.3
HARA	Preliminary	1%	35.6 ± 0.6	0.0 ± 0.0	35.9 ± 2.0
HARA	Preliminary	3%	37.6 ± 1.2	0.7 ± 1.2	36.3 ± 1.5
HARA	Preliminary	5%	36.7 ± 2.3	1.3 ± 2.3	35.9 ± 2.8

Table 66: Preliminary predictions with five training labels per class and the original validation sets. Preliminary and clean downstream accuracy use all test nodes; attacked accuracy uses 100 fixed targets and five shared Nettack edits. Accuracy columns are percentages, Classes and Empty are counts. Values are mean and sample standard deviation across three joint mask/training replicates.

Dataset	Method	Prelim. acc.	Classes	Empty	Clean test	Attacked
Actor	GCN+CLASP	22.1 ± 0.9	5.0 ± 0.0	0.0 ± 0.0	21.1 ± 0.4	21.0 ± 4.0
Actor	HARA	20.4 ± 2.3	5.0 ± 0.0	0.0 ± 0.0	22.2 ± 2.8	19.3 ± 2.5
Roman-Empire	GCN+CLASP	21.2 ± 0.8	18.0 ± 0.0	0.0 ± 0.0	11.4 ± 0.5	8.7 ± 4.5
Roman-Empire	HARA	21.1 ± 1.1	18.0 ± 0.0	0.0 ± 0.0	26.8 ± 1.8	22.7 ± 3.2
Amazon-Ratings	GCN+CLASP	24.3 ± 3.1	5.0 ± 0.0	0.0 ± 0.0	27.7 ± 5.7	27.0 ± 1.0
Amazon-Ratings	HARA	24.1 ± 3.5	5.0 ± 0.0	0.0 ± 0.0	33.4 ± 0.4	36.3 ± 2.1

F.3 DENOMINATORS AND LOCAL ATTACK COMPARISON

Table 67: Actual initially correct ASR denominators for reduced training labels. Each entry gives model seeds 0/1/2 in that order. The denominator is the same at all budgets for a fitted model.

Method	Train	Cora	Actor	Roman-Empire	Amazon-Ratings
GCN	100%	73/74/78	34/32/30	48/50/48	43/42/42
RUNG-MCP	100%	81/84/75	30/31/33	63/62/64	41/42/42
HARA-Mean	100%	70/72/68	35/36/35	77/80/76	39/42/37
GCN+Global	100%	66/62/67	20/23/29	30/24/31	36/44/39
GCN+CLASP	100%	65/58/64	27/23/28	25/25/27	36/39/43
HARA	100%	75/73/71	34/38/36	71/69/72	41/36/37
GCN	50%	74/76/81	27/27/33	45/46/49	41/40/43
RUNG-MCP	50%	73/78/74	30/27/28	60/62/63	39/39/39
HARA-Mean	50%	68/68/74	31/29/35	73/75/72	37/39/39
GCN+Global	50%	57/65/57	20/21/27	30/25/33	38/34/34
GCN+CLASP	50%	63/60/65	25/27/27	29/27/27	34/43/38
HARA	50%	68/71/76	29/30/34	66/68/72	37/42/39
GCN	25%	64/69/69	24/27/38	42/45/45	38/42/43
RUNG-MCP	25%	66/65/71	31/25/29	59/60/60	40/43/42
HARA-Mean	25%	65/64/70	22/32/29	70/76/73	40/41/39
GCN+Global	25%	50/45/60	23/29/28	23/23/27	42/36/37
GCN+CLASP	25%	52/50/55	26/26/25	30/22/23	37/38/37
HARA	25%	64/60/73	28/31/31	58/61/69	37/40/34
GCN	10%	46/49/61	22/27/28	36/32/37	31/40/37
RUNG-MCP	10%	59/51/62	26/23/27	55/62/58	34/37/35
HARA-Mean	10%	67/46/60	21/32/29	63/66/62	37/37/40
GCN+Global	10%	11/23/46	27/33/29	25/23/22	35/37/38
GCN+CLASP	10%	31/18/45	26/25/28	23/23/22	36/42/38
HARA	10%	52/59/59	22/30/37	60/62/58	37/38/39

Table 68: Actual initially correct ASR denominators for model specific local evasion. Each entry gives model seeds 0/1/2 in that order. The denominator is the same at all budgets for a fitted model.

Method	CiteSeer	Squirrel-F	Actor
GCN	76/71/74	30/34/33	34/32/30
RUNG-MCP	70/71/72	28/30/29	30/31/33
HARA-Mean	61/67/63	32/35/40	35/36/35
GCN+Global	57/61/59	31/32/32	20/23/29
GCN+CLASP	56/63/57	33/31/34	27/23/28
HARA	73/67/66	36/35/32	34/38/36

Table 69: Actual initially correct ASR denominators for global PR-BCD evasion. Each entry gives model seeds 0/1/2 in that order. The denominator is the same at all budgets for a fitted model.

Method	Cora	CiteSeer	Squirrel-F	Actor
GCN	798/804/813	696/683/683	147/147/158	465/464/467
RUNG-MCP	831/821/804	704/697/700	144/146/144	480/480/490
HARA-Mean	790/772/785	602/623/580	158/158/175	521/565/533
GCN+Global	702/659/707	519/559/526	136/144/143	401/385/410
GCN+CLASP	701/656/708	535/534/512	147/153/148	415/402/387
HARA	786/759/753	670/607/602	173/174/175	544/534/526

Table 70: Actual initially correct ASR denominators with five training labels per class. Each entry gives model seeds 0/1/2, out of the same 100 targets. Validation labels remain available.

Method	Actor	Roman-Empire	Amazon-Ratings
GCN	11/24/19	19/24/20	26/15/23
RUNG-MCP	21/24/17	30/37/31	37/25/26
HARA-Mean	21/21/21	29/26/24	37/28/33
GCN+Global	22/18/24	9/10/14	34/29/26
GCN+CLASP	21/18/24	9/15/8	35/28/25
HARA	17/18/23	21/28/27	33/30/28

Table 71: Actor: shared Nettack and model specific local attacks on the same full label fitted checkpoints and the same 100 targets, at budget five. The portfolio chooses an evaluated graph separately for each target from shared, model specific and clean graphs. It is feasible for separate targeted attacks; it does not assert simultaneous errors on one global graph. Accuracy and ASR are percentages, mean and sample standard deviation over three fitted models.

Method	Shared acc.	Model specific acc.	Portfolio acc.	Portfolio ASR
GCN	1.0 ± 0.0	2.0 ± 0.0	1.0 ± 0.0	96.9 ± 0.2
RUNG-MCP	19.0 ± 4.6	4.0 ± 1.0	4.0 ± 1.0	87.1 ± 3.8
HARA-Mean	32.7 ± 1.5	25.3 ± 1.5	25.3 ± 1.5	28.3 ± 4.7
GCN+Global	19.0 ± 1.7	1.7 ± 0.6	1.7 ± 0.6	92.9 ± 2.8
GCN+CLASP	23.0 ± 4.0	2.3 ± 1.2	2.3 ± 1.2	91.3 ± 3.8
HARA	31.0 ± 4.6	26.7 ± 4.7	26.7 ± 4.7	26.2 ± 9.1

F.4 GRAPH EDIT CONDITION AND GLOBAL BUDGET CURVES

Table 72: Applicability of the graph edit sufficient condition at the first one hop call in layers 0 and 1. Counts pool dependent bucket observations across three fitted models, restricted to initially correct targets with accepted edits. The budget condition uses the clean bucket size m, actual undirected edit count K, and total reassigned surviving directed edges R. Observed majority checks the exact retained member count. The sufficient condition can fail even when retained majority holds; neither condition certifies classification.

Dataset	Method	Layer	Nonempty	$m > 2K + R$	Observed majority
CiteSeer	GCN+CLASP	0	6336	6294	6336
CiteSeer	GCN+CLASP	1	6336	6294	6336
CiteSeer	HARA	0	7416	7272	7416
CiteSeer	HARA	1	7416	7272	7416
Squirrel-F	GCN+CLASP	0	2225	2179	2225
Squirrel-F	GCN+CLASP	1	2225	2179	2225
Squirrel-F	HARA	0	2575	2561	2575
Squirrel-F	HARA	1	2575	2561	2575
Actor	GCN+CLASP	0	1875	1617	1875
Actor	GCN+CLASP	1	1875	1617	1875
Actor	HARA	0	2550	2058	2550
Actor	HARA	1	2550	2058	2550

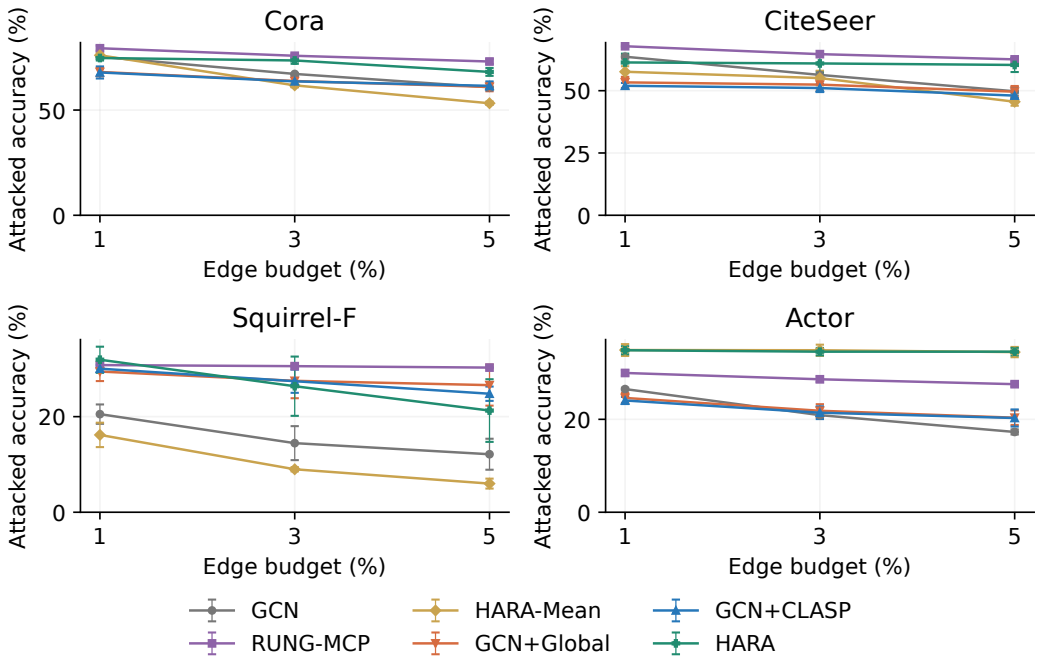


Figure 5: Global model specific PR-BCD evasion under the fixed training recipe. Each point selects one feasible whole graph from the current and smaller budgets, including the clean graph. Points and error bars are mean and sample standard deviation over three joint attack/training replicates. Accuracy uses all test nodes; panels have separate scales.

G Shared Attack Comparisons and Diagnostics

Under the same transferred Nettack attacks, nine methods are evaluated on seven datasets, with three model seeds and the same 100 target nodes per dataset. Each method's conditional RAUC uses its own initially correct targets. Averaging datasets equally, HARA has a descriptive mean RAUC of 0.126, compared with 0.130 for RUNG-MCP and 0.133 for GCN+CLASP. GCN+CLASP has the lowest per dataset mean RAUC on Cora, Chameleon-F, Squirrel-F, and Amazon-Ratings. HARA, HARA-Median, and RUNG-MCP each have the lowest value on one other dataset. The SoftMedian and HARA-Median labels identify the capped soft medoid implementations specified in Appendix H.

Table 73: Descriptive equal dataset means in the nine method shared Nettack study. Each dataset first averages three seeds. Lower ASR and RAUC are better; wins count the lowest mean RAUC among these nine methods.

Method	Clean Acc.	ASR@1	ASR@3	ASR@5	RAUC	Wins
GCN	0.483	0.607	0.880	0.935	0.721	0
RUNG-MCP	0.495	0.043	0.144	0.299	0.130	1
RUNG-L1	0.498	0.103	0.334	0.467	0.258	0
SoftMedian	0.468	0.143	0.354	0.519	0.289	0
EvenNet	0.525	0.340	0.578	0.657	0.465	0
HARA-Mean	0.525	0.194	0.437	0.558	0.345	0
HARA-Median	0.517	0.155	0.326	0.440	0.265	1
GCN+CLASP	0.414	0.054	0.164	0.256	0.133	4
HARA	0.513	0.062	0.159	0.218	0.126	1

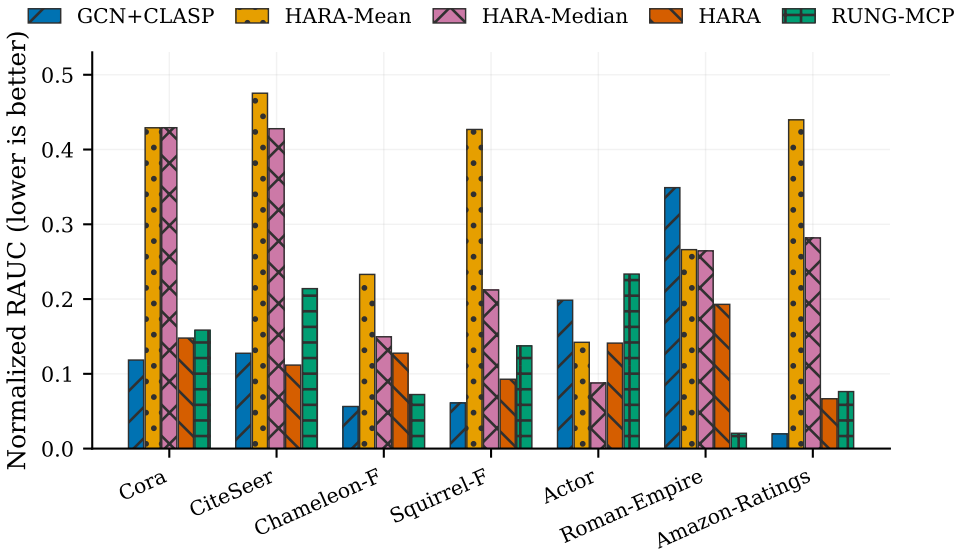


Figure 6: Module and host comparisons under shared Nettack perturbations with fixed configurations. Lower normalized RAUC is better. Each dataset averages three model seeds, with conditional ASR defined over each method's initially correct targets. Table 75 gives the numerical values, and Appendix J reports accuracy and paired outcomes on common initially correct targets.

GCN+CLASP has lower conditional RAUC than GCN on all seven datasets. HARA also has lower conditional RAUC than HARA-Mean on all seven, with aggregation, refinement count, and anchoring changing together. The matched global scale controls in the stress studies isolate class pair grouping within each host.

The paired comparison restricts evaluation to targets initially classified correctly by all nine methods. At budget five, these sets have per seed sizes $49/45/50$, $39/45/38$, $14/16/17$, $17/19/18$, $5/5/2$, $9/10/10$, and $23/20/12$ in dataset order. Actor and Roman-Empire thus have particularly small paired sets.

Table 74: HARA conditional ASR and normalized RAUC in the shared attack study. Values are mean and sample standard deviation across three model seeds; $n_0/n_1/n_2$ are the actual denominators of initially correct targets.

Dataset	Clean correct $n_0/n_1/n_2$	ASR@1	ASR@3	ASR@5	RAUC
Cora	75/73/71	0.036 ± 0.021	0.160 ± 0.032	0.365 ± 0.024	0.148 ± 0.007
CiteSeer	73/67/66	0.039 ± 0.022	0.149 ± 0.038	0.202 ± 0.056	0.112 ± 0.025
Chameleon-F	38/32/37	0.065 ± 0.014	0.177 ± 0.018	0.187 ± 0.003	0.128 ± 0.010
Squirrel-F	36/35/32	0.038 ± 0.042	0.124 ± 0.141	0.159 ± 0.223	0.093 ± 0.112
Actor	34/38/36	0.131 ± 0.049	0.160 ± 0.067	0.189 ± 0.096	0.141 ± 0.060
Roman-Empire	75/72/72	0.077 ± 0.020	0.251 ± 0.087	0.347 ± 0.153	0.193 ± 0.070
Amazon-Ratings	40/40/37	0.050 ± 0.087	0.092 ± 0.088	0.075 ± 0.066	0.067 ± 0.068

Table 75: Per dataset mean normalized RAUC for module and host comparisons under shared attacks. Bold marks the descriptive minimum among all nine methods.

Dataset	GCN	GCN+CLASP	HARA-Mean	HARA-Median	HARA	RUNG-MCP
Cora	0.670	**0.118**	0.429	0.429	0.148	0.158
CiteSeer	0.745	0.128	0.475	0.428	**0.112**	0.214
Chameleon-F	0.781	**0.056**	0.233	0.150	0.128	0.072
Squirrel-F	0.727	**0.061**	0.427	0.212	0.093	0.137
Actor	0.815	0.199	0.142	**0.088**	0.141	0.233
Roman-Empire	0.729	0.349	0.266	0.264	0.193	**0.021**
Amazon-Ratings	0.579	**0.020**	0.440	0.282	0.067	0.076

G.1 Inference time label interventions

The eight label interventions change inference time label assignments while keeping the weights of HARA and its preliminary classifier fixed. The assignments use recomputation, fixed clean predictions, random labels, an oracle diagnostic, and four corruption levels. Each mode defines its own initially correct target set for conditional ASR. At budget five, recomputed versus fixed clean ASR is 0.365 versus 0.433 on Cora and 0.189 versus 0.140 on Actor. Across all eight label modes, ASR ranges from 0.318 to 0.512 on Cora and from 0.140 to 0.208 on Actor. Recomputing predictions lowers ASR on Cora and raises it on Actor relative to fixed clean predictions.

The separate 21 independently fitted preliminary classifier diagnostic checkpoints have test accuracies from 0.29 to 0.75. Table 76 reports their clean bucket support. These checkpoints measure standalone provider behavior, so their accuracies are not measurements of the providers inside the fitted HARA models used for the label interventions.

Table 76: Test accuracy and ordered clean class pair support for the independently fitted preliminary classifier diagnostics. Counts refer to graph wide buckets, rather than buckets of every downstream model.

Dataset	Preliminary test acc.	Empty buckets	Nonempty ≤ 5	Minimum nonempty
Cora	0.751 to 0.753	0	0	7 to 9
CiteSeer	0.651 to 0.656	0	0	13 to 14
Chameleon-F	0.363 to 0.369	0	0	37
Squirrel-F	0.371 to 0.373	0	0	346
Actor	0.290 to 0.294	0	0 to 1	5 to 7
Roman-Empire	0.365 to 0.367	139	23 to 27	1
Amazon-Ratings	0.394 to 0.395	16	2	3

G.2 Clean bucket support

Using the provider checkpoints from the corresponding fitted HARA models, the support diagnostic assigns 1,079 initially correct budget five target observations to five prescribed support regimes. Roman-Empire has $3/6$ attack successes at support zero and $3/3$ at support 1 to 5. Chameleon-F

has $6/9$ at support 6 to 50 and $5/57$ at support 51 to 500. These counts pool dependent observations across three fits. The association between support and attack success varies by dataset; the study does not intervene on support or test a pooling rule.

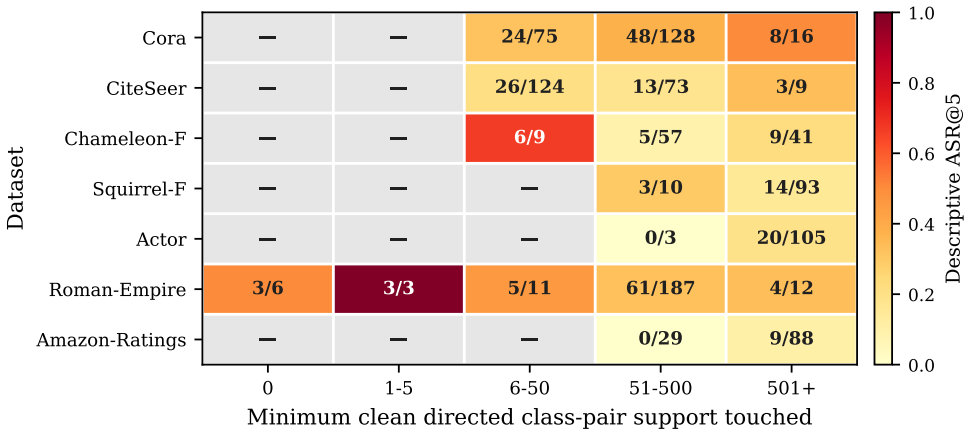


Figure 7: Shared attack outcomes grouped by the minimum clean directed class pair support touched by each budget five attack in the fixed configuration HARA study. Cells show attack successes over target counts, pooled across three fitted models; gray cells have no targets. The colors encode the corresponding descriptive ASR.

G.3 HARA ONLY LOCAL TOP 20 AUDIT

The local top 20 audit contains 2,700 target budget records on CiteSeer, Squirrel-F, and Actor. Using separately fitted runs for the two attack protocols, adaptive versus shared Nettack ASR at budget five is 0.452 versus 0.202, 0.441 versus 0.159, and 0.272 versus 0.189, respectively. Across seeds, accepted action instances contain 375 additions and 764 deletions on CiteSeer, 606 additions and 65 deletions on Squirrel-F, and 332 additions and 387 deletions on Actor.

The top 20 audit samples from the expanded two hop support. When the cap is active, sampling can change the weighted proposal base point. Each reported attack success is verified by an exact discrete evaluation. The finite search leaves optimality and possible gradient masking unresolved. The separate six method expanded support study uses the sample preserving relaxation described in Appendix B.3.

H Implementation and Protocol Details

H.1 Evaluated configurations

The shared attack study fixes hidden width 64 and dropout 0.5 for all methods where applicable. Adam uses learning rate 0.01 and weight decay 5×10^{-4}. Training runs for at most 200 epochs, with early stopping by validation loss and patience 50. HARA uses two layers, CLASP aggregation in the one hop and two hop channels, $\gamma_\star = 1.25$, two anchored refinements, anchor coefficient 0.15, and adaptive mixing. GCN+CLASP uses the same $\gamma_\star$ and provider but no channel separation, two hop graph, anchored iteration, or adaptive mixer. HARA-Mean and HARA-Median retain the channels and mixer but use one unanchored refinement ($T = 1$, $\alpha = 0$), whereas HARA uses $T = 2$ and $\alpha = 0.15$. Comparing these configurations changes aggregation, refinement, and anchoring together.

RUNG-MCP and RUNG-L1 use 10 propagation steps, $\widehat{\lambda} = 0.9$, and $\gamma = 6.0$. EvenNet (Lei et al., 2022) uses 10 layers and $\alpha = 0.1$.

The implementation labeled SoftMedian in the tables uses a soft medoid neighborhood rule (Geisler et al., 2020), with temperature one and a deterministic cap of 16 neighbors. HARA-Median uses the same rule in its graph channels. The coordinatewise Soft Median of Geisler et al. (2021) is a different method outside this comparison.

H.2 Preliminary classifier state machine

For every dataset and model seed, the provider follows four explicit states: `unfitted`, `fitted`, `checkpointed`, and `frozen`. Downstream training is rejected until a checkpoint has been loaded and every preliminary parameter has `requires_grad=False`. Formal prediction mode is `perturbed_recompute`; `fixed_clean`, random labels, oracle labels, and controlled corruption are diagnostic modes. Stored diagnostics include the checkpoint hash, parameter change norm, train/validation/test accuracy, prediction histogram, ordered bucket histogram, and freeze state.

H.3 Weighted graph invariants

The interface `forward(x, edge_index, edge_weight=None)` is shared by training and attack code. Unit tests compare omitted weights with unit weights. On uncapped toy graphs, they check zero weight support and candidate derivatives. Controlled toy cases compare gradient directions with one sided finite differences. Additional tests check deterministic two hop construction and cache identity for different graphs of the same size. Graph statistics in the mixer use the same weights as message aggregation. Appendix B.3 specifies the capped two hop extension used in the attacks.

H.4 Target manifest and attack artifacts

Each dataset has a manifest of 100 test nodes, selected independently of the models and stratified by class and degree. All methods use the same manifest and shared attack artifact. Each target record identifies the dataset, model, seed, budget, attack artifact, and flips. It also stores clean and final predictions, initial correctness, attack success, preliminary label changes, and validity. These records supply the aggregates and tables. Attacks set `perturb_structure=True` and `perturb_features=False`, allowing structural edits while holding features fixed.

I Metric and Evidence Semantics

Conditional ASR measures the fraction of initially correct targets misclassified after attack, using a separate denominator for each model. Overall error measures the fraction of all manifest targets misclassified after attack. Its denominator includes initially correct and initially incorrect targets. The paired comparison fixes the population to targets initially classified correctly by all nine methods. RAUC integrates the conditional budget curve, so comparisons across models also depend on the initially correct population and clean accuracy.

Clopper and Pearson binomial intervals are reported for each combination of dataset, seed, method, and budget. The binomial model does not account for dependence between graph targets. Three seeds measure variation across fitted models, with target and budget observations nested within each fit.

J Complete Shared Attack Results

Table 77: Per dataset normalized RAUC for the nine evaluated methods. Values are mean ± sample standard deviation over seeds; lower is better.

Dataset	GCN	RUNG-MCP	RUNG-L1	SoftMedian	EvenNet	HARA-Mean	HARA-Median	GCN+CLASP	HARA
Cora	0.670 ± 0.008	0.158 ± 0.012	0.334 ± 0.009	0.461 ± 0.001	0.552 ± 0.013	0.429 ± 0.068	0.429 ± 0.007	0.118 ± 0.021	0.148 ± 0.007
CiteSeer	0.745 ± 0.002	0.214 ± 0.020	0.405 ± 0.009	0.601 ± 0.012	0.641 ± 0.009	0.475 ± 0.032	0.428 ± 0.055	0.128 ± 0.034	0.112 ± 0.025
Chameleon-F	0.781 ± 0.029	0.072 ± 0.015	0.085 ± 0.017	0.154 ± 0.035	0.509 ± 0.035	0.233 ± 0.064	0.150 ± 0.055	0.056 ± 0.031	0.128 ± 0.010
Squirrel-F	0.727 ± 0.017	0.137 ± 0.016	0.183 ± 0.026	0.085 ± 0.023	0.710 ± 0.023	0.427 ± 0.063	0.212 ± 0.057	0.061 ± 0.031	0.093 ± 0.112
Actor	0.815 ± 0.005	0.233 ± 0.009	0.534 ± 0.006	0.294 ± 0.048	0.207 ± 0.044	0.142 ± 0.038	0.088 ± 0.074	0.199 ± 0.018	0.141 ± 0.060
Roman-Empire	0.729 ± 0.024	0.021 ± 0.007	0.174 ± 0.010	0.283 ± 0.020	0.160 ± 0.014	0.266 ± 0.033	0.264 ± 0.019	0.349 ± 0.023	0.193 ± 0.070
Amazon-Ratings	0.579 ± 0.045	0.076 ± 0.024	0.091 ± 0.025	0.143 ± 0.023	0.473 ± 0.011	0.440 ± 0.071	0.282 ± 0.069	0.020 ± 0.025	0.067 ± 0.068

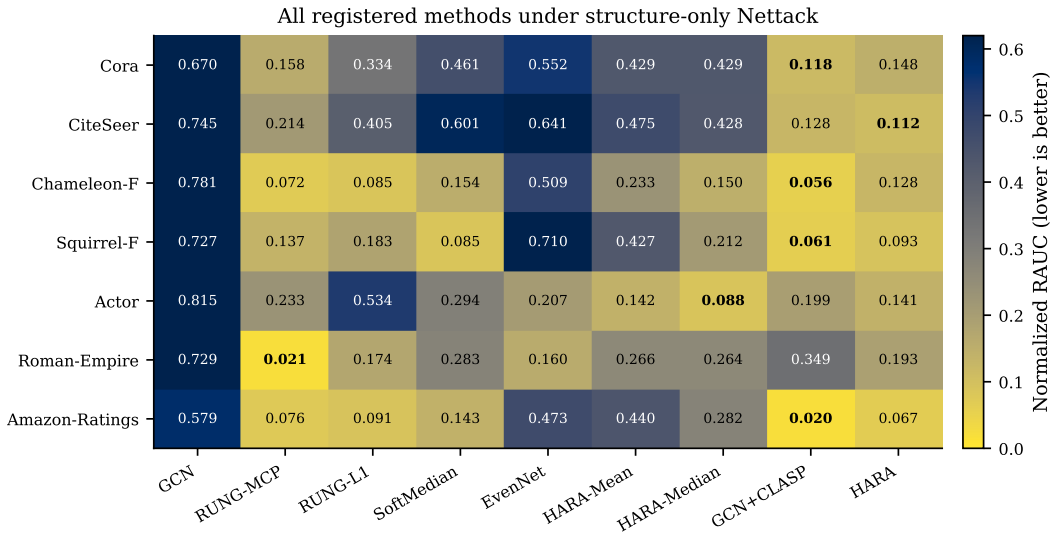


Figure 8: Normalized RAUC for all nine methods under shared Nettack attacks with fixed configurations. Entries average three seeds on each of seven datasets. Bold annotations mark the descriptive minimum within each dataset; lower is better.

K Complete Neutral Attack Results

This section reports every registered method and dataset. Conditional ASR, end to end error, and RAUC are averaged over three independently trained seeds; denominators of initially correct targets are shown in seed order. Pooled target counts are descriptive.

K.1 CORA

Table 78: Complete registered method results on Cora. Lower ASR, end to end error, and RAUC are better.

Method	Clean acc.	$n_0/n_1/n_2$	ASR@1	ASR@3	ASR@5	E2E@5	RAUC
GCN	0.805 ± 0.008	73/74/78	0.418 ± 0.013	0.871 ± 0.011	0.982 ± 0.007	0.987 ± 0.006	0.670 ± 0.008
RUNG-MCP	0.812 ± 0.010	78/79/77	0.038 ± 0.000	0.192 ± 0.020	0.350 ± 0.030	0.490 ± 0.030	0.158 ± 0.012
RUNG-L1	0.822 ± 0.006	78/80/81	0.100 ± 0.032	0.435 ± 0.011	0.649 ± 0.028	0.720 ± 0.026	0.334 ± 0.009
SoftMedian	0.768 ± 0.004	72/71/69	0.170 ± 0.025	0.604 ± 0.027	0.844 ± 0.015	0.890 ± 0.010	0.461 ± 0.001
EvenNet	0.821 ± 0.003	78/77/76	0.281 ± 0.011	0.705 ± 0.036	0.926 ± 0.008	0.943 ± 0.006	0.552 ± 0.013
HARA-Mean	0.782 ± 0.009	70/72/68	0.185 ± 0.063	0.537 ± 0.098	0.794 ± 0.051	0.857 ± 0.032	0.429 ± 0.068
HARA-Median	0.733 ± 0.011	69/69/67	0.146 ± 0.013	0.546 ± 0.008	0.834 ± 0.007	0.887 ± 0.006	0.429 ± 0.007
GCN+CLASP	0.688 ± 0.028	65/58/64	0.015 ± 0.015	0.122 ± 0.018	0.324 ± 0.047	0.550 ± 0.000	0.118 ± 0.021
HARA	0.766 ± 0.018	75/73/71	0.036 ± 0.021	0.160 ± 0.032	0.365 ± 0.024	0.510 ± 0.017	0.148 ± 0.007

To interpret these results, read clean accuracy and overall error alongside conditional ASR. The first two measure utility on the full target manifest. ASR measures failures among each seed's initially correct targets.

Table 79: Results for every registered method and seed on Cora. Parentheses contain seed level conditional ASR.

Method	Seed	Clean acc.	Clean n	Success@1/n (ASR)	Success@3/n (ASR)	Success@5/n (ASR)	RAUC
GCN	0	0.798	73	30/73 (0.411)	64/73 (0.877)	72/73 (0.986)	0.671
	1	0.804	74	32/74 (0.432)	65/74 (0.878)	73/74 (0.986)	0.678
	2	0.813	78	32/78 (0.410)	67/78 (0.859)	76/78 (0.974)	0.662
RUNG-MCP	0	0.823	78	3/78 (0.038)	16/78 (0.205)	30/78 (0.385)	0.171
	1	0.803	79	3/79 (0.038)	16/79 (0.203)	26/79 (0.329)	0.158
	2	0.809	77	3/77 (0.039)	13/77 (0.169)	26/77 (0.338)	0.147
RUNG-L1	0	0.828	78	5/78 (0.064)	33/78 (0.423)	53/78 (0.679)	0.324
	1	0.816	80	9/80 (0.113)	35/80 (0.438)	50/80 (0.625)	0.334
	2	0.822	81	10/81 (0.123)	36/81 (0.444)	52/81 (0.642)	0.343
SoftMedian	0	0.771	72	13/72 (0.181)	42/72 (0.583)	62/72 (0.861)	0.460
	1	0.768	71	10/71 (0.141)	45/71 (0.634)	59/71 (0.831)	0.462
	2	0.764	69	13/69 (0.188)	41/69 (0.594)	58/69 (0.841)	0.462
EvenNet	0	0.823	78	21/78 (0.269)	58/78 (0.744)	73/78 (0.936)	0.565
	1	0.818	77	22/77 (0.286)	54/77 (0.701)	71/77 (0.922)	0.551
	2	0.822	76	22/76 (0.289)	51/76 (0.671)	70/76 (0.921)	0.539
HARA-Mean	0	0.790	70	17/70 (0.243)	43/70 (0.614)	58/70 (0.829)	0.484
	1	0.772	72	14/72 (0.194)	41/72 (0.569)	59/72 (0.819)	0.450
	2	0.785	68	8/68 (0.118)	29/68 (0.426)	50/68 (0.735)	0.353
HARA-Median	0	0.734	69	11/69 (0.159)	38/69 (0.551)	58/69 (0.841)	0.436
	1	0.722	69	10/69 (0.145)	38/69 (0.551)	57/69 (0.826)	0.429
	2	0.743	67	9/67 (0.134)	36/67 (0.537)	56/67 (0.836)	0.422
GCN+CLASP	0	0.701	65	2/65 (0.031)	9/65 (0.138)	24/65 (0.369)	0.138
	1	0.656	58	0/58 (0.000)	6/58 (0.103)	16/58 (0.276)	0.097
	2	0.708	64	1/64 (0.016)	8/64 (0.125)	21/64 (0.328)	0.120
HARA	0	0.786	75	3/75 (0.040)	13/75 (0.173)	28/75 (0.373)	0.156
	1	0.759	73	4/73 (0.055)	9/73 (0.123)	28/73 (0.384)	0.142
	2	0.753	71	1/71 (0.014)	13/71 (0.183)	24/71 (0.338)	0.145

K.2 CITESEER

Table 80: Complete registered method results on CiteSeer. Lower ASR, end to end error, and RAUC are better.

Method	Clean acc.	$n_0/n_1/n_2$	ASR@1	ASR@3	ASR@5	E2E@5	RAUC
GCN	0.687 ± 0.008	76/71/74	0.607 ± 0.014	0.932 ± 0.002	0.950 ± 0.009	0.963 ± 0.006	0.745 ± 0.002
RUNG-MCP	0.701 ± 0.007	72/73/74	0.032 ± 0.016	0.214 ± 0.026	0.593 ± 0.034	0.703 ± 0.021	0.214 ± 0.020
RUNG-L1	0.701 ± 0.003	72/73/74	0.078 ± 0.022	0.557 ± 0.020	0.794 ± 0.003	0.850 ± 0.000	0.405 ± 0.009
SoftMedian	0.653 ± 0.005	69/72/71	0.301 ± 0.029	0.821 ± 0.011	0.910 ± 0.007	0.937 ± 0.006	0.601 ± 0.012
EvenNet	0.698 ± 0.006	75/75/75	0.458 ± 0.020	0.809 ± 0.015	0.902 ± 0.008	0.927 ± 0.006	0.641 ± 0.009
HARA-Mean	0.602 ± 0.022	61/67/63	0.273 ± 0.032	0.615 ± 0.070	0.738 ± 0.021	0.827 ± 0.015	0.475 ± 0.032
HARA-Median	0.601 ± 0.007	62/68/67	0.218 ± 0.018	0.555 ± 0.099	0.701 ± 0.061	0.793 ± 0.064	0.428 ± 0.055
GCN+CLASP	0.527 ± 0.013	56/63/57	0.033 ± 0.032	0.156 ± 0.053	0.277 ± 0.114	0.527 ± 0.081	0.128 ± 0.034
HARA	0.626 ± 0.038	73/67/66	0.039 ± 0.022	0.149 ± 0.038	0.202 ± 0.056	0.443 ± 0.029	0.112 ± 0.025

To interpret these results, read clean accuracy and overall error alongside conditional ASR. The first two measure utility on the full target manifest. ASR measures failures among each seed's initially correct targets.

Table 81: Results for every registered method and seed on CiteSeer. Parentheses contain seed level conditional ASR.

Method	Seed	Clean acc.	Clean n	Success@1/n (ASR)	Success@3/n (ASR)	Success@5/n (ASR)	RAUC
GCN	0	0.696	76	45/76 (0.592)	71/76 (0.934)	73/76 (0.961)	0.743
	1	0.683	71	44/71 (0.620)	66/71 (0.930)	67/71 (0.944)	0.746
	2	0.683	74	45/74 (0.608)	69/74 (0.932)	70/74 (0.946)	0.745
RUNG-MCP	0	0.708	72	1/72 (0.014)	14/72 (0.194)	40/72 (0.556)	0.193
	1	0.695	73	3/73 (0.041)	15/73 (0.205)	44/73 (0.603)	0.215
	2	0.699	74	3/74 (0.041)	18/74 (0.243)	46/74 (0.622)	0.234
RUNG-L1	0	0.705	72	7/72 (0.097)	41/72 (0.569)	57/72 (0.792)	0.415
	1	0.699	73	6/73 (0.082)	39/73 (0.534)	58/73 (0.795)	0.397
	2	0.699	74	4/74 (0.054)	42/74 (0.568)	59/74 (0.797)	0.403
SoftMedian	0	0.657	69	19/69 (0.275)	56/69 (0.812)	63/69 (0.913)	0.590
	1	0.655	72	24/72 (0.333)	60/72 (0.833)	65/72 (0.903)	0.614
	2	0.647	71	21/71 (0.296)	58/71 (0.817)	65/71 (0.915)	0.599
EvenNet	0	0.702	75	33/75 (0.440)	60/75 (0.800)	67/75 (0.893)	0.631
	1	0.692	75	34/75 (0.453)	62/75 (0.827)	68/75 (0.907)	0.648
	2	0.701	75	36/75 (0.480)	60/75 (0.800)	68/75 (0.907)	0.645
HARA-Mean	0	0.602	61	17/61 (0.279)	41/61 (0.672)	44/61 (0.721)	0.497
	1	0.623	67	16/67 (0.239)	36/67 (0.537)	51/67 (0.761)	0.439
	2	0.580	63	19/63 (0.302)	40/63 (0.635)	46/63 (0.730)	0.490
HARA-Median	0	0.600	62	13/62 (0.210)	38/62 (0.613)	45/62 (0.726)	0.453
	1	0.595	68	14/68 (0.206)	30/68 (0.441)	43/68 (0.632)	0.365
	2	0.608	67	16/67 (0.239)	41/67 (0.612)	50/67 (0.746)	0.466
GCN+CLASP	0	0.535	56	2/56 (0.036)	10/56 (0.179)	16/56 (0.286)	0.139
	1	0.534	63	4/63 (0.063)	6/63 (0.095)	10/63 (0.159)	0.089
	2	0.512	57	0/57 (0.000)	11/57 (0.193)	22/57 (0.386)	0.154
HARA	0	0.670	73	3/73 (0.041)	14/73 (0.192)	19/73 (0.260)	0.141
	1	0.607	67	4/67 (0.060)	8/67 (0.119)	10/67 (0.149)	0.096
	2	0.602	66	1/66 (0.015)	9/66 (0.136)	13/66 (0.197)	0.098

K.3 CHAMELEON-F

Table 82: Complete registered method results on Chameleon-F. Lower ASR, end to end error, and RAUC are better.

Method	Clean acc.	$n_0/n_1/n_2$	ASR@1	ASR@3	ASR@5	E2E@5	RAUC
GCN	0.348 ± 0.009	32/35/34	0.781 ± 0.055	0.902 ± 0.064	0.931 ± 0.015	0.977 ± 0.006	0.781 ± 0.029
RUNG-MCP	0.309 ± 0.006	28/30/28	0.023 ± 0.020	0.082 ± 0.022	0.163 ± 0.026	0.753 ± 0.021	0.072 ± 0.015
RUNG-L1	0.311 ± 0.003	29/30/28	0.023 ± 0.020	0.115 ± 0.020	0.161 ± 0.025	0.750 ± 0.017	0.085 ± 0.017
SoftMedian	0.374 ± 0.006	37/40/37	0.068 ± 0.071	0.184 ± 0.019	0.298 ± 0.081	0.727 ± 0.040	0.154 ± 0.035
EvenNet	0.333 ± 0.023	29/34/30	0.313 ± 0.028	0.655 ± 0.048	0.762 ± 0.054	0.927 ± 0.015	0.509 ± 0.035
HARA-Mean	0.380 ± 0.028	41/32/40	0.095 ± 0.048	0.282 ± 0.082	0.459 ± 0.210	0.727 ± 0.090	0.233 ± 0.064
HARA-Median	0.371 ± 0.021	44/34/43	0.118 ± 0.062	0.171 ± 0.070	0.229 ± 0.058	0.663 ± 0.050	0.150 ± 0.055
GCN+CLASP	0.369 ± 0.022	34/38/43	0.045 ± 0.044	0.069 ± 0.037	0.076 ± 0.035	0.610 ± 0.017	0.056 ± 0.031
HARA	0.365 ± 0.023	38/32/37	0.065 ± 0.014	0.177 ± 0.018	0.187 ± 0.003	0.670 ± 0.030	0.128 ± 0.010

To interpret these results, read clean accuracy and overall error alongside conditional ASR. The first two measure utility on the full target manifest. ASR measures failures among each seed's initially correct targets.

Table 83: Results for every registered method and seed on Chameleon-F. Parentheses contain seed level conditional ASR.

Method	Seed	Clean acc.	Clean n	Success@1/n (ASR)	Success@3/n (ASR)	Success@5/n (ASR)	RAUC
GCN	0	0.346	32	23/32 (0.719)	30/32 (0.938)	30/32 (0.938)	0.778
	1	0.358	35	28/35 (0.800)	29/35 (0.829)	32/35 (0.914)	0.754
	2	0.341	34	28/34 (0.824)	32/34 (0.941)	32/34 (0.941)	0.812
RUNG-MCP	0	0.302	28	1/28 (0.036)	3/28 (0.107)	5/28 (0.179)	0.089
	1	0.313	30	1/30 (0.033)	2/30 (0.067)	4/30 (0.133)	0.063
	2	0.313	28	0/28 (0.000)	2/28 (0.071)	5/28 (0.179)	0.064
RUNG-L1	0	0.307	29	1/29 (0.034)	4/29 (0.138)	5/29 (0.172)	0.100
	1	0.313	30	0/30 (0.000)	3/30 (0.100)	4/30 (0.133)	0.067
	2	0.313	28	1/28 (0.036)	3/28 (0.107)	5/28 (0.179)	0.089
SoftMedian	0	0.369	37	1/37 (0.027)	7/37 (0.189)	14/37 (0.378)	0.159
	1	0.380	40	6/40 (0.150)	8/40 (0.200)	12/40 (0.300)	0.185
	2	0.374	37	1/37 (0.027)	6/37 (0.162)	8/37 (0.216)	0.116
EvenNet	0	0.313	29	10/29 (0.345)	20/29 (0.690)	23/29 (0.793)	0.538
	1	0.358	34	10/34 (0.294)	23/34 (0.676)	27/34 (0.794)	0.518
	2	0.330	30	9/30 (0.300)	18/30 (0.600)	21/30 (0.700)	0.470
HARA-Mean	0	0.408	41	3/41 (0.073)	9/41 (0.220)	17/41 (0.415)	0.193
	1	0.352	32	2/32 (0.062)	12/32 (0.375)	22/32 (0.688)	0.306
	2	0.380	40	6/40 (0.150)	10/40 (0.250)	11/40 (0.275)	0.200
HARA-Median	0	0.380	44	7/44 (0.159)	11/44 (0.250)	13/44 (0.295)	0.207
	1	0.346	34	5/34 (0.147)	5/34 (0.147)	7/34 (0.206)	0.144
	2	0.385	43	2/43 (0.047)	5/43 (0.116)	8/43 (0.186)	0.098
GCN+CLASP	0	0.346	34	3/34 (0.088)	3/34 (0.088)	2/34 (0.059)	0.074
	1	0.369	38	0/38 (0.000)	1/38 (0.026)	2/38 (0.053)	0.021
	2	0.391	43	2/43 (0.047)	4/43 (0.093)	5/43 (0.116)	0.074
HARA	0	0.391	38	2/38 (0.053)	7/38 (0.184)	7/38 (0.184)	0.126
	1	0.358	32	2/32 (0.062)	5/32 (0.156)	6/32 (0.188)	0.119
	2	0.346	37	3/37 (0.081)	7/37 (0.189)	7/37 (0.189)	0.138

K.4 SQUIRREL-F

Table 84: Complete registered method results on Squirrel-F. Lower ASR, end to end error, and RAUC are better.

Method	Clean acc.	$n_0/n_1/n_2$	ASR@1	ASR@3	ASR@5	E2E@5	RAUC
GCN	0.339 ± 0.014	30/34/33	0.721 ± 0.038	0.824 ± 0.024	0.908 ± 0.051	0.970 ± 0.017	0.727 ± 0.017
RUNG-MCP	0.325 ± 0.003	29/29/31	0.045 ± 0.021	0.169 ± 0.035	0.281 ± 0.027	0.787 ± 0.015	0.137 ± 0.016
RUNG-L1	0.327 ± 0.003	29/29/32	0.055 ± 0.018	0.243 ± 0.037	0.343 ± 0.032	0.803 ± 0.006	0.183 ± 0.026
SoftMedian	0.324 ± 0.003	30/30/29	0.045 ± 0.019	0.090 ± 0.018	0.179 ± 0.067	0.747 ± 0.023	0.085 ± 0.023
EvenNet	0.345 ± 0.010	36/32/32	0.728 ± 0.046	0.809 ± 0.026	0.840 ± 0.006	0.947 ± 0.006	0.710 ± 0.023
HARA-Mean	0.368 ± 0.018	32/35/38	0.179 ± 0.089	0.557 ± 0.118	0.752 ± 0.010	0.843 ± 0.012	0.427 ± 0.063
HARA-Median	0.368 ± 0.003	32/36/33	0.138 ± 0.013	0.225 ± 0.077	0.403 ± 0.146	0.740 ± 0.070	0.212 ± 0.057
GCN+CLASP	0.336 ± 0.007	33/31/34	0.062 ± 0.034	0.071 ± 0.046	0.070 ± 0.076	0.670 ± 0.030	0.061 ± 0.031
HARA	0.390 ± 0.003	36/35/32	0.038 ± 0.042	0.124 ± 0.141	0.159 ± 0.223	0.657 ± 0.047	0.093 ± 0.112

To interpret these results, read clean accuracy and overall error alongside conditional ASR. The first two measure utility on the full target manifest. ASR measures failures among each seed's initially correct targets.

Table 85: Results for every registered method and seed on Squirrel-F. Parentheses contain seed level conditional ASR.

Method	Seed	Clean acc.	Clean n	Success@1/n (ASR)	Success@3/n (ASR)	Success@5/n (ASR)	RAUC
GCN	0	0.330	30	21/30 (0.700)	24/30 (0.800)	28/30 (0.933)	0.717
	1	0.330	34	26/34 (0.765)	28/34 (0.824)	32/34 (0.941)	0.747
	2	0.355	33	23/33 (0.697)	28/33 (0.848)	28/33 (0.848)	0.718
RUNG-MCP	0	0.328	29	1/29 (0.034)	6/29 (0.207)	9/29 (0.310)	0.155
	1	0.324	29	2/29 (0.069)	4/29 (0.138)	8/29 (0.276)	0.131
	2	0.324	31	1/31 (0.032)	5/31 (0.161)	8/31 (0.258)	0.126
RUNG-L1	0	0.330	29	1/29 (0.034)	6/29 (0.207)	9/29 (0.310)	0.155
	1	0.324	29	2/29 (0.069)	7/29 (0.241)	10/29 (0.345)	0.186
	2	0.328	32	2/32 (0.062)	9/32 (0.281)	12/32 (0.375)	0.206
SoftMedian	0	0.324	30	1/30 (0.033)	3/30 (0.100)	7/30 (0.233)	0.097
	1	0.328	30	2/30 (0.067)	3/30 (0.100)	6/30 (0.200)	0.100
	2	0.321	29	1/29 (0.034)	2/29 (0.069)	3/29 (0.103)	0.059
EvenNet	0	0.357	36	28/36 (0.778)	30/36 (0.833)	30/36 (0.833)	0.733
	1	0.339	32	22/32 (0.688)	25/32 (0.781)	27/32 (0.844)	0.688
	2	0.339	32	23/32 (0.719)	26/32 (0.812)	27/32 (0.844)	0.709
HARA-Mean	0	0.360	32	6/32 (0.188)	22/32 (0.688)	24/32 (0.750)	0.481
	1	0.355	35	3/35 (0.086)	16/35 (0.457)	26/35 (0.743)	0.357
	2	0.389	38	10/38 (0.263)	20/38 (0.526)	29/38 (0.763)	0.442
HARA-Median	0	0.364	32	4/32 (0.125)	7/32 (0.219)	14/32 (0.438)	0.212
	1	0.371	36	5/36 (0.139)	11/36 (0.306)	19/36 (0.528)	0.269
	2	0.369	33	5/33 (0.152)	5/33 (0.152)	8/33 (0.242)	0.155
GCN+CLASP	0	0.330	33	2/33 (0.061)	4/33 (0.121)	5/33 (0.152)	0.097
	1	0.344	31	3/31 (0.097)	1/31 (0.032)	0/31 (0.000)	0.042
	2	0.333	34	1/34 (0.029)	2/34 (0.059)	2/34 (0.059)	0.044
HARA	0	0.389	36	3/36 (0.083)	10/36 (0.278)	15/36 (0.417)	0.219
	1	0.389	35	0/35 (0.000)	0/35 (0.000)	1/35 (0.029)	0.006
	2	0.393	32	1/32 (0.031)	3/32 (0.094)	1/32 (0.031)	0.053

K.5 ACTOR

Table 86: Complete registered method results on Actor. Lower ASR, end to end error, and RAUC are better.

Method	Clean acc.	$n_0/n_1/n_2$	ASR@1	ASR@3	ASR@5	E2E@5	RAUC
GCN	0.306 ± 0.001	34/32/30	0.781 ± 0.014	0.969 ± 0.002	0.969 ± 0.002	0.990 ± 0.000	0.815 ± 0.005
RUNG-MCP	0.316 ± 0.004	34/30/29	0.119 ± 0.019	0.269 ± 0.030	0.452 ± 0.013	0.803 ± 0.023	0.233 ± 0.009
RUNG-L1	0.309 ± 0.006	28/29/26	0.352 ± 0.062	0.685 ± 0.060	0.771 ± 0.014	0.933 ± 0.006	0.534 ± 0.006
SoftMedian	0.284 ± 0.001	31/29/25	0.242 ± 0.071	0.349 ± 0.060	0.410 ± 0.037	0.770 ± 0.010	0.294 ± 0.048
EvenNet	0.370 ± 0.014	38/37/32	0.159 ± 0.025	0.274 ± 0.065	0.248 ± 0.084	0.713 ± 0.049	0.207 ± 0.044
HARA-Mean	0.357 ± 0.012	36/36/35	0.150 ± 0.034	0.159 ± 0.043	0.168 ± 0.056	0.673 ± 0.015	0.142 ± 0.038
HARA-Median	0.366 ± 0.022	31/31/40	0.072 ± 0.075	0.091 ± 0.088	0.150 ± 0.081	0.673 ± 0.051	0.088 ± 0.074
GCN+CLASP	0.264 ± 0.006	27/23/33	0.084 ± 0.009	0.264 ± 0.007	0.338 ± 0.070	0.783 ± 0.023	0.199 ± 0.018
HARA	0.352 ± 0.006	34/38/36	0.131 ± 0.049	0.160 ± 0.067	0.189 ± 0.096	0.690 ± 0.046	0.141 ± 0.060

To interpret these results, read clean accuracy and overall error alongside conditional ASR. The first two measure utility on the full target manifest. ASR measures failures among each seed's initially correct targets.

Table 87: Results for every registered method and seed on Actor. Parentheses contain seed level conditional ASR.

Method	Seed	Clean acc.	Clean n	Success@1/n (ASR)	Success@3/n (ASR)	Success@5/n (ASR)	RAUC
GCN	0	0.306	34	27/34 (0.794)	33/34 (0.971)	33/34 (0.971)	0.821
	1	0.305	32	25/32 (0.781)	31/32 (0.969)	31/32 (0.969)	0.816
	2	0.307	30	23/30 (0.767)	29/30 (0.967)	29/30 (0.967)	0.810
RUNG-MCP	0	0.316	34	4/34 (0.118)	9/34 (0.265)	15/34 (0.441)	0.229
	1	0.320	30	3/30 (0.100)	9/30 (0.300)	14/30 (0.467)	0.243
	2	0.312	29	4/29 (0.138)	7/29 (0.241)	13/29 (0.448)	0.228
RUNG-L1	0	0.314	28	9/28 (0.321)	20/28 (0.714)	22/28 (0.786)	0.539
	1	0.302	29	9/29 (0.310)	21/29 (0.724)	22/29 (0.759)	0.534
	2	0.309	26	11/26 (0.423)	16/26 (0.615)	20/26 (0.769)	0.527
SoftMedian	0	0.285	31	9/31 (0.290)	12/31 (0.387)	14/31 (0.452)	0.332
	1	0.283	29	8/29 (0.276)	11/29 (0.379)	11/29 (0.379)	0.310
	2	0.283	25	4/25 (0.160)	7/25 (0.280)	10/25 (0.400)	0.240
EvenNet	0	0.382	38	7/38 (0.184)	10/38 (0.263)	8/38 (0.211)	0.203
	1	0.375	37	5/37 (0.135)	8/37 (0.216)	7/37 (0.189)	0.165
	2	0.354	32	5/32 (0.156)	11/32 (0.344)	11/32 (0.344)	0.253
HARA-Mean	0	0.350	36	6/36 (0.167)	7/36 (0.194)	8/36 (0.222)	0.172
	1	0.372	36	4/36 (0.111)	4/36 (0.111)	4/36 (0.111)	0.100
	2	0.351	35	6/35 (0.171)	6/35 (0.171)	6/35 (0.171)	0.154
HARA-Median	0	0.348	31	2/31 (0.065)	3/31 (0.097)	5/31 (0.161)	0.090
	1	0.361	31	0/31 (0.000)	0/31 (0.000)	2/31 (0.065)	0.013
	2	0.391	40	6/40 (0.150)	7/40 (0.175)	9/40 (0.225)	0.160
GCN+CLASP	0	0.270	27	2/27 (0.074)	7/27 (0.259)	7/27 (0.259)	0.178
	1	0.264	23	2/23 (0.087)	6/23 (0.261)	9/23 (0.391)	0.209
	2	0.258	33	3/33 (0.091)	9/33 (0.273)	12/33 (0.364)	0.209
HARA	0	0.358	34	6/34 (0.176)	8/34 (0.235)	10/34 (0.294)	0.206
	1	0.351	38	3/38 (0.079)	4/38 (0.105)	4/38 (0.105)	0.087
	2	0.346	36	5/36 (0.139)	5/36 (0.139)	6/36 (0.167)	0.131

K.6 ROMAN-EMPIRE

Table 88: Complete registered method results on Roman-Empire. Lower ASR, end to end error, and RAUC are better.

Method	Clean acc.	$n_0/n_1/n_2$	ASR@1	ASR@3	ASR@5	E2E@5	RAUC
GCN	0.463 ± 0.004	47/45/49	0.532 ± 0.037	0.930 ± 0.030	0.986 ± 0.012	0.993 ± 0.006	0.729 ± 0.024
RUNG-MCP	0.581 ± 0.003	63/64/63	0.005 ± 0.009	0.016 ± 0.016	0.063 ± 0.001	0.397 ± 0.006	0.021 ± 0.007
RUNG-L1	0.599 ± 0.002	64/65/65	0.072 ± 0.017	0.206 ± 0.022	0.350 ± 0.016	0.550 ± 0.010	0.174 ± 0.010
SoftMedian	0.431 ± 0.005	40/40/41	0.132 ± 0.015	0.264 ± 0.056	0.686 ± 0.013	0.837 ± 0.012	0.283 ± 0.020
EvenNet	0.672 ± 0.004	71/73/72	0.088 ± 0.015	0.185 ± 0.019	0.296 ± 0.017	0.483 ± 0.006	0.160 ± 0.014
HARA-Mean	0.764 ± 0.008	75/80/79	0.150 ± 0.051	0.341 ± 0.037	0.423 ± 0.034	0.520 ± 0.035	0.266 ± 0.033
HARA-Median	0.759 ± 0.005	77/76/72	0.213 ± 0.032	0.333 ± 0.024	0.337 ± 0.027	0.463 ± 0.012	0.264 ± 0.019
GCN+CLASP	0.282 ± 0.001	26/25/25	0.119 ± 0.042	0.448 ± 0.029	0.671 ± 0.043	0.867 ± 0.006	0.349 ± 0.023
HARA	0.698 ± 0.006	75/72/72	0.077 ± 0.020	0.251 ± 0.087	0.347 ± 0.153	0.503 ± 0.115	0.193 ± 0.070

To interpret these results, read clean accuracy and overall error alongside conditional ASR. The first two measure utility on the full target manifest. ASR measures failures among each seed's initially correct targets.

Table 89: Results for every registered method and seed on Roman-Empire. Parentheses contain seed level conditional ASR.

Method	Seed	Clean acc.	Clean n	Success@1/n (ASR)	Success@3/n (ASR)	Success@5/n (ASR)	RAUC
GCN	0	0.459	47	27/47 (0.574)	45/47 (0.957)	47/47 (1.000)	0.755
	1	0.463	45	23/45 (0.511)	42/45 (0.933)	44/45 (0.978)	0.722
	2	0.468	49	25/49 (0.510)	44/49 (0.898)	48/49 (0.980)	0.708
RUNG-MCP	0	0.581	63	0/63 (0.000)	0/63 (0.000)	4/63 (0.063)	0.013
	1	0.578	64	0/64 (0.000)	2/64 (0.031)	4/64 (0.062)	0.025
	2	0.583	63	1/63 (0.016)	1/63 (0.016)	4/63 (0.063)	0.024
RUNG-L1	0	0.600	64	4/64 (0.062)	12/64 (0.188)	22/64 (0.344)	0.163
	1	0.596	65	4/65 (0.062)	15/65 (0.231)	22/65 (0.338)	0.178
	2	0.600	65	6/65 (0.092)	13/65 (0.200)	24/65 (0.369)	0.182
SoftMedian	0	0.428	40	6/40 (0.150)	8/40 (0.200)	27/40 (0.675)	0.260
	1	0.427	40	5/40 (0.125)	12/40 (0.300)	28/40 (0.700)	0.297
	2	0.437	41	5/41 (0.122)	12/41 (0.293)	28/41 (0.683)	0.290
EvenNet	0	0.669	71	5/71 (0.070)	12/71 (0.169)	20/71 (0.282)	0.145
	1	0.677	73	7/73 (0.096)	15/73 (0.205)	23/73 (0.315)	0.174
	2	0.671	72	7/72 (0.097)	13/72 (0.181)	21/72 (0.292)	0.160
HARA-Mean	0	0.760	75	11/75 (0.147)	23/75 (0.307)	32/75 (0.427)	0.252
	1	0.773	80	8/80 (0.100)	27/80 (0.338)	31/80 (0.388)	0.243
	2	0.760	79	16/79 (0.203)	30/79 (0.380)	36/79 (0.456)	0.304
HARA-Median	0	0.760	77	19/77 (0.247)	27/77 (0.351)	27/77 (0.351)	0.284
	1	0.762	76	14/76 (0.184)	26/76 (0.342)	27/76 (0.355)	0.263
	2	0.753	72	15/72 (0.208)	22/72 (0.306)	22/72 (0.306)	0.246
GCN+CLASP	0	0.282	26	2/26 (0.077)	11/26 (0.423)	17/26 (0.654)	0.323
	1	0.282	25	3/25 (0.120)	12/25 (0.480)	16/25 (0.640)	0.356
	2	0.283	25	4/25 (0.160)	11/25 (0.440)	18/25 (0.720)	0.368
HARA	0	0.702	75	7/75 (0.093)	21/75 (0.280)	26/75 (0.347)	0.209
	1	0.691	72	4/72 (0.056)	11/72 (0.153)	14/72 (0.194)	0.117
	2	0.700	72	6/72 (0.083)	23/72 (0.319)	36/72 (0.500)	0.253

K.7 AMAZON-RATINGS

Table 90: Complete registered method results on Amazon-Ratings. Lower ASR, end to end error, and RAUC are better.

Method	Clean acc.	$n_0/n_1/n_2$	ASR@1	ASR@3	ASR@5	E2E@5	RAUC
GCN	0.432 ± 0.004	41/42/43	0.411 ± 0.087	0.730 ± 0.048	0.818 ± 0.076	0.923 ± 0.032	0.579 ± 0.045
RUNG-MCP	0.419 ± 0.006	42/40/39	0.041 ± 0.037	0.065 ± 0.036	0.189 ± 0.032	0.673 ± 0.006	0.076 ± 0.024
RUNG-L1	0.421 ± 0.006	41/43/39	0.041 ± 0.016	0.097 ± 0.048	0.202 ± 0.042	0.673 ± 0.006	0.091 ± 0.025
SoftMedian	0.440 ± 0.003	45/42/43	0.046 ± 0.021	0.168 ± 0.043	0.307 ± 0.045	0.700 ± 0.020	0.143 ± 0.023
EvenNet	0.434 ± 0.002	40/41/44	0.351 ± 0.023	0.608 ± 0.020	0.624 ± 0.034	0.840 ± 0.010	0.473 ± 0.011
HARA-Mean	0.425 ± 0.003	39/42/37	0.325 ± 0.073	0.571 ± 0.091	0.570 ± 0.071	0.803 ± 0.025	0.440 ± 0.071
HARA-Median	0.420 ± 0.005	42/44/38	0.181 ± 0.071	0.358 ± 0.093	0.422 ± 0.067	0.743 ± 0.015	0.282 ± 0.069
GCN+CLASP	0.434 ± 0.002	40/38/37	0.018 ± 0.031	0.018 ± 0.031	0.035 ± 0.016	0.603 ± 0.015	0.020 ± 0.025
HARA	0.397 ± 0.023	40/40/37	0.050 ± 0.087	0.092 ± 0.088	0.075 ± 0.066	0.613 ± 0.015	0.067 ± 0.068

To interpret these results, read clean accuracy and overall error alongside conditional ASR. The first two measure utility on the full target manifest. ASR measures failures among each seed's initially correct targets.

Table 91: Results for every registered method and seed on Amazon-Ratings. Parentheses contain seed level conditional ASR.

Method	Seed	Clean acc.	Clean n	Success@1/n (ASR)	Success@3/n (ASR)	Success@5/n (ASR)	RAUC
GCN	0	0.434	41	13/41 (0.317)	29/41 (0.707)	32/41 (0.780)	0.534
	1	0.434	42	18/42 (0.429)	33/42 (0.786)	38/42 (0.905)	0.624
	2	0.427	43	21/43 (0.488)	30/43 (0.698)	33/43 (0.767)	0.579
RUNG-MCP	0	0.424	42	3/42 (0.071)	4/42 (0.095)	9/42 (0.214)	0.102
	1	0.421	40	0/40 (0.000)	3/40 (0.075)	8/40 (0.200)	0.070
	2	0.412	39	2/39 (0.051)	1/39 (0.026)	6/39 (0.154)	0.056
RUNG-L1	0	0.424	41	2/41 (0.049)	6/41 (0.146)	9/41 (0.220)	0.117
	1	0.425	43	1/43 (0.023)	4/43 (0.093)	10/43 (0.233)	0.091
	2	0.413	39	2/39 (0.051)	2/39 (0.051)	6/39 (0.154)	0.067
SoftMedian	0	0.443	45	3/45 (0.067)	9/45 (0.200)	15/45 (0.333)	0.167
	1	0.441	42	1/42 (0.024)	5/42 (0.119)	14/42 (0.333)	0.121
	2	0.437	43	2/43 (0.047)	8/43 (0.186)	11/43 (0.256)	0.140
EvenNet	0	0.436	40	13/40 (0.325)	25/40 (0.625)	26/40 (0.650)	0.478
	1	0.433	41	15/41 (0.366)	24/41 (0.585)	24/41 (0.585)	0.461
	2	0.433	44	16/44 (0.364)	27/44 (0.614)	28/44 (0.636)	0.482
HARA-Mean	0	0.428	39	12/39 (0.308)	20/39 (0.513)	20/39 (0.513)	0.400
	1	0.421	42	11/42 (0.262)	22/42 (0.524)	23/42 (0.548)	0.398
	2	0.426	37	15/37 (0.405)	25/37 (0.676)	24/37 (0.649)	0.522
HARA-Median	0	0.425	42	6/42 (0.143)	11/42 (0.262)	16/42 (0.381)	0.224
	1	0.414	44	6/44 (0.136)	16/44 (0.364)	17/44 (0.386)	0.264
	2	0.421	38	10/38 (0.263)	17/38 (0.447)	19/38 (0.500)	0.358
GCN+CLASP	0	0.436	40	0/40 (0.000)	0/40 (0.000)	1/40 (0.025)	0.005
	1	0.434	38	0/38 (0.000)	0/38 (0.000)	1/38 (0.026)	0.005
	2	0.431	37	2/37 (0.054)	2/37 (0.054)	2/37 (0.054)	0.049
HARA	0	0.410	40	0/40 (0.000)	4/40 (0.100)	5/40 (0.125)	0.065
	1	0.412	40	6/40 (0.150)	7/40 (0.175)	4/40 (0.100)	0.135
	2	0.370	37	0/37 (0.000)	0/37 (0.000)	0/37 (0.000)	0.000

L NINE WAY COMMON CLEAN PAIRED RESULTS

For each dataset and seed, the intersection contains targets classified correctly by all nine registered methods. The tables therefore hold the evaluated target population fixed. Together with method specific conditional ASR, they separate a shared target comparison from each model's full initially correct set.

L.1 CORA

Table 92: Nine way common clean paired conditional ASR on Cora. Denominators are ordered by seed.

Method	Common $n_0/n_1/n_2$	Paired ASR@1	Paired ASR@3	Paired ASR@5
GCN	49/45/50	0.236 ± 0.012	0.799 ± 0.018	0.972 ± 0.011
RUNG-MCP	49/45/50	0.000 ± 0.000	0.062 ± 0.033	0.179 ± 0.044
RUNG-L1	49/45/50	0.000 ± 0.000	0.200 ± 0.054	0.443 ± 0.066
SoftMedian	49/45/50	0.061 ± 0.034	0.486 ± 0.036	0.797 ± 0.036
EvenNet	49/45/50	0.056 ± 0.031	0.548 ± 0.041	0.888 ± 0.019
HARA-Mean	49/45/50	0.083 ± 0.018	0.396 ± 0.067	0.722 ± 0.029
HARA-Median	49/45/50	0.049 ± 0.025	0.459 ± 0.030	0.783 ± 0.045
GCN+CLASP	49/45/50	0.007 ± 0.012	0.110 ± 0.019	0.296 ± 0.064
HARA	49/45/50	0.000 ± 0.000	0.090 ± 0.020	0.243 ± 0.022

L.2 CITESEER

Table 93: Nine way common clean paired conditional ASR on CiteSeer. Denominators are ordered by seed.

Method	Common $n_0/n_1/n_2$	Paired ASR@1	Paired ASR@3	Paired ASR@5
GCN	39/45/38	0.465 ± 0.051	0.892 ± 0.022	0.917 ± 0.020
RUNG-MCP	39/45/38	0.016 ± 0.014	0.105 ± 0.047	0.466 ± 0.027
RUNG-L1	39/45/38	0.024 ± 0.022	0.428 ± 0.031	0.696 ± 0.033
SoftMedian	39/45/38	0.162 ± 0.035	0.761 ± 0.022	0.885 ± 0.015
EvenNet	39/45/38	0.262 ± 0.023	0.702 ± 0.064	0.843 ± 0.026
HARA-Mean	39/45/38	0.212 ± 0.025	0.529 ± 0.075	0.639 ± 0.022
HARA-Median	39/45/38	0.145 ± 0.058	0.433 ± 0.106	0.603 ± 0.080
GCN+CLASP	39/45/38	0.023 ± 0.022	0.119 ± 0.084	0.235 ± 0.104
HARA	39/45/38	0.016 ± 0.014	0.092 ± 0.035	0.148 ± 0.013

L.3 CHAMELEON-F

Table 94: Nine way common clean paired conditional ASR on Chameleon-F. Denominators are ordered by seed.

Method	Common $n_0/n_1/n_2$	Paired ASR@1	Paired ASR@3	Paired ASR@5
GCN	14/16/17	0.873 ± 0.059	0.960 ± 0.035	0.960 ± 0.035
RUNG-MCP	14/16/17	0.000 ± 0.000	0.040 ± 0.035	0.084 ± 0.030
RUNG-L1	14/16/17	0.000 ± 0.000	0.040 ± 0.035	0.084 ± 0.030
SoftMedian	14/16/17	0.021 ± 0.036	0.060 ± 0.059	0.166 ± 0.090
EvenNet	14/16/17	0.172 ± 0.045	0.728 ± 0.072	0.859 ± 0.124
HARA-Mean	14/16/17	0.084 ± 0.030	0.230 ± 0.085	0.490 ± 0.175
HARA-Median	14/16/17	0.064 ± 0.006	0.105 ± 0.029	0.084 ± 0.030
GCN+CLASP	14/16/17	0.000 ± 0.000	0.024 ± 0.041	0.000 ± 0.000
HARA	14/16/17	0.000 ± 0.000	0.045 ± 0.039	0.021 ± 0.036

L.4 SQUIRREL-F

Table 95: Nine way common clean paired conditional ASR on Squirrel-F. Denominators are ordered by seed.

Method	Common $n_0/n_1/n_2$	Paired ASR@1	Paired ASR@3	Paired ASR@5
GCN	17/19/18	0.833 ± 0.009	0.833 ± 0.009	0.927 ± 0.085
RUNG-MCP	17/19/18	0.000 ± 0.000	0.131 ± 0.040	0.279 ± 0.067
RUNG-L1	17/19/18	0.000 ± 0.000	0.203 ± 0.024	0.370 ± 0.018
SoftMedian	17/19/18	0.000 ± 0.000	0.056 ± 0.003	0.167 ± 0.097
EvenNet	17/19/18	0.833 ± 0.009	0.833 ± 0.009	0.833 ± 0.009
HARA-Mean	17/19/18	0.074 ± 0.032	0.562 ± 0.180	0.870 ± 0.065
HARA-Median	17/19/18	0.000 ± 0.000	0.090 ± 0.109	0.349 ± 0.267
GCN+CLASP	17/19/18	0.038 ± 0.033	0.038 ± 0.033	0.038 ± 0.033
HARA	17/19/18	0.020 ± 0.034	0.077 ± 0.090	0.098 ± 0.170

L.5 ACTOR

Table 96: Nine way common clean paired conditional ASR on Actor. Denominators are ordered by seed.

Method	Common $n_0/n_1/n_2$	Paired ASR@1	Paired ASR@3	Paired ASR@5
GCN	5/5/2	0.133 ± 0.115	0.700 ± 0.173	0.700 ± 0.173
RUNG-MCP	5/5/2	0.067 ± 0.115	0.333 ± 0.306	0.333 ± 0.306
RUNG-L1	5/5/2	0.067 ± 0.115	0.267 ± 0.231	0.367 ± 0.153
SoftMedian	5/5/2	0.300 ± 0.173	0.000 ± 0.000	0.000 ± 0.000
EvenNet	5/5/2	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
HARA-Mean	5/5/2	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
HARA-Median	5/5/2	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
GCN+CLASP	5/5/2	0.067 ± 0.115	0.067 ± 0.115	0.067 ± 0.115
HARA	5/5/2	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

L.6 ROMAN-EMPIRE

Table 97: Nine way common clean paired conditional ASR on Roman-Empire. Denominators are ordered by seed.

Method	Common $n_0/n_1/n_2$	Paired ASR@1	Paired ASR@3	Paired ASR@5
GCN	9/10/10	0.307 ± 0.089	0.933 ± 0.058	1.000 ± 0.000
RUNG-MCP	9/10/10	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
RUNG-L1	9/10/10	0.000 ± 0.000	0.033 ± 0.058	0.137 ± 0.055
SoftMedian	9/10/10	0.274 ± 0.045	0.274 ± 0.045	0.752 ± 0.177
EvenNet	9/10/10	0.000 ± 0.000	0.000 ± 0.000	0.104 ± 0.006
HARA-Mean	9/10/10	0.037 ± 0.064	0.137 ± 0.055	0.244 ± 0.077
HARA-Median	9/10/10	0.000 ± 0.000	0.137 ± 0.055	0.137 ± 0.055
GCN+CLASP	9/10/10	0.033 ± 0.058	0.348 ± 0.083	0.585 ± 0.103
HARA	9/10/10	0.000 ± 0.000	0.070 ± 0.061	0.141 ± 0.122

L.7 AMAZON-RATINGS

Table 98: Nine way common clean paired conditional ASR on Amazon-Ratings. Denominators are ordered by seed.

Method	Common $n_0/n_1/n_2$	Paired ASR@1	Paired ASR@3	Paired ASR@5
GCN	23/20/12	0.243 ± 0.069	0.516 ± 0.164	0.665 ± 0.152
RUNG-MCP	23/20/12	0.000 ± 0.000	0.000 ± 0.000	0.149 ± 0.018
RUNG-L1	23/20/12	0.000 ± 0.000	0.000 ± 0.000	0.149 ± 0.018
SoftMedian	23/20/12	0.000 ± 0.000	0.028 ± 0.048	0.152 ± 0.061
EvenNet	23/20/12	0.138 ± 0.059	0.424 ± 0.151	0.452 ± 0.103
HARA-Mean	23/20/12	0.180 ± 0.018	0.376 ± 0.116	0.393 ± 0.124
HARA-Median	23/20/12	0.073 ± 0.020	0.268 ± 0.031	0.288 ± 0.105
GCN+CLASP	23/20/12	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
HARA	23/20/12	0.033 ± 0.058	0.079 ± 0.075	0.046 ± 0.044

M HARA SEED LEVEL NEUTRAL ATTACK RESULTS

The following long table exposes every HARA seed before any cross seed averaging. Each success count uses the displayed initially correct denominator.

Table 99: HARA seed level clean utility, attack outcomes, and RAUC.

Dataset	Seed	Clean acc.	Clean n	Success@1/n	Success@3/n	Success@5/n	RAUC
Cora	0	0.786	75	3/75 (0.040)	13/75 (0.173)	28/75 (0.373)	0.156
Cora	1	0.759	73	4/73 (0.055)	9/73 (0.123)	28/73 (0.384)	0.142
Cora	2	0.753	71	1/71 (0.014)	13/71 (0.183)	24/71 (0.338)	0.145
CiteSeer	0	0.670	73	3/73 (0.041)	14/73 (0.192)	19/73 (0.260)	0.141
CiteSeer	1	0.607	67	4/67 (0.060)	8/67 (0.119)	10/67 (0.149)	0.096
CiteSeer	2	0.602	66	1/66 (0.015)	9/66 (0.136)	13/66 (0.197)	0.098
Chameleon-F	0	0.391	38	2/38 (0.053)	7/38 (0.184)	7/38 (0.184)	0.126
Chameleon-F	1	0.358	32	2/32 (0.062)	5/32 (0.156)	6/32 (0.188)	0.119
Chameleon-F	2	0.346	37	3/37 (0.081)	7/37 (0.189)	7/37 (0.189)	0.138
Squirrel-F	0	0.389	36	3/36 (0.083)	10/36 (0.278)	15/36 (0.417)	0.219
Squirrel-F	1	0.389	35	0/35 (0.000)	0/35 (0.000)	1/35 (0.029)	0.006
Squirrel-F	2	0.393	32	1/32 (0.031)	3/32 (0.094)	1/32 (0.031)	0.053
Actor	0	0.358	34	6/34 (0.176)	8/34 (0.235)	10/34 (0.294)	0.206
Actor	1	0.351	38	3/38 (0.079)	4/38 (0.105)	4/38 (0.105)	0.087
Actor	2	0.346	36	5/36 (0.139)	5/36 (0.139)	6/36 (0.167)	0.131
Roman-Empire	0	0.702	75	7/75 (0.093)	21/75 (0.280)	26/75 (0.347)	0.209
Roman-Empire	1	0.691	72	4/72 (0.056)	11/72 (0.153)	14/72 (0.194)	0.117
Roman-Empire	2	0.700	72	6/72 (0.083)	23/72 (0.319)	36/72 (0.500)	0.253
Amazon-Ratings	0	0.410	40	0/40 (0.000)	4/40 (0.100)	5/40 (0.125)	0.065
Amazon-Ratings	1	0.412	40	6/40 (0.150)	7/40 (0.175)	4/40 (0.100)	0.135
Amazon-Ratings	2	0.370	37	0/37 (0.000)	0/37 (0.000)	0/37 (0.000)	0.000

N COMPLETE PRELIMINARY LABEL SENSITIVITY

The downstream HARA model is held fixed while the attack time label provider is intervened on. Oracle is therefore a diagnostic inference time input. Success/n is retained for transparent counting.

N.1 CORA

Table 100: Full attack time preliminary label sensitivity on Cora.

Mode	ASR@1	ASR@3	ASR@5	Success@1/n	Success@3/n	Success@5/n
Perturbed recomputation	0.036 ± 0.021	0.160 ± 0.032	0.365 ± 0.024	8/219	35/219	80/219
Fixed clean	0.082 ± 0.024	0.205 ± 0.012	0.433 ± 0.070	18/219	45/219	95/219
Random	0.045 ± 0.015	0.170 ± 0.047	0.332 ± 0.035	10/222	38/222	74/222
Oracle diagnostic	0.127 ± 0.028	0.277 ± 0.059	0.512 ± 0.029	27/213	59/213	109/213
Noise 10%	0.032 ± 0.008	0.163 ± 0.041	0.377 ± 0.010	7/220	36/220	83/220
Noise 20%	0.032 ± 0.008	0.174 ± 0.036	0.362 ± 0.034	7/218	38/218	79/218
Noise 30%	0.049 ± 0.006	0.191 ± 0.053	0.318 ± 0.040	11/223	43/223	71/223
Noise 50%	0.053 ± 0.013	0.185 ± 0.059	0.337 ± 0.051	12/225	42/225	76/225

N.2 ACTOR

Table 101: Full attack time preliminary label sensitivity on Actor.

Mode	ASR@1	ASR@3	ASR@5	Success@1/n	Success@3/n	Success@5/n
Perturbed recomputation	0.131 ± 0.049	0.160 ± 0.067	0.189 ± 0.096	14/108	17/108	20/108
Fixed clean	0.104 ± 0.067	0.140 ± 0.036	0.140 ± 0.036	11/108	15/108	15/108
Random	0.135 ± 0.043	0.191 ± 0.078	0.200 ± 0.082	15/111	21/111	22/111
Oracle diagnostic	0.091 ± 0.055	0.159 ± 0.028	0.208 ± 0.073	10/108	17/108	22/108
Noise 10%	0.115 ± 0.030	0.137 ± 0.082	0.176 ± 0.100	12/105	14/105	18/105
Noise 20%	0.096 ± 0.047	0.136 ± 0.092	0.165 ± 0.121	10/105	14/105	17/105
Noise 30%	0.092 ± 0.042	0.148 ± 0.090	0.185 ± 0.132	10/109	16/109	20/109
Noise 50%	0.088 ± 0.052	0.107 ± 0.045	0.145 ± 0.103	9/104	11/104	15/104

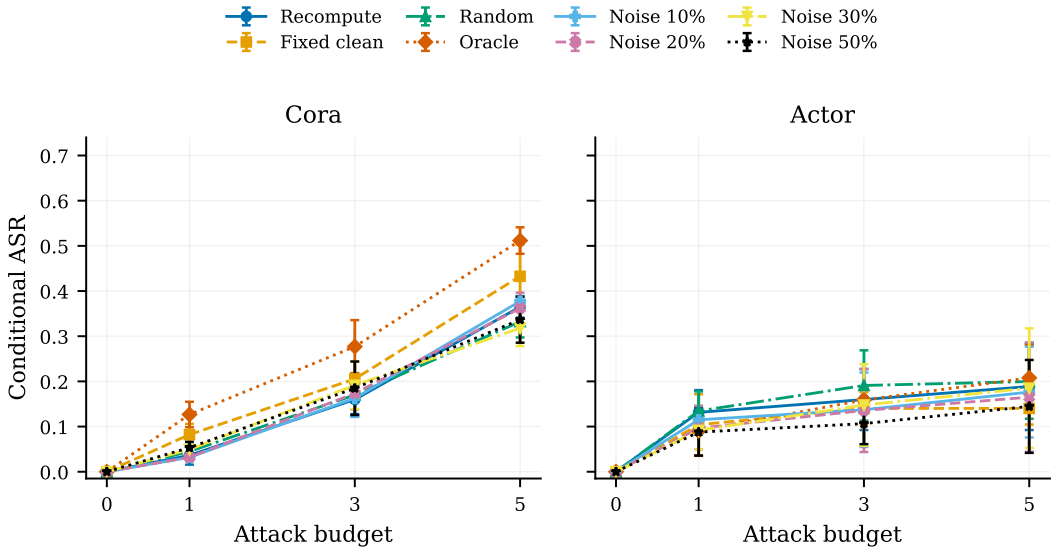


Figure 9: Sensitivity to preliminary label assignments under shared attacks on Cora and Actor. The eight inference modes use fixed HARA models, including recomputed predictions, fixed clean predictions, random labels, oracle labels, and four corruption levels. Error bars are sample standard deviations across three model seeds.

O PRELIMINARY CHECKPOINTS AND CLEAN CLASS PAIR BUCKETS

All 21 preliminary checkpoints updated during training and were then frozen. Accuracy is evaluated on the registered split; bucket counts are computed from the predicted ordered class pairs on the clean graph.

Table 102: Per seed preliminary classifier accuracy and bucket support.

Dataset	Seed	Train acc.	Val acc.	Test acc.	Empty	Nonempty ≤ 5	Min. nonempty
Cora	0	1.000	0.742	0.751	0	0	7
Cora	1	1.000	0.736	0.753	0	0	9
Cora	2	1.000	0.742	0.753	0	0	9
CiteSeer	0	1.000	0.662	0.656	0	0	13
CiteSeer	1	1.000	0.656	0.651	0	0	14
CiteSeer	2	1.000	0.658	0.653	0	0	13
Chameleon-F	0	0.768	0.423	0.369	0	0	37
Chameleon-F	1	0.773	0.423	0.363	0	0	37
Chameleon-F	2	0.773	0.426	0.363	0	0	37
Squirrel-F	0	0.738	0.368	0.373	0	0	346
Squirrel-F	1	0.739	0.367	0.371	0	0	346
Squirrel-F	2	0.738	0.368	0.373	0	0	346
Actor	0	0.426	0.300	0.291	0	1	5
Actor	1	0.425	0.297	0.294	0	0	7
Actor	2	0.426	0.299	0.290	0	0	6
Roman-Empire	0	0.392	0.369	0.367	139	23	1
Roman-Empire	1	0.393	0.370	0.365	139	27	1
Roman-Empire	2	0.393	0.369	0.366	139	27	1
Amazon-Ratings	0	0.391	0.389	0.395	16	2	3
Amazon-Ratings	1	0.391	0.390	0.395	16	2	3
Amazon-Ratings	2	0.390	0.391	0.394	16	2	3

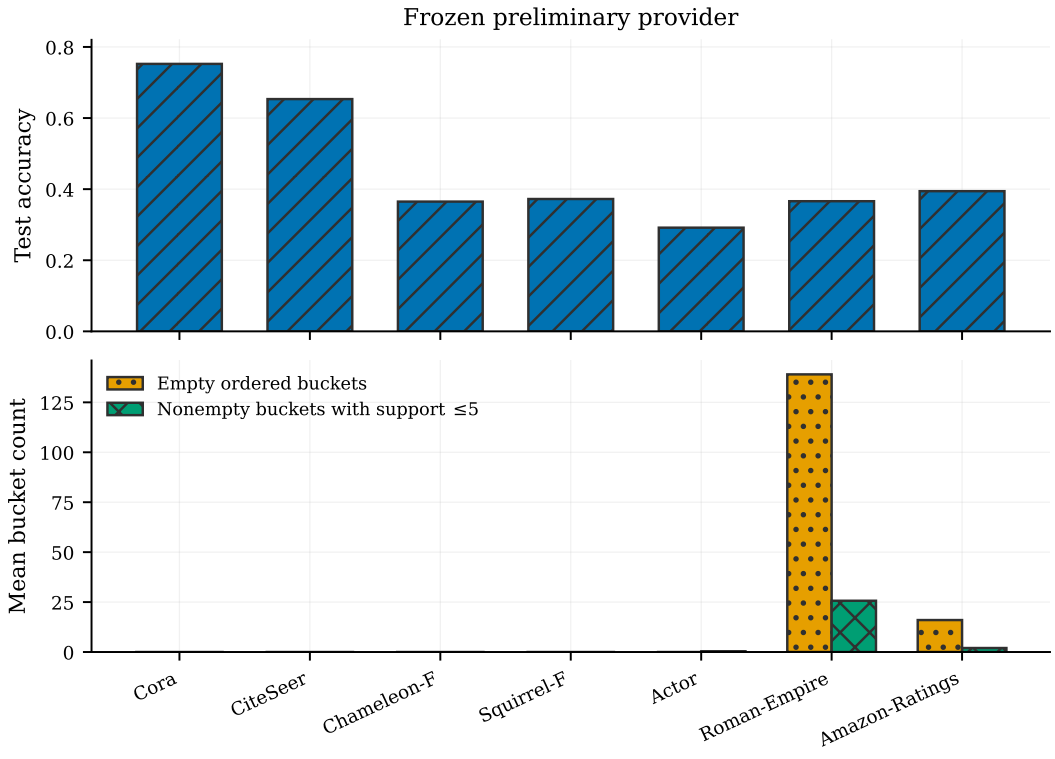


Figure 10: Mean test accuracy and clean ordered bucket census for the independently fitted preliminary classifier diagnostics across three seeds. The lower panel counts empty buckets and nonempty buckets with support at most five. These diagnostics describe the support available for class pair calibration.

P Full Seed by Support Diagnostic

Support is the minimum clean directed class pair bucket size touched by the registered budget five attack. Zero target cells are retained to expose where a bin is absent. Preliminary change rates provide target level descriptive associations.

P.1 CORA

Table 103: All seed by support cells on Cora. N/A denotes cells with no targets.

Seed	Support bin	Targets	Success	ASR@5	Prelim. changes	Change rate
0	0	0	0	N/A	0	N/A
0	1-5	0	0	N/A	0	N/A
0	6-50	25	7	0.280	23	0.920
0	51-500	43	16	0.372	40	0.930
0	501+	7	5	0.714	0	0.000
1	0	0	0	N/A	0	N/A
1	1-5	0	0	N/A	0	N/A
1	6-50	26	9	0.346	24	0.923
1	51-500	43	18	0.419	40	0.930
1	501+	4	1	0.250	1	0.250
2	0	0	0	N/A	0	N/A
2	1-5	0	0	N/A	0	N/A
2	6-50	24	8	0.333	23	0.958
2	51-500	42	14	0.333	39	0.929
2	501+	5	2	0.400	1	0.200

P.2 CITESEER

Table 104: All seed by support cells on CiteSeer. N/A denotes cells with no targets.

Seed	Support bin	Targets	Success	ASR@5	Prelim. changes	Change rate
0	0	0	0	N/A	0	N/A
0	1-5	0	0	N/A	0	N/A
0	6-50	42	14	0.333	40	0.952
0	51-500	29	5	0.172	27	0.931
0	501+	2	0	0.000	0	0.000
1	0	0	0	N/A	0	N/A
1	1-5	0	0	N/A	0	N/A
1	6-50	41	7	0.171	39	0.951
1	51-500	21	2	0.095	19	0.905
1	501+	5	1	0.200	1	0.200
2	0	0	0	N/A	0	N/A
2	1-5	0	0	N/A	0	N/A
2	6-50	41	5	0.122	39	0.951
2	51-500	23	6	0.261	22	0.957
2	501+	2	2	1.000	1	0.500

P.3 CHAMELEON-F

Table 105: All seed by support cells on Chameleon-F. N/A denotes cells with no targets.

Seed	Support bin	Targets	Success	ASR@5	Prelim. changes	Change rate
0	0	0	0	N/A	0	N/A
0	1-5	0	0	N/A	0	N/A
0	6-50	5	3	0.600	5	1.000
0	51-500	18	1	0.056	18	1.000
0	501+	15	3	0.200	9	0.600
1	0	0	0	N/A	0	N/A
1	1-5	0	0	N/A	0	N/A
1	6-50	1	1	1.000	1	1.000
1	51-500	20	4	0.200	20	1.000
1	501+	11	1	0.091	7	0.636
2	0	0	0	N/A	0	N/A
2	1-5	0	0	N/A	0	N/A
2	6-50	3	2	0.667	3	1.000
2	51-500	19	0	0.000	19	1.000
2	501+	15	5	0.333	11	0.733

P.4 SQUIRREL-F

Table 106: All seed by support cells on Squirrel-F. N/A denotes cells with no targets.

Seed	Support bin	Targets	Success	ASR@5	Prelim. changes	Change rate
0	0	0	0	N/A	0	N/A
0	1-5	0	0	N/A	0	N/A
0	6-50	0	0	N/A	0	N/A
0	51-500	4	3	0.750	4	1.000
0	501+	32	12	0.375	29	0.906
1	0	0	0	N/A	0	N/A
1	1-5	0	0	N/A	0	N/A
1	6-50	0	0	N/A	0	N/A
1	51-500	2	0	0.000	2	1.000
1	501+	33	1	0.030	28	0.848
2	0	0	0	N/A	0	N/A
2	1-5	0	0	N/A	0	N/A
2	6-50	0	0	N/A	0	N/A
2	51-500	4	0	0.000	4	1.000
2	501+	28	1	0.036	25	0.893

P.5 ACTOR

Table 107: All seed by support cells on Actor. N/A denotes cells with no targets.

Seed	Support bin	Targets	Success	ASR@5	Prelim. changes	Change rate
0	0	0	0	N/A	0	N/A
0	1-5	0	0	N/A	0	N/A
0	6-50	0	0	N/A	0	N/A
0	51-500	1	0	0.000	1	1.000
0	501+	33	10	0.303	24	0.727
1	0	0	0	N/A	0	N/A
1	1-5	0	0	N/A	0	N/A
1	6-50	0	0	N/A	0	N/A
1	51-500	1	0	0.000	1	1.000
1	501+	37	4	0.108	29	0.784
2	0	0	0	N/A	0	N/A
2	1-5	0	0	N/A	0	N/A
2	6-50	0	0	N/A	0	N/A
2	51-500	1	0	0.000	1	1.000
2	501+	35	6	0.171	29	0.829

P.6 ROMAN-EMPIRE

Table 108: All seed by support cells on Roman-Empire. N/A denotes cells with no targets.

Seed	Support bin	Targets	Success	ASR@5	Prelim. changes	Change rate
0	0	2	1	0.500	2	1.000
0	1-5	1	1	1.000	1	1.000
0	6-50	4	2	0.500	4	1.000
0	51-500	64	20	0.312	64	1.000
0	501+	4	2	0.500	3	0.750
1	0	2	1	0.500	2	1.000
1	1-5	1	1	1.000	1	1.000
1	6-50	3	0	0.000	3	1.000
1	51-500	62	12	0.194	60	0.968
1	501+	4	0	0.000	3	0.750
2	0	2	1	0.500	2	1.000
2	1-5	1	1	1.000	1	1.000
2	6-50	4	3	0.750	4	1.000
2	51-500	61	29	0.475	60	0.984
2	501+	4	2	0.500	3	0.750

P.7 AMAZON-RATINGS

Table 109: All seed by support cells on Amazon-Ratings. N/A denotes cells with no targets.

Seed	Support bin	Targets	Success	ASR@5	Prelim. changes	Change rate
0	0	0	0	N/A	0	N/A
0	1-5	0	0	N/A	0	N/A
0	6-50	0	0	N/A	0	N/A
0	51-500	10	0	0.000	0	0.000
0	501+	30	5	0.167	14	0.467
1	0	0	0	N/A	0	N/A
1	1-5	0	0	N/A	0	N/A
1	6-50	0	0	N/A	0	N/A
1	51-500	9	0	0.000	8	0.889
1	501+	31	4	0.129	13	0.419
2	0	0	0	N/A	0	N/A
2	1-5	0	0	N/A	0	N/A
2	6-50	0	0	N/A	0	N/A
2	51-500	10	0	0.000	1	0.100
2	501+	27	0	0.000	11	0.407

Q COMPLETE ADAPTIVE CANDIDATE SEARCH AUDIT

The aggregate table reports the same conditional and end to end metrics as the neutral protocol. The seed level audit then records accepted additions/deletions, discrete exact improvement ranges, preliminary label changes, and terminal states.

Table 110: Adaptive HARA aggregate outcomes at all registered budgets.

Dataset	Budget	$n_0/n_1/n_2$	ASR	Success/n	End to end error	RAUC
CiteSeer	1	73/67/66	0.306 ± 0.033	63/206	0.523 ± 0.032	0.347 ± 0.028
	3	73/67/66	0.413 ± 0.041	85/206	0.597 ± 0.038	0.347 ± 0.028
	5	73/67/66	0.452 ± 0.016	93/206	0.623 ± 0.031	0.347 ± 0.028
Squirrel-F	1	36/35/32	0.215 ± 0.089	22/103	0.730 ± 0.040	0.310 ± 0.101
	3	36/35/32	0.394 ± 0.140	41/103	0.793 ± 0.042	0.310 ± 0.101
	5	36/35/32	0.441 ± 0.174	46/103	0.810 ± 0.052	0.310 ± 0.101
Actor	1	34/38/36	0.158 ± 0.033	17/108	0.697 ± 0.023	0.206 ± 0.064
	3	34/38/36	0.262 ± 0.092	28/108	0.733 ± 0.046	0.206 ± 0.064
	5	34/38/36	0.272 ± 0.099	29/108	0.737 ± 0.049	0.206 ± 0.064

Table 111: Per seed adaptive search audit. Exact improvement and gradient columns show the minimum and maximum among accepted steps.

Dataset	Seed	Records	Add/del	Accepted	Exact improvement range	$\|g\|$ range	Prelim. changed	Prediction changed
CiteSeer	0	300	135/269	404	1.34e-05 to 6.83e+00	4.40e-03 to 1.80e+11	60	85
	1	300	103/282	385	2.38e-07 to 6.25e+00	3.92e-03 to 1.70e+11	48	73
	2	300	137/213	350	6.83e-03 to 6.29e+00	1.37e-02 to 7.60e+11	37	83
Squirrel-F	0	300	196/37	233	6.91e-05 to 1.23e+00	2.10e-04 to 1.03e+09	17	52
	1	300	234/17	251	9.97e-04 to 1.19e+00	1.25e-03 to 1.19e+09	23	28
	2	300	176/11	187	7.21e-04 to 1.69e+00	2.41e-04 to 4.18e+08	17	29
Actor	0	300	85/127	212	5.92e-05 to 1.10e+00	5.34e-03 to 5.02e+08	49	26
	1	300	121/147	268	8.42e-06 to 3.21e-01	2.00e-04 to 2.49e+08	60	17
	2	300	126/113	239	4.19e-04 to 2.61e+00	1.83e-03 to 2.92e+09	69	31

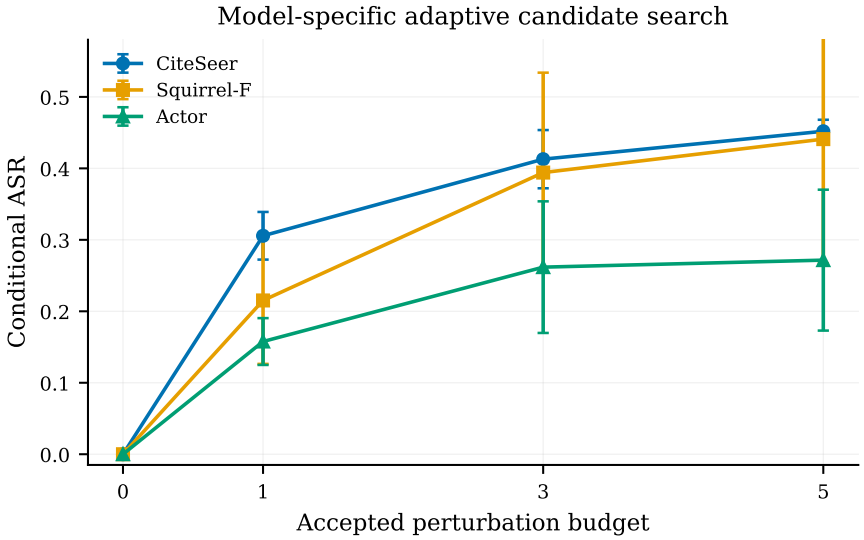


Figure 11: HARA conditional ASR under the bounded local top 20 candidate search with fixed configurations. Exact trials recompute preliminary labels and graph operations before accepting an edit. Error bars are sample standard deviations across three model seeds.

R Dataset Statistics

Table 112: Datasets in the shared attack evaluation. Edge counts are directed entries in the processed PyG graph; h is edge homophily.

Dataset	Nodes	Directed edges	Features	Classes	h
Cora	2,708	10,556	1,433	7	0.810
CiteSeer	3,327	9,104	3,703	6	0.736
Chameleon-F	890	17,758	2,325	5	0.238
Squirrel-F	2,223	94,136	2,089	5	0.208
Actor	7,600	53,411	932	5	0.218
Roman-Empire	22,662	65,854	300	18	0.047
Amazon-Ratings	24,492	186,100	300	5	0.380