Joint Certification for Attributed Graphs: Beyond Topology-Only Robustness

Blaise Delattre[1]* **Hengyu Wu**[1]*
Wei Yang Bryan Lim[2] **Yang Cao**[1]

[1]Department of Computer Science, School of Computing
Institute of Science Tokyo, Tokyo, Japan
[2]College of Computing and Data Science
Nanyang Technological University, Singapore
bldelattre@gmail.com

Abstract

Graph neural networks for bot and fraud detection operate on attributed graphs, where adversaries can jointly modify local edges and continuous node features. We study node-level certification under a mixed threat model with up to d edge flips incident to the target node and a Frobenius feature budget ϵ on a prescribed node set. We instantiate hybrid randomized smoothing for this graph setting and compare two sound certificates: a Sequential NP composition baseline for the joint topology–continuous-feature threat model, and a coupled Hybrid NP certificate. Under homogeneous edge smoothing, the outer Neyman–Pearson optimization over adversarial adjacency matrices reduces exactly to an $O(d^2)$ search over edge-deletion and edge-addition counts. Across PubMed, Amazon Fraud, YelpChi, and MGTAB, the Hybrid NP certificate yields non-trivial certified regions under joint edge and global feature perturbations. We also construct joint attacks that change estimated smoothed predictions on topology-only certified nodes across PubMed, Amazon Fraud, MGTAB and YelpChi, illustrating the robustness gap when continuous feature perturbations are allowed. The results support joint discrete–continuous certification as the appropriate robustness notion for attributed graphs with continuous features. Code is available online.

1 Introduction

Graph neural networks make predictions by recursively aggregating neighbor features along edges, coupling topology and node attributes by construction: changing edges alters *which* features are aggregated, while perturbing features corrupts *what* is passed along the remaining edges. In realistic deployments this coupling is precisely what an adversary may exploit. In fraud detection on transaction and review graphs (Cheng et al., 2024; Dou et al., 2020), an attacker controlling a small set of accounts can simultaneously rewire connections to legitimate users and manipulate continuous transaction or behavioral attributes. In social media bot detection (Wang et al., 2023; Cresci et al., 2017), coordinated accounts may alter both local interactions and account-level signals. In citation graphs (Zügner and Günnemann, 2019), both links and document embeddings may be manipulated. A meaningful robustness guarantee for attributed graphs should therefore account for topology and continuous features *simultaneously*, as illustrated in Figure 1.

A joint threat model for the target node. We consider node-level robustness for a target v on a fixed vertex set. The adversary may flip up to d edges incident to v, including insertions connecting v

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

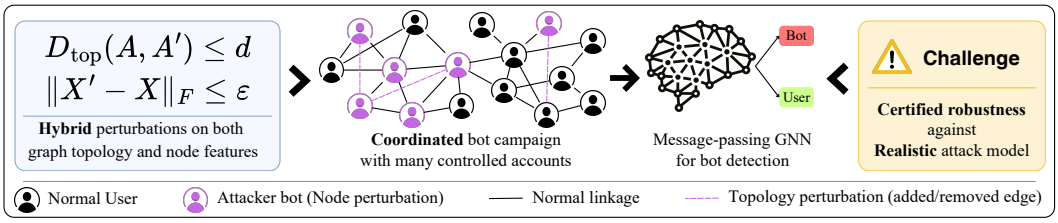


Figure 1: Motivating challenge for certified robustness under realistic attributed-graph attacks. In social-media bot detection, coordinated attackers may perturb many controlled accounts and rewire local connections simultaneously. This motivates certifying robustness under a joint budget with up to d edge flips incident to the target node and continuous feature perturbations on a prescribed node set, with $\|X' - X\|_F \leq \varepsilon$.

to any existing node, and perturb continuous features on a prescribed node set $\mathcal{C} \subseteq V$. Formally,

$$\mathcal{B}^{\mathcal{C}}_{d,\varepsilon}(A, X; v) = \{(A', X + \Delta) : A' \in \mathcal{A}_d(A; v),\ \text{supp}_{\text{row}}(\Delta) \subseteq \mathcal{C},\ \|\Delta\|_F \leq \varepsilon\}. \tag{1}$$

Section 3 defines the structural support and the matching smoothing channel. Taking $\mathcal{C} = V$ gives the global Frobenius model, which covers perturbations concentrated on a few nodes or spread across many; for s affected nodes, the root-mean-square node-wise ℓ_2 magnitude is at most $\varepsilon/\sqrt{s}$. The main certification curves use this global feature model; $\mathcal{C} = \{v\}$ gives the target-feature special case. The noise support is fixed under both clean and adversarial inputs and contains every attackable feature row. An insertion can expose previously unreachable existing nodes; non-incident edge flips and new-vertex injection are outside the evaluated structural threat model.

Gap in existing certificates. Graph robustness certificates cover several threat geometries. Topology-only smoothing holds features fixed (Wang et al., 2021; Scholten et al., 2022), while sparsity-aware smoothing also jointly certifies structure and binary attributes (Bojchevski et al., 2020). GNNCert and AGNNCert use node-level corruption budgets rather than continuous total-norm budgets (Xia et al., 2024; Li and Wang, 2025); these geometries are complementary. ELEGANT combines attribute and structural smoothing to certify a group-fairness event (Dong et al., 2026). Our goal is a tractable coupled NP bound for prediction invariance under edge flips and continuous feature changes. We build on the abstract hybrid framework of Delattre et al. (2026) and make explicit the support conditions and count reduction needed for this graph setting.

Joint attacks on topology-certified nodes. The distinction is not only formal. We construct joint attacks that combine edge flips with feature-level projected gradient descent (Madry et al., 2018), applied to nodes certified by topology-only smoothing. Across PubMed (Yang et al., 2016), MGTAB (Shi et al., 2025), Amazon Fraud (Dou et al., 2020), and YelpChi (Mukherjee et al., 2013), we find changes of estimated smoothed predictions on topology-certified nodes once continuous feature perturbations are allowed. These attacks occur at budgets for which our joint certificate gives no robustness guarantee, supporting the need to certify topology and continuous features simultaneously.

Contributions.

- **Explicit structural and feature supports.** We formulate a mixed threat model combining edge flips incident to the target node with continuous feature changes on a prescribed node set, and prove the support conditions needed for a valid product-channel certificate.
- **Tractable Hybrid NP certification.** We separate soundness for a fixed classifier from exactness of the NP count reduction. Homogeneous edge smoothing reduces the outer NP search to $\mathcal{O}(d^2)$ count pairs. We prove maximality among bounds uniformly valid for tests with the same clean mass and channel laws, and dominance over both sound sequential orders.
- **Threat-model mismatch of topology-only certificates.** We attack nodes certified by topology-only smoothing at $d = 1$ using joint edge–feature perturbations. The attacks flip estimated smoothed predictions at budgets where the joint certificate gives no guarantee, empirically demonstrating that topology-only certificates can overestimate robustness for attributed graphs with continuous features.

2 Related Work

Randomized smoothing. Gaussian smoothing yields closed-form ℓ_2 certificates through a Neyman–Pearson analysis (Cohen et al., 2019). Discrete and hierarchical variants address categorical or structured perturbations (Bojchevski et al., 2020; Scholten et al., 2023). Our product-space NP formulation follows the hybrid framework of Delattre et al. (2026). We make its graph instantiation tractable through homogeneous edge smoothing, explicit support conditions and evaluation on the sampled computation graph.

Graph robustness and joint discrete certificates. Nettack perturbs both graph structure and attributes (Zügner et al., 2018). Deterministic feature certificates handle bounded changes to binary attributes (Zügner and Günnemann, 2019), while topology smoothing certifies edge flips for node classification (Wang et al., 2021), community detection (Jia et al., 2020) and graph classification (Gao et al., 2020). Gray-box message interception exploits message-passing structure to tighten certificates (Scholten et al., 2022). Sparsity-aware smoothing (Bojchevski et al., 2020) also certifies structure and binary attributes jointly, using different addition and deletion probabilities. For binary X, X', the identity $\|X' - X\|_F^2 = \|X' - X\|_0$ maps a d_f-flip budget to radius $\sqrt{d_f}$ within the binary domain. The corresponding continuous Frobenius ball also includes nonbinary perturbations, so it is a larger threat set; comparisons must also account for structural support and channels.

GNNCert (Xia et al., 2024) and AGNNCert (Li and Wang, 2025) use deterministic divide-and-conquer certificates for discrete edge and node corruptions. GNNCert addresses graph classification; AGNNCert extends this family to node classification and broader perturbations. Their node-count feature guarantees allow unbounded changes on a bounded number of nodes, whereas a total-norm guarantee bounds the energy on a prescribed support. These geometries are complementary: neither universally contains the other. With at most k changed nodes and per-node norm at most δ, a global Frobenius certificate applies when $\sqrt{k}\delta \leq \varepsilon$; this comparison requires global feature smoothing and does not follow from target-only noise.

Joint smoothing for fairness. ELEGANT (Dong et al., 2026) first Gaussian-smooths a group-bias indicator over node attributes and then Bernoulli-smooths that classifier over graph structure. It therefore already combines continuous attribute and discrete structural guarantees through a nested construction. Its certified object is a group-fairness event, whereas ours is invariance of one target prediction. Changing its indicator to a target-class event would define an adapted method, not a faithful numerical fairness baseline. This prior compositional construction motivates treating Sequential NP as a sound comparison baseline, with the coupled NP relaxation as the object of our analysis.

Continuous attributes. Real-valued signals occur in transaction and behavioral statistics for fraud detection (Weber et al., 2019; Dou et al., 2020; Liu et al., 2021, 2020), sensor measurements in traffic forecasting (Li et al., 2018), and molecular representations (Reiser et al., 2022). These applications motivate continuous feature budgets alongside structural changes (Sun et al., 2023). Their certified coverage depends on the chosen supports and smoothing channel, as formalized in Section 3.

3 Hybrid certificate for attributed graphs

3.1 Threat model and locality

Assumptions. We consider an L-layer message-passing GNN on a fixed vertex set V. The topology channel independently corrupts a fixed candidate slot set S_v with homogeneous deletion and addition probabilities $p^-, p^+ \in (0, 1)$. The feature channel adds independent isotropic Gaussian noise with $\sigma > 0$ on a nonempty, fixed node set $\mathcal{C} \subseteq V$. Both supports are held fixed under clean and adversarial inputs, and features may be perturbed only on $\mathcal{C}$. These conditions ensure absolute continuity of the adversarial channel with respect to the clean channel.

Structural support. For an undirected graph, each slot is an unordered pair represented once by (i, j) with $i < j$. The evaluated structural support is the full target-incident row,

$$S_v = \{(\min(v, j), \max(v, j)) : j \in V \setminus \{v\}\}.$$

An insertion may connect v to any existing node, including one outside its clean L-hop neighborhood. This support excludes non-incident edge flips and injection of new vertices. The theory also applies

to other fixed slot sets under the same product channel. Define

$$\mathcal{A}_d(\mathbf{A}; v) = \Big\{\mathbf{A}' : \textstyle\sum_{(i,j)\in S_v} |A_{ij} - A'_{ij}| \le d, \quad A'_{ij} = A_{ij} \text{ outside } S_v\Big\}. \tag{2}$$

For a fixed feature support, the joint threat set is

$$\mathcal{B}^{\mathcal{C}}_{d,\varepsilon}(\mathbf{A}, \mathbf{X}; v) = \{(\mathbf{A}', \mathbf{X} + \Delta) : \mathbf{A}' \in \mathcal{A}_d(\mathbf{A}; v),\ \mathrm{supp}_{\mathrm{row}}(\Delta) \subseteq \mathcal{C},\ \|\Delta\|_F \le \varepsilon\}. \tag{3}$$

The global Frobenius model corresponds to $\mathcal{C} = V$; with $\mathcal{C} = \{v\}$ only the target's feature vector may change. We omit the superscript when the support is understood.

Lemma 1 (Message-passing locality at a realized topology)**.** *Fix a realized topology B and its propagation weights. The prediction at v depends on features only through its computation support $C_v(B)$, contained in the L-hop neighborhood for L one-hop layers. Thus $f_v(B, Y) = \bar{f}_v(B, P_B Y)$ for a map $\bar{f}_v$, where P_B selects these feature rows.*

The support is determined after sampling B and can grow following insertions. Local evaluation must retain required normalization weights, including boundary degree information. Locality reduces inference work; the count reduction instead uses homogeneity on the fixed slot set.

3.2 Two sound certificates for the joint threat model

Write $Z_A \sim p_1(\cdot \mid \mathbf{A})$ and $Z_X = \mathbf{X} + \iota_{\mathcal{C}}\eta$, where $\eta \sim \mathcal{N}(0, \sigma^2 I_{|\mathcal{C}|\times p})$ is independent of Z_A and $\iota_{\mathcal{C}}$ embeds its rows in the full matrix. For a fixed measurable score $h : \Omega \to [0,1]$, let $g_h(\mathbf{A}, \mathbf{X}) = \mathbb{E}[h(Z_A, Z_X)]$, $p_A = g_h(\mathbf{A}, \mathbf{X}) \in (0,1)$, and

$$p^h_{\mathrm{adv}}(d, \varepsilon) := \inf_{(\mathbf{A}', \mathbf{X}') \in \mathcal{B}^{\mathcal{C}}_{d,\varepsilon}} g_h(\mathbf{A}', \mathbf{X}').$$

This true worst-case score depends on the identities of modified edges and on the fixed classifier.

Sequential NP: composition baseline. Let $T_d(p)$ be the topology-only NP lower-bound map at clean mass p on S_v, and $G_\varepsilon(p) = \Phi(\Phi^{-1}(p) - \varepsilon/\sigma)$, with continuous endpoint conventions at $p = 0, 1$. The two orders are

$$p^{T\to X}_{\mathrm{seq}} = G_\varepsilon(T_d(p_A)), \qquad p^{X\to T}_{\mathrm{seq}} = T_d(G_\varepsilon(p_A)).$$

Both are sound, as is their pointwise maximum $p^{\max}_{\mathrm{seq}}$ (Proposition 5). Figure 2 uses the baseline $p_{\mathrm{seq}} = p^{T\to X}_{\mathrm{seq}}$. Its effective mass $p_{\mathrm{eff}}(d) := T_d(p_A)$ is a computable lower bound, rather than the true score $p^h_{\mathrm{adv}}(d, 0)$.

Hybrid NP: coupled likelihood-ratio certificate. We specialize the product-space NP framework of Delattre et al. (2026). Let P be the clean product measure on $(Z_A, (Z_X)_{\mathcal{C}})$ and $Q_{\mathbf{A}',\mathbf{X}'}$ the corresponding adversarial measure. Define

$$V_{\mathrm{NP}}(\mathbf{A}', \mathbf{X}') = \inf_{\substack{a:\Omega\to[0,1] \text{ measurable} \\ \mathbb{E}_P[a] = p_A}} \mathbb{E}_{Q_{\mathbf{A}',\mathbf{X}'}}[a], \qquad V_{\mathrm{NP}}(\mathbf{A}'; \varepsilon) = \inf_{\substack{\mathrm{supp}_{\mathrm{row}}(\mathbf{X}'-\mathbf{X})\subseteq\mathcal{C} \\ \|\mathbf{X}'-\mathbf{X}\|_F \le \varepsilon}} V_{\mathrm{NP}}(\mathbf{A}', \mathbf{X}').$$

Write $\mathcal{H}$ for the measurable $[0,1]$-valued test class. These programs optimize over all tests with the given clean mass, rather than the predictions of a particular GNN.

Corollary 1 (Specialized hybrid NP value)**.** *Fix $\mathbf{A}'$ and $r = \|\mathbf{X}' - \mathbf{X}\|_F > 0$ on the fixed feature support. Set $\gamma_1(Z_A) = p_1(Z_A \mid \mathbf{A}')/p_1(Z_A \mid \mathbf{A})$ and*

$$F(t; r) = \sum_{Z_A} p_1(Z_A \mid \mathbf{A}) \Phi\left(\frac{r}{2\sigma} + \frac{\log t - \log \gamma_1(Z_A)}{r/\sigma}\right).$$

There is a unique $t^\star(r) > 0$ with $F(t^\star(r); r) = p_A$, and

$$V_{\mathrm{NP}}(\mathbf{A}', \mathbf{X}') = V(\mathbf{A}'; r) = \sum_{Z_A} p_1(Z_A \mid \mathbf{A}') \Phi\left(-\frac{r}{2\sigma} + \frac{\log t^\star(r) - \log \gamma_1(Z_A)}{r/\sigma}\right).$$

The value is non-increasing in r, so $V_{\mathrm{NP}}(\mathbf{A}'; \varepsilon) = V(\mathbf{A}'; \varepsilon)$. At $r = 0$, use the discrete topology NP value, with randomization on a threshold atom as necessary. For every fixed score h,

$$p^h_{\mathrm{adv}}(d, \varepsilon) \ge \underline{p}_{\mathrm{hyb}}(d, \varepsilon) := \inf_{\mathbf{A}' \in \mathcal{A}_d(\mathbf{A}; v)} V_{\mathrm{NP}}(\mathbf{A}'; \varepsilon).$$

The characterization and radius monotonicity are justified in Appendix E.1. For multiclass classification, use $h = h_{c_A} := \mathbf{1}\{f_v(Z_A, Z_X) = c_A\}$, where c_A is the clean smoothed prediction. The sufficient rule is $\underline{p}_{\mathrm{hyb}} > 1/2$; Appendix D.3 also gives a top-two-class condition.

Proposition 1 (Axis consistency)**.** *With clean mass p_A from the same hybrid channel,*

$$\underline{p}_{\mathrm{hyb}}(d, 0) = T_d(p_A) = \underline{p}_{\mathrm{topo}}^{\bar{h}^X}(d), \qquad \underline{p}_{\mathrm{hyb}}(0, \varepsilon) = G_\varepsilon(p_A),$$

where $\bar{h}^X$ is the feature-marginalized score. The maps act on corresponding marginalized scores and match standard single-channel smoothing when the other channel is degenerate.

The formal statement and proof are in Appendix E.3.

Lemma 2 (Fixed feature support)**.** *If the same $\mathcal{C}$ is noised under both centers and $\mathrm{supp}_{\mathrm{row}}(\mathbf{X}' - \mathbf{X}) \subseteq \mathcal{C}$, the two scores are obtained from product measures on the same space by the same classifier map. Corollary 1 therefore applies, including when inserted edges expose previously unseen nodes. For $\mathcal{C} = V$, sampling noise only on $C_v(B)$ after drawing B is an exact marginalization of the global feature channel.*

Proofs and support-specific cases are in Appendix H.1. The projected norm inequality alone does not establish this result. Target-only noise certifies changes only to X_v, and a noise support recomputed from each input center can destroy absolute continuity. **Feature support of the main curves.** The main certification curves use global Gaussian feature noise, $\mathcal{C} = V$. Their Frobenius guarantee covers perturbations on any graph node, including the target, with one total norm budget. Appendix I.1 compares this channel with target-only smoothing on Amazon.

3.3 Count reduction under homogeneous edge smoothing

Topology kernel. Each existing slot in $S = S_v$ is deleted with probability p^- and each empty slot added with probability p^+, independently; entries outside S remain fixed. This is the sparsity-aware binary channel of Bojchevski et al. (2020). The distribution of $\log \gamma_1$ depends only on the adversary's deletion and addition counts (d^-, d^+), as shown by the explicit atoms in Appendix F.

In the main experiments, a symmetric overwrite kernel keeps each slot's value with probability q and otherwise draws a $\mathrm{Bernoulli}(1/2)$ bit. Equivalently,

$$p^- = p^+ = (1 - q)/2, \qquad \rho = \Pr[Z_{A,e} = A_e] = (1 + q)/2.$$

Homogeneity across slots, rather than equality of p^- and p^+, enables the reduction. Asymmetric choices can reduce the expected clean-graph corruption $m_1 p^- + m_0 p^+$ on sparse graphs.

Theorem 1 (Soundness and exact NP count reduction)**.** *Let m_1, m_0 count existing and empty slots in S. Under the fixed supports and homogeneous product channel above,*

$$\inf_{\mathbf{A}' \in \mathcal{A}_d(\mathbf{A}; v)} V_{\mathrm{NP}}(\mathbf{A}'; \varepsilon) = \underline{p}_{\mathrm{hyb}}(d, \varepsilon) = \min_{\substack{d^- + d^+ \le d \\ 0 \le d^- \le m_1,\ 0 \le d^+ \le m_0}} V(d^-, d^+; \varepsilon).$$

For every fixed h of clean mass p_A, $p_{\mathrm{adv}}^h(d, \varepsilon) \ge \underline{p}_{\mathrm{hyb}}(d, \varepsilon)$.

The proof in Appendix E.1 separates two facts: the likelihood-ratio law is determined by counts, and every feasible count pair is realized by an admissible adjacency. Edge identities remain relevant to h, but not to this NP relaxation.

Proposition 2 (Maximality given the clean mass and channel laws)**.** *Fix the channels, supports and budgets. If $b(p_A)$ lower-bounds the adversarial mass of every measurable test $a : \Omega \to [0, 1]$ with $\mathbb{E}_P[a] = p_A$, at every admissible input, then $b(p_A) \le \underline{p}_{\mathrm{hyb}}(d, \varepsilon)$. In particular,*

$$p_{\mathrm{seq}}^{\max} \le \underline{p}_{\mathrm{hyb}}(d, \varepsilon) \le p_{\mathrm{adv}}^h(d, \varepsilon).$$

The proof is in Appendix H.2. This optimality is within the stated information class; it does not imply tightness for a particular GNN. Identity-aware improvements require additional classifier information beyond its scalar clean mass.

Relation to existing smoothing frameworks. The product-space NP formulation is inherited from Delattre et al. (2026); sparse discrete smoothing is established by Bojchevski et al. (2020). Here we make the graph instantiation explicit: fixed structural and feature supports, a tractable count reduction of the outer NP search, and sound composition baselines in both orders. Locality justifies efficient evaluation at each sampled topology. Maximality identifies the information needed to tighten the count-based certificate for a particular classifier.

3.4 Algorithm and complexity

Algorithm 1 estimates p_A after independent class selection, applies a binomial lower confidence bound, and evaluates the NP value on the feasible count grid. For $\varepsilon > 0$, each pair requires scalar root finding in $\log t$; at $\varepsilon = 0$ it uses the discrete NP calculation. The outer search has $O(d^2)$ pairs, with $O(d^3 \log(1/\delta_t))$ scalar CDF evaluations in total at root tolerance δ_t. Monte Carlo cost also depends on the candidate-slot count and realized computation support, which may grow after insertions. The $O(d^2)$ claim concerns the outer NP search, not the full sampling cost.

Sparse-continuous budgets. The fixed-support formulation allows a prescribed attackable node set. An isotropic channel does not exploit an unknown sparse support inside that set; mixed support–norm budgets require further analysis.

4 Experiments

We evaluate (i) whether the joint certificate is non-vacuous on attributed-graph benchmarks, (ii) how much Hybrid NP improves over the Sequential NP composition baseline, and (iii) whether topology-only certificates overestimate robustness once the adversary is allowed to perturb continuous node features.

4.1 Comparison protocol and baselines

Table 1: Threat-model coverage of graph robustness certificates. Numerical comparisons require matching structural supports, feature domains and certified events. For Gaussian RS and our certificates, continuous perturbations are restricted to the prescribed feature support $\mathcal{C}$. ELEGANT certifies a group-fairness event.

Method	Structure threat	Feature threat	Guarantee	Comparison scope
Topology-only RS (Wang et al., 2021; Jia et al., 2020; Gao et al., 2020)	$\mathcal{A}_d(A; v)$	none	probabilistic	$\epsilon = 0$ axis
Gaussian RS (Cohen et al., 2019)	none	$\|X' - X\|_F \le \epsilon$	probabilistic	$d = 0$ axis
GNNCert (Xia et al., 2024)/AGNNCert (Li and Wang, 2025)	bounded edges	k nodes, arbitrary replacement	deterministic	structure axis; AGNNCert feature budgets (Appendix I.2)
Sparsity-aware RS (Bojchevski et al., 2020)	edge additions/deletions	binary attribute flips	probabilistic	joint discrete budgets
ELEGANT (Dong et al., 2026)	edge flips	continuous attributes	probabilistic	group-fairness event
Sequential NP (ours)	$\mathcal{A}_d(A; v)$	$\|X' - X\|_F \le \epsilon$	probabilistic	sound composition baseline
Hybrid NP (ours)	$\mathcal{A}_d(A; v)$	$\|X' - X\|_F \le \epsilon$	probabilistic	proposed joint certificate

Threat-model coverage. Recent joint graph certificates such as GNNCert (Xia et al., 2024) and AGNNCert (Li and Wang, 2025) certify simultaneous structure and feature perturbations, but their feature threat model is discrete: the adversary may arbitrarily replace the feature vectors of up to k nodes,

$$\mathcal{B}^{\mathrm{disc}}_{d,k}(A, X) = \{(A', X') : \ A' \text{ differs from } A \text{ in at most } d \text{ edges}, \ |\{i : X'_i \neq X_i\}| \le k\}.$$

Our setting uses the fixed-support joint threat set in (3). The main certification curves use $\mathcal{C} = V$, allowing feature changes on any node, including the target, under one total Frobenius budget. Appendix I.2 additionally evaluates global feature certificates ($\mathcal{C} = V$) at fixed topology. The relationship between node-count and total-norm feature budgets depends on the admissible support

Table 2: Summary of the main hybrid-certification settings used in Figure 2. The parameter q is the overwrite-keep probability of the topology smoothing kernel; its effective retention probability is $\rho = (1+q)/2$. The parameters L, N, and α denote the number of message-passing layers, Monte Carlo samples per node, and Clopper–Pearson failure probability, respectively. The ϵ ranges give approximate horizontal-axis limits, including a 5% margin beyond the last sampled ϵ with positive certified accuracy. All four datasets use global feature smoothing, $\mathcal{C} = V$.

Dataset	Task	Base model	L	q	ρ	σ	σ_{train}	Certified nodes	d	ϵ range	N	α
YelpChi (Mukherjee et al., 2013)	Fraud detection	BWGNN	2	0.6	0.8	0.005	0.005	500	$\{0,\dots,5\}$	$[0, 0.0166]$	10 000	10^{-3}
Amazon Fraud (Dou et al., 2020)	Fraud detection	BWGNN	2	0.5	0.75	0.1	0.5	500	$\{0,\dots,5\}$	$[0, 0.3324]$	10 000	10^{-3}
MGTAB (Shi et al., 2025)	Bot detection	2-layer GCN	2	0.4	0.7	0.4	0.4	400	$\{0,\dots,8\}$	$[0, 1.3296]$	10 000	10^{-3}
PubMed (Yang et al., 2016)	Paper classification	2-layer GCN	2	0.7	0.85	0.01	0.04	1 000	$\{0,1,2,3\}$	$[0, 0.0327]$	10 000	10^{-3}

and feature domain. For global real-valued feature perturbations, the two threat sets are incomparable when $1 \le k < |V|$ and $\varepsilon > 0$. On a common real-valued domain, target-only bounded perturbations are contained in an arbitrary row-replacement threat with $k \ge 1$ when the target is included. We compare methods along their corresponding axes. At $\varepsilon = 0$, our method is compared to topology-only graph certificates. At $d = 0$, it is compared to Gaussian feature smoothing. For the mixed budget $d > 0, \varepsilon > 0$, we compare Hybrid NP with the sound Sequential NP baseline from Section 3.2, using the same structural support, feature support, and joint smoothing channel. Table 1 summarizes the threat-model coverage of each baseline.

Sequential NP and Hybrid NP are both instantiated in this work for the same joint threat model. Sequential NP is the natural sound composition baseline; Hybrid NP uses the coupled likelihood ratio and is the main certificate. Proposition 1 formalizes these axis comparisons: on the $\varepsilon = 0$ axis, Hybrid NP reduces to topology-only graph randomized smoothing applied to the feature-marginalized score; on the $d = 0$ axis, it reduces to Gaussian randomized smoothing applied to the topology-marginalized score.

Experimental protocol. We evaluate stratified test targets selected before correctness filtering. Independent pilot draws select a class; N Monte Carlo draws per node provide simultaneous one-sided Clopper–Pearson bounds at overall failure probability α. We use the top-two-class condition in Appendix D.3. Table 2 reports N, α, and message-passing depth L. Certified accuracy is the fraction of all evaluated targets whose selected class is correct and certified at the stated joint budget. All feature radii ϵ are measured after the preprocessing used by the corresponding base model. Split, optimizer, checkpoint, and certification details are listed in Appendix C.

Topology-only smoothing looks strong at $\varepsilon = 0$. On the four datasets considered here, topology-only randomized smoothing (Wang et al., 2021) yields high certified accuracy that decays slowly with the edge-flip budget K, and remains non-vacuous up to $K \approx 9$–14 (Figure 5 in Appendix A.2). This appearance is misleading under the joint threat model: Section 4.3 shows that even at $K = 1$, a joint attacker allowed a small ℓ_2-bounded feature perturbation changes the estimated smoothed predictions of a substantial fraction of topology-certified nodes.

4.2 Main hybrid certification results

We evaluate Hybrid NP and Sequential NP on four attributed-graph benchmarks spanning fraud detection (YelpChi, Amazon Fraud), bot detection (MGTAB), and citation classification (PubMed). For the fraud datasets we use BWGNN (Tang et al., 2022) as the base detector; for MGTAB and PubMed we use a standard two-layer GCN as in Wang et al. (2021). Per-dataset configurations are summarized in Table 2; training and certification details are given in Appendix C.

YelpChi. On YelpChi ($q = 0.6$, $\rho = 0.8$, $\sigma = 0.005$, $\sigma_{\text{train}} = 0.005$, 500 stratified nodes; Figure 2a), Hybrid NP remains non-vacuous for all $d \in \{0,\dots,5\}$. The gain over Sequential NP is largest at high topology budgets: for $d = 5$, the sequential certificate collapses near $\epsilon \simeq 0.002$, while Hybrid NP remains non-zero until $\epsilon \simeq 0.006$.

Amazon Fraud. On Amazon Fraud ($q = 0.5$, $\rho = 0.75$, $\sigma = 0.1$, $\sigma_{\text{train}} = 0.5$, 500 stratified nodes; Figure 2b), Hybrid NP extends the positive certified region for every $d \ge 1$. At $d = 2$ and $\epsilon \approx 0.10$, it certifies 58.2% of targets versus 46.6% for Sequential NP. At $d = 5$ and the same feature radius, Hybrid NP retains 25.8% coverage while Sequential NP is vacuous.

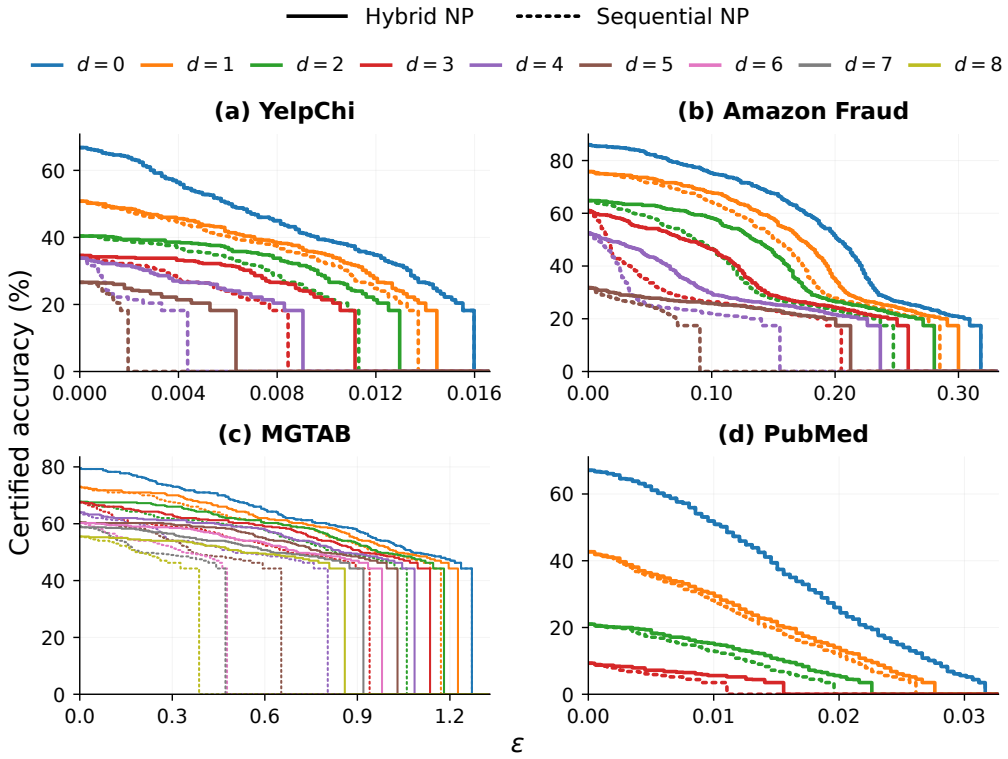


Figure 2: Certified accuracy (%) versus ϵ for Hybrid NP (solid) and Sequential NP (dotted) certificates. Colors indicate topology budget d. All panels use global Gaussian feature smoothing ($\mathcal{C} = V$), $N = 10{,}000$ draws per target, and the total Frobenius feature budget ϵ.

MGTAB. On MGTAB ($q = 0.4$, $\rho = 0.7$, $\sigma = 0.4$, $\sigma_{\text{train}} = 0.4$, 400 stratified nodes; Figure 2c), the certificate remains non-vacuous over a larger continuous budget on the standardized feature scale: the $d = 0$ Hybrid NP certificate persists up to $\epsilon \approx 1.27$, and all nine budgets $d \in \{0, \ldots, 8\}$ are non-trivial at $\epsilon = 0$. Hybrid NP extends the positive certified region for every non-zero topology budget.

PubMed. On PubMed ($q = 0.7$, $\rho = 0.85$, $\sigma = 0.01$, $\sigma_{\text{train}} = 0.04$, 1 000 nodes; Figure 2d), all curves decrease smoothly with ϵ. For $d = 0$, the two methods coincide by construction. For $d \geq 1$, Hybrid NP never falls below Sequential NP on the evaluated grid. At $d = 3$, the last positive sampled radius is about 0.0151 for Hybrid NP and 0.0106 for Sequential NP.

Hybrid vs. Sequential NP. Across all four datasets in Figure 2, Hybrid NP and Sequential NP coincide at $d = 0$ by construction. At $\epsilon = 0$, the joint certificate reduces to the topology-axis lower bound $\underline{p}_{\text{hyb}}(d, 0) = T_d(p_A) \leq p^h_{\text{adv}}(d, 0)$. For $d \geq 1$ and $\epsilon > 0$, Hybrid NP yields larger certified regions. On YelpChi at $d = 5$, Sequential NP collapses near $\epsilon \simeq 0.002$ while Hybrid NP remains non-zero until $\epsilon \simeq 0.006$. This reflects the benefit of ordering samples by the coupled topology–feature likelihood ratio rather than composing the two channels sequentially.

Appendix A.1 studies the smoothing parameters σ and q. Appendix I adds baseline comparisons, both sequential orders, accuracy costs, seed and architecture sensitivity, and runtime diagnostics, with explicit attack supports and evaluation denominators.

4.3 Topology-Only Failure Rate Under Joint Attacks

We evaluate how often nodes certified by the topology-only certificate at edge budget $d = 1$ become vulnerable once the adversary is allowed to perturb both graph structure and continuous node features. We emphasize that this is not a soundness violation of the topology-only certificate—it is sound

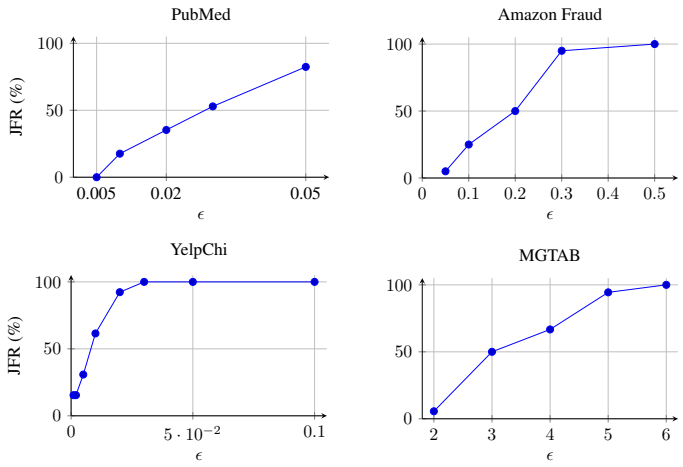


Figure 3: Joint-attack failure rate (JFR) of topology-only certified nodes under $(1, \epsilon)$ attacks. For each dataset, the curve reports the percentage of nodes certified by the topology-only certificate at $d = 1$ whose estimated smoothed prediction is flipped by a joint attack with one edge flip and feature budget ϵ. Each clean and attacked label is estimated from 200 fresh Monte Carlo votes.

for its own threat model. We measure changes of their estimated smoothed labels under the *joint* threat model that the topology-only certificate does not cover. Specifically, we build attack based on Nettack (Zügner et al., 2018) and strengthen the feature perturbation component using a PGD-based attack (Madry et al., 2018).

For each dataset, we define the evaluated target set

$$N_{\text{topo}}(1) = \{v : v \text{ is evaluated and topology-certified at } d = 1\},$$

and attack every node in this set under the joint threat model $(1, \epsilon)$: one edge flip plus an ℓ_2-bounded feature perturbation of radius ϵ. The joint-attack failure rate (JFR) is

$$\text{JFR}(\epsilon) = \frac{\#\{v \in N_{\text{topo}}(1) : \widehat{g}_{200}(x_v, A_v) \neq \widehat{g}_{200}(x'_v, A'_v)\}}{|N_{\text{topo}}(1)|},$$

where $\widehat{g}_{200}$ is the majority-vote estimate of the smoothed label from 200 fresh Monte Carlo draws per input, and (x'_v, A'_v) is the attacked sample. JFR counts changes of these estimated labels, without a high-confidence statistical test of a flip of the exact smoothed classifier. The joint attack precomputes one Nettack-style edge flip and optimizes the base-model loss with projected ℓ_2 feature updates. The attack hyperparameters and evaluation budgets are reported in Appendix C.3.

Figure 3 reports the failure rate curves on PubMed, Amazon Fraud, YelpChi, and MGTAB. Across these datasets, the fraction of topology-only certified nodes whose estimated smoothed prediction is flipped increases rapidly with the feature budget ϵ. On PubMed, the failure rate rises from 0 at $\epsilon = 0.005$ to 0.824 at $\epsilon = 0.05$. On Amazon Fraud, it rises from 0.05 at $\epsilon = 0.05$ to 1.0 at $\epsilon = 0.5$. On YelpChi, it already reaches 0.615 at $\epsilon = 0.01$ and 1.0 by $\epsilon = 0.03$. On MGTAB, it starts to rise at $\epsilon = 2.0$ and reaches 1.0 at $\epsilon = 6.0$. These estimated-label flips illustrate the robustness gap when continuous feature perturbations are allowed.

Control on hybrid-certified nodes. On PubMed and the fraud datasets, we also run the same joint $(1, \epsilon)$ attack on the subset of nodes certified by Hybrid NP at the corresponding budget, using the target-only control channel in Appendix C.3. In the tested cases, we observe no flips of the estimated smoothed prediction. This experiment is not used to establish soundness, which follows from the certificate; it checks that the attack exposes a genuine threat-model gap for topology-only certification rather than a generic failure of the smoothed classifier.

Feature scale and attack-budget calibration. The feature scales differ substantially across datasets, so the ℓ_2 attack radii used in the JFR experiments should be interpreted relative to each dataset's effective feature geometry. PubMed uses sparse, row-normalized bag-of-words features, with a median node norm of only 0.19, so even $\epsilon = 0.05$ remains small in per-coordinate terms. Amazon Fraud has much lower feature dimension and substantially larger per-feature mass, making the same

Table 3: Dataset scales used to calibrate JFR attack budgets. Feature norms, largest absolute coordinates, and binary degrees are global medians over dataset nodes, rather than medians over the selected attack targets. MGTAB features use train-split standardization; its degree uses the undirected edge-type-0 projection.

Dataset	Dim	Median $\|x\|_2$	Median $\max_i \|x_i\|$	Median deg	Attack ϵ range
PubMed	500	0.190	0.095	2	$[0.005, 0.05]$
Amazon Fraud	25	0.539	0.473	63	$[0.05, 0.5]$
YelpChi	32	0.198	0.049	58	$[0.001, 0.1]$
MGTAB (std.)	788	21.167	2.707	7	$[2.0, 6.0]$

absolute ℓ_2 budget more visible; accordingly, we use a larger but still graded range up to $\epsilon = 0.5$. YelpChi has dense but small-magnitude features, so the relevant radii are much smaller, reaching total failure already at low ϵ. MGTAB differs most from the other datasets: after train-split standardization, following common preprocessing for numerical social-media user attributes (Feng et al., 2021), it has 788 dense features and median node norm 21.167. Thus $\epsilon = 4$ is about 19% of that norm, a moderate relative perturbation. Table 3 summarizes the dataset scales and JFR attack ranges.

All reported ϵ values are measured in the post-preprocessing feature space used by the model and attack pipeline. In relative terms, the largest tested radii correspond to approximately 26%, 93%, 51%, and 28% of the median node ℓ_2 norm on PubMed, Amazon Fraud, YelpChi, and MGTAB, respectively.

5 Conclusion

We studied joint node-level certification under target-incident edge flips and continuous feature changes on a prescribed node set. Under homogeneous edge smoothing, the outer optimization of the NP relaxation reduces exactly to an $\mathcal{O}(d^2)$ search over deletion and addition counts. Its value is a sound lower bound for every fixed classifier, and maximal among bounds uniformly valid for all tests with the same clean mass and channel laws. Both sequential orders and their maximum are sound baselines dominated by this joint bound.

Our main certification experiments show non-trivial regions under target-incident edge flips and global Frobenius feature budgets, using global Gaussian feature smoothing. Joint attacks expose the mismatch between topology-only guarantees and a joint threat.

Limitations. The evaluated structural support permits connections to any existing node, but excludes non-incident edge changes and new-vertex injection. The noise support must be fixed and contain every attackable feature coordinate. The NP bound can be conservative for a particular GNN; its information-class optimality does not establish model-specific tightness. Frobenius budgets also do not exploit an unknown sparse perturbation support. Edge-dependent probabilities require richer statistics than aggregate deletion and addition counts.

Extensions. The same product-space construction applies to graph classification with graph-level scores and a fixed admissible slot set, and to link prediction with the computation support of both endpoints at each realized topology. Larger supports increase clean noise and sampling cost; these extensions require their own empirical evaluation. Typed slots with a common homogeneous kernel retain the count argument (Appendix H.4), whereas relation-specific kernels require counts for each relation. Our fraud experiments use simplified graph views. For graph-aware language models, application of the certificate would require a well-defined input channel, admissible support and evaluable class event satisfying the same assumptions. Prompt-template, serialization and retrieval changes are outside the graph perturbation model considered here.

Acknowledgments and Disclosure of Funding

This work is supported by JST PRESTO JPMJPR23P5, JST CREST JPMJCR21M2, JST NEXUS JPMJNX25C4. The authors declare no competing interests related to this work.

References

Aleksandar Bojchevski, Johannes Gasteiger, and Stephan Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Dawei Cheng, Yao Zou, Sheng Xiang, and Changjun Jiang. Graph neural networks for financial fraud detection: A review. *arXiv preprint*, 2024.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017.

Blaise Delattre, Hengyu Wu, Paul Caillon, Wei Yang Bryan Lim, and Yang Cao. Certified robustness under heterogeneous perturbations via hybrid randomized smoothing. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Yushun Dong, Binchi Zhang, Hanghang Tong, and Jundong Li. Certified defense on the fairness of graph neural networks. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2026.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S. Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020.

Shangbin Feng, Herun Wan, Ningnan Wang, and Minnan Luo. BotRGCN: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2021.

Zhidong Gao, Rui Hu, and Yanmin Gong. Certified robustness of graph classification against topology attack with randomized smoothing. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, 2020.

Jinyuan Jia, Binghui Wang, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of community detection against adversarial structural perturbation via randomized smoothing. In *Proceedings of The Web Conference (WWW)*, 2020.

Jiate Li and Binghui Wang. AGNNCert: Defending graph neural networks against arbitrary perturbations with deterministic certification. In *34th USENIX Security Symposium (USENIX Security 25)*, 2025.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In *Proceedings of The Web Conference 2021*, 2021.

Zhiwei Liu, Yingtong Dou, Philip S. Yu, Yutong Deng, and Hao Peng. Alleviating the inconsistency problem of applying graph neural network to fraud detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. What Yelp Fake Review Filter Might Be Doing? *Proceedings of the International AAAI Conference on Web and Social Media*, 2013.

Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry. *Communications Materials*, 2022.

Yan Scholten, Jan Schuchardt, Simon Geisler, Aleksandar Bojchevski, and Stephan Günnemann. Randomized message-interception smoothing: Gray-box certificates for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yan Scholten, Jan Schuchardt, Aleksandar Bojchevski, and Stephan Günnemann. Hierarchical randomized smoothing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shuhao Shi, Kai Qiao, Zihao Liu, Jie Yang, Chen Chen, Jian Chen, and Bin Yan. MGTAB: A multi-relational graph-based Twitter account detection benchmark. *Neurocomputing*, 2025.

Lichao Sun, Yingtong Dou, Carl Yang, Kai Zhang, Ji Wang, Philip S. Yu, Lifang He, and Bo Li. Adversarial attack and defense on graph data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of graph neural networks against adversarial structural perturbation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021.

Lanjun Wang, Xinran Qiao, Yanwei Xie, Weizhi Nie, Yongdong Zhang, and Anan Liu. My brother helps me: Node injection based adversarial attack on social bot detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

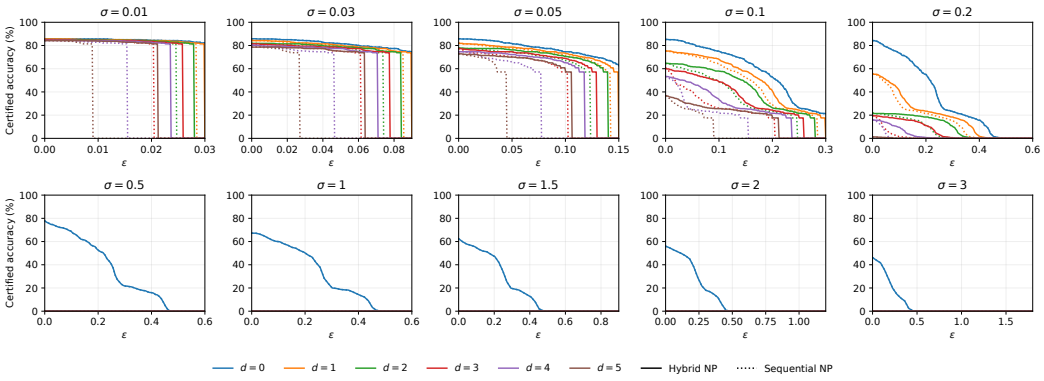
Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I. Weidele, Claudio Bellei, Tom Robinson, and Charles E. Leiserson. Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint*, 2019.

Zaishuo Xia, Han Yang, Binghui Wang, and Jinyuan Jia. GNNCert: Deterministic certification of graph neural networks against adversarial perturbations. In *The Twelfth International Conference on Learning Representations*, 2024.

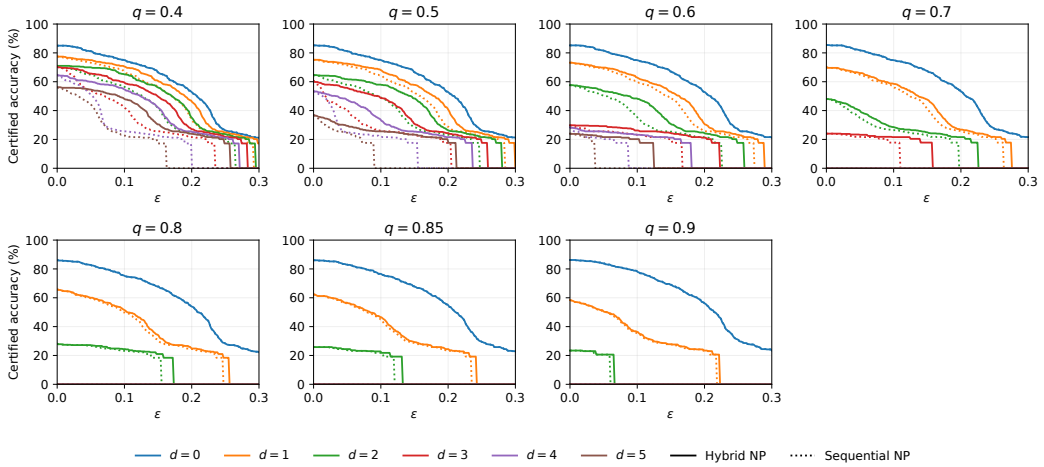
Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Daniel Zügner and Stephan Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 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 and Data Mining*, 2018.



(a) Amazon Fraud diagnostic sweep: varying σ with fixed $q = 0.5$ (effective retention $\rho = 0.75$).



(b) Amazon Fraud: varying q with fixed $\sigma = 0.1$.

Figure 4: Amazon Fraud ablations for an order-2 BWGNN with row-normalized adjacency and target-only feature smoothing ($\mathcal{C} = \{v\}$). Certified accuracy is the percentage of 500 common test targets that are correctly classified and certified, using $N_0 = 50$, $N = 10\,000$ independent estimation draws per target, and $\alpha = 10^{-3}$. Colors denote $d \in \{0, \ldots, 5\}$; solid and dotted curves show Hybrid NP and Sequential NP, respectively.

A Additional Experiments

A.1 Ablations

Ablation Study We use the Amazon graph and BWGNN checkpoint of Section 4.2, with target-only feature smoothing for these parameter sweeps. We sweep the feature smoothing noise σ and the overwrite-keep probability q on Amazon Fraud (Figure 4). The two parameters control, respectively, the feature radius/coverage tradeoff and the topology budget that admits a non-trivial certificate. Larger σ can widen the certified ϵ-radius but may reduce the clean smoothed margin, while larger q (equivalently larger $\rho = (1 + q)/2$) changes the topology likelihood-ratio atoms and hence the certified discrete budget. At $q = 0.5$, every tested $\sigma \geq 0.5$ gives zero certified accuracy for $d \geq 1$, even at $\epsilon = 0$; positive coverage remains on the feature-only axis $d = 0$.

A.2 Topology-only baseline

Figure 5 reports the topology-only baseline (Wang et al., 2021) at $p = \rho = 0.7$ ($q = 0.4$) across the four datasets used in the rest of the paper. At K=0, certified accuracy matches clean accuracy (87.6% on Amazon, 69.0% on YelpChi, 72.0% on PubMed and 87.0% on MGTAB), and decays slowly

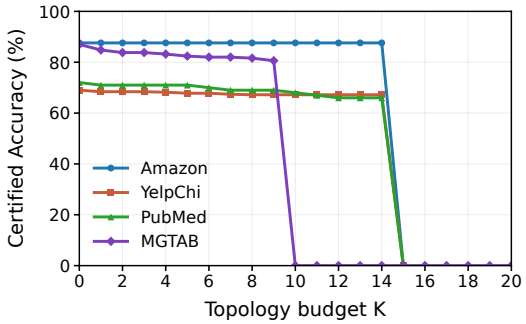


Figure 5: Topology-only certified accuracy as a function of the edge-flip budget K at effective topology retention $p = \rho = 0.7$ (overwrite-keep probability $q = 0.4$), for the four benchmarks: Amazon, YelpChi, PubMed, and MGTAB. Each curve is obtained from the Neyman–Pearson threshold on the smoothed classifier; the abrupt drop to 0 marks the budget at which the threshold $\underline{p_A}$ exceeds the maximum empirical lower confidence bound across all certified nodes.

as the attacker's topology budget grows. The certificate is non-vacuous up to $K{=}14$ on Amazon, YelpChi and PubMed, and up to $K{=}9$ on MGTAB; beyond these budgets every node falls below the Neyman–Pearson threshold and the certificate becomes vacuous. While these curves look reassuring, they only account for adversarial edge flips: Section 4.3 of the main paper shows that even at $K{=}1$ a joint attacker allowed an ℓ_2-bounded feature perturbation changes the estimated smoothed predictions of a substantial fraction of the nodes counted as certified here.

B Algorithms

Algorithm 1 gives the sufficient top-class Hybrid NP certificate. Sequential NP replaces the joint NP calculation by $G_\varepsilon(T_d(p_A^{\mathrm{lb}}))$, or by the maximum of both orders in Proposition 5. The topology map is a lower bound on the feature-marginalized score, not the exact worst case of the fixed GNN.

C Experimental Details

C.1 Main certification protocol

All main certified-accuracy curves use global Gaussian feature smoothing ($\mathcal{C} = V$), $N = 10\,000$ independent estimation draws per node, and overall failure probability $\alpha = 10^{-3}$. An independent pilot uses $N_0 = 50$ on Amazon/YelpChi, $N_0 = 100$ on MGTAB, and $N_0 = 10$ on PubMed. For class set $\mathcal{Y}$, we allocate failure probability $\alpha/|\mathcal{Y}|$ to the selected class's one-sided Clopper–Pearson lower bound and to each other class's upper bound. The largest competing upper bound enters the top-two condition in Appendix D.3; the selected class is excluded from the competitors, including at pilot ties. The same probability bounds are reused across all budgets and radii. We report one checkpoint per dataset, trained with random seed 42. Table 4 summarizes the training settings. Evaluation uses an NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of GPU memory. The 1,000-node PubMed evaluation takes approximately 40 minutes, including CPU certificate computation.

Dataset	model	split	seeds	N	α	optimizer	epochs	learning rate	checkpoint metric	certified nodes
YelpChi	BWGNN	CARE/DGFraud stratified split	42	10 000	10^{-3}	Adam	200	10^{-2}	validation balanced accuracy	500
Amazon Fraud	BWGNN	CARE/DGFraud stratified split	42	10 000	10^{-3}	Adam	200	5×10^{-3}	validation balanced accuracy	500
MGTAB	2-layer GCN	default train/validation/test split	42	10 000	10^{-3}	Adam	200	10^{-2}	validation AUC	400
PubMed	2-layer GCN	public Planetoid split	42	10 000	10^{-3}	Adam	200	10^{-2}	validation accuracy	1 000

Table 4: Reproducibility details for the main certified-accuracy curves. The certified-node count is the number of evaluated test nodes used to compute certified accuracy.

Algorithm 1 Hybrid Randomized Smoothing Certificate

Require: Graph $(\mathbf{A}, \mathbf{X})$, target v, fixed supports $(S_v, \mathcal{C})$, channel (p^-, p^+, σ), budgets (d, ε), pilot size N_0, estimation size N, failure level α
Ensure: Lower bound $\underline{p}(d, \varepsilon)$ and a label or abstention

Count existing and empty slots m_1, m_0 in S_v
Select candidate class c_A by votes from N_0 pilot channel draws
for $i = 1$ to N **do**
    Draw $Z_A^{(i)} \sim p_1(\cdot \mid \mathbf{A})$, independently of the pilot
    Compute $R_i = \mathcal{C} \cap C_v(Z_A^{(i)})$ and retain propagation weights
    Draw $\eta_i \sim \mathcal{N}(0, \sigma^2 I_{|R_i| \times p})$ independently
    $y_i \leftarrow f_v(Z_A^{(i)}, \mathbf{X} + \iota_{R_i} \eta_i)$
end for
$n_A \leftarrow \sum_{i=1}^{N} \mathbf{1}\{y_i = c_A\}$
$p_A^{\mathrm{lb}} \leftarrow \mathrm{ClopperPearsonLower}(n_A, N, \alpha)$
if $p_A^{\mathrm{lb}} = 0$ **then**
    return abstain and the bound 0
end if
$\underline{p}(d, \varepsilon) \leftarrow 1$
for $d^- = 0$ to $\min(d, m_1)$ **do**
    for $d^+ = 0$ to $\min(d - d^-, m_0)$ **do**
        Compute the topology likelihood-ratio atoms and weights
        if $\varepsilon > 0$ **then**
            Solve $F(t^\star; \varepsilon) = p_A^{\mathrm{lb}}$ and evaluate $V(d^-, d^+; \varepsilon)$
        else
            Evaluate the discrete NP value $V(d^-, d^+; 0)$, including boundary randomization
        end if
        $\underline{p}(d, \varepsilon) \leftarrow \min\{\underline{p}(d, \varepsilon), V(d^-, d^+; \varepsilon)\}$
    end for
end for
return c_A if $\underline{p}(d, \varepsilon) > 1/2$, otherwise abstain; also return the bound

C.2 Dataset and model configurations

All topology smoothing parameters in this appendix follow the overwrite-keep convention $q \in [0, 1]$ of Section 3.3; the corresponding effective retention is $\rho = (1 + q)/2$. The main curves use $\mathcal{C} = V$. Parameter sweeps and ablations in Appendix A.1 use target-only feature smoothing.

For YelpChi, we use a BWGNN model with order 2, hidden dimension 64, row adjacency normalization, and no dropout. The graph is constructed using the CARE-style static fraud graph with top-50 neighbors. The checkpoint is trained with Gaussian feature noise of standard deviation $\sigma_{\mathrm{train}} = 0.005$. The main curve uses $q = 0.6$ and $\sigma = 0.005$. Parameter sweeps evaluate overwrite-keep probabilities

$$q \in \{0.60, 0.70, 0.80, 0.85, 0.90, 0.95\}$$

and feature-noise levels

$$\sigma \in \{0.001, 0.002, 0.005, 0.010\}.$$

For each setting, we certify up to 500 stratified randomly selected nodes, with continuous radius grid $\varepsilon \in [0, 0.03]$ using 200 grid points and discrete topology budgets $d \in \{0, 1, 2, 3, 4, 5\}$.

For Amazon Fraud, we use the same BWGNN-row architecture with order 2, hidden dimension 64, and CARE-style static graph construction with top-50 neighbors. The checkpoint used for the main certified-accuracy curve is trained with $\sigma_{\mathrm{train}} = 0.5$ and certified with $q = 0.5$ (effective retention $\rho = 0.75$) and $\sigma = 0.1$. Figure 4 varies q at fixed $\sigma = 0.1$ over

$$q \in \{0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90\}$$

and varies σ at fixed $q = 0.5$ over

$$\sigma \in \{0.01, 0.03, 0.05, 0.10, 0.20, 0.50, 1.00, 1.50, 2.00, 3.00\}.$$

Both sweeps use the same 500 stratified test targets, $N_0 = 50$, $N = 10\,000$, $\alpha = 10^{-3}$, and $d \in \{0, \ldots, 5\}$, with the probability bounds of Appendix C.1. For each σ in the sweep, 200 radius-grid points span $\varepsilon \in [0, 3\sigma]$. For the main certified-accuracy curve, we certify 500 stratified randomly selected nodes and evaluate topology budgets $d \in \{0, \ldots, 5\}$, with additional radii above 0.3 evaluated from the same probability bounds. The displayed axis ends at approximately 0.3324. The $q = 0.5$ ablation is a diagnostic smoothing-parameter sweep; it is not the degeneracy point of the topology likelihood ratio under the overwrite parameterization (which corresponds to $q = 0$).

For MGTAB, we train a two-layer GCN on the default train/validation/test split with train-split feature standardization. The checkpoint is trained and certified with $\sigma_{\text{train}} = \sigma = 0.4$ and overwrite-keep probability $q = 0.4$, corresponding to effective topology retention $\rho = 0.7$. We certify 400 stratified test nodes over $d \in \{0, \ldots, 8\}$ with 200 grid points in $\varepsilon \in [0, 1.00]$ and additional radii above 1 in Figure 2. Certification at every radius uses the same per-node Monte Carlo estimates and confidence bounds. The last positive sampled radius is approximately 1.2663; the displayed axis ends at approximately 1.3296, adding a 5% margin. These values are grid-based endpoints rather than exact continuous robustness thresholds.

For PubMed, we train a two-layer GCN on the public Planetoid split. The checkpoint is trained with $\sigma_{\text{train}} = 0.04$ and certified with $\sigma = 0.01$ and overwrite-keep probability $q = 0.7$ (effective retention $\rho = 0.85$). We certify 1 000 test nodes over $d \in \{0, 1, 2, 3\}$ with 200 radius-grid points over $\varepsilon \in [0, 0.1]$. Figure 2 displays the positive extent with a 5% margin, ending near 0.0327.

Noise mismatch on PubMed. For PubMed, the checkpoint is trained with $\sigma_{\text{train}} = 0.04$ and certified with $\sigma = 0.01$. We use the larger training noise as a regularizer for the smoothed classifier, while the smaller certification noise improves the certified radius–accuracy tradeoff on the evaluated grid. All reported certificates use the test-time value $\sigma = 0.01$; therefore the certificate remains valid for the stated smoothing distribution.

Across datasets, we use weighted negative log-likelihood for the fraud datasets to handle class imbalance, select checkpoints by the validation metric in Table 4, and apply independent Gaussian feature noise to every graph feature coordinate for the main curves.

C.3 Attack settings

The joint attacks in Figure 3 use the `NettackL2PGDv2` backend with one edge flip precomputed before feature optimization. The GCN datasets use the Nettack GCN surrogate; the fraud BWGNNs use direct greedy candidate scoring. Only the target node's feature row is modified (`feature_attack_scope=local`), so its ℓ_2 norm equals the total Frobenius norm. The untargeted objective is the base model's negative log-likelihood, without expectation-over-transformation sampling. Feature PGD uses 20 steps, normalized gradients, step size $\varepsilon/10$, and a single run starting at zero perturbation, with no additional random restarts. Each update is projected onto the radius-ε ball and clipped to the clean feature matrix's minimum and maximum values.

Training and attack seed are 42. We select correctly classified, topology-certified targets at $d = 1$, ranked by decreasing certified edge budget and confidence bound, then by node index, with a cap of 25 per dataset. The numbers of attacked targets and corresponding norm grids are:

Dataset	Attacked targets	ε grid
PubMed	17	$0.005, 0.01, 0.02, 0.03, 0.05$
Amazon Fraud	20	$0.05, 0.1, 0.2, 0.3, 0.5$
YelpChi	13	$0.001, 0.002, 0.005, 0.01, 0.02, 0.03, 0.05, 0.1$
MGTAB	18	$2, 3, 4, 5, 6$

Topology certification uses $N_0 = 50$, $N = 200$, and $\alpha = 10^{-3}$, with no confidence-budget rescaling. Its overwrite-keep probability is $q = 0.7$ on PubMed, $q = 0.6$ on both fraud datasets, and $q = 0.4$ on MGTAB. Clean and attacked topology-smoothed predictions are each estimated from 200 fresh votes in batches of 100. Success is a change of these estimated labels; it is not a base-model prediction flip or a high-confidence statistical test of a flip. The Hybrid control uses independent pilot/estimation budgets $N_0 = 50$, $N = 200$, $\alpha = 10^{-3}$, estimation batches of 64, and 200 evaluation votes. Its (q, σ) values are $(0.7, 0.01)$ on PubMed, $(0.5, 0.1)$ on Amazon, $(0.6, 0.005)$ on YelpChi, and $(0.4, 1.8)$ on MGTAB, with Gaussian noise on the target row.

On MGTAB, the attack changes the estimated smoothed prediction for 1, 9, 12, and 18 of the 18 targets at $\varepsilon = 2, 3, 4, 6$, respectively.

D Multiclass Certification

Throughout the theoretical appendix, $\mathcal{B}_{d,\varepsilon}$ denotes the fixed-support threat set $\mathcal{B}^{\mathcal{C}}_{d,\varepsilon}$ of (3); all feature changes are supported on the same set $\mathcal{C}$ used by the smoothing channel.

The binary formulation of Corollary 1 extends naturally to multiclass classification via the standard "top-class margin" reduction.

D.1 Multiclass smoothed classifier

Definition 1 (Multiclass smoothed classifier)**.** *Let $f : \{0,1\}^{n\times n} \times \mathbb{R}^{n\times p} \to [C]$ be a base classifier with C classes. Under the hybrid smoothing channel $p(\cdot \mid \mathbf{A}, \mathbf{X})$, define the per-class smoothed scores*

$$g_c(\mathbf{A}, \mathbf{X}) := \mathbb{P}\big(f(\mathbf{Z}_A, \mathbf{Z}_X) = c\big), \qquad c \in [C],$$

and the smoothed classifier

$$\hat{f}(\mathbf{A}, \mathbf{X}) := \operatorname{argmax}_{c\in[C]} g_c(\mathbf{A}, \mathbf{X}),$$

with ties broken by a fixed rule (e.g., smallest index). Let $c_A := \hat{f}(\mathbf{A}, \mathbf{X})$ denote the predicted class on the clean input.

Definition 2 (Top-class and runner-up probabilities)**.** *Define*

$$p_A := g_{c_A}(\mathbf{A}, \mathbf{X}) = \max_{c\in[C]} g_c(\mathbf{A}, \mathbf{X}), \qquad p_B := \max_{c\neq c_A} g_c(\mathbf{A}, \mathbf{X}).$$

Note that $p_A + p_B \le 1$ and $p_A \ge p_B$ by definition.

D.2 Certification via the top-class event

The key observation is that $\hat{f}(\mathbf{A}', \mathbf{X}') = c_A$ whenever $g_{c_A}(\mathbf{A}', \mathbf{X}') > \max_{c\neq c_A} g_c(\mathbf{A}', \mathbf{X}')$. A sufficient condition is $g_{c_A}(\mathbf{A}', \mathbf{X}') > \frac{1}{2}$, since all class probabilities sum to 1. The binary NP machinery applies directly to the event $\{f(\mathbf{Z}_A, \mathbf{Z}_X) = c_A\}$.

Theorem 2 (Multiclass hybrid certificate (sufficient condition))**.** *Under the same setup as Corollary 1, define the binary indicator $h_{c_A}(\mathbf{Z}_A, \mathbf{Z}_X) := \mathbf{1}\{f(\mathbf{Z}_A, \mathbf{Z}_X) = c_A\}$. Applying Corollary 1 to h_{c_A} with clean mass p_A yields a lower bound $\underline{p}_{\mathrm{hyb}}(d, \varepsilon)$ on $\inf_{(\mathbf{A}',\mathbf{X}')\in\mathcal{B}_{d,\varepsilon}} g_{c_A}(\mathbf{A}', \mathbf{X}')$.*

The smoothed classifier $\hat{f}$ is certifiably (d, ε)-robust if

$$\underline{p}_{\mathrm{hyb}}(d, \varepsilon) > \tfrac{1}{2}. \tag{4}$$

Proof. If $\underline{p}_{\mathrm{hyb}}(d, \varepsilon) > \frac{1}{2}$, then for every $(\mathbf{A}', \mathbf{X}') \in \mathcal{B}_{d,\varepsilon}$, $g_{c_A}(\mathbf{A}', \mathbf{X}') \ge \underline{p}_{\mathrm{hyb}}(d, \varepsilon) > \frac{1}{2}$. Since $\sum_c g_c(\mathbf{A}', \mathbf{X}') = 1$, no other class $c \neq c_A$ can have $g_c(\mathbf{A}', \mathbf{X}') \ge \frac{1}{2}$, so c_A remains the argmax. □

D.3 Tighter certificate using top-two classes

The $\frac{1}{2}$-threshold can be sharpened by exploiting the runner-up probability.

Theorem 3 (Multiclass hybrid certificate (top-two sufficient condition))**.** *In addition to the lower bound $\underline{p}_{\mathrm{hyb}}(d, \varepsilon)$ on the top class from Theorem 2, define an* upper bound *on the runner-up class under adversarial perturbation.*

For any $(\mathbf{A}', \mathbf{X}') \in \mathcal{B}_{d,\varepsilon}$, define the NP upper *value for the runner-up:*

$$U_{\mathrm{NP}}(\mathbf{A}', \mathbf{X}') := \sup_{\substack{h\in\mathcal{H}:\\ \mathbb{E}_{p(\cdot\mid\mathbf{A},\mathbf{X})}[h]=p_B}} \mathbb{E}_{p(\cdot\mid\mathbf{A}',\mathbf{X}')}[h].$$

By Neyman–Pearson, the supremum is achieved by an upper-tail test, with boundary randomization when the ratio law is atomic. At a positive feature shift, it is an indicator $h^\star = \mathbf{1}\{\gamma \geq s^\star\}$. Define

$$\overline{p}_B(d,\varepsilon) := \sup_{(\mathbf{A}',\mathbf{X}')\in\mathcal{B}_{d,\varepsilon}} U_{\mathrm{NP}}(\mathbf{A}',\mathbf{X}').$$

Then the smoothed classifier is certifiably (d,ε)-robust if

$$\underline{p}_{\mathrm{hyb}}(d,\varepsilon) > \overline{p}_B(d,\varepsilon). \tag{5}$$

Proof. Fix any $(\mathbf{A}',\mathbf{X}') \in \mathcal{B}_{d,\varepsilon}$. By Neyman–Pearson (applied to the event $\{f = c_A\}$ with clean mass p_A), $g_{c_A}(\mathbf{A}',\mathbf{X}') \geq V_{\mathrm{NP}}(\mathbf{A}',\mathbf{X}') \geq \underline{p}_{\mathrm{hyb}}(d,\varepsilon)$. By the "reverse" Neyman–Pearson (applied to the event $\{f = c_B\}$ for any $c_B \neq c_A$ with clean mass at most p_B), $g_{c_B}(\mathbf{A}',\mathbf{X}') \leq U_{\mathrm{NP}}(\mathbf{A}',\mathbf{X}') \leq \overline{p}_B(d,\varepsilon)$. Here the NP upper value is non-decreasing in the prescribed clean mass, so the same bound covers every competitor with clean mass at most p_B. Hence $g_{c_A}(\mathbf{A}',\mathbf{X}') > g_{c_B}(\mathbf{A}',\mathbf{X}')$ for all $c_B \neq c_A$, and c_A remains the prediction. □

Remark 1 (Computing the upper bound $\overline{p}_B$)**.** *For $\varepsilon > 0$, the NP upper value U_{NP} has the same structure as V_{NP} but with the likelihood-ratio test reversed ($\gamma \geq s$ instead of $\gamma \leq t$). Under the i.i.d. topology kernel, define the "reverse capacity map"*

$$\bar{F}(s;\varepsilon) := \sum_{k=0}^{d_-+d_+} w_k^{\mathrm{cln}}\ \bar{\Phi}\left(\frac{\varepsilon}{2\sigma} + \frac{\log s - \lambda_k}{\varepsilon/\sigma}\right),$$

where $\bar{\Phi} = 1 - \Phi$. Solve $\bar{F}(s^\star;\varepsilon) = p_B$ and evaluate

$$U(d_-,d_+;\varepsilon) = \sum_{k=0}^{d_-+d_+} w_k^{\mathrm{adv}}\ \bar{\Phi}\left(\frac{\varepsilon}{2\sigma} + \frac{\log s^\star(\varepsilon) - \lambda_k}{\varepsilon/\sigma} - \frac{\varepsilon}{\sigma}\right).$$

The worst case over (d_-,d_+) is a maximization (not minimization): $\overline{p}_B(d,\varepsilon) = \max_{d_-+d_+\leq d} U(d_-,d_+;\varepsilon)$. The cost is identical to the lower bound: $O(d^2)$ evaluations. At $\varepsilon = 0$, use the discrete upper-tail NP test with boundary randomization. Equivalently, $U = 1 - V|_{p_A = 1-p_B}$, which also shows that the upper value is non-decreasing in the feature radius.

D.4 Finite-sample multiclass certificate

In practice, p_A and p_B are estimated from N Monte Carlo samples $\{f(\mathbf{Z}_A^{(i)},\mathbf{Z}_X^{(i)})\}_{i=1}^N$. An independent pilot selects a candidate class c_A before these estimation draws. Conditional on the pilot, define $p_A = g_{c_A}(\mathbf{A},\mathbf{X})$ and $p_B = \max_{c\neq c_A} g_c(\mathbf{A},\mathbf{X})$. The following bounds apply even if the pilot selected the wrong class; a successful certificate proves that its candidate is the smoothed winner.

Proposition 3 (Finite-sample valid certificate)**.** *Let $\hat{n}_{c_A} = \#\{i : f(\mathbf{Z}_A^{(i)},\mathbf{Z}_X^{(i)}) = c_A\}$ be the count for the top class and $\hat{n}_{c_B} = \max_{c\neq c_A}\#\{i : f(\mathbf{Z}_A^{(i)},\mathbf{Z}_X^{(i)}) = c\}$ be the count for the runner-up. Define*

$$p_A^{\mathrm{lb}} := \mathrm{BinomLB}(\hat{n}_{c_A},\, N,\, 1-\alpha/2), \qquad p_B^{\mathrm{ub}} := \mathrm{BinomUB}(\hat{n}_{c_B},\, N,\, 1-\alpha/2),$$

where BinomLB *and* BinomUB *are one-sided Clopper–Pearson confidence bounds. Then with probability at least $1-\alpha$ over the Monte Carlo samples:*

(i) $p_A^{\mathrm{lb}} \leq p_A$ and $p_B^{\mathrm{ub}} \geq p_B$;

(ii) replacing p_A by p_A^{lb} in Theorem 2 yields a valid $(1-\alpha)$-confidence certificate;

(iii) replacing p_A by p_A^{lb} and p_B by p_B^{ub} in Theorem 3 yields a valid $(1-\alpha)$-confidence certificate.

Proof. (i) Condition on the independent pilot. The candidate is now fixed, so its count is binomial with probability p_A. Also fix a class $c_\star \neq c_A$ attaining the true maximum p_B. Its estimation count is binomial with probability p_B, and is no larger than $\hat{n}_{c_B}$. The Clopper–Pearson upper endpoint is non-decreasing in its count, so the upper endpoint based on $\hat{n}_{c_B}$ covers p_B whenever that based on

the count for $c_\star$ does. A union bound over these two events, each at failure level $\alpha/2$, proves the claimed joint coverage. Averaging over the pilot preserves it.

(ii) Since $V(d_-, d_+; \varepsilon)$ (from Corollary 2) is non-decreasing in p_A (the NP-optimal score against a fixed adversary decreases when the clean mass decreases), substituting the conservative $p_A^{\mathrm{lb}} \le p_A$ yields $V(d_-, d_+; \varepsilon)|_{p_A^{\mathrm{lb}}} \le V(d_-, d_+; \varepsilon)|_{p_A}$, hence $\underline{p}_{\mathrm{hyb}}|_{p_A^{\mathrm{lb}}} \le \underline{p}_{\mathrm{hyb}}|_{p_A} \le p_{\mathrm{adv}}(d, \varepsilon)$. The bound remains sound.

(iii) Similarly, $U(d_-, d_+; \varepsilon)$ is non-decreasing in p_B. Substituting $p_B^{\mathrm{ub}} \ge p_B$ gives $\overline{p}_B|_{p_B^{\mathrm{ub}}} \ge \overline{p}_B|_{p_B} \ge \sup g_{c_B}(\mathbf{A}', \mathbf{X}')$. Hence the condition $\underline{p}_{\mathrm{hyb}}|_{p_A^{\mathrm{lb}}} > \overline{p}_B|_{p_B^{\mathrm{ub}}}$ implies the true condition $g_{c_A}(\mathbf{A}', \mathbf{X}') > g_{c_B}(\mathbf{A}', \mathbf{X}')$ for all adversaries. □

Remark 2 (Sample complexity)**.** *For the certificate to be non-vacuous, p_A^{lb} must exceed $\frac{1}{2}$ (or $\overline{p}_B$ for the top-two condition). By standard Clopper–Pearson asymptotics, this requires $N = \Omega(1/\delta^2)$ samples to resolve a gap of $\delta = p_A - \frac{1}{2}$ at confidence level $1 - \alpha$. In practice, $N \in [10^4, 10^5]$ suffices for typical margins. Each sample requires one forward pass of the base GNN on a noisy input $(\mathbf{Z}_A, \mathbf{Z}_X)$, with cost depending on the realized computation support and its required propagation weights (Lemma 1).*

E Proofs

E.1 Proof of Theorem 1

Proof. We establish the NP characterization, radius monotonicity, count sufficiency, and then soundness for a fixed score.

NP characterization. On the common product space with fixed feature support $\mathcal{C}$, set $\Delta = \mathbf{X}' - \mathbf{X}$ and $r = \|\Delta\|_F$. The likelihood ratio of the adversarial and clean measures is

$$\Lambda(Z_A, (Z_X)_{\mathcal{C}}) = \gamma_1(Z_A) \exp\left(\frac{\langle (Z_X - \mathbf{X})_{\mathcal{C}}, \Delta_{\mathcal{C}} \rangle}{\sigma^2} - \frac{r^2}{2\sigma^2} \right).$$

For $r > 0$, the log-Gaussian term has law $\mathcal{N}(-r^2/(2\sigma^2), r^2/\sigma^2)$ under P and $\mathcal{N}(r^2/(2\sigma^2), r^2/\sigma^2)$ under Q, independently of the topology factor. The lower-tail NP test has clean mass $F(t; r)$ and adversarial mass $V(\mathbf{A}'; r)$ from Corollary 1. Since the topology ratio is positive and finite, F is continuous and strictly increasing from zero to one, giving a unique threshold for $p_A \in (0, 1)$. At $r = 0$, use $\mathbf{1}\{\gamma_1 < t\} + \kappa \mathbf{1}\{\gamma_1 = t\}$ to attain clean mass p_A.

Radius monotonicity. By isotropy, rotate and standardize the feature block so that the two Gaussian laws are $\mathcal{N}(0, I)$ and $\mathcal{N}((r/\sigma)e_1, I)$. For $0 \le r_1 \le r_2$ with $r_2 > 0$, the common random map

$$Y \longmapsto (r_1/r_2)Y + \sqrt{1 - (r_1/r_2)^2}\, W, \qquad W \sim \mathcal{N}(0, I),$$

turns the radius-r_2 experiment into the radius-r_1 experiment while retaining Z_A. Pulling back any radius-r_1 test through this map gives a $[0, 1]$-valued radius-r_2 test with identical masses. Thus the NP infimum cannot increase with r. The feature ball's infimum is consequently attained at $r = \varepsilon$ (or $r = 0$ when $\varepsilon = 0$).

Count sufficiency. The topology ratio factors over S_v. Unmodified slots contribute one; on modified slots, homogeneous independent noise gives the log-ratio law of Proposition 4. Under both centers, this law depends only on (d^-, d^+). The independent Gaussian factor depends only on r, so $V_{\mathrm{NP}}(\mathbf{A}', \mathbf{X}') = V(d^-, d^+; r)$, including the randomized topology test at $r = 0$. Every admissible adjacency induces feasible counts. Conversely, each feasible pair is realized by deleting any d^- existing slots and adding any d^+ empty ones. Hence

$$\inf_{\mathbf{A}' \in \mathcal{A}_d} V_{\mathrm{NP}}(\mathbf{A}'; \varepsilon) = \min_{\substack{d^- + d^+ \le d \\ d^- \le m_1,\ d^+ \le m_0}} V(d^-, d^+; \varepsilon) = \underline{p}_{\mathrm{hyb}}(d, \varepsilon).$$

Soundness. A fixed score h with $\mathbb{E}_P[h] = p_A$ is feasible in the NP program for every admissible pair. Therefore

$$g_h(\mathbf{A}', \mathbf{X}') \ge V_{\mathrm{NP}}(\mathbf{A}', \mathbf{X}') \ge V_{\mathrm{NP}}(\mathbf{A}'; \varepsilon) \ge \underline{p}_{\mathrm{hyb}}(d, \varepsilon).$$

Taking the infimum over the admissible pairs proves soundness. No step identifies this relaxed infimum with the true worst case of a fixed GNN. □

E.2 Proof of Theorem 4

Proof. The proof concerns product-form tests, not the implemented sequential baseline. Let the joint likelihood ratio be $\gamma = \gamma_1\gamma_2$. Since γ_2 is the exponential of a Gaussian random variable, its distribution is continuous. Conditional on any atom v of γ_1, the distribution of γ is $v\gamma_2$, also continuous. Because γ_1 has at most countably many atoms, $\mathbb{P}(\gamma = t) = 0$ for every t.

By the Neyman–Pearson lemma with a continuous likelihood ratio, the optimal test $h^\star = \mathbf{1}\{\gamma \le t^\star\}$ is the unique minimizer of adversarial expectation subject to the clean-mass constraint $\mathbb{E}_{p(\cdot|\mathbf{A},\mathbf{X})}[h] = p_A$, up to null sets.

Suppose $h^\star$ were product-form: $\mathbf{1}\{\gamma_1\gamma_2 \le t^\star\} = \mathbf{1}\{\gamma_1 \le t_1\}\mathbf{1}\{\gamma_2 \le t_2\}$ a.s. for some (t_1, t_2). By Assumption 1, γ_1 takes two values $v_1 < v_2$ with positive probability. On $\{\gamma_1 = v_1\}$ the NP region requires $\gamma_2 \le t^\star/v_1$, whereas on $\{\gamma_1 = v_2\}$ it requires $\gamma_2 \le t^\star/v_2$. These thresholds differ, and the interval $(t^\star/v_2, t^\star/v_1]$ has positive γ_2 probability. A product test must use the same γ_2 threshold for both topology atoms, so it cannot agree with $h^\star$ a.s.

Thus the unique NP minimizer is not in the product-form test class. Any product-form test satisfying the same clean-mass constraint differs from $h^\star$ on a positive-measure set and has strictly larger adversarial expectation. For the finite atomic topology channel, there are finitely many lower-tail topology sets. For each feasible set, the clean-mass constraint fixes the Gaussian threshold, so the product-form infimum is attained among finitely many tests. Taking this minimum yields $V_{\text{sep}} > V_{\text{NP}}$. □

E.3 Axis consistency: formal statement and proof (Proposition 1)

We restate the axis-consistency result of the body in formal form, then prove it.

Proposition (Axis consistency of the top-class certificate, formal)**.** *Consider the hybrid smoothing channel $p(Z_A, Z_X \mid \mathbf{A}, \mathbf{X}) = p_1(Z_A \mid \mathbf{A})\, p_2(Z_X \mid \mathbf{X})$ with Gaussian noise on the fixed support $\mathcal{C}$ and identical constant features outside it. Fix the clean smoothed prediction c_A and the binary top-class event h_{c_A}, and let $p_A = \mathbb{E}[h_{c_A}(Z_A, Z_X)]$ under the clean hybrid distribution. Let $\underline{p}_{\text{hyb}}(d, \varepsilon)$ denote the Hybrid NP lower-bound value for this event. Then:*

- Topology-only axis ($\varepsilon = 0$). *With the feature-marginalized score $\bar{h}^X_{c_A}(Z_A) := \mathbb{E}_{Z_X \sim p_2(\cdot|\mathbf{X})}[h_{c_A}(Z_A, Z_X)]$,*

$$\underline{p}_{\text{hyb}}(d, 0) = \underline{p}^{\bar{h}^X}_{\text{topo}}(d),$$

  where $\underline{p}^{\bar{h}^X}_{\text{topo}}(d)$ is the topology-only Neyman–Pearson randomized-smoothing lower bound over $\mathbf{A}' \in \mathcal{A}_d(\mathbf{A}; v)$ for $\bar{h}^X_{c_A}$. When p_1 is the homogeneous binary flip channel on S_v, this is the topology-only graph randomized-smoothing certificate of Wang et al. (2021) applied to $\bar{h}^X_{c_A}$. If additionally $p_2 = \delta_{\mathbf{X}}$, the reduction is exactly the standard topology-only certificate for $\mathbf{A}' \mapsto f_v(\mathbf{A}', \mathbf{X})$.

- Feature-only axis ($d = 0$). *With the topology-marginalized score $\bar{h}^A_{c_A}(Z_X) := \mathbb{E}_{Z_A \sim p_1(\cdot|\mathbf{A})}[h_{c_A}(Z_A, Z_X)]$,*

$$\underline{p}_{\text{hyb}}(0, \varepsilon) = \underline{p}^{\bar{h}^A}_{\text{Gauss}}(\varepsilon) := \Phi\Big(\Phi^{-1}(p_A) - \frac{\varepsilon}{\sigma}\Big),$$

  the Gaussian randomized-smoothing lower bound of Cohen et al. (2019) applied to $\bar{h}^A_{c_A}$. If additionally $p_1 = \delta_{\mathbf{A}}$, this is exactly the standard Gaussian certificate for $\mathbf{X} \mapsto f_v(\mathbf{A}, \mathbf{X})$.

Proof of Proposition 1. First set $\varepsilon = 0$. Every admissible adversarial input has the form $(\mathbf{A}', \mathbf{X})$ with $\mathbf{A}' \in \mathcal{A}_d(\mathbf{A}; v)$. For any fixed $\mathbf{A}'$, the joint likelihood ratio between the adversarial and clean smoothing distributions is

$$\frac{p_1(Z_A \mid \mathbf{A}')\, p_2(Z_X \mid \mathbf{X})}{p_1(Z_A \mid \mathbf{A})\, p_2(Z_X \mid \mathbf{X})} = \frac{p_1(Z_A \mid \mathbf{A}')}{p_1(Z_A \mid \mathbf{A})} =: \gamma_A(Z_A).$$

The joint likelihood ratio is therefore $\sigma(Z_A)$-measurable, so the Neyman–Pearson optimal lower-tail test on the joint product space is also $\sigma(Z_A)$-measurable, of the form $\mathbf{1}\{\gamma_A(Z_A) \le t\}$, up to possible randomization at the threshold.

By Fubini, the clean mass of the top-class event can be written as

$$p_A = \mathbb{E}_{p_1(\cdot|\mathbf{A})\,p_2(\cdot|\mathbf{X})}[h_{c_A}(Z_A, Z_X)] = \mathbb{E}_{p_1(\cdot|\mathbf{A})}[\bar{h}^X_{c_A}(Z_A)],$$

and similarly, for any admissible $\mathbf{A}'$,

$$\mathbb{E}_{p_1(\cdot|\mathbf{A}')\,p_2(\cdot|\mathbf{X})}[h_{c_A}(Z_A, Z_X)] = \mathbb{E}_{p_1(\cdot|\mathbf{A}')}[\bar{h}^X_{c_A}(Z_A)].$$

Thus the joint NP value on the product space equals the topology-only NP value on the marginal Z_A-space for the feature-marginalized score $\bar{h}^X_{c_A}$. Taking the worst case over $\mathbf{A}' \in \mathcal{A}_d(\mathbf{A}; v)$ gives $\underline{p}_{\mathrm{hyb}}(d, 0) = \underline{p}^{\bar{h}^X}_{\mathrm{topo}}(d)$. If $p_2 = \delta_{\mathbf{X}}$, then $\bar{h}^X_{c_A}(Z_A)$ reduces to $\mathbf{1}\{f_v(Z_A, \mathbf{X}) = c_A\}$, yielding the standard topology-only top-class-event certificate.

Now set $d = 0$. Every admissible adversarial input has the form $(\mathbf{A}, \mathbf{X}')$ with $\|\mathbf{X}' - \mathbf{X}\|_F \le \varepsilon$. For any fixed $\mathbf{X}'$, the joint likelihood ratio is

$$\frac{p_1(Z_A \mid \mathbf{A})\, p_2(Z_X \mid \mathbf{X}')}{p_1(Z_A \mid \mathbf{A})\, p_2(Z_X \mid \mathbf{X})} = \frac{p_2(Z_X \mid \mathbf{X}')}{p_2(Z_X \mid \mathbf{X})} =: \gamma_X(Z_X),$$

so the joint likelihood ratio is $\sigma(Z_X)$-measurable and the NP-optimal joint test is also $\sigma(Z_X)$-measurable. By Fubini,

$$p_A = \mathbb{E}_{p_1(\cdot|\mathbf{A})\,p_2(\cdot|\mathbf{X})}[h_{c_A}(Z_A, Z_X)] = \mathbb{E}_{p_2(\cdot|\mathbf{X})}[\bar{h}^A_{c_A}(Z_X)],$$

and likewise, under $\mathbf{X}'$,

$$\mathbb{E}_{p_1(\cdot|\mathbf{A})\,p_2(\cdot|\mathbf{X}')}[h_{c_A}(Z_A, Z_X)] = \mathbb{E}_{p_2(\cdot|\mathbf{X}')}[\bar{h}^A_{c_A}(Z_X)].$$

The joint NP value on the product space therefore equals the Gaussian NP value on the marginal Z_X-space for the topology-marginalized score $\bar{h}^A_{c_A}$.

For isotropic Gaussian smoothing, the Neyman–Pearson Gaussian smoothing lower bound at radius $r = \|\mathbf{X}' - \mathbf{X}\|_F$ is $\Phi(\Phi^{-1}(p_A) - r/\sigma)$. Minimizing over $\|\mathbf{X}' - \mathbf{X}\|_F \le \varepsilon$ gives

$$\underline{p}^{\bar{h}^A}_{\mathrm{Gauss}}(\varepsilon) = \Phi\Big(\Phi^{-1}(p_A) - \frac{\varepsilon}{\sigma}\Big).$$

If $p_1 = \delta_{\mathbf{A}}$, then $\bar{h}^A_{c_A}(Z_X)$ reduces to $\mathbf{1}\{f_v(\mathbf{A}, Z_X) = c_A\}$, yielding the standard Gaussian top-class-event certificate. □

F Explicit certificate under the homogeneous i.i.d. topology kernel

We now specialize Corollary 1 to the homogeneous per-edge corruption kernel of Section 3.3, giving closed-form expressions for the capacity map and the NP value in terms of flip counts.

F.1 Atoms of the topology likelihood ratio

Proposition 4 (Atomic decomposition of $\log \gamma_1$)**.** *Under the homogeneous i.i.d. topology kernel on a candidate set S, let $(\mathbf{A}_{\mathrm{adv}})$ be an adversary with deletion count $d_- := \#\{(i,j) \in S : A_{ij} = 1,\, A_{\mathrm{adv},ij} = 0\}$ and addition count $d_+ := \#\{(i,j) \in S : A_{ij} = 0,\, A_{\mathrm{adv},ij} = 1\}$. Define the per-coordinate log-ratios*

$$\alpha := \log \frac{1 - p_+}{p_-}, \qquad \beta := \log \frac{p_+}{1 - p_-}.$$

Then

$$\log \gamma_1(\mathbf{Z}_A) \;=\; (\alpha - \beta)\, K \;+\; d_-\beta \;-\; d_+\alpha, \tag{6}$$

where $K := K_- + K_+$ and

$$K_- := \#\big\{(i,j) \in S_- : Z_{A,ij} = 0\big\}, \qquad K_+ := \#\big\{(i,j) \in S_+ : Z_{A,ij} = 1\big\},$$

and $S_- := \{(i,j) \in S : A_{ij} = 1,\, A_{\mathrm{adv},ij} = 0\}$, $S_+ := \{(i,j) \in S : A_{ij} = 0,\, A_{\mathrm{adv},ij} = 1\}$.

Under the clean distribution $p_1(\cdot \mid \mathbf{A})$:

$$K_- \sim \mathrm{Bin}(d_-,\, p_-), \qquad K_+ \sim \mathrm{Bin}(d_+,\, p_+), \qquad K_- \perp\!\!\!\perp K_+. \tag{7}$$

Under the adversarial distribution $p_1(\cdot \mid \mathbf{A}_{\mathrm{adv}})$:

$$K_- \sim \mathrm{Bin}(d_-,\, 1 - p_+), \qquad K_+ \sim \mathrm{Bin}(d_+,\, 1 - p_-), \qquad K_- \perp\!\!\!\perp K_+. \tag{8}$$

In particular, $\log \gamma_1$ takes at most $d_- + d_+ + 1$ distinct values, indexed by $k \in \{0, 1, \ldots, d_- + d_+\}$.

Proof. Partition S into S_-, S_+, and $S_0 := S \setminus (S_- \cup S_+)$. On S_0, we have $A_{ij} = A_{\mathrm{adv},ij}$, so each per-coordinate ratio $p(Z_{A,ij} \mid A_{\mathrm{adv},ij})/p(Z_{A,ij} \mid A_{ij}) = 1$. Hence

$$\log \gamma_1(\mathbf{Z}_A) = \sum_{(i,j)\in S_-} \log \frac{p(Z_{A,ij} \mid 0)}{p(Z_{A,ij} \mid 1)} + \sum_{(i,j)\in S_+} \log \frac{p(Z_{A,ij} \mid 1)}{p(Z_{A,ij} \mid 0)}.$$

For a deletion site $(i,j) \in S_-$ (where $A_{ij} = 1$, $A_{\mathrm{adv},ij} = 0$):

$$\log \frac{p(Z_{A,ij} \mid 0)}{p(Z_{A,ij} \mid 1)} = \begin{cases} \alpha = \log \frac{1-p_+}{p_-} & \text{if } Z_{A,ij} = 0, \\ \beta = \log \frac{p_+}{1-p_-} & \text{if } Z_{A,ij} = 1. \end{cases}$$

For an addition site $(i,j) \in S_+$ (where $A_{ij} = 0$, $A_{\mathrm{adv},ij} = 1$):

$$\log \frac{p(Z_{A,ij} \mid 1)}{p(Z_{A,ij} \mid 0)} = \begin{cases} -\alpha & \text{if } Z_{A,ij} = 0, \\ -\beta & \text{if } Z_{A,ij} = 1. \end{cases}$$

Thus, writing each indicator in terms of K_- and K_+:

$$\begin{aligned} \log \gamma_1 &= K_-\alpha + (d_- - K_-)\beta + K_+(-\beta) + (d_+ - K_+)(-\alpha) \\ &= (\alpha - \beta)(K_- + K_+) + d_-\beta - d_+\alpha \\ &= (\alpha - \beta)K + d_-\beta - d_+\alpha. \end{aligned}$$

The distributional claims follow from the independence of the per-edge noise and the marginal distributions: under $p_1(\cdot \mid \mathbf{A})$, $Z_{A,ij} \sim \mathrm{Bernoulli}(1-p_-)$ for existing edges and $Z_{A,ij} \sim \mathrm{Bernoulli}(p_+)$ for non-edges. Under $p_1(\cdot \mid \mathbf{A}_{\mathrm{adv}})$, the roles are swapped on $S_- \cup S_+$. □

F.2 Closed-form capacity map and NP value

Using the atomic decomposition above, we can write the capacity map F and the NP value V as finite sums.

Corollary 2 (Explicit hybrid capacity map and certificate)**.** *For $\varepsilon > 0$, let $\lambda_k := (\alpha - \beta)k + d_-\beta - d_+\alpha$ denote the k-th atom of $\log \gamma_1$, for $k = 0, 1, \ldots, d_- + d_+$. Define the clean and adversarial weights*

$$w_k^{\mathrm{cln}} := \mathbb{P}(K = k \mid \mathbf{A}) = \sum_{j=\max(0,k-d_+)}^{\min(k,d_-)} \binom{d_-}{j} p_-^j (1-p_-)^{d_- - j} \binom{d_+}{k-j} p_+^{k-j} (1-p_+)^{d_+ - (k-j)}, \tag{9}$$

$$\begin{aligned} w_k^{\mathrm{adv}} &:= \mathbb{P}(K = k \mid \mathbf{A}_{\mathrm{adv}}) \\ &= \sum_{j=\max(0,k-d_+)}^{\min(k,d_-)} \binom{d_-}{j} (1-p_+)^j \, p_+^{d_- - j} \binom{d_+}{k-j} (1-p_-)^{k-j} \, p_-^{d_+ - (k-j)}. \end{aligned} \tag{10}$$

Then the hybrid capacity map (Corollary 1) becomes

$$\boxed{F(t;\varepsilon) = \sum_{k=0}^{d_-+d_+} w_k^{\mathrm{cln}} \, \Phi\left(\frac{\varepsilon}{2\sigma} + \frac{\log t - \lambda_k}{\varepsilon/\sigma} \right),} \tag{11}$$

and the NP value at these flip counts is

$$\boxed{V(d_-, d_+; \varepsilon) = \sum_{k=0}^{d_-+d_+} w_k^{\mathrm{adv}} \, \Phi\left(\frac{\varepsilon}{2\sigma} + \frac{\log t^\star(\varepsilon) - \lambda_k}{\varepsilon/\sigma} - \frac{\varepsilon}{\sigma} \right),} \tag{12}$$

where $t^\star(\varepsilon) > 0$ is the unique solution to $F(t^\star(\varepsilon); \varepsilon) = p_A$.

The certified worst-case lower bound under joint budget (d, ε) is

$$\underline{p}_{\mathrm{hyb}}(d, \varepsilon) = \min_{\substack{d_-+d_+ \le d \\ 0 \le d_- \le m_1,\ 0 \le d_+ \le m_0}} V(d_-, d_+; \varepsilon), \tag{13}$$

where $m_1 = \#\{(i,j) \in S : A_{ij} = 1\}$ and $m_0 = \#\{(i,j) \in S : A_{ij} = 0\}$.

Proof. Substitute the atomic decomposition of Proposition 4 into the capacity map and NP value of Corollary 1. The sum over all $\mathbf{Z}_A$ collapses to a sum over the sufficient statistic K, weighted by the corresponding binomial-convolution probabilities. At $\varepsilon = 0$, sort the topology ratio atoms, accumulate clean mass from the smallest ratios, and randomize at the boundary atom to attain p_A. The corresponding adversarial mass defines $V(d_-, d_+; 0)$ without divisions by ε. □

Remark 3 (Computational cost)**.** *Evaluating $F(t; \varepsilon)$ and $V(d_-, d_+; \varepsilon)$ costs $O(d)$ Φ-evaluations each. The root-finding for $t^\star$ requires $O(\log(1/\delta_t))$ iterations of bisection (in $\log t$), where δ_t is the desired precision. The outer minimization* (13) *iterates over at most $\frac{1}{2}(d+1)(d+2)$ pairs (d_-, d_+), i.e., $O(d^2)$ evaluations. Thus the total certificate computation (excluding Monte Carlo estimation of p_A) costs $O(d^3 \log(1/\delta_t))$ Φ-evaluations, which is negligible compared to the N forward passes needed to estimate p_A.*

G Additional Discussion on Separate Certificates

We distinguish the sound sequential maps from marginal rectangle bounds and product-form restrictions, which need not be sound for the joint threat. For a binary class indicator $f : \Omega \to \{0, 1\}$, let $g(\mathbf{A}, \mathbf{X}) := \mathbb{E}[f(\mathbf{Z}_A, \mathbf{Z}_X)]$ be a smoothed classifier built from independent channels $p_1(\mathbf{Z}_A \mid \mathbf{A})$ and $p_2(\mathbf{Z}_X \mid \mathbf{X})$. We write $p_A := g(\mathbf{A}, \mathbf{X})$ for the clean score and $p_{\mathrm{adv}}(d, \varepsilon) := \inf_{(\mathbf{A}', \mathbf{X}') \in \mathcal{B}_{d,\varepsilon}} g(\mathbf{A}', \mathbf{X}')$ for the joint worst-case score. A bound $\underline{p}(d, \varepsilon)$ is *sound* if $\underline{p}(d, \varepsilon) \le p_{\mathrm{adv}}(d, \varepsilon)$.

Sequential NP. For a clean mass p, let $T_d(p) = \min_{d^- + d^+ \le d} V(d^-, d^+; 0)|_{p_A = p}$, with the usual feasibility constraints on the counts, and $G_\varepsilon(p) = \Phi(\Phi^{-1}(p) - \varepsilon/\sigma)$. The effective mass used by the implemented order is $p_{\mathrm{eff}}(d) = T_d(p_A) \le p^h_{\mathrm{adv}}(d, 0)$. Both maps are non-decreasing: for the NP map, scaling a test of clean mass p' by p/p' gives a feasible test at any $p \le p'$; apply this argument to each count pair and take the minimum.

Proposition 5 (Both sequential orders are sound)**.** *For the same clean product channel and fixed supports,*

$$p^{T \to X}_{\mathrm{seq}} = G_\varepsilon(T_d(p_A)), \qquad p^{X \to T}_{\mathrm{seq}} = T_d(G_\varepsilon(p_A))$$

are both lower bounds on $p^h_{\mathrm{adv}}(d, \varepsilon)$. Their maximum is also sound and satisfies $p^{\max}_{\mathrm{seq}} \le \underline{p}_{\mathrm{hyb}}(d, \varepsilon)$. The two orders need not coincide.

Proof. At fixed $\mathbf{A}'$, apply the Gaussian NP map to the bounded test obtained by averaging h over $p_1(\cdot \mid \mathbf{A}')$. For every admissible $\mathbf{X}'$, this gives

$$g_h(\mathbf{A}', \mathbf{X}') \ge G_\varepsilon(g_h(\mathbf{A}', \mathbf{X})) \ge G_\varepsilon(T_d(p_A)).$$

The second inequality uses topology NP soundness and monotonicity. For the reverse order, at a fixed admissible $\mathbf{X}'$ average h over its feature channel to obtain a $[0, 1]$-valued topology test. Topology NP soundness and Gaussian soundness at $\mathbf{A}$ give

$$g_h(\mathbf{A}', \mathbf{X}') \ge T_d(g_h(\mathbf{A}, \mathbf{X}')) \ge T_d(G_\varepsilon(p_A)).$$

Taking the infimum over admissible pairs proves both bounds and their maximum. These arguments hold for every test of clean mass p_A; Proposition 2 therefore proves dominance.

For non-coincidence, take one existing slot with deletion probability $1/4$, addition probability $1/4$, and $d = 1$. Its topology NP map is $T_1(p) = p/3$ for $p \le 3/4$ and $T_1(p) = 3p - 2$ otherwise. At $p_A = 3/4$ and any $\varepsilon > 0$, $G_\varepsilon(T_1(p_A)) = G_\varepsilon(1/4)$ while $T_1(G_\varepsilon(p_A)) = G_\varepsilon(3/4)/3$. Their right derivatives at $\varepsilon/\sigma = 0$ are respectively $-\phi(\Phi^{-1}(1/4))$ and $-\phi(\Phi^{-1}(3/4))/3$, which differ. Thus the orders differ for sufficiently small positive radii. Here ϕ is the standard normal density. □

For the implemented $T \to X$ order, the certified feature radius at threshold τ is

$$\varepsilon_{\mathrm{seq}}(d) = \sigma[\Phi^{-1}(T_d(p_A)) - \Phi^{-1}(\tau)]_+.$$

The dominance result is non-strict: a universal strict improvement over Sequential NP is not asserted. The product-form test result below concerns a different, generally unsound restricted relaxation.

Rectangle composition. The simplest approach takes $\underline{p}_{\mathrm{rect}}(d,\varepsilon) := \min\{p_{\mathrm{adv}}(d,0),\, p_{\mathrm{adv}}(0,\varepsilon)\}$, evaluating each axis with the other held at its clean value. Under the joint budget, the interaction through message passing can cause $g(\mathbf{A}',\mathbf{X}')$ to drop below both marginal worst cases. The two axis guarantees alone therefore do not justify the rectangle bound. It is valid only under the weaker *either/or* model where the adversary perturbs topology **or** features but not both.

Separable Neyman–Pearson. A more refined approach restricts the NP test class to product-form regions. Let $\mathcal{H}_{\mathrm{sep}}$ denote tests $h(\mathbf{Z}_A,\mathbf{Z}_X) = \mathbf{1}\{\gamma_1(\mathbf{Z}_A) \le t_1\} \cdot \mathbf{1}\{\gamma_2(\mathbf{Z}_X) \le t_2\}$, and define

$$V_{\mathrm{sep}}(\mathbf{A}_{\mathrm{adv}},\mathbf{X}_{\mathrm{adv}}) := \inf_{\substack{h\in\mathcal{H}_{\mathrm{sep}}:\\ \mathbb{E}_{p(\cdot\mid\mathbf{A},\mathbf{X})}[h]=p_A}} \mathbb{E}_{p(\cdot\mid\mathbf{A}_{\mathrm{adv}},\mathbf{X}_{\mathrm{adv}})}[h].$$

The thresholds (t_1,t_2) are coupled through the clean-mass constraint, so the program minimizes over this restricted test class. Since $\mathcal{H}_{\mathrm{sep}} \subseteq \mathcal{H}$, we have $V_{\mathrm{sep}} \ge V_{\mathrm{NP}}$ pointwise. A classifier's acceptance event $\{f(\mathbf{Z}_A,\mathbf{Z}_X)=1\}$ that factorizes as $\{\gamma_1 \le t_1\} \cap \{\gamma_2 \le t_2\}$ would be a feasible test for this restricted program, but a general GNN class event need not have that form. Thus the restriction does not give a uniformly sound lower bound. In particular, choosing the joint NP-optimal event as the class event gives adversarial mass V_{NP}, which is strictly below V_{sep} under the assumptions of Theorem 4.

The gap is structurally unavoidable. The gap above is not an artifact of a particular instance. Under mild non-degeneracy, no product-form test can match the joint NP optimum—the gap is a structural consequence of the coupled likelihood ratio.

Assumption 1 (Non-degenerate topology likelihood ratio)**.** *$\gamma_1(\mathbf{Z}_A)$ takes at least two distinct values with positive probability under $p_1(\cdot \mid \mathbf{A})$.*

This holds whenever $d^- + d^+ \ge 1$ and $p_- + p_+ \ne 1$. When $p_- + p_+ = 1$, all log-ratio atoms collapse to zero. In the overwrite parameterization used in experiments, where $p^- = p^+ = (1-q)/2$, the condition $p^- + p^+ = 1$ corresponds to $q = 0$. All main settings use $q > 0$.

Theorem 4 (Non-composability of product-form tests under coupled likelihood ratios)**.** *Let $(\mathbf{A}_{\mathrm{adv}},\mathbf{X}_{\mathrm{adv}})$ be an adversary with $\|\mathbf{X}_{\mathrm{adv}} - \mathbf{X}\|_F = \varepsilon > 0$ and $d^- + d^+ \ge 1$. Under Assumption 1, for every $p_A \in (0,1)$:*

$$V_{\mathrm{sep}}(\mathbf{A}_{\mathrm{adv}},\mathbf{X}_{\mathrm{adv}}) \;>\; V_{\mathrm{NP}}(\mathbf{A}_{\mathrm{adv}},\mathbf{X}_{\mathrm{adv}}).$$

That is, no product-form test can achieve the joint NP optimum: the adaptive threshold structure $\gamma_2 \le t^\star/\gamma_1$ cannot be reproduced by any fixed threshold $\gamma_2 \le t_2$.

Proof in Appendix E.2.

Remark 4. *The product-form theorem above explains why naively composing independent acceptance regions can be misleading. It is not used to claim a universal strict separation between Hybrid NP and the implemented Sequential NP baseline.*

H Feature support and optimality of the NP relaxation

H.1 Fixed feature support and exact local sampling

Let $\mathcal{C} \subseteq V$ be nonempty and fixed once for the target instance, and let $R \subseteq \mathcal{C}$ be the attackable node set. The same $\mathcal{C}$ and structural slot set S_v are used at every admissible center. The feature channel is $Z_X = \mathbf{X} + \iota_{\mathcal{C}}\eta$, with $\eta \sim \mathcal{N}(0,\sigma^2 I_{|\mathcal{C}|\times p})$ independent of the topology draw. For $\Delta = \mathbf{X}' - \mathbf{X}$ with $\mathrm{supp}_{\mathrm{row}}(\Delta) \subseteq R$, the block $\mathbf{X}_{\mathcal{C}^c}$ is identical under both centers.

Proof of Lemma 2. On the common space $(B, Y_{\mathcal{C}})$, the clean and adversarial input measures are

$$P = p_1(\cdot \mid \mathbf{A}) \otimes \mathcal{N}(\mathbf{X}_{\mathcal{C}},\sigma^2 I), \qquad Q = p_1(\cdot \mid \mathbf{A}') \otimes \mathcal{N}(\mathbf{X}_{\mathcal{C}} + \Delta_{\mathcal{C}},\sigma^2 I).$$

Both scores use the same map $a(B,Y_{\mathcal{C}}) = h(B,\iota_{\mathcal{C}}Y_{\mathcal{C}} + \iota_{\mathcal{C}^c}\mathbf{X}_{\mathcal{C}^c})$. Even if B makes nodes outside $\mathcal{C}$ visible, their features remain the same constants in this map. The likelihood ratio is

$$\frac{dQ}{dP}(B,Y_{\mathcal{C}}) = \gamma_1(B)\exp\!\left(\frac{\langle Y_{\mathcal{C}} - \mathbf{X}_{\mathcal{C}},\Delta_{\mathcal{C}}\rangle}{\sigma^2} - \frac{\|\Delta_{\mathcal{C}}\|_F^2}{2\sigma^2}\right).$$

Since $\|\Delta_{\mathcal{C}}\|_F = \|\Delta\|_F \le \varepsilon$, this is precisely the product experiment in Corollary 1. Pulling a classifier event back through the common map gives a feasible NP test with the same clean and adversarial masses, so the bound applies.

For the local implementation, fix a realized topology B and retain all propagation weights required by Lemma 1. With global Gaussian noise η_V, locality gives

$$f_v(B, \mathbf{X} + \eta_V) = \bar{f}_v(B, P_B\mathbf{X} + P_B\eta_V), \qquad P_B\eta_V \sim \mathcal{N}(0, \sigma^2 I_{|C_v(B)|\times p}).$$

The feature noise is independent of the topology. Conditional on each B, sampling just this Gaussian marginal therefore gives the same distribution of predictions as sampling the full matrix. Averaging over either clean or adversarial topology draws preserves the equality. The ambient global product channel supplies the common sample space; the variable computation scope is only an implementation of its marginal. □

Target, global, and clean-frozen supports. With $\mathcal{C} = R = \{v\}$, the certificate permits only changes to X_v, although the topology may connect v to any existing node. With $\mathcal{C} = V$, it covers a Frobenius perturbation over any node support, including newly visible nodes. A set chosen from the clean graph and then frozen under both centers is also valid, but only for feature perturbations inside that fixed set. These channels can give different clean masses and must be evaluated separately.

In contrast, recomputing the noise set from each input center, such as $\mathcal{C}(\mathbf{A}) = \mathcal{N}_L(v; \mathbf{A})$ and $\mathcal{C}(\mathbf{A}') = \mathcal{N}_L(v; \mathbf{A}')$, need not be valid. A row noised only at one center is Gaussian in one measure and deterministic in the other, so the product likelihood-ratio argument fails. This differs from selecting $C_v(B)$ *after* a common topology draw as an exact marginal implementation of the global channel.

H.2 Maximality within the clean-mass information class

Proof of Proposition 2. Fix an admissible $(\mathbf{A}', \mathbf{X}')$, write $Q = Q_{\mathbf{A}',\mathbf{X}'}$, and let $\Lambda = dQ/dP$. The generalized NP lemma supplies an optimal test

$$a^\star = \mathbf{1}\{\Lambda < t\} + \kappa \mathbf{1}\{\Lambda = t\}, \qquad 0 \le \kappa \le 1,$$

calibrated to $\mathbb{E}_P[a^\star] = p_A$, with $\mathbb{E}_Q[a^\star] = V_{\mathrm{NP}}(\mathbf{A}', \mathbf{X}')$. Validity of $b(p_A)$ for every such test implies $b(p_A) \le V_{\mathrm{NP}}(\mathbf{A}', \mathbf{X}')$. Taking the infimum over all admissible pairs and applying Theorem 1 gives $b(p_A) \le \underline{p}_{\mathrm{hyb}}(d, \varepsilon)$. Both sequential maps are uniformly sound on this test class by Proposition 5, so their maximum satisfies the same inequality. □

Atomic and continuous cases. At a positive feature shift, the log-Gaussian likelihood ratio has a continuous distribution independent of the finite topology factor, and an indicator test attains the optimum. At zero shift, the ratio is atomic and boundary randomization may be needed. This is why the proposition is stated for $[0, 1]$-valued tests. In the present channel, the nonempty Gaussian block with $\sigma > 0$ can also realize that randomization deterministically: at zero shift it has the same law under both measures and is independent of Λ. Choose an event H in that block of probability κ, and use $\mathbf{1}\{\Lambda < t\} + \mathbf{1}\{\Lambda = t\}\mathbf{1}_H$. It has the same two masses as $a^\star$. Thus maximality also holds over unrestricted binary tests on this product space. It does not assert attainability by a particular architecture or trained GNN.

Interpreting conservatism. The count reduction is exact for the NP relaxation, which uses only the clean mass and the channel laws. A fixed GNN can have a larger worst-case score, because its class event need not be NP-optimal. Finite-sample confidence corrections introduce further conservatism. An evaluated identity-specific adversary does not solve the full adversarial search; likewise, the first successful attack radius is an upper bound on the minimal adversarial radius. Such diagnostics can measure observed gaps, but cannot establish tightness by themselves.

H.3 Homogeneous asymmetric kernels

Corollary 3 (Asymmetry preserves the count reduction)**.** *Theorem 1 and the explicit NP formulas hold for any homogeneous pair $p^-, p^+ \in (0, 1)$, without requiring $p^- = p^+$.*

Proof. Proposition 4 and Corollary 2 derive the likelihood-ratio law for a general pair (p^-, p^+). Their arguments use slotwise independence and homogeneity, and the realizability of count pairs is independent of the probabilities. Consequently the proof of Theorem 1 is unchanged. □

For a perturbation with at least one edge flip and positive channel probabilities, non-degeneracy of the topology ratio is equivalent to $p^- + p^+ \neq 1$, or $\alpha \neq \beta$ in Proposition 4. At $p^- + p^+ = 1$ the topology channel is independent of its center; the count reduction still holds, but topology provides no likelihood-ratio separation. Kernel selection changes the clean noise and hence p_A; it must not be confused with changing the count arithmetic at fixed p_A.

H.4 Extension to typed slots

Consider R relation types and represent each target-incident typed slot by $(\min(v, j), \max(v, j), \rho)$, with $j \neq v$ and $\rho \in [R]$. Model these slots as independently allowed binary coordinates, with a common homogeneous channel (p^-, p^+). Then $m_1 = \sum_{\rho=1}^{R} \deg_\rho(v)$ and $m_0 = R(n-1) - m_1$. Retyping an edge deletes one typed coordinate and inserts another, using two units of the flip budget.

Corollary 4 (Count reduction for homogeneous typed slots)**.** *Under this typed-slot model and the fixed feature-support assumptions, Theorem 1 holds with the typed counts* (m_1, m_0)*. The outer NP search remains* $O(d^2)$*.*

Proof. The likelihood-ratio calculation in Proposition 4 uses only binary slot values, independent corruption and homogeneous probabilities. The typed slots therefore have the same ratio atoms and binomial laws for (d^-, d^+). Every feasible pair is realized by flipping the specified numbers of existing and empty typed coordinates, so the count-reduction and soundness proofs apply unchanged. □

If the channel parameters depend on the relation, the ratio law depends on the vector $(d_1^-, d_1^+, \ldots, d_R^-, d_R^+)$ rather than just its two totals. The outer search then requires relation-specific counts. Additional constraints, such as permitting only one relation per node pair, can also restrict count feasibility; the unconstrained grid is still a sound relaxation but need not be exact for that smaller threat set. The multi-relation result is theoretical; the empirical evaluations use single-relation graphs.

I Supplementary comparisons and empirical diagnostics

I.1 Protocols, feature support, and denominators

These experiments supplement the main results with baseline comparisons, accuracy costs, sensitivity analyses, and computational diagnostics. Table 5 summarizes four supplementary evaluations with $N = 10{,}000$ actual Monte Carlo draws per node, following $N_0 = 100$ class-selection draws. They use stratified test-node sampling, node-selection seed 0, training and Monte Carlo seed 42, and the default confidence level $1 - \alpha$ with $\alpha = 10^{-3}$. Exceptions are stated below. Their configurations are independent of Figure 2: in particular, supplementary Amazon uses $q = 0.7$, $\sigma = \sigma_{\text{train}} = 0.02$, and 1,000 nodes, whereas the main Amazon configuration uses $q = 0.5$, $\sigma = 0.1$, $\sigma_{\text{train}} = 0.5$, and 500 nodes. The supplementary models are the two-layer GCN on PubMed/MGTAB and order-2, row-normalized BWGNN on Amazon/YelpChi. MGTAB features use the train-split standardization described in Appendix C.

Table 5: Supplementary $N = 10{,}000$ runs. n counts all sampled test nodes; n_{corr} counts correct smoothed predictions. All four runs use target-only Gaussian feature noise, seed 42, $N_0 = 100$, and $\alpha = 10^{-3}$.

Dataset	n	n_{corr}	q	σ	σ_{train}	a_{smooth}
PubMed	1000	713	0.7	0.01	0.04	71.30%
Amazon Fraud	1000	881	0.7	0.02	0.02	88.10%
YelpChi	500	355	0.6	0.005	0.005	71.00%
MGTAB	400	319	0.4	0.4	0.4	79.75%

For a sampled population of n test nodes, let n_{corr} be the number whose smoothed prediction is correct and n_{cert} the number that are both correct and certified at the stated budget. We distinguish

$$\text{CA}_{\text{all}} = \frac{n_{\text{cert}}}{n}, \qquad \text{CA}_{\text{corr}} = \frac{n_{\text{cert}}}{n_{\text{corr}}}, \qquad a_{\text{smooth}} = \frac{n_{\text{corr}}}{n}, \qquad \text{CA}_{\text{all}} = a_{\text{smooth}}\,\text{CA}_{\text{corr}}.$$

The conditional rate is defined when $n_{\text{corr}} > 0$. The factor $\text{CA}_{\text{corr}}/\text{CA}_{\text{all}}$, when nonzero, is about 1.40, 1.14, 1.41, and 1.25 on PubMed, Amazon, YelpChi, and MGTAB, respectively. Every table states its denominator. Accuracy values are reported as percentages; the formulas above use the underlying fractions.

The feature support for the four evaluations in Table 5 is $R = \{v\}$: Gaussian noise and the certified feature perturbation are confined to the target node. The Frobenius norm then equals the target-vector ℓ_2 norm. Certificates on a larger fixed support require the corresponding smoothing distribution (Appendix H.1). A matched Amazon pilot with 25 nodes, $N_0 = 50$, and $N = 200$ produced identical predicted labels under target-only and global feature noise, with 24 correct predictions in each case. Their vote counts differ, so equal predicted labels do not establish equivalent certificates. The global-feature comparison in Table 9 uses probability estimates obtained under global Gaussian noise.

Joint certification with global feature support. Figure 6 compares target-only ($\mathcal{C} = \{v\}$) and global ($\mathcal{C} = V$) Gaussian feature smoothing on 500 common, class-balanced Amazon Fraud test targets. Both evaluations use the BWGNN checkpoint and graph of Figure 2, the target-incident edge support S_v, $q = 0.5$ ($\rho = 0.75$), $\sigma = 0.1$, $N_0 = 50$, $N = 10{,}000$ independent Monte Carlo draws per target, and $\alpha = 10^{-3}$. The binary certificate requires the lower bound on the selected class probability to exceed $1/2$. Under $\mathcal{C} = V$, the joint guarantee covers up to d edge flips on S_v and feature perturbations on any graph node, including the target, with total Frobenius norm at most ε.

Hybrid NP certifies at least as many targets as Sequential NP at every evaluated joint budget under both feature supports. The maximum absolute difference in certified accuracy between the two supports is 6.8 percentage points for each method over the evaluated radius grid. At $d = 0, \varepsilon = 0$, certified accuracy is 85.2% for target-only smoothing and 86.0% for global smoothing. At $d = 2$ and $\varepsilon \approx 0.10$, the global rates are 58.2% for Hybrid NP and 46.6% for Sequential NP; the target-only rates are 58.0% and 49.6%, respectively. The advantage of joint NP certification therefore persists with global feature support, while the change of smoothing distribution also changes the certified accuracy values.

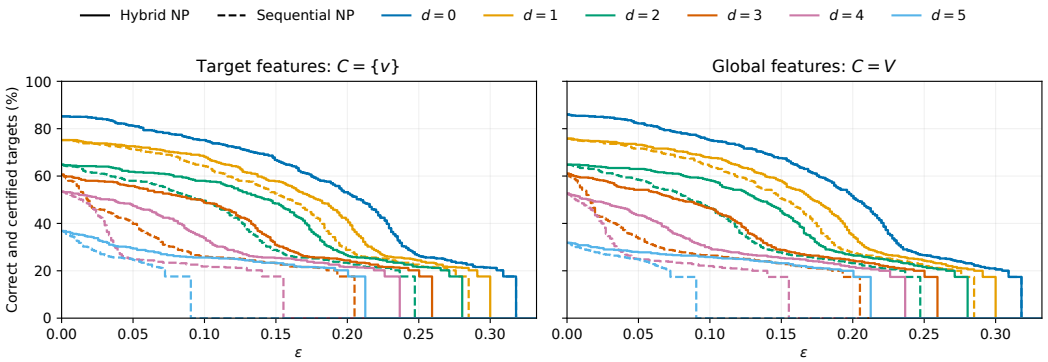


Figure 6: Joint certification under target-only and global Gaussian feature smoothing on Amazon Fraud. The two panels use the same checkpoint, graph, 500 test targets, Monte Carlo budgets, and $d \in \{0, \ldots, 5\}$. Colors indicate d; solid and dashed lines denote Hybrid NP and Sequential NP. Certified accuracy is the number of correct and certified targets divided by 500.

I.2 Native baseline comparisons and threat-set geometry

Structure-only axis. Table 6 reports AGNNCert (Li and Wang, 2025), topology-only randomized smoothing (TopoRS), and Hybrid NP at $\epsilon = 0$. AGNNCert certifies its deterministic partition-voting ensemble. The edge-centric implementation partitions unordered node pairs; the node-centric implementation hashes outgoing directed edges by source node. Their native perturbation conventions and voted predictors differ from the randomized classifiers with target-incident edge slots used by TopoRS and Hybrid NP. The latter also retains Gaussian feature noise when evaluated at $\epsilon = 0$, so it does not deploy the same classifier as TopoRS. The table therefore describes native guarantees, with different supports, channels, and, on PubMed, sampled populations. It is not a controlled ranking on an identical attack set and predictor. GNNCert (Xia et al., 2024) addresses graph classification; AGNNCert provides the node-classification comparison for this divide-and-conquer family.

Table 6: Native structure-axis guarantees at $d = 1$, $\epsilon = 0$; entries are $\mathrm{CA}_{\mathrm{all}}$. AGNNCert uses $T = 10$ and 300 nodes; TopoRS uses $N = 1{,}000$, $q = 0.7$, $\sigma = 0$, and 300 nodes. Hybrid NP uses $N = 1{,}000$, $q = 0.7$, $\sigma = 0.01$, and 300 nodes on Cora/Citeseer, but 1,000 on PubMed. The ensembles and edge supports differ, as detailed in the text.

Dataset	AGNNCert-E	AGNNCert-N	TopoRS	Hybrid NP
PubMed	74.0%	74.0%	52.3%	33.1%
Cora	70.0%	71.3%	36.0%	21.3%
Citeseer	55.3%	58.3%	35.0%	4.3%

Binary attributes. Table 7 evaluates sparsity-aware randomized smoothing (Bojchevski et al., 2020) on Cora. Its binary channel uses $(p_-^A, p_+^A) = (0.4, 0)$ and $(p_-^X, p_+^X) = (0.6, 0.01)$, with $N = 10{,}000$, $\alpha = 0.01$, and 1,000 test nodes. The two Hybrid NP runs use 200 nodes, $N = 1{,}000$, $\alpha = 10^{-3}$, target-only feature noise with $\sigma = 0.5$ or 1, and topology noise $(p_-, p_+) = (0.4, 10^{-6})$. Both evaluations restrict the adversarial structural budget to deletions, although the Hybrid sampler has a small nonzero insertion probability.

For binary features, $\|X' - X\|_F^2$ equals the number of flipped entries. Thus $\epsilon = \sqrt{d_X}$ embeds d_X binary flips into a continuous ball *on the same feature support*. Here the discrete method permits global feature flips, while Hybrid NP perturbs only the target features; neither numerical column is a matched-support superiority test. The Gaussian ball additionally includes nonbinary changes on its own support.

Partition voting and vote robustness. Table 8 reports AGNNCert-N structure-only runs at two partition counts T. For an evaluated node v, let $c_a(v)$ and $c_b(v)$ be the two largest class-vote counts, with corresponding class indices $y_a(v)$ and $y_b(v)$, ordered by decreasing count and then increasing class index. With ties in the final prediction resolved toward the smaller class index, the vote budget

Table 7: Illustrative native-budget comparison on binary Cora features ($\mathrm{CA}_{\mathrm{all}}$). The discrete method allows global binary feature flips; Hybrid NP allows continuous changes on the target features. $\epsilon = \sqrt{d_X}$ is a magnitude calibration, not an equality of these attack sets. d_A is a global edge-deletion budget for the discrete method and a target-incident deletion budget for Hybrid NP.

d_A	d_X	ϵ	Sparsity-aware RS	Hybrid $\sigma = 0.5$	Hybrid $\sigma = 1$
0	0	0.00	80.8%	81.5%	72.0%
0	1	1.00	41.5%	9.5%	22.5%
0	2	1.41	5.3%	0.0%	9.0%
0	4	2.00	0.0%	0.0%	4.5%
0	9	3.00	0.0%	0.0%	0.0%
1	0	0.00	47.0%	53.5%	25.5%
1	1	1.00	21.4%	2.5%	8.5%
1	2	1.41	1.6%	0.0%	5.5%
1	4	2.00	0.0%	0.0%	0.0%
1	9	3.00	0.0%	0.0%	0.0%

is

$$M_v = \left\lfloor \frac{c_a(v) - c_b(v) - \mathbf{1}[y_a(v) > y_b(v)]}{2} \right\rfloor .$$

Up to M_v individual subgraph votes can be changed without changing the aggregated prediction. A changed vote can reduce the winning margin by two, which explains the factor $1/2$; the indicator accounts for the tie rule. Thus M_v is a target-dependent tolerance to vote changes, whereas T is the number of votes cast. Translating this vote budget into an attack budget requires bounding the number of affected subgraphs. Increasing T from 200 to 300 raises the mean vote budget on both datasets; clean voting accuracy increases on PubMed and remains near 89% on Amazon. These structure-only evaluations do not establish robustness to arbitrary feature replacement on M_v nodes.

Table 8: AGNNCert-N structure-only runs. T is the number of graph partitions; the shared GNN casts one vote per nonempty subgraph. n is the number of evaluated nodes. Clean voting accuracy is the percentage correctly classified by the aggregated vote. The last column is $\bar{M} = n^{-1} \sum_v M_v$, averaged over all n evaluated nodes, including incorrect predictions; it is not a uniform per-node guarantee.

Dataset	T	n	Clean voting accuracy (%)	Mean M_v (votes)
PubMed	200	300	67.0%	98.45
PubMed	300	300	71.0%	148.43
Amazon Fraud	200	200	89.5%	99.48
Amazon Fraud	300	200	89.0%	149.44

Fixed-topology multi-node feature budgets. We compare feature-only guarantees on Amazon Fraud, Cora, Citeseer, PubMed, YelpChi, and MGTAB, with fixed topology and no structural smoothing. Within each dataset, both methods use the same 200 stratified test targets, selected before correctness filtering. Sampling is class-balanced on Cora, Citeseer, PubMed, YelpChi, and MGTAB. Both methods use order-2, row-normalized BWGNNs on Amazon/YelpChi and two-layer, row-normalized GCNs on the other datasets. Their checkpoints, training objectives, and deployed predictors are distinct. AGNNCert uses $T = 300$ source-hash partitions; Hybrid uses independent Gaussian noise on all graph feature coordinates, with the dataset-specific σ in Table 9, $N_0 = 1{,}000$, $N = 10{,}000$ actual draws, and $\alpha = 10^{-3}$ per target. Training and Monte Carlo seed are 42 and node-selection seed is 0. These global-noise statistics are independent of the target-only runs above and of Figure 2. Cora uses raw binary bag-of-words features; Citeseer/PubMed use row-normalized attributes and MGTAB uses train-split standardization, so absolute radii should be interpreted within each dataset.

Here k_v counts the maximum number of modified feature rows for target v, whereas ϵ is the total norm across those rows. For a target v, the common comparison set is

$$\mathcal{U}^{\neg v}_{k_v,\epsilon} = \left\{ \Delta X : \ \Delta X_{v,:} = 0, \quad \left| \{ u : \Delta X_{u,:} \neq 0 \} \right| \leq k_v, \quad \|\Delta X\|_F \leq \epsilon \right\}.$$

The Hybrid global Gaussian certificate covers this set for every k_v when the estimated radius exceeds ϵ; it also covers target-inclusive perturbations at the same total norm. For this fixed-graph AGNNCert

implementation, each non-target feature row influences at most one subgraph vote. Thus $k_v \leq M_v$ is a sufficient count-only certificate, allowing arbitrary feature magnitudes on these rows. M_v counts tolerable changed votes, while k_v counts attacked nodes. We report the four rules

$$k_v \in \{\lfloor M_v/2 \rfloor,\ M_v,\ M_v + 1,\ 2M_v\},$$

using each target's own M_v for both methods. The observed M_v ranges are 137–150 on Amazon, 139–150 on Cora, 142–150 on Citeseer, 135–150 on PubMed, 116–150 on YelpChi, and 16–150 on MGTAB. On Amazon, the four k_v ranges are therefore 68–75, 137–150, 138–151, and 274–300. Its 20 jointly correct attack targets all have $M_v = 150$, yielding the tested budgets 75, 150, 151, and 300 on that subset. On Cora, Citeseer, PubMed, YelpChi, and MGTAB, each attack subset contains 20 jointly correct targets selected by seeded class-balanced sampling; certificate coverage uses all 200 test targets.

Table 9 identifies a favorable regime for the global norm certificate on all six datasets: at $k_v = M_v + 1$ and $k_v = 2M_v$, Hybrid retains positive coverage for each displayed norm ($\epsilon/\sigma = 0.5, 1, 2.5$), while the AGNNCert count-only condition certifies no targets. At $k_v = \lfloor M_v/2 \rfloor$ and $k_v = M_v$, AGNNCert certifies all of its correct predictions. Hybrid nevertheless has higher coverage on Cora (68.0% versus 60.5%) and Citeseer (51.0% versus 43.0%) at $\epsilon = 0.5\sigma$, and on PubMed at both $\epsilon = 0.5\sigma$ (71.0% versus 66.5%) and $\epsilon = \sigma$ (68.5% versus 66.5%). AGNNCert has higher coverage at these two node-budget rules on Amazon, YelpChi, and MGTAB for every displayed norm. The comparison concerns the reported certificates of distinct predictors on a common, target-dependent threat set; it does not isolate certificate strength for an identical classifier. At larger norms, Hybrid coverage can also vanish; the table does not imply robustness at arbitrary ϵ.

The count-only AGNNCert column is a sufficient guarantee rather than an exhaustive support-aware certificate. In particular, all tested non-target supports on Amazon's 20 attack targets affect at most 148 votes. AGNNCert therefore certifies these particular supports even when they contain 300 selected nodes. This does not turn them into a count-only guarantee for arbitrary 300-node supports. Conversely, including the target can affect all T votes, so this AGNNCert vote-budget argument gives no positive uniform target-inclusive row budget. Perturbed rows are selected from the target's influential two-hop pool; an empirical attack is evaluated only when that pool contains enough rows for its node budget. The table reports certificates on all 200 targets per dataset, separately from empirical attacks on conditional 20-target subsets; Amazon's attack subset contains only class-0 targets.

Fairness certificates. ELEGANT (Dong et al., 2026) certifies a group-fairness event through continuous-feature and discrete-structure smoothing. Its certified rate is not a per-node prediction-invariance rate. We include it in the methodological comparison (Table 1), rather than assign its fairness guarantee a numerical prediction-robustness score. Replacing its fairness event by a target-class event would define an adapted method.

I.3 Accuracy costs and feature-radius calibration

Table 10 separates training with noise from smoothing at deployment and from certification. In this checkpoint comparison, noise training changes accuracy by $+0.2$, -2.1, $+1.6$, and $+0.25$ percentage points on PubMed, Amazon, YelpChi, and MGTAB. Relative to the noise-trained base model, deployment smoothing changes it by -7.6, $+0.1$, -2.0, and -5.5 percentage points, respectively. Certification then retains only predictions whose probability bounds establish invariance at the stated joint budget. These are separate costs; a conditional certified fraction alone would hide the loss in prediction accuracy.

Table 11 expresses a coverage threshold relative to the feature scale. At $d = 0$, define r_{25} as the largest evaluated radius at which at least 25% of the correctly classified smoothed nodes remain certified. Normalization by the median test-node norm helps interpret the target-feature budget. Normalization by $\|X\|_F$ provides a second scale reference; it does not enlarge the certified support. The JFR attack grids mostly extend beyond r_{25}, so failure at these larger radii does not contradict the joint certificate. The MGTAB value reaches the supplementary grid cap of 1.0 and is a lower limit, not an estimate of the exact cutoff.

Table 9: Certified coverage for node-dependent feature budgets on 200 common test targets per dataset, with fixed topology and no edge perturbations. Entries give correct-and-certified percentages. The adversary may modify at most k_v arbitrary non-target feature rows with total $\|\Delta X\|_F \leq \epsilon$. Each rule uses the AGNNCert vote tolerance M_v of that target. Hybrid uses global Gaussian noise of scale σ; its column applies to all four k_v rules. All Gaussian estimates use $N_0 = 1{,}000$, $N = 10{,}000$ actual draws, and $\alpha = 10^{-3}$ per target. AGNNCert uses the sufficient count-only condition $k_v \leq M_v$, independent of ϵ. Zero under this condition does not imply observed prediction errors or exclude stronger guarantees for particular supports.

Dataset	σ	ϵ	Hybrid	AGNNCert ($T = 300$)			
			all four k_v rules	$k_v = \lfloor M_v/2 \rfloor$	$k_v = M_v$	$k_v = M_v + 1$	$k_v = 2M_v$
Amazon Fraud	0.02	0.01	89.0%	89.5%	89.5%	0.0%	0.0%
		0.02	72.5%	89.5%	89.5%	0.0%	0.0%
		0.05	67.5%	89.5%	89.5%	0.0%	0.0%
Cora	0.5	0.25	68.0%	60.5%	60.5%	0.0%	0.0%
		0.5	49.0%	60.5%	60.5%	0.0%	0.0%
		1.25	9.0%	60.5%	60.5%	0.0%	0.0%
Citeseer	0.01	0.005	51.0%	43.0%	43.0%	0.0%	0.0%
		0.01	42.5%	43.0%	43.0%	0.0%	0.0%
		0.025	17.5%	43.0%	43.0%	0.0%	0.0%
PubMed	0.01	0.005	71.0%	66.5%	66.5%	0.0%	0.0%
		0.01	68.5%	66.5%	66.5%	0.0%	0.0%
		0.025	49.0%	66.5%	66.5%	0.0%	0.0%
YelpChi	0.005	0.0025	72.0%	76.0%	76.0%	0.0%	0.0%
		0.005	64.0%	76.0%	76.0%	0.0%	0.0%
		0.0125	40.0%	76.0%	76.0%	0.0%	0.0%
MGTAB	0.4	0.2	82.0%	87.0%	87.0%	0.0%	0.0%
		0.4	79.0%	87.0%	87.0%	0.0%	0.0%
		1	71.5%	87.0%	87.0%	0.0%	0.0%

Table 10: Four stages on the same sampled test population: standard training without noise, the noise-trained base model without deployment smoothing, the deployed smoothed classifier, and Hybrid NP certified accuracy $\mathrm{CA}_{\mathrm{all}}$ at $d = 1$. The last two columns give the evaluated ϵ and certified rate. Amazon and MGTAB use the closest evaluated grid point to σ.

Dataset	Standard base	Noise-trained base	Smoothed	ϵ	$\mathrm{CA}_{\mathrm{all}}$
PubMed	78.70%	78.90%	71.30%	0.01	25.70%
Amazon Fraud	90.10%	88.00%	88.10%	0.025	67.40%
YelpChi	71.40%	73.00%	71.00%	0.005	41.80%
MGTAB	85.00%	85.25%	79.75%	0.416667	67.50%

Table 11: Hybrid NP feature-radius calibration at $d = 0$. r_{25} is the largest evaluated radius with $\mathrm{CA}_{\mathrm{corr}} \geq 25\%$, not a percentile of per-node radii. Norm ratios are percentages. MGTAB reaches the evaluated grid cap, so its radius and ratios are lower limits. The last column counts JFR attack-grid radii no larger than r_{25}.

Dataset	r_{25}	$100 r_{25}/\mathrm{med}\|x_v\|_2$	$100 r_{25}/\|X\|_F$	JFR grid
PubMed	0.021875	11.40	0.0777	3/5
Amazon Fraud	0.0625	11.62	0.0883	1/5
YelpChi	0.013333	6.74	0.0313	4/8
MGTAB	≥ 1.0	4.72	0.0354	0/5

I.4 Sequential order and seed sensitivity

Table 12 compares both sound sequential orders using the same probability estimates as Hybrid NP. The displayed feature radii are evaluated grid points near 1.25σ (about 1.31σ on PubMed and 1.33σ on YelpChi). Topology followed by features is stronger at these points, but Hybrid NP still gives a positive gap over that order on all four datasets. The absolute gap varies substantially: on supplementary Amazon it is only about 0.45–0.57 percentage points, so order sensitivity should not be conflated with a uniformly large advantage of coupled certification.

Table 12: Both sequential orders on the $N = 10{,}000$ supplementary runs. Rates are $\mathrm{CA}_{\mathrm{corr}}$; T and F denote topology and features. At every displayed point, T→F is the stronger sequential order. The final column is Hybrid NP minus the stronger order, in percentage points. To obtain $\mathrm{CA}_{\mathrm{all}}$ in percent, multiply the displayed rate by the corresponding a_{smooth} in Table 5 and divide by 100.

Dataset	d	ϵ	T→F	F→T	Hybrid	Δ (pp)
PubMed	1	0.013125	27.9%	13.6%	29.9%	1.96
PubMed	2	0.013125	11.9%	0.0%	15.3%	3.37
Amazon Fraud	1	0.025	75.9%	73.9%	76.5%	0.57
Amazon Fraud	2	0.025	73.4%	0.0%	73.9%	0.45
YelpChi	1	0.00666667	47.0%	34.6%	52.4%	5.35
YelpChi	2	0.00666667	33.5%	0.0%	39.7%	6.20
MGTAB	1	0.5	77.4%	73.4%	80.9%	3.45
MGTAB	2	0.5	73.0%	67.4%	76.2%	3.13

Five independently trained seeds on PubMed and YelpChi give the results in Table 13. The mean smoothed accuracy is $(64.0 \pm 4.6)\%$ on PubMed and $(69.8 \pm 0.9)\%$ on YelpChi. The paired per-seed gaps remain positive at the selected nonzero operating points; on YelpChi at $d = 3$, the sequential rate is zero for all five seeds. The reported deviations measure variability across seeds. They do not give node-bootstrap confidence intervals or establish Monte Carlo convergence. In particular, confidence in each node certificate and uncertainty in an aggregate test-node rate are different quantities. Larger N can tighten probability intervals, but these separate $N = 1{,}000$ and $N = 10{,}000$ runs are not a proof that every realized certified region grows monotonically.

Table 13: Five-seed sensitivity (seeds 0–4), $N = 1{,}000$, $\alpha = 10^{-3}$, and 500 sampled nodes per seed. Accuracy entries are mean $\pm$ sample standard deviation of $\mathrm{CA}_{\mathrm{corr}}$ in percent; paired per-seed gaps and their deviations are in percentage points. They are not bootstrap confidence intervals. Smoothing uses $(q, \sigma) = (0.7, 0.01)$ on PubMed and $(0.6, 0.005)$ on YelpChi.

Dataset	d	ϵ	Hybrid	Sequential max	Gap (pp)
PubMed	1	0.013125	$23.6 \pm 3.7\%$	$21.4 \pm 3.9\%$	2.2 ± 0.2
YelpChi	1	0.00833333	$37.5 \pm 5.0\%$	$30.1 \pm 4.9\%$	7.3 ± 1.4
YelpChi	3	0.000833333	$24.8 \pm 5.2\%$	$0.0 \pm 0.0\%$	24.8 ± 5.2

I.5 Asymmetric kernels and model architecture

Table 14 instantiates the asymmetric edge channel covered by Corollary 3. On this Amazon sweep, $(p_-, p_+) = (0.9, 0.001)$ reduces expected insertions from about 1,781 to 12 per draw relative to $(0.15, 0.15)$, and retains nonzero certification at $d = 3$. It also deletes 90% of the eligible existing edges in each draw. The intermediate kernels include zero-coverage settings, despite identical smoothed accuracy, illustrating that accuracy alone does not determine the certificate. These dataset-specific results do not establish a universal kernel choice or robustness improvement without a smoothing cost.

The 40-node PubMed pilot in Table 15 applies the same certificate to two- and three-layer GCNs and a two-layer GraphSAGE. All three yield nonzero joint certified rates at $(d, \epsilon) = (1, \sigma)$. The small pilot has limited precision: one node changes an all-node rate by 2.5 percentage points. Evaluation uses the fast local path for the two-layer GCN and exact fallback evaluation for the other models; their much larger runtimes reflect this implementation difference. The pilot demonstrates applicability, rather than comparable throughput or a consistent accuracy ranking across architectures.

Table 14: Amazon Fraud asymmetric-kernel sweep: $N = 1{,}000$, 500 nodes, target feature noise $\sigma = 0.02$, and $\alpha = 10^{-3}$. All rows have smoothed accuracy 88.8%. E_- and E_+ are expected deletions and insertions per draw; certification columns are $\mathrm{CA}_{\mathrm{all}}$ at the specified (d, ϵ).

p_-	p_+	E_-	E_+	$(1, 0)$	$(1, 0.02)$	$(3, 0)$	$(3, 0.02)$
0.15	0.15	10.1	1781.4	70.0%	68.0%	0.0%	0.0%
0.15	0.01	10.1	118.8	0.0%	0.0%	0.0%	0.0%
0.15	0.03	10.1	356.3	67.0%	0.0%	0.0%	0.0%
0.3	0.01	20.2	118.8	0.0%	0.0%	0.0%	0.0%
0.5	0.001	33.7	11.9	0.0%	0.0%	0.0%	0.0%
0.9	0.001	60.6	11.9	88.0%	70.8%	82.6%	69.4%

Table 15: PubMed architecture pilot: 40 sampled nodes, $q = 0.7$, $\sigma = 0.01$, $N = 1{,}000$, and $\alpha = 10^{-3}$. Clean accuracy is evaluated on the full test set; smoothed and certified rates use the 40-node pilot. Certification columns are $\mathrm{CA}_{\mathrm{all}}$. H denotes Hybrid NP; S denotes Sequential NP. Times are seconds per sampled node, including certification.

Architecture	Clean	Smoothed	H $(1, 0)$	H $(1, \sigma)$	S $(1, \sigma)$	s/node
GCN (2-layer)	78.9%	60.0%	27.5%	15.0%	12.5%	1.26
GCN (3-layer)	78.5%	40.0%	27.5%	20.0%	20.0%	74.98
GraphSAGE (2-layer)	78.1%	65.0%	32.5%	12.5%	7.5%	85.90

I.6 Runtime and candidate-slot diagnostics

Table 16 separates prediction sampling from certificate arithmetic. On the evaluated $N = 10{,}000$ grids, Hybrid NP arithmetic takes about 0.013–0.362 seconds per node, compared with 4.76–7.14 seconds for the shared Monte Carlo pass. Mean total-time overhead relative to TopoRS is about 1.00–1.06. These measurements apply to the stated models, GPU, and budget grid; the $O(d^2)$ count search removes an edge-identity search from certificate arithmetic, not graph-dependent prediction costs.

Table 16: Mean seconds per node on an NVIDIA RTX A6000, 15 nodes per dataset, $N_0 = 100$, $N = 10{,}000$, and 25 feature-radius grid points for each topology budget up to $d_{\max}$. MC is the shared prediction-sampling pass. Arithmetic excludes MC. The ratio is the mean per-node Hybrid/TopoRS total-time ratio, not a ratio of the displayed means.

Dataset	$d_{\max}$	MC	Topo arith.	Hybrid arith.	Hybrid total	Ratio
PubMed	1	7.140	0.002	0.014	7.154	1.002
PubMed	3	6.968	0.017	0.096	7.064	1.011
PubMed	5	6.934	0.060	0.265	7.199	1.029
Amazon Fraud	1	4.765	0.002	0.013	4.778	1.002
Amazon Fraud	3	4.762	0.018	0.106	4.868	1.018
Amazon Fraud	5	4.762	0.084	0.362	5.124	1.057

Table 17 changes candidate-slot counts only in the certificate calculation, keeping the full-sampler class-probability estimates fixed. The homogeneous-channel count reduction yields identical displayed rates, while the hypothetical number of sampled insertions changes markedly. This separates the combinatorial certificate from the smoothing cost. A genuine restricted-sampler experiment would change the distribution of noisy graphs and would require new probability estimates; this diagnostic does not establish unchanged accuracy or certification for that new sampler.

I.7 Empirical gaps and limits of the diagnostics

Evaluated edge identities. On 50 correctly classified PubMed nodes, a structure-budget-1 diagnostic enumerates incident-edge deletions and samples 60 insertion candidates per node. It uses 1,000 clean draws and 500 draws per edited graph, with Bonferroni correction at total level $\alpha = 10^{-3}$ over the evaluated edits for each node. The gap between the minimum estimated winning-class probability over these edits and the count-based NP lower bound has median 0.394, interquartile range $[0.145, 0.553]$, and maximum 0.702. Using the simultaneous lower confidence bounds for the

Table 17: PubMed candidate-slot arithmetic with fixed probabilities estimated from the full sampler ($n = 1{,}000$, $N = 1{,}000$, $q = 0.7$, $\sigma = 0.01$). The insertion expectations are hypothetical for each restricted slot set. The certified column is $\mathrm{CA}_{\mathrm{all}}(d = 1, \epsilon = \sigma)$. Only the candidate-slot counts in the certificate calculation vary; the sampling distribution is fixed.

Slot set	Mean slots	Expected insertions	$\mathrm{CA}_{\mathrm{all}}$
full ($n - 1$)	19716.0	2956.7	21.4%
3-hop envelope	406.7	60.3	21.4%
2-hop + pool(50)	109.7	15.8	21.4%
2-hop envelope	59.7	8.3	21.4%

evaluated probabilities gives a median gap of 0.317. This diagnoses slack relative to the *tested* edits. Because untested insertions could lower the true minimum, it does not give a lower bound on the true classifier-specific conservatism. Even with exact probabilities, minimizing over a subset gives an upper bound on the full minimum. It is also distinct from the classifier-agnostic optimality statement in Appendix H.

Adaptive attacks. A separate PubMed diagnostic tests 40 nodes at $d = 1$, using an expectation-over-transformation loss with 8 samples per step, 40 optimization steps, 2 restarts, and 1,000 fresh votes to confirm each successful attack. The feature-radius grid is $\{0.005, 0.01, 0.015, 0.02, 0.03, 0.05, 0.075, 0.1, 0.15, 0.2\}$. All 40 attacks succeed within this grid; the median first-successful-radius to certified-radius ratio is 4.425 and the median difference is 0.0581. No observed successful radius is below its certified radius. The certificate uses target-only feature noise, whereas this attack permits local feature changes on a union of neighborhoods. Consequently, these ratios compare different supports and do not estimate the gap to the minimum adversarial radius inside the certified target-only threat set. Attack outcomes also concern finite-vote estimates of the smoothed predictor; they do not replace the probabilistic soundness argument.

The sampled nodes in these diagnostics are restricted to correct predictions, so their descriptive statistics should not be interpreted as all-test-node certified accuracy.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the formal threat model, the proposed Hybrid NP certificate, the $\mathcal{O}(d^2)$ count reduction for Hybrid NP under homogeneous edge smoothing, and the empirical evaluation showing non-trivial joint certified regions and failure cases for topology-only certificates under joint attacks; all are matched by Sections 3 and 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations paragraph in Section 5 discusses the Frobenius feature model, the homogeneous edge-smoothing assumption, and the projected graph views used for some fraud datasets.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions are listed at the start of Section 3; Lemma 1, Corollary 1 and Proposition 1 are stated formally and proved in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The certification algorithm is given in Algorithm 1; Table 2 and Appendix C list base models, smoothing parameters, sample counts, and budgets used for every experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used in the paper (PubMed, Amazon Fraud, YelpChi, and MGTAB) are publicly available. The supplemental material includes the code, configuration files, and instructions needed to run the training and certification pipelines and reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Smoothing parameters $(p, \sigma, \sigma_{\text{train}})$, base models, certified-node counts, and budget grids are listed in Table 2; training details are given in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The certificates use one-sided Clopper–Pearson probability bounds. Table 13 reports means and standard deviations across five training seeds on PubMed and YelpChi. Several other experiments use one training seed, and aggregate test-node confidence intervals are not provided.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C reports the compute environment used for the hybrid certification experiments, including the GPU worker type, GPU memory, and representative execution time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only publicly released benchmark graphs and standard GNN training, and contributes a defensive certification procedure.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The contribution is a robustness certificate, primarily a defensive tool. We also describe a joint attack used to expose threat-model mismatch on topology-only certified nodes; we apply it only to public benchmark graphs and discuss it as a diagnostic, not a deployable exploit.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper releases neither high-risk models nor scraped datasets; it studies certified robustness of small GNNs on public benchmarks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original sources for all datasets, base models, and baseline methods used. All assets are used consistently with their permitted research-use conditions.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new datasets or models.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subjects research, so no IRB review is applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core methodology of this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.