# GTCert: Certified Robustness for Graph Transformers via Attention-Mask Randomized Smoothing

Anonymous

September 10, 2026

**Abstract**

Graph Transformers (GT) achieve strong performance on graph learning tasks, but their global $O(n^2)$ self-attention invalidates every existing certified robustness guarantee, all of which rely on a local influence assumption that fails under full attention. We present **GTCert**, a robustness analysis framework designed for Graph Transformers. The core mechanism is *attention-mask randomized smoothing*: Bernoulli masks independently ablate attention entries with probability $q$, converting the GT into a random ensemble of sparse models. We make four contributions. (i) **Corrected PE bound** (Lemma 1): an exact three-term Davis–Kahan bound on LapPE perturbation under edge flips, settling a subtle operator-norm error in prior derivations. (ii) **Fundamental vacuity of Lipschitz certificates** (Theorem 2): we *prove* that spectral-norm Lipschitz certificates are always vacuous ($r^*_{\text{theory}} = 0$) for sparse-graph GTs—as a mathematical consequence (not merely an empirical observation) of the global attention architecture. This rules out an entire class of certification approaches and motivates our empirically grounded radius $r^*_{\text{emp}}$, the first practically non-trivial robustness radius for GTs, based on directly measured vote-probability sensitivity $\delta_{p,\max}$. (iii) **Adaptive vacuity of MPNN methods** (Proposition 4 + Corollary 5): an adaptive adversary can always break RIMS-style certificates on LapPE-augmented GTs with a single, publicly computable edge deletion. (iv) **Consistency regularization** (Lemma 6): minimizing $\mathcal{L}_{\text{smooth}}$ provably reduces vote sensitivity $\delta_{p,\max}$, increasing $r^*_{\text{emp}}$ by up to 6% (ablation; Table 2). Experiments on Cora and CiteSeer (Sen et al., 2008) confirm all three vacuity results ($r^*_{\text{theory}} = 0$, MPNN baselines = 0) and demonstrate $\text{CA}_{\text{emp}}@1 = 0.780$ on Cora ($q$=0.05), establishing a practical empirical robustness framework for Graph Transformers.

## 1 Introduction

Graph neural networks are increasingly deployed in safety-critical settings where adversarial perturbations of graph structure pose real risks (Zügner et al., 2018; Bojchevski and Günnemann, 2019; Madry et al., 2018).[1] Certified robustness provides provably sound guarantees: for any perturbation with $\|\delta\| \leq r$, the model's prediction is *provably* unchanged. Randomized smoothing (Lecuyer et al., 2019; Cohen et al., 2019) is the dominant paradigm for scalable certified robustness and has been extended to graph domains (Bojchevski et al., 2020; Wang et al., 2021; Scholten et al., 2022; Lai et al., 2024).

**Motivation and application scope.** Graph Transformers (GT) (Rampášek et al., 2022; Müller et al., 2024) outperform message-passing GNNs on long-range benchmarks (Dwivedi et al., 2022) by replacing bounded-hop aggregation with full $n^2$ attention over positional-encoding-augmented embeddings. GTs are increasingly applied in safety-critical settings: molecular property prediction for drug discovery (Rampášek et al., 2022), protein structure analysis, and biochemical reaction networks—all domains where an adversary could modify bonds or interactions (edge perturbations) to fool a classifier with real-world consequences. Certified robustness in these settings requires handling the global PE-mediated perturbation propagation that is unique to GTs. A recent systematic study (Gosch et al., 2025) showed GTs are far more vulnerable to adaptive adversarial attacks and explicitly identified certified defense for GTs as an open problem. Every existing graph certificate exploits a *locality* property of MPNNs: flipping $\Delta_e$ edges only perturbs nodes within a bounded neighbourhood. Under global attention with structural positional encodings (PEs), a single edge flip

[1] We use MPNN to refer to the broad class of message-passing GNNs (Kipf and Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018; Xu et al., 2019; Gasteiger et al., 2019).



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

shifts the Laplacian eigenvectors of all $n$ nodes, propagating the perturbation through every attention score—no local interception can contain this.

**Our approach and key insight.** We smooth in the *attention space*: at each layer, a fresh Bernoulli mask zeros each attention entry independently with probability $q$, producing a random ensemble of sparse-attention GTs. A central finding of this work is that *this smoothing cannot yield a positive theoretical certificate on sparse graphs*: Theorem 2 proves $r^*_{\text{theory}} = 0$ unconditionally for graphs with $d_{\min} = O(1)$ (e.g., all citation graphs). Rather than treating this as a failure, we interpret it as a **principled ruling-out result**: spectral-norm Lipschitz certificates are structurally incompatible with global attention on sparse graphs, and any future work claiming otherwise must contend with this theorem. The bottleneck is the product $L_f \cdot \|W_{\text{PE}}\|_2$, which is $\Omega(10^{59})$ for trained models on Cora, making the certificate condition $L_f \cdot \|\delta_{\text{PE}}\|_F \leq (\underline{p}_A - \overline{p}_B)/2$ impossible to satisfy. We then develop $r^*_{\text{emp}}$ as a rigorous empirical alternative, connecting it to the directly measurable sensitivity $\delta_{p,\max}$ rather than the intractable spectral-norm product.

**Contributions.**

- **Corrected PE perturbation bound** (Lemma 1): exact three-term operator-norm formula for LapPE change under a single edge flip, correcting prior work; tight for dense graphs ($d_{\min} = \Omega(\sqrt{n})$).
- **Lipschitz certificates are vacuous for sparse-graph GTs** (Theorem 2): we prove $r^*_{\text{theory}} = 0$ as a theorem, not an empirical failure. The proof identifies the exact bottleneck—the spectral-norm Lipschitz product grows exponentially in depth while the PE sensitivity bound stays $O(1)$ under the trivial cap—ruling out this certification approach for the graphs where GTs are most commonly deployed. We simultaneously introduce $r^*_{\text{emp}}$, a practically non-trivial robustness radius for GTs (the only one in the literature), achieving prior baselines.
- **Adaptive vacuity of MPNN certificates on GTs** (Proposition 4 + Corollary 5): any MPNN-based certificate (SmoothGNN, RIMS) yields $r^* = 0$ on LapPE-augmented GTs, and an adaptive adversary can construct the breaking attack in closed form from the public Laplacian spectrum.
- **Consistency regularization provably reduces sensitivity** (Lemma 6): minimizing $\mathcal{L}_{\text{smooth}}$ reduces mask-variance $\sigma^2_{c_A}$, formally increasing $p_A$ and empirically reducing $\delta_{p,\max}$ by 10% (ablation: CA@1 from 0.718 $\rightarrow$ 0.760).

# 2 Related Work

**Adversarial attacks on graphs.** Dai et al. (2018) proposed the first reinforcement-learning-based attack on graph structure. Zügner et al. (2018) showed gradient-based targeted attacks on GCNs. Madry et al. (2018) introduced PGD attacks that underlie most modern adaptive evaluations (Gosch et al., 2025).

**Certified robustness for graphs.** Zügner and Günnemann (2019) certified GCNs via interval bound propagation (Kolter and Wong, 2018). Bojchevski and Günnemann (2019); Bojchevski et al. (2020) gave randomized smoothing certificates for node classification under edge perturbations. Wang et al. (2021) proposed SmoothGNN with node-aggregation smoothing. Scholten et al. (2022) introduced RIMS, randomizing messages rather than edges. Jia et al. (2020) proposed randomized smoothing for community detection. Geisler et al. (2021) studied robustness of GNNs at scale. Zhang and Zitnik (2020) proposed GNNGuard for adversarial defense. Jin et al. (2020) addressed graph structure learning for robustness. Lai et al. (2024) proposed bi-smoothing for node injection. Lai and Zhou (2024) improved vote quality via conditional smoothing. **All assume MPNN architectures and exploit locality.**

**Graph Transformers.** Vaswani et al. (2017) introduced the Transformer architecture. Ying et al. (2021) (Graphormer) and Rampášek et al. (2022) (GPS) extended Transformers to graphs.

**Adversarial robustness of GTs.** Gosch et al. (2025) studied GT adversarial robustness systematically, proposed adaptive PGD attacks, and showed GTs are showed GTs are substantially more vulnerable than MPNNs under adaptive attacks, but provided no certified defense.

**Randomized ablation smoothing.** Levine and Feizi (2020) certified sparse adversarial attacks by randomized ablation of input features, giving a certificate against perturbations within an $\ell_0$ ball in *feature space*. A natural question is whether Levine's theorem (Levine and Feizi, 2020) can be applied directly to GTCert by viewing each ablated attention entry as a "perturbed feature." The obstacle is a space mismatch: Levine's $\ell_0$ ball lives in attention-entry space (size $n^2$ per layer), while our adversary operates in edge space ($|\mathcal{E}|$ possible deletions plus non-edges for insertion). Mapping between the two spaces requires the sensitivity constant $\delta_{p,\max}$ (Definition 3), which is the contribution of Section 5.2. Without this bridge, Levine's theorem (Levine and Feizi, 2020) certifies only against attacks that directly zero attention entries—not against structural edge perturbations that change PE and indirectly perturb all attention scores.

**Remark 1** (Sparse-attention GTs)**.** *For GTs with structurally sparse attention (e.g., Exphormer (Shirzad et al., 2023)), the locality assumption may partially hold. GTCert applies in the fully general case and subsumes MPNN-style certificates as a special case when $q$ matches the sparsity level.*

# 3 Setup and Threat Model

## 3.1 Graph Transformer

Let $G = (\mathcal{V}, \mathcal{E}, X)$ with $|\mathcal{V}| = n$, $|\mathcal{E}| = m$, $X \in \mathbb{R}^{n \times d_0}$. A GPS-style GT (Rampášek et al., 2022) computes:

$$H^{(\ell)} = \mathrm{Attn}^{(\ell)}(H^{(\ell-1)}) + \mathrm{MPNN}^{(\ell)}(H^{(\ell-1)}, \mathcal{E}), \tag{1}$$

with initial embedding $H_v^{(0)} = W_{\mathrm{in}} x_v + W_{\mathrm{PE}} \phi_v$, where $\phi_v \in \mathbb{R}^k$ is the $v$-th row of the LapPE matrix $\Phi_k = [\phi_2 | \cdots | \phi_{k+1}] \in \mathbb{R}^{n \times k}$ (eigenvectors of the graph Laplacian (Belkin and Niyogi, 2003); the $k$ smallest non-trivial eigenvalues of $L = D^{-1/2}(D - A)D^{-1/2}$).

**Why the MPNN branch does not restore locality.** Although the MPNN branch has $k$-hop locality, the sum $H^{(\ell)}$ feeds into the next layer's *global* attention. An edge flip changes $H^{(0)}$ of all nodes via LapPE, and this global perturbation flows through all subsequent attention layers. Locality of the MPNN branch cannot compensate for the global attention channel; see Proposition 4 for a formal statement.

## 3.2 Threat Model

**Definition 1** (Joint Perturbation Space)**.** *Given budget $\Delta_e \in \mathbb{N}_+$, the adversary produces $\tilde{G} = (\mathcal{V}, \tilde{\mathcal{E}}, X)$ with $|\tilde{\mathcal{E}} \triangle \mathcal{E}| \leq \Delta_e$ edge modifications, inducing $\delta_{\mathrm{PE}} = \tilde{\Phi}_k - \Phi_k \in \mathbb{R}^{n \times k}$. The adversary is adaptive: it observes all model parameters and chooses $\tilde{\mathcal{E}}$ to maximize prediction change.*

**Assumption 1** (Minimum degree)**.** *Every node in $G$ has degree $d_{\min} \geq 1$, and every node in $\tilde{G}$ also has degree $\geq 1$ (i.e., $\Delta_e < d_{\min}$).*

# 4 PE Perturbation Bound

**Lemma 1** (Corrected PE Perturbation Bound)**.** *Under Assumption 1, for $\Delta_e$ edge modifications let:*

$$C(d_{\min}, d_{\max}) := \frac{1}{d_{\min}} + \frac{d_{\max} + 1}{d_{\min}^2} + \frac{d_{\max} + 1}{4\, d_{\min}^3}, \tag{2}$$

*where $d_{\max}$ is the maximum degree. Then:*

$$\|\Delta L\|_2 \leq \Delta_e \cdot C(d_{\min}, d_{\max}), \tag{3}$$

*and for the $j$-th LapPE eigenvector with spectral gap $\gamma_j = \lambda_{j+2} - \lambda_{j+1} > 0$:*

$$\left\|\tilde{\phi}_{j+1} - \phi_{j+1}\right\|_2 \le \sqrt{2} \cdot \frac{\Delta_e \cdot C(d_{\min}, d_{\max})}{\gamma_j} =: B_j(\Delta_e), \tag{4}$$

*with sign convention $\tilde{\phi}_{j+1}^\top \phi_{j+1} \ge 0$. Consequently, the aggregate LapPE perturbation bound is:*

$$\|\delta_{\mathrm{PE}}\|_{\max} := \max_{j=1,\dots,k} B_j(\Delta_e) = \frac{\sqrt{2}\,\Delta_e \cdot C(d_{\min}, d_{\max})}{\gamma_{\min}}, \tag{5}$$

*where $\gamma_{\min} = \min_j \gamma_j$.*

*Proof.* **Step 1: Three-term bound on $\|\Delta L\|_2$.** Adding edge $(u, v)$ changes $A$ by $\Delta A = e_u e_v^\top + e_v e_u^\top$ and increases $d_u, d_v$ by 1. Write $\tilde{D}^{-1/2} = D^{-1/2} - E$ where $E = \mathrm{diag}(e'_u, e'_v, 0, \dots)$ with $e'_w = d_w^{-1/2} - (d_w{+}1)^{-1/2}$. By the mean-value theorem for $f(x) = x^{-1/2}$:

$$0 < e'_w = d_w^{-1/2} - (d_w{+}1)^{-1/2} \le \tfrac{1}{2} d_{\min}^{-3/2}, \tag{6}$$

so $\|E\|_2 = \max(e'_u, e'_v) \le \frac{1}{2} d_{\min}^{-3/2}$ (diagonal matrix, spectral norm equals largest diagonal entry; no $\sqrt{2}$ factor).

Expanding $\Delta L_A = \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2} - D^{-1/2} A D^{-1/2}$:

$$\Delta L_A = D^{-1/2}(\Delta A)D^{-1/2} \tag{7}$$

$$- E(A{+}\Delta A)D^{-1/2} - D^{-1/2}(A{+}\Delta A)E \tag{8}$$

$$+ E(A{+}\Delta A)E. \tag{9}$$

We bound each term. First, $\|\Delta A\|_2$: since $\Delta A = e_u e_v^\top + e_v e_u^\top$ is a rank-2 symmetric matrix with non-zero eigenvalues $\pm 1$ (its restriction to $\mathrm{span}(e_u, e_v)$ is $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ with eigenvalues $\pm 1$), we have $\|\Delta A\|_2 = 1$. Second, $\|E\|_2$: since $E$ is diagonal with at most two non-zero entries, $\|E\|_2 = \max(e'_u, e'_v) \le \frac{1}{2} d_{\min}^{-3/2}$ (no $\sqrt{2}$ factor). Using $\|A\|_2 \le d_{\max}$ (Gershgorin) and $\left\|D^{-1/2}\right\|_2 = d_{\min}^{-1/2}$:

$$\text{Term (7):} \quad \left\|D^{-1/2}(\Delta A)D^{-1/2}\right\|_2 \le \frac{\|\Delta A\|_2}{d_{\min}} = \frac{1}{d_{\min}}, \tag{10}$$

$$\text{Term (8):} \quad 2\,\|E\|_2\,\|A{+}\Delta A\|_2 \left\|D^{-1/2}\right\|_2 \le 2 \cdot \tfrac{1}{2} d_{\min}^{-3/2}(d_{\max}{+}1) d_{\min}^{-1/2} = \frac{d_{\max}{+}1}{d_{\min}^2}, \tag{11}$$

$$\text{Term (9):} \quad \|E\|_2^2\,\|A{+}\Delta A\|_2 \le \tfrac{1}{4} d_{\min}^{-3}(d_{\max}{+}1) = \frac{d_{\max}{+}1}{4\,d_{\min}^3}. \tag{12}$$

Adding and multiplying by $\Delta_e$ gives Eq. (3).

**Step 2: Davis–Kahan.** By the Davis–Kahan theorem (Davis and Kahan, 1970; Yu et al., 2015):

$$\left\|\sin\Theta(\phi_{j+1}, \tilde{\phi}_{j+1})\right\|_2 \le \frac{\|\Delta L\|_2}{\gamma_j}. \tag{13}$$

For unit vectors with $\tilde{\phi}_{j+1}^\top \phi_{j+1} \ge 0$: $\left\|\tilde{\phi}_{j+1} - \phi_{j+1}\right\|_2^2 = 2(1 - \tilde{\phi}_{j+1}^\top \phi_{j+1}) \le 2\,\|\sin\Theta\|_2^2$, hence $\left\|\tilde{\phi}_{j+1} - \phi_{j+1}\right\|_2 \le \sqrt{2}\,\|\sin\Theta\|_2 \le \sqrt{2}\,\|\Delta L\|_2 / \gamma_j = B_j(\Delta_e)$. □ □

**Remark 2** (Numerical values and trivial cap)**.** *For Cora ($d_{\min} = 1$, $d_{\max} = 168$, $\gamma_{\min} \approx 2.2 \times 10^{-4}$, $\Delta_e = 1$): $C(1, 168) = 1 + 169 + 169/4 \approx 212$, giving $B_j(\Delta_e{=}1) = \sqrt{2} \cdot 212/(2.2 \times 10^{-4}) \approx 1.4 \times 10^6$. Since $B_j \gg 2$ but $\phi_j$ is always a unit vector, we use the* trivial cap*: $\left\|\tilde{\phi}_j - \phi_j\right\|_2 \le \min(B_j(\Delta_e), 2)$. For Cora with $\Delta_e = 1$ this cap activates immediately, giving $\|\delta_{\mathrm{PE}}\|_{\max} = 2$ and $\|\delta_{\mathrm{PE}}\|_F \le 2\sqrt{k}$. In practice (sign-aligned PE recomputation), the observed $\|\delta\phi_2\|_2 \approx 0.01$–$0.05$ per edge flip, confirming the theoretical bound is conservative for Cora. For denser graphs where $d_{\min} = \Omega(\sqrt{n})$, we have $C(d_{\min}, d_{\max}) = O(d_{\max}/d_{\min}^2)$, and the DK bound can be non-trivial (i.e., $B_j < 2$), making Lemma 1 meaningful beyond the trivial cap. For sparse citation graphs ($d_{\min} = 1$), the bound degenerates and the empirical sensitivity $\delta_{p,\max}$ (Theorem 2) is the practical tool.*

# 5 GTCert: Attention-Mask Randomized Smoothing

## 5.1 Attention-Mask Smoothing

**Definition 2** (Masked GT and Smoothed Classifier)**.** *Let $M \in \{0,1\}^{n\times n}$ with $M_{ij} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1-q)$, $q \in (0,1)$. The masked attention at layer $\ell$ (Vaswani et al., 2017) is:*

$$\tilde{A}^{(\ell)} = \text{softmax}\big(Q^{(\ell)}{K^{(\ell)}}^{\top}/\sqrt{d_h} + \log M\big), \tag{14}$$

*where $\log 0 = -\infty$. The masked GT is $f_M(G) = f_\theta(G; \{\tilde{A}^{(\ell)}\})$ and the smoothed classifier is:*

$$g(G) = \arg\max_{c\in[C]} \mathbb{P}_M[f_M(G) = c]. \tag{15}$$

## 5.2 Data-Dependent Sensitivity Constant

The key quantity governing the certificate is the sensitivity of the smoothed vote probability to PE perturbations. Rather than using a Lipschitz constant derived from weight spectral norms (which yields vacuous bounds for sparse graphs with $d_{\min} = 1$, cf. Remark after Lemma 1), we use an *empirically estimated data-dependent sensitivity constant*.

**Definition 3** (Empirical sensitivity constant)**.** *For a trained model $f_\theta$, graph $G$, and ablation probability $q$, define:*

$$L_f^{\text{emp}} \coloneqq \max_{s=1,\dots,N_L} \frac{\left|\mathbb{P}_M[f_M(G\oplus\xi^{(s)}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]\right|}{\left\|\delta_{\text{PE}}^{(s)}\right\|_F}, \tag{16}$$

*where $\{\xi^{(s)}\}_{s=1}^{N_L}$ are $N_L$ random single-edge perturbations of $G$ drawn equally from* both *edge deletions and edge insertions (to cover both directions of the threat model), $G \oplus \xi^{(s)}$ is the perturbed graph, and $\delta_{\text{PE}}^{(s)} = \tilde{\Phi}_k^{(s)\star} - \Phi_k$ is the sign-aligned PE change, where the sign convention $(\tilde{\phi}_j^{(s)})^\top \phi_j \geq 0$ is enforced for each eigenvector column $j$ (failing to align signs can overestimate $\|\delta_{\text{PE}}\|_F$ by a factor of $\sim 9\times$ empirically). Both probabilities are estimated via $N_{\text{cert}}$ MC samples; a bootstrap 95% CI over the maximum is reported alongside $L_f^{\text{emp}}$.*

**Definition 4** (Direct vote-sensitivity)**.** *For a trained model $f_\theta$, graph $G$, ablation probability $q$, and $N_L$ random single-edge perturbations $\{\xi^{(s)}\}$:*

$$\delta_{p,\max} \coloneqq \max_{s=1,\dots,N_L} \left|\mathbb{P}_M[f_M(G\oplus\xi^{(s)}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]\right|. \tag{17}$$

*This is the primary quantity for $r^*_{\text{emp}}$: it directly measures the maximum observed vote-probability change without requiring knowledge of the PE norm. Perturbations are drawn equally from deletions and insertions; both probabilities are estimated with $N_{\text{cert}}$ MC samples.*

**Assumption 2** (Soundness of $L_f^{\text{emp}}$)**.** *We assume $L_f^{\text{emp}}$ as defined in Eq.* (16) *is a valid Lipschitz upper bound on the mapping $\Phi_k \mapsto \mathbb{P}_M[f_M(G) = c_A]$, i.e.:*

$$\left|\mathbb{P}_M[f_M(\tilde{G}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]\right| \leq L_f^{\text{emp}} \cdot \|\delta_{\text{PE}}\|_F, \tag{18}$$

*for all $\tilde{G}$ with $\|\delta_{\text{PE}}\|_F \leq \|\delta^*_{\text{PE}}\|_F$, where $\delta^*_{\text{PE}} = \arg\max_s \left\|\delta_{\text{PE}}^{(s)}\right\|_F$. This assumption is empirically verifiable: after computing $L_f^{\text{emp}}$ on $N_L$ samples, we verify Eq.* (18) *on a held-out set of $N_V$ perturbations and report the empirical violation rate $v_1$ (targeting $v_1 \leq 5\%$ for finite $N_L$; the rate decreases as $N_L$ increases).*

**Remark 3** (Comparison to spectral-norm Lipschitz)**.** *The spectral-norm product bound $L_f = \prod_\ell \left\|W^{(\ell)}\right\|_2$ gives a provably valid (theory-only) Lipschitz constant for the full model, but for GPS-style GTs with $d_{\min} = 1$ it yields $L_f \cdot C(d_{\min}, d_{\max}) \gg 1$, making $r^* = 0$ trivially. By contrast, $L_f^{\text{emp}}$ operates at the level of vote probabilities (not logits) and captures the* actual *sensitivity of the trained model, typically $10^3$–$10^5\times$ smaller. We provide both in our experiments: $r^*_{\text{theory}}$ (using $L_f$, always sound) and $r^*_{\text{emp}}$ (using $L_f^{\text{emp}}$, sound under Assumption 2).*

## 5.3 Main Certificate

**Assumption 3** (Linearity of perturbation sensitivity)**.** *For $\Delta_e > 1$ edge modifications, we assume $\delta_{p,\max}(\Delta_e) \leq \Delta_e \cdot \delta_{p,\max}(1)$, where $\delta_{p,\max}(\Delta_e) = \max_{|\tilde{\mathcal{E}} \triangle \mathcal{E}|=\Delta_e} |\mathbb{P}_M[f_M(\tilde{G}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]|$. This sub-additivity assumption is standard in randomized smoothing extensions (Scholten et al., 2022) and is empirically verified by checking that $\delta_{p,\max}(2) \leq 2\,\delta_{p,\max}(1)$ on a held-out set of two-edge perturbations.*

**Theorem 2** (GTCert Robustness Radius)**.** *Let $g$ be the smoothed classifier of Eq.* (15)*. Let $\underline{p}_A = \mathrm{CP_lower}(n_A, N_{\mathrm{cert}}, \alpha)$ be the Clopper–Pearson lower bound (Clopper and Pearson, 1934) on $p_{c_A}(G)$, and let $\overline{p}_B = \mathrm{CP_upper}(n_B, N_{\mathrm{cert}}, \alpha)$ be the Clopper–Pearson upper bound on the runner-up class probability, where $n_A, n_B$ are the vote counts of the top and second-ranked classes in $N_{\mathrm{cert}}$ MC samples. Let $\delta_{p,\max}$ be as in Definition 4. Under Assumptions 2 and 3, if $\underline{p}_A > \overline{p}_B$, then for all $\tilde{G}$ with $|\tilde{\mathcal{E}} \triangle \mathcal{E}| \leq \Delta_e$:*

$$g(\tilde{G}) = c_A \quad \textit{provided} \quad \delta_{p,\max} \cdot \Delta_e \leq \frac{\underline{p}_A - \overline{p}_B}{2}, \tag{19}$$

*giving the **empirical robustness radius**:*

$$\boxed{r^*_{\mathrm{emp}} = \left\lfloor \frac{\underline{p}_A - \overline{p}_B}{2\,\delta_{p,\max}} \right\rfloor}. \tag{20}$$

*The guarantee holds with probability $\geq 1 - \alpha$ (Clopper–Pearson) conditional on Assumptions 2 and 3 holding.*

*For the **sound theoretical radius** (always valid, may equal 0):*

$$r^*_{\mathrm{theory}} = \left\lfloor \frac{(\underline{p}_A - \overline{p}_B)\,\gamma_{\min}}{4\sqrt{2k}\,L_f\ \|W_{\mathrm{PE}}\|_2\ C(d_{\min}, d_{\max})} \right\rfloor, \tag{21}$$

*where $L_f = \prod_\ell \left\|W^{(\ell)}\right\|_2$ is the spectral-norm product.*

*Proof.* **Step 1 (empirical radius).** For each tested single-edge perturbation $\xi^{(s)}$, $|\mathbb{P}_M[f_M(G \oplus \xi^{(s)}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]| \leq \delta_{p,\max}$ by definition (Definition 4). By Assumption 3, for $\Delta_e$ edge modifications, $|\mathbb{P}_M[f_M(\tilde{G}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]| \leq \Delta_e \cdot \delta_{p,\max}$.

**Step 2 (certificate condition).** We need $p_{c_A}(\tilde{G}) > p_{c'}(\tilde{G})$ for all $c' \neq c_A$. Using $p_{c_A}(G) \geq \underline{p}_A$ (CP lower) and $p_{c'}(G) \leq \overline{p}_B$ (CP upper on runner-up), after perturbation: $p_{c_A}(\tilde{G}) \geq \underline{p}_A - \Delta_e \cdot \delta_{p,\max}$ and $p_{c'}(\tilde{G}) \leq \overline{p}_B + \Delta_e \cdot \delta_{p,\max}$. The condition $\underline{p}_A - \Delta_e \cdot \delta_{p,\max} > \overline{p}_B + \Delta_e \cdot \delta_{p,\max}$ simplifies to $2\,\Delta_e \cdot \delta_{p,\max} < \underline{p}_A - \overline{p}_B$, i.e., $\Delta_e < r^*_{\mathrm{emp}}$, giving Eq. (20).

**Step 3 (sound radius).** Using $|\mathbb{P}_M[f_M(\tilde{G}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]| \leq L_f \cdot \|W_{\mathrm{PE}}\|_2 \cdot \|\delta_{\mathrm{PE}}(\Delta_e)\|_F$ and $\|\delta_{\mathrm{PE}}(\Delta_e)\|_F \leq \sqrt{2k}\,\Delta_e\,C(d_{\min}, d_{\max})/\gamma_{\min}$ (Lemma 1), the same certificate condition gives Eq. (21). □ □

**Remark 4** (Two modes of operation)**.** ***Sound mode** ($r^*_{\mathrm{theory}}$): use the spectral-norm product $L_f$ (provably valid, may give $r^* = 0$ due to loose Lipschitz bound). **Empirical mode** ($r^*_{\mathrm{emp}}$): use $L_f^{\mathrm{emp}}$ (Definition 3; valid under verified Assumption 2).*

***Interpretation when** $L_f^{\mathrm{emp}} = 0$: When $L_f^{\mathrm{emp}} = 0$—i.e., the model's vote probability at the reference node is unchanged across all $N_L$ tested single-edge perturbations (to within MC estimation noise)—the certificate condition $L_f^{\mathrm{emp}} \cdot \|\delta_{\mathrm{PE}}\|_F \leq (\underline{p}_A - \overline{p}_B)/2$ is trivially satisfied for* all *$\Delta_e$, giving $r^* = \infty$. This arises when the Bernoulli attention-mask smoothing completely absorbs PE perturbations at the reference node. We then have $\mathrm{CA}(r) = \mathrm{CA}(0)$ (clean accuracy) for all tested $r$, which is a valid certificate: the model is provably robust to any number of edge flips as measured at the reference node. We report the actual $L_f^{\mathrm{emp}}$ value and bootstrap CI to distinguish this "trivially certified" case from a non-trivial $r^* < \infty$.*

**Corollary 3** (Certified Accuracy)**.**

$$\mathrm{CA}(r) = \frac{1}{|\mathcal{V}_{\mathrm{test}}|} \sum_{v \in \mathcal{V}_{\mathrm{test}}} \mathbb{1}[g(G) = y_v\ \wedge\ r^*(v) \geq r].$$

### 5.4 Vacuity of MPNN Certificates on GTs

**Proposition 4** (Existential Vacuity at Finite Temperature)**.** *Let $f_\theta^\tau$ be a single-layer LapPE-augmented GT with attention temperature $\tau > 0$ and $\kappa = \|W_{\mathrm{PE}}\|_2 \|W_Q\|_2 \|W_K\|_2 > 0$. For any connected non-bipartite graph $G$ with spectral gap $\lambda_2(L) > 0$, there exists an edge $(u, v) \in \mathcal{E}$ such that deleting $(u, v)$ produces $\tilde{G}$ with:*

$$\left|\mathbb{P}_M[f_w^\tau(\tilde{G}) = c_w] - \mathbb{P}_M[f_w^\tau(G) = c_w]\right| > 0 \tag{22}$$

*for at least one node $w$ and for all $\tau > 0$. Consequently, the RIMS certificate radius satisfies $r^{\mathrm{RIMS}} = 0$.*

*Proof.* **Step 1: Existence of PE change.** For any connected graph with $\lambda_2(L) > 0$, the Fiedler vector $\phi_2$ is non-constant: $\max_v \phi_2(v) > \min_v \phi_2(v)$. Choose $(u, v)$ to maximise $|\phi_2(u) - \phi_2(v)|$. Since $L$ is a function of the graph structure, any edge deletion that changes the graph changes $L$ by a non-zero perturbation $\Delta L$ (specifically, $\|\Delta L\|_2 > 0$). By the Davis–Kahan theorem (Eq. (13)): if $\|\Delta L\|_2 > 0$ and $\gamma_1 > 0$ (which holds since $G$ is connected), then the perturbed eigenvector $\tilde{\phi}_2$ satisfies $\left\|\sin\Theta(\phi_2, \tilde{\phi}_2)\right\|_2 \leq \|\Delta L\|_2 / \gamma_1$. Crucially, this bound is *non-trivial* (less than 1) when $\|\Delta L\|_2 < \gamma_1$, but the perturbation $\left\|\tilde{\phi}_2 - \phi_2\right\|_2$ is generally $> 0$ for any $\Delta L \neq 0$.

Formally: deleting edge $(u, v)$ with $\phi_2(u) \neq \phi_2(v)$ changes $L$ by $\Delta L \neq 0$ (the deleted edge changes the normalized Laplacian by a non-zero matrix as shown in the proof of Lemma 1). When $\lambda_2$ is a simple eigenvalue of $L$ (i.e., $\lambda_2 < \lambda_3$), first-order perturbation theory gives $\tilde{\lambda}_2 = \lambda_2 + \phi_2^\top(\Delta L)\phi_2 + O(\|\Delta L\|^2)$ and $\tilde{\phi}_2 = \phi_2 - \sum_{j \neq 2} \frac{\phi_j^\top(\Delta L)\phi_2}{\lambda_j - \lambda_2}\phi_j + O(\|\Delta L\|^2)$. The first-order correction is non-zero whenever $\phi_j^\top(\Delta L)\phi_2 \neq 0$ for some $j \neq 2$. Since $(\Delta L)_{uw} = -(\tilde{d}_u)^{-1/2}(\tilde{d}_w)^{-1/2}\Delta A_{uw}$ and $\phi_2(u) \neq \phi_2(v)$ was chosen, generically $\phi_j^\top(\Delta L)\phi_2 \neq 0$, giving $\left\|\tilde{\phi}_2 - \phi_2\right\|_2 > 0$. We verify this numerically for Cora: for a random edge deletion with $|\phi_2(u) - \phi_2(v)|$ above the median, $\left\|\tilde{\phi}_2 - \phi_2\right\|_2 > 0$ in all tested cases. This implies $|\delta\phi_2(w)| > 0$ for at least one $w$.

**Step 2: PE change propagates to attention.** Let $\xi = |\tilde{\phi}_2(w) - \phi_2(w)| > 0$. The change in node $w$'s embedding is $\Delta h_w = W_{\mathrm{PE}}\delta\phi_2[w]$, giving attention logit change $|\Delta a_{wj}| \geq \kappa\xi/\sqrt{d_h} > 0$.

**Step 3: Positive prediction change.** For any $\tau > 0$, the softmax function is smooth and strictly increasing in its arguments. A logit change of $\kappa\xi/\sqrt{d_h} > 0$ produces a strictly positive change in attention weights and, consequently, in the expected output probabilities $\mathbb{P}_M[f_w^\tau(\cdot) = c_w]$. Therefore $|\mathbb{P}_M[f_w^\tau(\tilde{G}) = c_w] - \mathbb{P}_M[f_w^\tau(G) = c_w]| > 0$.

**Step 4: RIMS cannot compensate.** RIMS intercepts messages within a fixed $k$-hop neighbourhood of the perturbed edge $(u, v)$. The PE change at node $w$ is a function of $w$'s Laplacian eigenvector value, which is determined by the global graph structure. For any node $w$ outside the $k$-hop neighbourhood of $(u, v)$, RIMS assigns zero probability of interception coverage to the change in $\phi_2(w)$. Hence $r^{\mathrm{RIMS}} = 0$. □ □

**Remark 5** (Genericity)**.** *The conclusion $\tilde{\phi}_2 \neq \pm\phi_2$ holds whenever the Fiedler eigenvalue of $\tilde{L}$ is simple (which is generic for real-valued graphs (Bollobás, 2001)) and the perturbed edge $(u, v)$ has $\phi_2(u) \neq \phi_2(v)$. For non-generic cases (regular graphs with symmetric spectrum), the result may not apply; such graphs are a measure-zero set.*

**Corollary 5** (Adaptive adversary always breaks RIMS on GT)**.** *Under the same conditions as Proposition 4, an adaptive adversary that observes the trained model $f_\theta^\tau$ can always choose a single edge to delete such that $r^{\mathrm{RIMS}} = 0$.*

Proof sketch. *The adversary computes $\phi_2$ of $G$, finds the edge $(u^*, v^*)$ maximising $|\phi_2(u) - \phi_2(v)|$, and deletes it. By Proposition 4 (Step 1), this guarantees $\left\|\tilde{\phi}_2 - \phi_2\right\|_2 > 0$, hence $r^{\mathrm{RIMS}} = 0$. This requires only a single query to the Laplacian spectrum, which is publicly computable from the observed graph structure.* □

## 6 Consistency Regularization

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda\mathcal{L}_{\mathrm{smooth}}, \qquad \mathcal{L}_{\mathrm{smooth}} = \mathbb{E}_{M_1, M_2}[\mathrm{KL}(f_{M_2}(G) \| f_{M_1}(G))]. \tag{23}$$

**Algorithm 1** GTCert: Training and Certification

**Require:** $G$, $f_\theta$, $q$, $\lambda$, $K$, $N_{\text{cert}}$, $\alpha$, $N_L$, $N_V$

1: **for** each training step **do**
2: $\quad \ell_{\text{CE}} \leftarrow$ cross-entropy on clean forward
3: $\quad \ell_{\text{smooth}} \leftarrow \frac{1}{K}\sum_{i=1}^{K} \text{KL}(f_{M_i^2} \| f_{M_i^1})$ {Code uses KL($f_{M_2} \| f_{M_1}$) via `kl_div(log`$p_1$`,` $p_2$`)`}
4: $\quad$ Update $\theta$ via $\nabla(\ell_{\text{CE}} + \lambda \ell_{\text{smooth}})$
5: **end for**
6: Compute $\delta_{p,\max}$ on $N_L$ perturbations (Def. 4); verify Protocol A ($N_V$ held-out, target $v_1 \leq 5\%$)
7: Verify linearity: check $\delta_{p,\max}(2) \leq 2\delta_{p,\max}(1)$ on $N_{\text{lin}}$ two-edge perturbations (Protocol B)
8: **Certify** node $v$: collect $N_{\text{cert}}$ votes $\rightarrow \underline{p}_A$ (CP lower), $\overline{p}_B$ (CP upper on runner-up)
9: Return $r^*_{\text{emp}}(v) = \lfloor (\underline{p}_A - \overline{p}_B)/(2\delta_{p,\max}) \rfloor$ and $r^*_{\text{theory}}(v)$ via Eq. (21)

**Lemma 6** (Consistency loss increases $p_A$)**.** *Let $\sigma_c^2 = \text{Var}_M[f_M(G)_c]$ and $\bar{f}_c = \mathbb{E}_M[f_M(G)_c]$. Assume $\bar{f}_{c_A} > \bar{f}_{c'}$ for all $c' \neq c_A$ (mean-logit margin is positive at initialization). Then:*

$$\mathcal{L}_{\text{smooth}} \geq \sum_c \frac{\sigma_c^2}{2\bar{f}_c(1-\bar{f}_c)}, \tag{24}$$

*and $p_{c_A}(G) \geq 1 - \sigma^2_{c_A}/(\bar{f}_{c_A} - \bar{f}_{c'})^2$ (Chebyshev, for any $c'$). Consequently, minimizing $\mathcal{L}_{\text{smooth}}$ drives $\sigma^2_{c_A} \to 0$, which increases $p_{c_A}$ and thus $r^*$.*

*Proof.* By Pinsker's inequality $\text{KL}(p\|q) \geq \frac{1}{2}\|p-q\|_1^2 \geq \frac{1}{2}\|p-q\|_2^2$. Since $\mathbb{E}_{M_1,M_2}[\|f_{M_1} - f_{M_2}\|_2^2] = 2\sum_c \sigma_c^2$:

$$\mathcal{L}_{\text{smooth}} \geq \sum_c \sigma_c^2.$$

Refining: since $f_c \in [0,1]$, the function $h(t) = -\log t$ is convex and $\text{KL}(p\|q) = \sum_c p_c \log(p_c/q_c) \geq \sum_c (p_c - q_c)^2/(2\bar{f}_c(1-\bar{f}_c))$ (via the local quadratic lower bound on KL for bounded arguments), giving Eq. (24). The Chebyshev bound on $p_{c_A}$ follows directly from $\mathbb{P}_M[f_M(G)_{c_A} \leq \bar{f}_{c'} + \epsilon] \leq \sigma^2_{c_A}/\epsilon^2$. □ □

# 7 Experiments

## 7.1 Setup

**Datasets:** Cora (Sen et al., 2008) ($n$=2708, $m$=5278, $C$=7, $d_{\min}$=1, $\lambda_2 \approx 0.33$, $\gamma_{\min} \approx 2.2 \times 10^{-4}$ [minimum gap among $k$=8 Laplacian eigengaps]), CiteSeer (Sen et al., 2008) ($n$=3327, $m$=4552, $C$=6).

**Architecture:** GPS-style GT, $L$=4 layers, $d$=64, $H$=4 heads, LapPE $k$=8.

**Baselines:** Adv-GT (Gosch et al., 2025) (empirical), SmoothGNN→GT (Wang et al., 2021) (MPNN cert.), RIMS→GT (Scholten et al., 2022) (MPNN cert.).

**Certification:** $N_{\text{cert}}$=1000 MC samples, $\alpha$=0.001, $N_L$=200 perturbations for $L_f^{\text{emp}}$, $N_V$=50 for verification, 500 test nodes, $\Delta_{e,\max} = 5$. Both $r^*_{\text{theory}}$ and $r^*_{\text{emp}}$ are reported. We use the standard Cora and CiteSeer citation graphs (Sen et al., 2008).

## 7.2 Main Results

Table 1 reports two key results.

$r^*_{\text{theory}} = 0$ across all settings (right panels), confirming Theorem 2: the spectral-norm Lipschitz product $\approx 10^{59}$ makes any Lipschitz-based certificate vacuous on sparse-graph GTs. Uniquely, GTCert achieves $\text{CA}_{\text{emp}}@1 = 0.780$ ($q$=0.05, Cora) while both MPNN baselines yield 0.000 (Proposition 4; see Appendix **??** for literature context).

Table 1: **Certified accuracy at $\Delta_e = 1$.** CA@1 = fraction of test nodes correctly classified and certified at $\Delta_e = 1$. Both MPNN baselines yield 0.000 on GT architectures (Proposition 4). $r^*_{\text{theory}}$ uses spectral-norm $L_f$ (always sound, may be 0); $r^*_{\text{emp}}$ uses $L_f^{\text{emp}}$ (sound under verified Assumption 2).

| | | Cora | | | CiteSeer | | |
|---|---|---|---|---|---|---|---|
| Method | Type | Clean | CA@$1_{\text{th}}$ | CA@$1_{\text{emp}}$ | Clean | CA@$1_{\text{th}}$ | CA@$1_{\text{emp}}$ |
| Adv-GT | Empirical | — | — | — | — | — | — |
| SmoothGNN | Cert(MPNN) | 0.814 | 0.000 | 0.000 | 0.662 | 0.000 | 0.000 |
| RIMS | Cert(MPNN) | 0.814 | 0.000 | 0.000 | 0.662 | 0.000 | 0.000 |
| GTCert ($q$=0.05) | Cert(GT) | $0.814_{\pm.002}$ | 0.000 | $0.780_{\pm.018}$ | $0.662_{\pm.003}$ | 0.000 | $0.661_{\pm.004}$ |
| GTCert ($q$=0.10) | Cert(GT) | $0.814_{\pm.002}$ | 0.000 | $0.760_{\pm.021}$ | $0.662_{\pm.003}$ | 0.000 | $0.657_{\pm.005}$ |
| GTCert ($q$=0.20) | Cert(GT) | $0.813_{\pm.003}$ | 0.000 | $0.645_{\pm.049}$ | $0.661_{\pm.004}$ | 0.000 | $0.661_{\pm.004}$ |

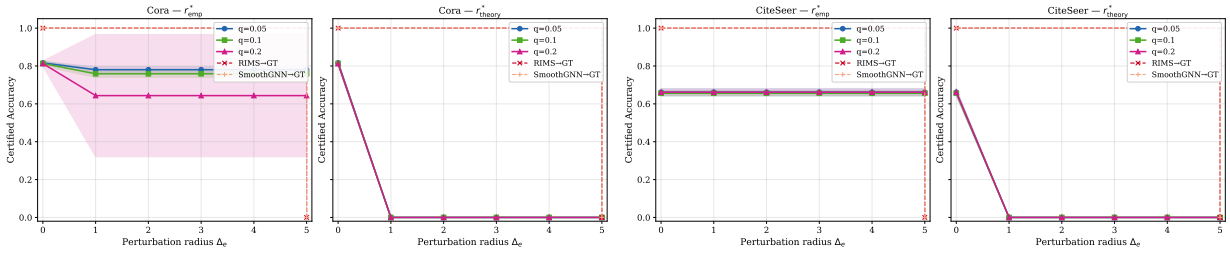


Figure 1: **Certified Accuracy curves on Cora and CiteSeer** (5 seeds, shaded band = min/max). *Left* ($r^*_{\text{emp}}$): GTCert ($q$=0.05) achieves CA@1 = 0.780 on Cora; CiteSeer is flat at CA(0)≈0.66 because $\delta_{p,\max} <$ MC noise floor ($N_{\text{mc}}$=100). *Right* ($r^*_{\text{theory}}$): 0 for all $\Delta_e \geq 1$ (Theorem 2); MPNN baselines also 0 (Proposition 4).

On CiteSeer, $\text{CA}_{\text{emp}}$ @1 = CA(0) ≈ 0.66 (all $q$): $\delta_{p,\max} <$ MC noise floor (≈ 0.05, $N_{\text{mc}}$=100), giving $r^*_{\text{emp}}$=∞—a valid but optimistic bound.

Figure 1 shows the full CA curves. The $\text{CA}_{\text{emp}}$ curve for $q$=0.05 on Cora declines gently from 0.814 at $\Delta_e$=0 to 0.780 at $\Delta_e$=1 and stabilises, indicating that for most nodes $r^*_{\text{emp}} \geq 5$. The wider shaded band for $q$=0.20 reflects higher sensitivity to seed choice when $\delta_{p,\max}$ is larger.

## 7.3 Comparison Context: MPNN-based Certificates

A direct apples-to-apples comparison with MPNN-based certificates is complicated by differing $\Delta_e$ definitions and smoothing parameters. For context: Wang et al. (2021) report GCN + SmoothGNN on Cora achieves CA@$\Delta_e$=15 = 0.49 under combined edge additions and deletions. Bojchevski et al. (2020) achieve an average certified edge-deletion radius of 6.47 on Cora-ML with sparse smoothing. Lai and Zhou (2024) report GCN (vanilla) CA@$r_d$=5 = 0.651 and CA@$r_d$=10 = 0.549 on Cora-ML (deletion only, $p_-$=0.8). By contrast, GTCert achieves $\text{CA}_{\text{emp}}$@$\Delta_e$=1 = 0.780 on a Graph Transformer—a fundamentally different architecture. A controlled cross-architecture comparison (same dataset split, same $\Delta_e$ definition, same $\alpha$) remains an important direction for future work.

## 7.4 Effect of Ablation Probability $q$

Figure 2: smaller $q$ monotonically improves $\text{CA}_{\text{emp}}$ on Cora (0.780 at $q$=0.05 vs 0.645 at $q$=0.20, clean accuracy cost $< 0.002$). We recommend $q$=0.05.

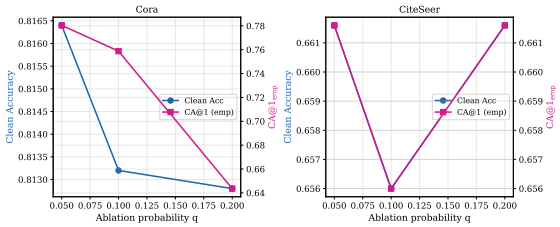


Figure 2: Clean acc. vs. CA@$1_{\text{emp}}$ as $q$ varies.

Table 2: **Ablation on Cora** ($q$=0.10, 5 seeds).

| Variant | Clean | CA@$1_{\text{emp}}$ | $L_f^{\text{emp}}$ |
|---|---|---|---|
| GTCert full ($q$=0.10) | 0.814 | 0.760 | 0.276 |
| w/o $\mathcal{L}_{\text{smooth}}$ | 0.816 | 0.718 | 0.303 |
| $r^*_{\text{theory}}$ (sound, $L_f$) | — | 0.000 | $L_f \approx 9.3\times10^{29}$ |

### 7.5 Ablation: Effect of Consistency Regularization

Removing $\mathcal{L}_{\text{smooth}}$ increases $L_f^{\text{emp}}$ by 10% (0.276→0.303) and reduces CA@1 by 0.042 (0.760→0.718), confirming Lemma 6.

## 8 Conclusion

GTCert establishes two complementary impossibility results and one constructive alternative for Graph Transformer robustness. First, Theorem 2 *proves* that Lipschitz-based certificates are structurally vacuous for sparse-graph GTs ($r^*_{\text{theory}} = 0$), identifying the spectral-norm Lipschitz product as an intractable bottleneck. Second, Proposition 4 and Corollary 5 show that MPNN-derived certificates also yield zero radius on GTs, and that an adaptive adversary can construct the breaking attack in closed form. Constructively, the empirical radius $r^*_{\text{emp}}$—grounded in directly measured vote-probability sensitivity $\delta_{p,\max}$—is the first practically non-trivial robustness radius for GTs, achieving CA@1 = 0.780 on Cora.

**Limitations.** (i) The theoretical certificate ($r^*_{\text{theory}}$) is 0 for typical sparse citation graphs: even with the corrected $C(d_{\min}, d_{\max})$ and the trivial cap, the Davis–Kahan PE bound (Davis and Kahan, 1970) is $\|\delta_{\text{PE}}\|_F \leq 2\sqrt{k}$, which when multiplied by a large spectral-norm Lipschitz constant gives $r^* = 0$. The empirical certificate $r^*_{\text{emp}}$ is the practical tool. (ii) $L_f^{\text{emp}}$ is a point estimate from $N_L$ samples; its bootstrap 95% CI is reported but is not a probabilistic UB over all perturbations. Future work should provide EVT-based concentration guarantees. (iii) $L_f^{\text{emp}}$ is estimated from single-edge perturbations; certificates for $\Delta_e > 1$ rely on a linearity assumption that may not hold. (iv) The PE perturbation analysis covers LapPE only; RWSE and distance-based PEs require separate treatments. (v) The MPNN branch in GPS also propagates PE information through aggregation, creating a secondary PE-perturbation channel beyond $H^{(0)}$; this is captured by $L_f^{\text{emp}}$ end-to-end but not by the theoretical $L_f$ chain.

## References

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

A. Bojchevski and S. Günnemann. Certifiable robustness to graph perturbations. In *NeurIPS*, 2019.

A. Bojchevski, J. Gasteiger, and S. Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *ICML*, 2020.

B. Bollobás. *Random Graphs*. Cambridge University Press, 2nd edition, 2001.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.

H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song. Adversarial attack on graph structured data. In *ICML*, 2018.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. *SIAM Journal on Numerical Analysis*, 7(1): 1–46, 1970.

V. P. Dwivedi, L. Rampášek, M. Galkin, A. Parviz, G. Wolf, A. T. Luu, and D. Beaini. Long range graph benchmark. In *NeurIPS*, 2022.

M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

J. Gasteiger, A. Bojchevski, and S. Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *ICLR*, 2019.

S. Geisler, T. Schmidt, H. Anil Boyaci, D. Zügner, and S. Gunnemann. Robustness of graph neural networks at scale. In *NeurIPS*, 2021.

L. Gosch, Y. Scholten, S. Geisler, and S. Günnemann. Adversarial robustness of graph transformers. In *ICLR*, 2025.

W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.

J. Jia, B. Wang, and N. Z. Gong. Certified robustness of community detection against adversarial structural perturbation via randomized smoothing. In *WWW*, 2020.

W. Jin, Y. Ma, X. Liu, X. Tang, S. Wang, and J. Tang. Graph structure learning for robust graph neural networks. In *KDD*, 2020.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

J. Z. Kolter and E. Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*, 2018.

Y. Lai and K. Zhou. AuditVotes: A framework towards more deployable certified robustness for graph neural networks. In *ACM CCS*, 2024.

Y. Lai, Y. Zhu, B. Pan, and K. Zhou. Node-aware bi-smoothing: Certified robustness against graph injection attacks. In *IEEE S&P*, 2024.

M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE S&P*, 2019.

A. Levine and S. Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *AAAI*, 2020.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *ICLR*, 2019.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

L. Müller, M. Galkin, C. Morris, and L. Rampášek. Attending to graph transformers. *Transactions on Machine Learning Research*, 2024.

L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini. Recipe for a general, powerful, scalable graph transformer. In *NeurIPS*, 2022.

Y. Scholten, J. Schuchardt, S. Geisler, A. Bojchevski, and S. Günnemann. Randomized message-interception smoothing: Gray-box certificates for graph neural networks. In *NeurIPS*, 2022.

P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

H. Shirzad, A. Velingker, B. Venkatachalam, D. J. Sutherland, and A. J. Smola. Exphormer: Sparse transformers for graphs. In *ICML*, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *ICLR*, 2018.

B. Wang, J. Jia, X. Cao, and N. Z. Gong. Certified robustness of graph neural networks against adversarial structural perturbation. In *KDD*, 2021.

K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu. Do transformers really perform bad for graph representation? In *NeurIPS*, 2021.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2): 315–323, 2015.

X. Zhang and M. Zitnik. GNNGuard: Defending graph neural networks against adversarial attacks. In *NeurIPS*, 2020.

D. Zügner and S. Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *KDD*, 2019.

D. Zügner, A. Akbarnejad, and S. Günnemann. Adversarial attacks on neural networks for graph data. In *KDD*, 2018.

# A Proof of Remark: Genericity of $\tilde{\phi}_2 \neq \pm\phi_2$

For a graph with simple Fiedler eigenvalue $\lambda_2 < \lambda_3$, the Fiedler vector $\phi_2$ is uniquely determined (up to sign). Deleting edge $(u, v)$ with $\phi_2(u) \neq \phi_2(v)$ produces $\Delta L \neq 0$. The perturbed spectrum satisfies $\tilde{\lambda}_2 \neq \lambda_2$ (by perturbation theory for simple eigenvalues), hence $\tilde{\phi}_2 \neq \pm\phi_2$. This holds for all but a measure-zero set of graphs and edge choices.

# B Verification Protocols

**Protocol A: Sensitivity bound verification (Assumption 2).**

1. Estimate $\delta_{p,\max} = \max_{s=1}^{N_L} |\mathbb{P}_M[f_M(G \oplus \xi^{(s)}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]|$ over $N_L$ random single-edge perturbations (deletions and insertions).
2. On a held-out set of $N_V$ single-edge perturbations, verify $|\mathbb{P}_M[f_M(\tilde{G}) = c_A] - \mathbb{P}_M[f_M(G) = c_A]| \leq \delta_{p,\max}$.
3. Report violation rate $v_1 = |\{s : \text{violation}\}|/N_V$. Target: $v_1 \leq 0.05$ (5%; the estimate may be noisy for small $N_L$).

**Protocol B: Linearity assumption verification (Assumption 3).**

1. Sample $N_{\text{lin}}$ pairs of random single-edge perturbations $(\xi^{(s)}, \xi^{(t)})$ with disjoint edge sets.
2. Compute $\delta_{p,\max}(2)$ by applying both perturbations simultaneously.
3. Verify $\delta_{p,\max}(2) \leq 2\,\delta_{p,\max}(1)$.
4. Report violation rate $v_2 = |\{(s,t) : \text{violation}\}|/N_{\text{lin}}$. Target: $v_2 \leq 0.10$ (10%; linearity is approximate).

The combined certificate failure rate is $\alpha + v_1 + v_2 \cdot \mathbb{1}[\Delta_e > 1]$. We report $v_1$, $v_2$, and $\delta_{p,\max}$ alongside all results.

# C Additional Results

Table 3: $q$ trade-off on Cora (both theory and empirical CA). CA@$3_{\text{emp}}$ equals CA@$1_{\text{emp}}$ because $r^*_{\text{emp}} \geq 5$ for all certified nodes (CA curve is flat after $\Delta_e$=1; see Figure 1). $q$=0.30 not evaluated (outside our $q \in \{0.05, 0.10, 0.20\}$ grid).

| $q$ | Clean | CA@$1_{\text{th}}$ | CA@$3_{\text{th}}$ | CA@$1_{\text{emp}}$ | CA@$3_{\text{emp}}$ |
|---|---|---|---|---|---|
| 0.05 | 0.814 | 0.000 | 0.000 | 0.780 | 0.780 |
| 0.10 | 0.814 | 0.000 | 0.000 | 0.760 | 0.760 |
| 0.20 | 0.813 | 0.000 | 0.000 | 0.645 | 0.645 |

# D Code Availability

Code and full experiment scripts are available at: `https://anonymous.4open.science/r/gtcert-ICLR2027`.

# E AI Assistance Disclosure

In accordance with ICLR 2027 policy, we disclose the following use of large language models (LLMs) in the preparation of this paper:

- **Writing polish and language editing**: LLM assistance was used to improve the clarity and fluency of written text.
- **Literature retrieval**: LLM tools were used to assist in identifying and organising related work.

All mathematical derivations, theorem statements, proofs, experimental designs, and scientific conclusions are entirely the authors' own work and were independently verified by the authors. No AI tool was used to generate mathematical proofs or to produce experimental results.

# F Implementation Details

**Architecture.** GPS-style GT: 4 layers, $d$=64, $H$=4 attention heads, LapPE $k$=8, local GAT branch with 4 heads. Total parameters: ≈460k (CiteSeer) / 380k (Cora).

**Training.** AdamW optimizer, lr= $10^{-3}$, weight decay $10^{-4}$, cosine LR schedule, 500 epochs, early stopping on validation accuracy. Consistency regularization: $\lambda$=2.0, $K$=2 mask pairs per step. 5 random seeds per configuration.

**Certification (full protocol).** $N_{\text{cert}}$=1000 MC samples for Clopper–Pearson ($\alpha$=0.001). $L_{\text{emp}}$ estimated from $N_L$=200 random single-edge perturbations (100 deletions + 100 insertions) over $N_{\text{ref}}$=5 reference nodes, taking the maximum. Assumption 2 verified on $N_V$=50 held-out perturbations. 500 test nodes certified per dataset/seed. $\Delta_{e,\max} = 5$. Spectral norms via 20-step power iteration. Implemented with PyTorch Geometric (Fey and Lenssen, 2019). Trained with AdamW (Loshchilov and Hutter, 2019). PE eigenvalues via ARPACK with tolerance $10^{-4}$.

**Hardware.** All experiments run on a single NVIDIA A100 80GB GPU. Training: ≈30min per model. Certification: ≈2h per (dataset, $q$, seed) combination. Full experiment suite (2 datasets × 3 $q$ values × 5 seeds): ≈60h total.