Exact Certification of (Graph) Neural Networks Against Label Poisoning

Mahalakshmi Sabanayagam*, Lukas Gosch*, Stephan Günnemann, Debarghya Ghoshdastidar
* Equal Contribution

First *EXACT* Certificate for Neural Networks Against Label Poisoning

- ➢ Leverages the **Neural Tangent Kernel** (NTK) to capture the training dynamics of wide NNs
- ➢ Finding: a **novel phenomenon** of robustness plateauing for intermediate perturbation budgets
- ➢ Well suited for **semi-supervised learning**, we focus on node classification in graphs

Label Poisoning

An adversary $\mathscr{A}$ can perturb a small *fraction* ϵ *of the training data labels* Y to induce misclassification of a classifier f_θ after training on $\tilde{Y}$.

$$\mathscr{A}(Y) = \left\{ \tilde{Y} \mid \left\| \tilde{Y} - Y \right\|_0 \leq \epsilon m,\ m = \text{No. of training data} \right\}$$

Robustness Certification

Prove that the prediction of f_θ for a given test point doesn't change for any $\tilde{Y} \in \mathscr{A}(Y)$ compared to training on the clean data labels Y

$$\min_{\tilde{Y},\theta} \mathscr{L}_{att}(\theta, \tilde{Y}) \quad \text{s. t.} \quad \tilde{Y} \in \mathscr{A}(Y) \ \wedge\ \theta \in \arg\min_{\theta'} \mathscr{L}_{trn}(\theta', \tilde{Y})$$

Bilevel problem!!
Unsolved for neural networks so far…
Unsolved even for classical models!

Is it even possible to derive a practically computable certificate for (G)NNs?

✓ Yes, for **sufficiently wide networks** by leveraging the NTK
✓ Yes, given sparse labels, e.g., **semi-supervised** learning settings

On Infinite-width NNs and the NTKs

When width W of a NN f_θ goes to infinity → training dynamics described by its NTK

- NTK Q_{ij} between samples i and j is $\mathbb{E}_\theta[\langle \nabla_\theta f_\theta(x_i), \nabla_\theta f_\theta(x_j) \rangle]$
- NTK readily available for different (G)NN architectures

On the Equivalence to Support Vector Machines

Train f_θ by optimizing a soft-margin loss with gradient descent

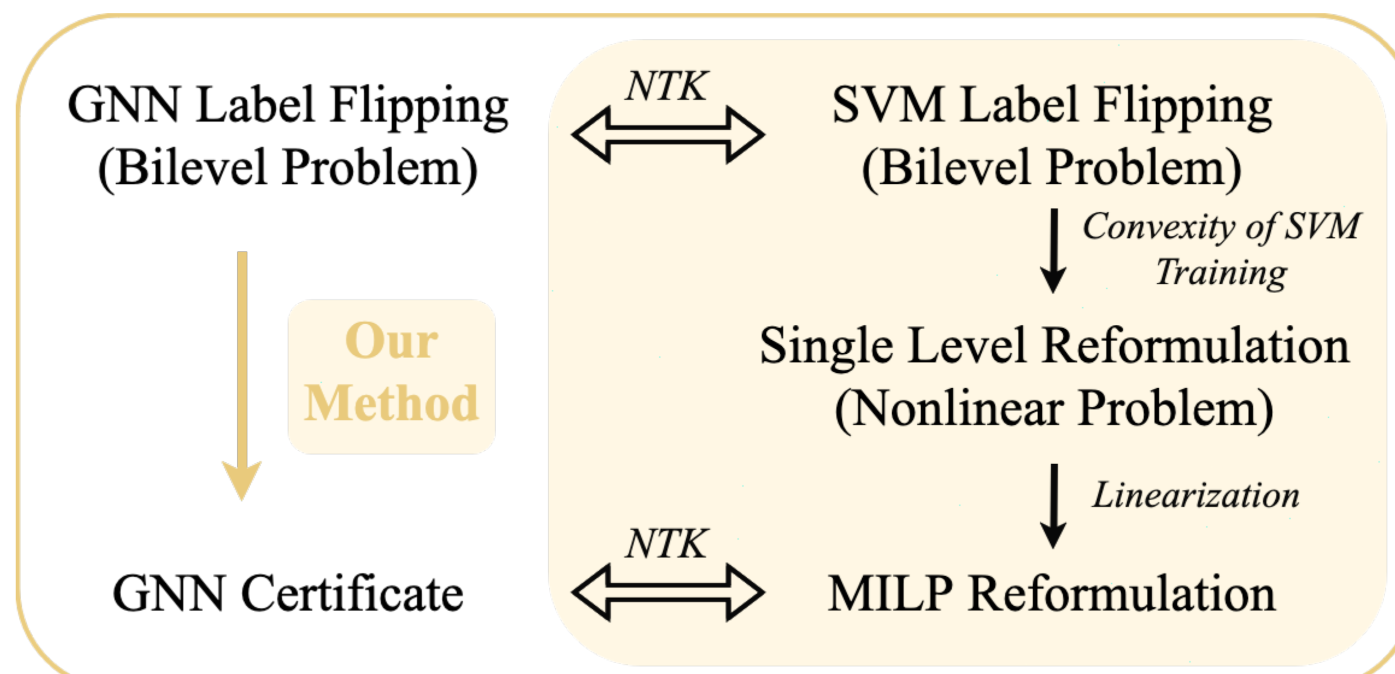
$$\mathscr{L}_{trn}(\theta, Y) = \min_\theta C \sum_{i=1}^{m} \max(0, 1 - y_i f_\theta(x_i)) + \frac{1}{2} \| W^L \|_2^2$$

When the width of f_θ goes to infinity → training dynamics equivalent to soft-margin SVM with f_θ's NTK as the kernel

Recap: Dual problem of an SVM

$$S(Y): \quad \min_\alpha - \sum_{i=1}^{m} \alpha_i + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j Q_{ij} \ \text{s.t.}\ 0 \leq \alpha_i \leq C \ \forall i \in [m]$$

LabelCert: Our Certification Framework



Sample-wise certificate: Guarantee if each test prediction is robust **independently**

$$P(Y): \quad \min_{\tilde{Y},\alpha} sgn(\hat{p}_t) \sum_{i \in V_L} \tilde{y}_i \alpha_i Q_{ti} \quad \text{s. t.} \quad \tilde{Y} \in \mathscr{A}(Y) \ \wedge\ \alpha \in S(\tilde{Y})$$

Original prediction: $\hat{p}_t$

Exact!

Collective certificate: Number of test predictions that are **simultaneously** robust

$$C(Y): \quad \max_{\tilde{Y},\alpha} \sum_{t \in \mathscr{T}} \mathbb{I}[sgn(\hat{p}_t) \neq sgn(p_t)] \quad \text{s. t.} \quad \tilde{Y} \in \mathscr{A}(Y) \ \wedge\ \alpha \in S(\tilde{Y})$$

New prediction: p_t

Exact!

MILP Reformulation: The Mathy-Gritty Details

Proposition: $\alpha \in S(\tilde{Y})$ can be replaced by its KKT conditions → **single-level problem**

KKT conditions

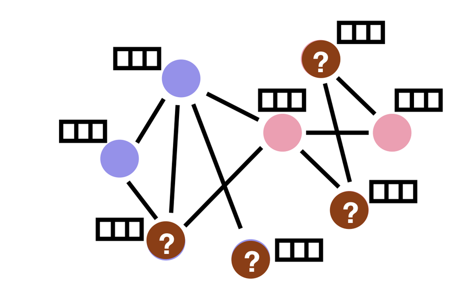
(Stationarity) $\forall i \in [m]: \ \sum_{j=1}^{m} \tilde{y}_i \tilde{y}_j \alpha_j \Theta_{ij} - 1 - u_i + v_i = 0$ → **linearise** product terms through new variables $z_i = \alpha_i \tilde{y}_i$ & $R_{ij} = \tilde{y}_i z_j$

(Complementary Slackness) $u_i \alpha_i = 0,\ v_i(C - \alpha_i) = 0$ → **linearise** using combinatorial structure and big-M constraints

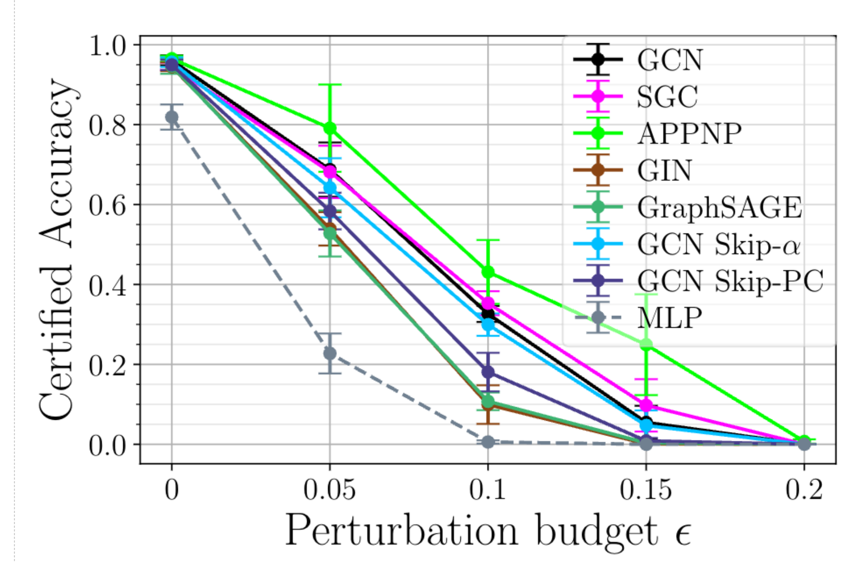
Experiments

Semi-supervised node classification using GNNs

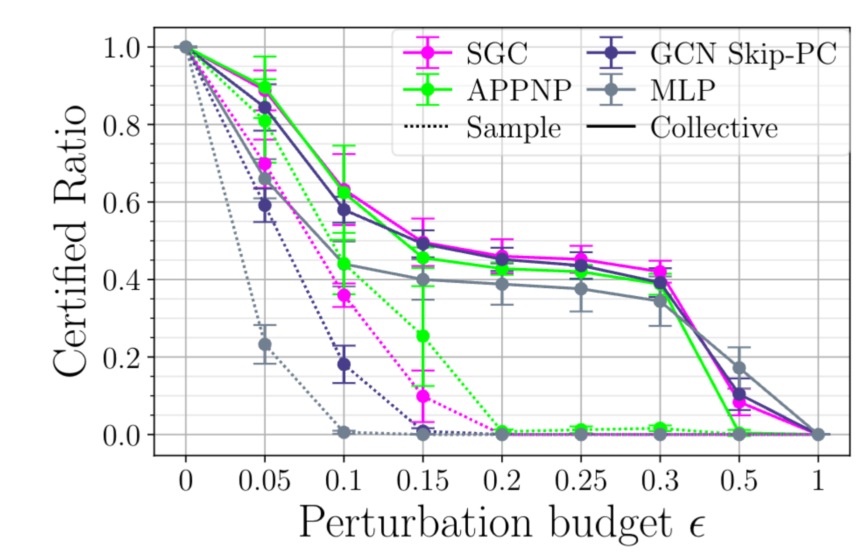
- Given a graph $G = (A, X)$ with node features X, and labels Y, label the unlabeled nodes in G
- Dataset: Cora-MLb (see our paper for others)



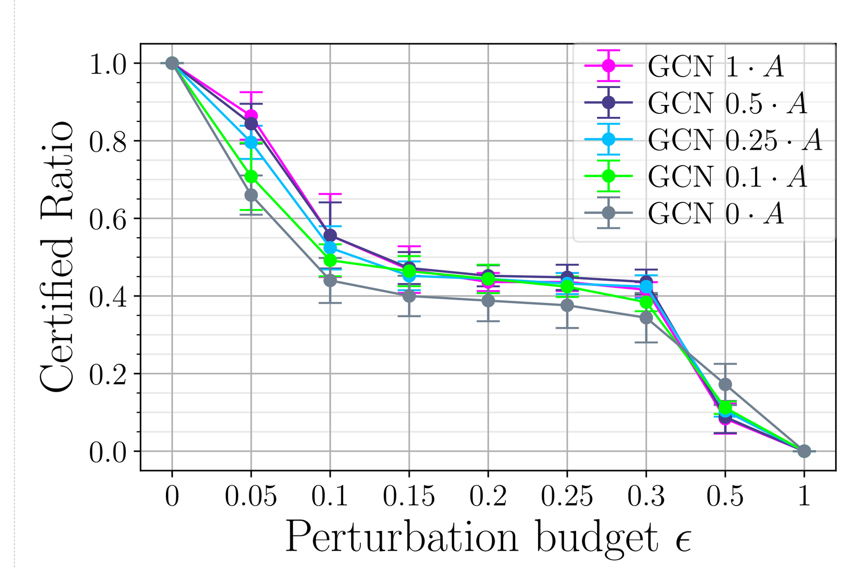
Sample-wise certificate



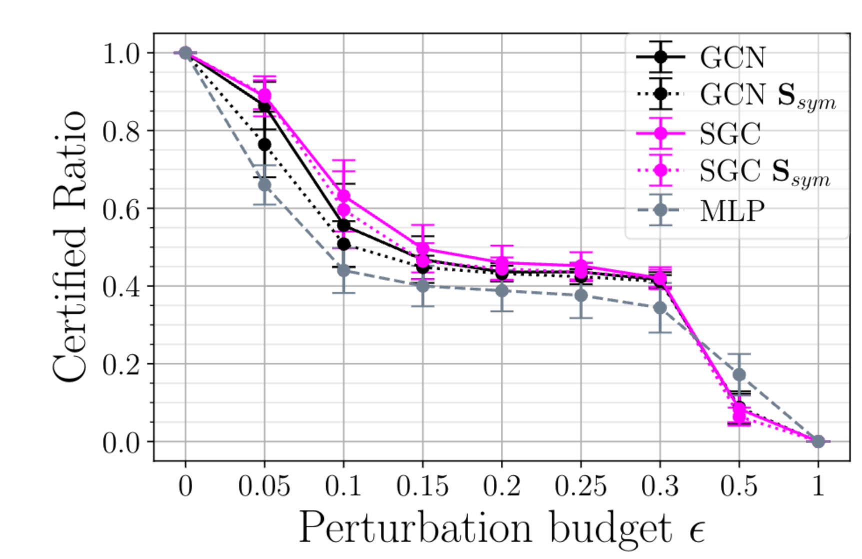
Collective certificate



Effect of graph structure



Effect of convolution S_{row} vs S_{sym}



Key Takeaways on Certifying (G)NNs Against Label Poisoning

- ➢ Phenomenon: **robustness plateaus** at intermediate perturbations
- ➢ Collective certificates **complement** sample-wise
- ➢ GNN robustness hierarchies are **strongly data dependent**
- ➢ On *graph structure:* **Increasing graph information, graph density and homophily help** robustness
- ➢ On *architecture choices:* **Linear activation helps, depth in skip-connection hurts** robustness

Open questions: extending to other data domains, scalability