Exact Verification of Graph Neural Networks with Incremental Constraint Solving

Minghao Liu, Chia-Hsuan Lu, Marta Kwiatkowska
University of Oxford, United Kingdom
{minghao.liu, chia-hsuan.lu, marta.kwiatkowska}@cs.ox.ac.uk

Graph Neural Networks (GNNs)

GNNs are a family of neural network architectures over **graph data**. GNNs have been deployed in real-world high-stakes applications, e.g.:

- **Financial networks**: Fraud detection, anti-money laundering…
- **Chemical compounds**: Molecular property prediction, drug discovery…
- **Social networks**: Friend recommendation, community detection…

Input: Attributed directed graph $G = \langle V, E, X \rangle$.

- V: a node set;
- $E \subseteq V \times V$: an edge set;
- $X \in \mathbb{R}^{|V| \times D}$: real attribute vectors for the nodes.

Message passing: The learning mechanism of GNNs.

- $\mathbf{h}_v^{(k)}$: real-valued embedding vector of each node $v \in V$ for the k-th layer of GNN ($\mathbf{h}_v^{(0)} = X_v$); assume there are K layers;
- We consider *GraphSAGE* architecture in this paper:

$$\mathbf{h}_v^{(k)} = \sigma\left(\mathbf{W}_1^{(k)} \cdot \mathbf{h}_v^{(k-1)} + \mathbf{W}_2^{(k)} \cdot \mathbf{aggr}\left(\left\{\mathbf{h}_u^{(k-1)} \mid u \in \mathcal{N}(v)\right\}\right) + \mathbf{b}_1^{(k)}\right)$$

ReLU activation; Aggregators: `sum`, `max`, `mean`

- Apply a classifier to $\mathbf{h}_v^{(K)}$ to obtain node/graph classification predictions.

Adversarial Robustness of GNNs

GNNs have been shown to be vulnerable to **adversarial attacks**.

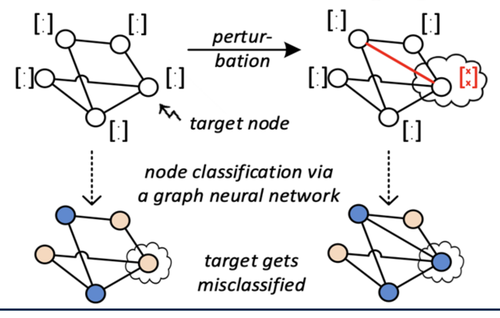
Admissible Graph Perturbation Space

- **Attribute perturbation**: Change the value of node attributes.
  - For every $v \in V$ and $1 \le i \le d_0$, $\epsilon_{v,i}^l \le \tilde{\mathbf{x}}_v[i] \le \epsilon_{v,i}^u$;
- **Structural perturbation**: Insert/remove edges in a fragile edge set F.
  - Only perturb fragile edges: $F \backslash E \subseteq \tilde{E} \subseteq E \cup F$;
  - Global budget: $E \backslash \tilde{E} + \tilde{E} \backslash E \le \Delta$;
  - Local budget: For every $v \in V$, $\left|\mathcal{N}(v) \backslash \tilde{\mathcal{N}}(v)\right| + \left|\tilde{\mathcal{N}}(v) \backslash \mathcal{N}(v)\right| \le \delta_v$;

Definition of GNN Robustness

Given a node-classification GNN f, an attributed directed graph G with admissible perturbation space $Q(G)$, a target node $t \in V$, and $f(G, t) = \hat{c}_t$, we say that f is adversarially **robust** for t with class $\hat{c}_t$ if and only if, for every perturbed graph $\tilde{G} \in Q(G)$, it holds that $f(\tilde{G}, t) = \hat{c}_t$.

- We consider both node- and graph-classification GNNs in this paper.



Shown in red: Demonstration of attribute and structural perturbations.

Contributions

We use formal verification to guarantee the robustness of GNNs.

- We introduce the first exact verification method for GNNs with two common aggregators: `max` and `mean`.
- We design specialised **tightened bound propagation** strategies and propose an **incremental constraint solving** algorithm.
- We implement *GNNev*, an **open-source exact verifier** for GNNs that supports new aggregators, attribute/structural perturbations, and edge additions/deletions.
- The performance is evaluated through extensive experiments on real-world graph datasets (e.g., fraud detection, biochemistry).

Methodology

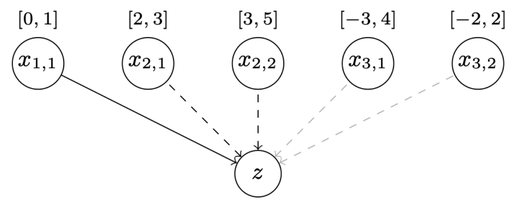
Bound Tightening for New Aggregators

- We consider the upper and lower bounds of $z \coloneqq \mathbf{aggr}(X_1 \cup X_2' \cup X_3')$.
  - X_1: Set of variables from non-fragile edges $(E \backslash F)$
  - X_2: Set of variables from fragile edges $(E \cap F)$, and $X_2' \subseteq X_2$;
  - X_3: Set of variables from fragile non-edges $(F \backslash E)$, and $X_3' \subseteq X_3$;
  - There is a constant integer s, such that $|X_2 \backslash X_2'| + |X_3'| \le s$.

Max aggregation: in $O(N \log s)$ time by case analysis.

Mean aggregation: improve from $O(s^2 + N \log s)$ to $O((s + N) \log s)$.

* N is the total size of the sets X_1, X_2, and X_3.



	Plain	Tightened
max (ub)	$\max(1,3,5,4,2) = \mathbf{5}$	$\max(1,5,4) = \mathbf{5}$
max (lb)	$\min(0,2,3,-3,-2) = -3$	$\max(0,2) = \mathbf{2}$
mean (ub)	$\frac{1+5+4}{3} = 3.33$	$\frac{1+5+3+4}{4} = \mathbf{3.25}$
mean (lb)	$\frac{0-3-2}{3} = -1.67$	$\frac{0+2+3-3}{4} = \mathbf{0.5}$

- The bounds for max and mean aggregations are **tight** (reachable).

Incremental Constraint Solving

- We encode the verification problems as MIP instances.
- Graphs are typically large, and solving efficiency is decreased.
- We utilise incremental solving to iteratively solve a series of relaxation problems (i.e., encode one new layer) and decide robust results early.
- Exactness is still maintained and performance is improved.

Implementation & Experiments

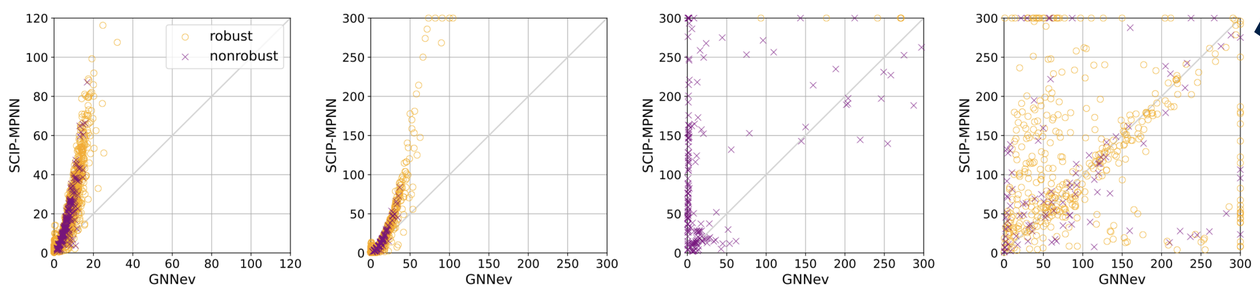
We implement our method as *GNNev* with underlying MIP solver *Gurobi*.

- Support 3 common aggregators, attribute/structural perturbations, and is easy to deploy (can load PyTorch Geometric models directly).

Datasets: 4 for node classification (Cora, CiteSeer, Amazon, Yelp) and 2 for graph classification (MUTAG, ENZYMES).

Experimental Results

- Comparison with baselines *SCIP-MPNN* on sum-aggregated GNNs showed that performance of *GNNev* is highly competitive, especially on node classification GNNs.
- Evaluation on all aggregators & perturbations showed that *GNNev* performs better for node classification, and max-aggregated instances are more challenging to scale up.
- Case study on fraud detection and biochemistry showed that *GNNev* can gain insight into the susceptibility of GNNs to adversarial attacks.



(a) Cora (b) CiteSeer (c) MUTAG (d) ENZYMES

FM 2026 TOKYO — The 27th International Symposium on Formal Methods

UNIVERSITY OF OXFORD

UKRI UK Research and Innovation

elsa European Lighthouse on Secure and Safe AI

Extended paper on arXiv

Source code on GitHub