Technical University of Munich
Department of Informatics
Data Analytics and Machine Learning Group

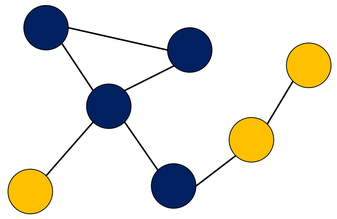
Collective Robustness Certificates

Exploiting Interdependence in Graph Neural Networks

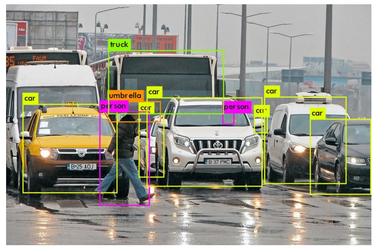
Jan Schuchardt, Aleksandar Bojchevski, Johannes Klicpera, Stephan Günnemann

Motivation

Many tasks involve **multiple predictions** based on a **single input …**



Multiple nodes classified in a single graph



Multiple objects detected in a single image

… that can be **jointly attacked.**

Existing robustness certificates only consider single predictions.

Research Question

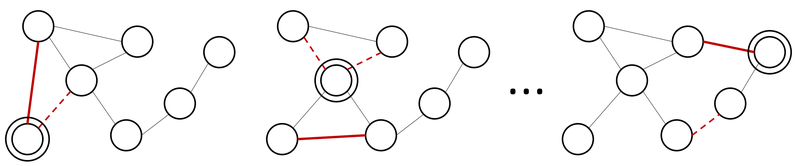
How can we certify the **collective adversarial robustness**

- of classification models
- based on Graph Neural Networks?

A naïve method

Certify each prediction independently!

But this assumes a different attack on each prediction.

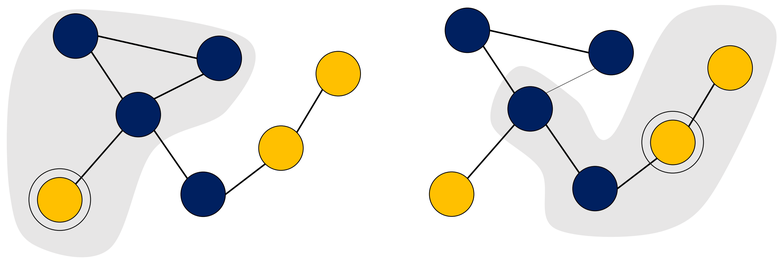


Our method

- Certifies collective adversarial robustness
- Combines single-prediction certificates
- Models a shared, persistent input

Ingredient 1: Locality

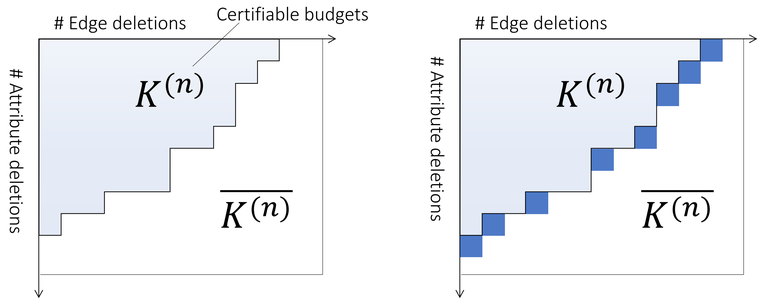
Predictions are based on local receptive fields (think GCN).



→ Not all perturbations affect all predictions

Ingredient 2: Linear certificate encoding

Enable evaluation of single-prediction certificates within LPs …

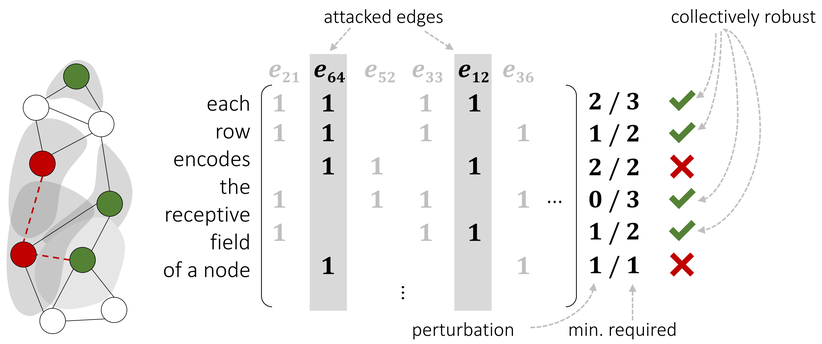


… by encoding their pareto front in the adversarial budget space.

→ Prediction n robust to $\boldsymbol{b}$ perturbations $\Leftarrow \neg\exists \boldsymbol{p} \in P^{(n)} : \forall d : b_d \geq p_d$.

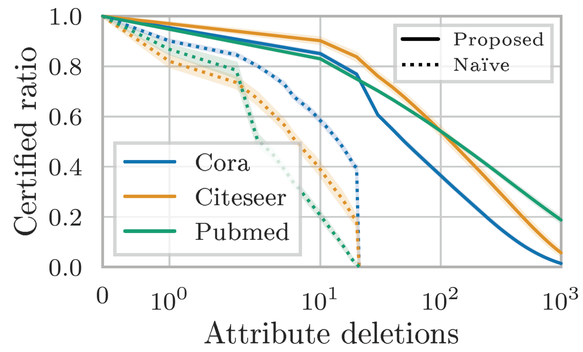
Combining Ingredients 1 & 2

- Optimize over a single perturbed graph
- Aggregate local perturbation $\boldsymbol{b}^{(n)}$ in each receptive field
- Evaluate single-prediction certificates based on $\boldsymbol{b}^{(n)}$

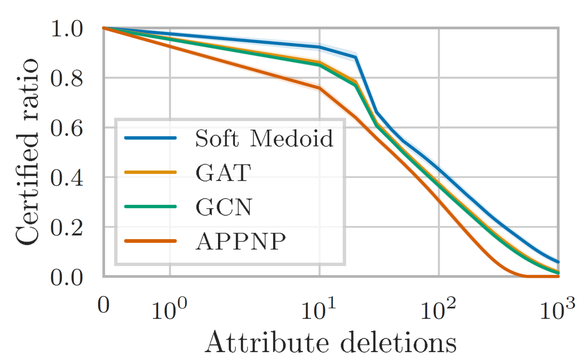


Our certificate …

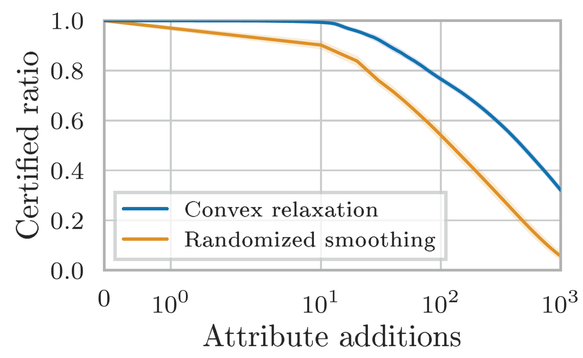
… yields orders of magnitude stronger robustness guarantees,



… is model-agnostic,



… compatible with any single-prediction certificate,



… and strengthened by any advance in classifier certification.