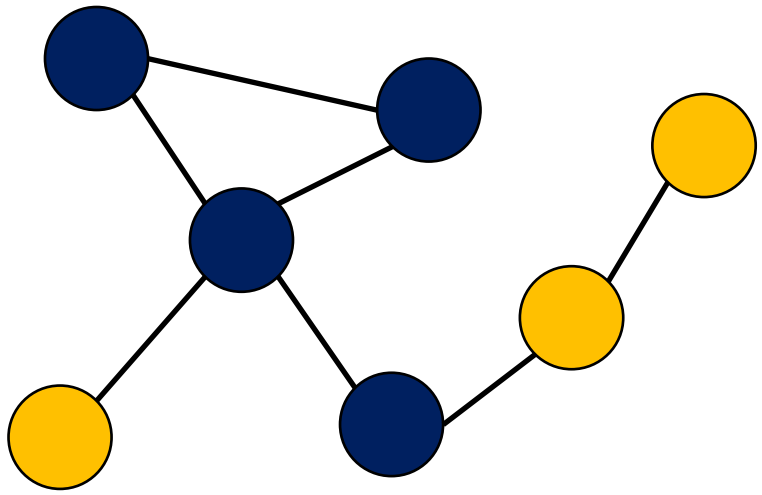
Technical University of Munich
Department of Informatics
Data Analytics and Machine Learning Group

Collective Robustness Certificates

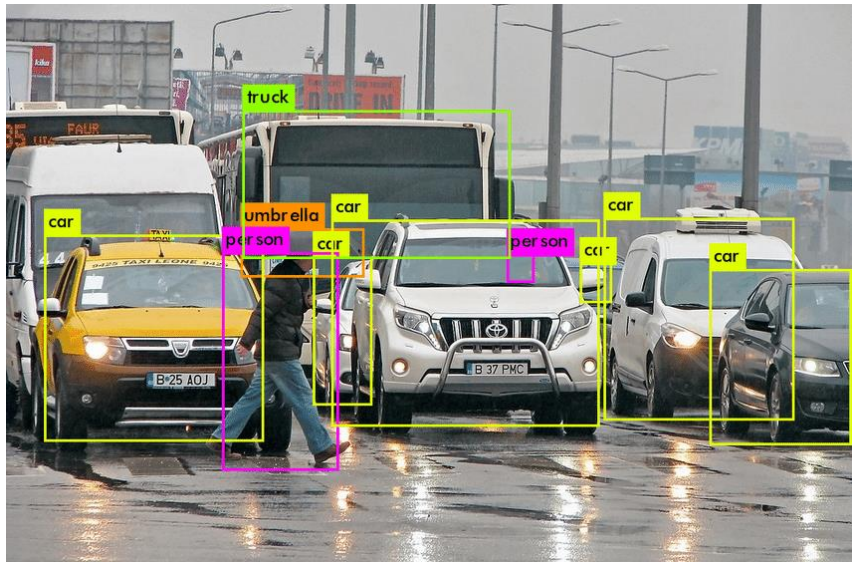
Exploiting Interdependence in Graph Neural Networks

Jan Schuchardt, Aleksandar Bojchevski, Johannes Klicpera, Stephan Günnemann

Motivation



Multiple nodes classified in a single graph



Multiple objects detected in a single image [1]

Adversarial robustness certification so far focused on single-prediction tasks

[1] A Convolutional Neural Network based Live Object Recognition System as Blind Aid. Kedar Potdar et. al. arXiv:1811.10399

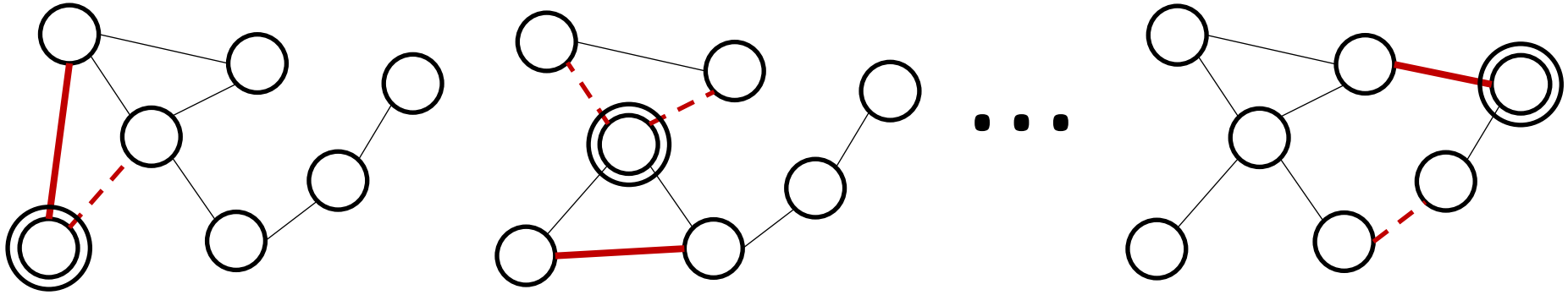
Research Question

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

- for classification models
- based on Graph Neural Networks

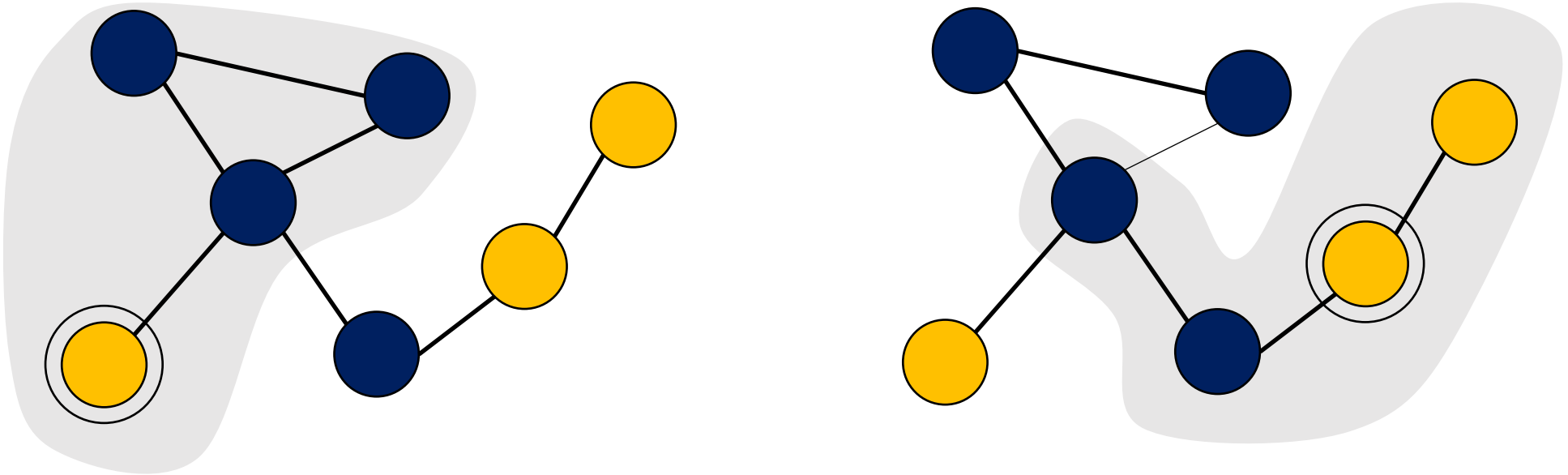
A naïve certificate

- Certify each prediction independently
- Assumes a different attack on each prediction



Our certificate

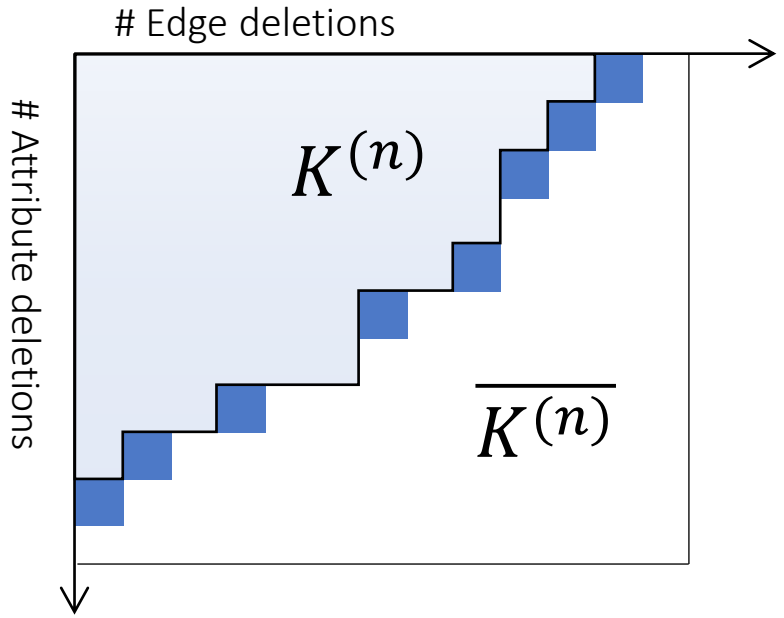
Ingredient 1: Locality



→ Not all perturbations affect all predictions

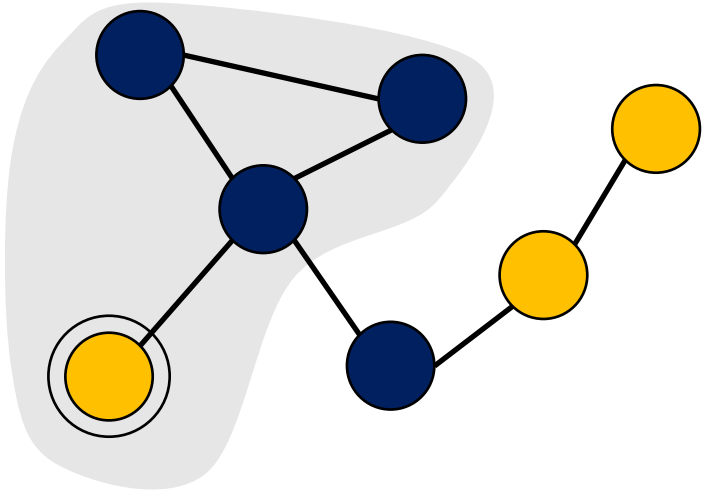
Ingredient 2: Linear certificate encoding

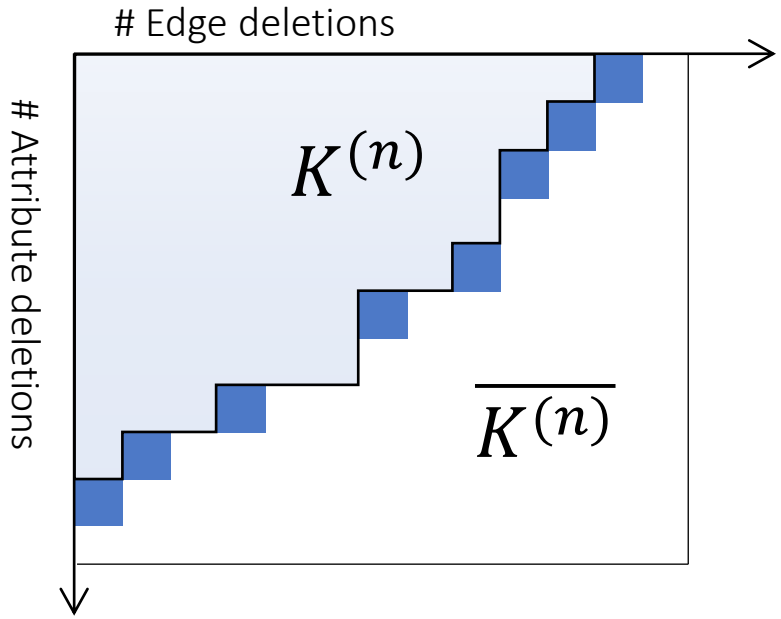
Evaluate single-prediction certificates via linear constraints …



… by encoding their pareto front

Combining Ingredients 1 & 2





Given a single perturbed graph:

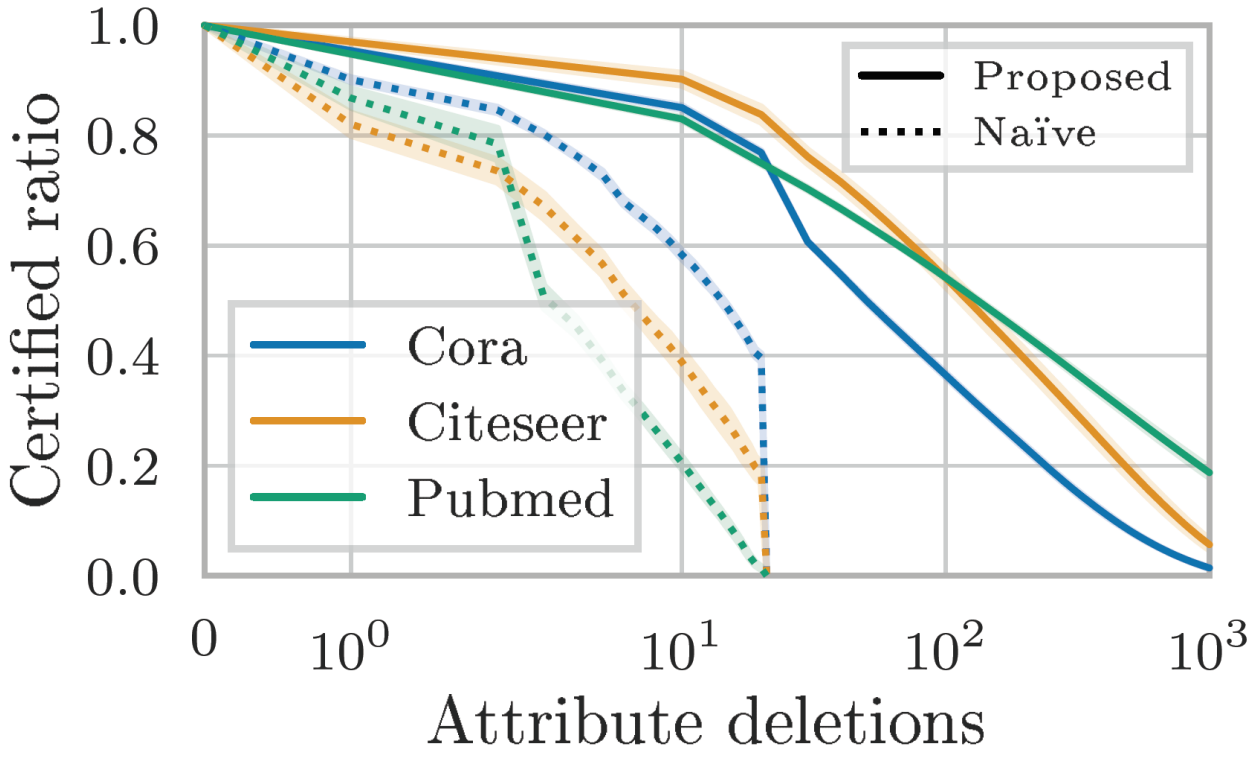
1. Aggregate local perturbation $\boldsymbol{b}^{(n)}$ in each receptive field
2. Evaluate single-prediction certificates based on $\boldsymbol{b}^{(n)}$

$\min_{G \in B}$

Results

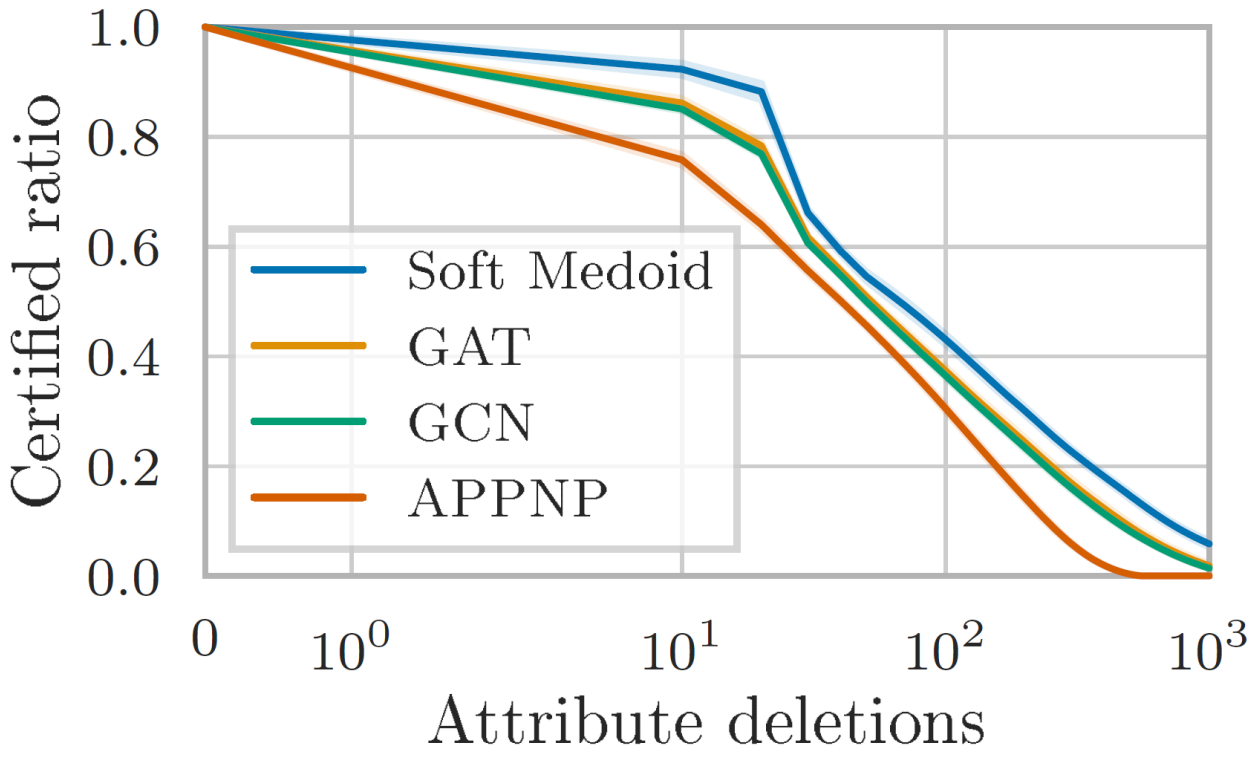
Our certificate …

- is orders of magnitudes stronger,



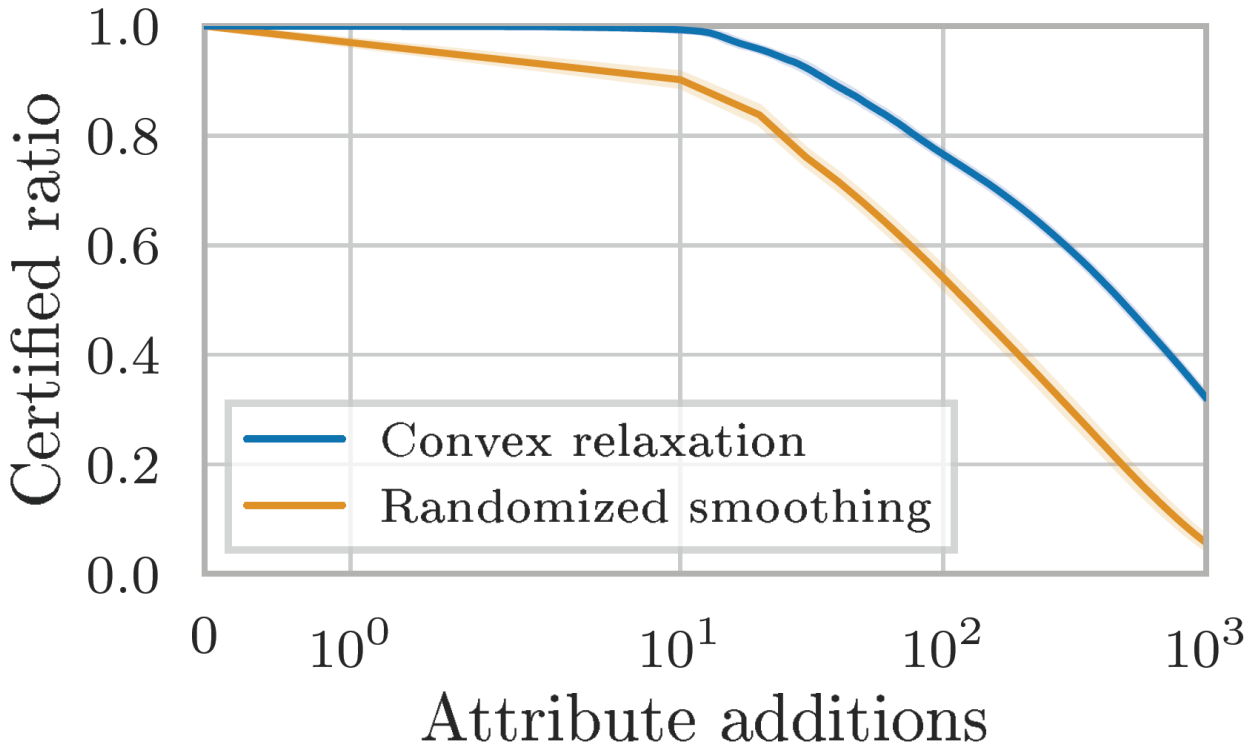
Our certificate …

- is orders of magnitudes stronger,
- model-agnostic,



Our certificate …

- is orders of magnitudes stronger,
- model-agnostic,
- compatible with any „base certificate“.



Our certificate …

- combines single-prediction certificates more intelligently
- by modelling **locality** and a **shared input**.

Our certificate …

- combines single-prediction certificates more intelligently
- by modelling **locality** and a **shared input**.

Poster Session 10

PDT: May 6, 2021, 1 a.m.

UTC: May 6, 2021, 8 a.m.