Collective Robustness Certificates
Exploiting Interdependence in Graph Neural Networks

Jan Schuchardt, Aleksandar Bojchevski, Johannes Klicpera, Stephan Günnemann
Motivation

Adversarial robustness certification so far focused on single-prediction tasks

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

Attack 1

Attack 2

Attack $N$
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 $b^{(n)}$ in each receptive field
2. Evaluate single-prediction certificates based on $b^{(n)}$
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.