Randomized Message-Interception Smoothing
Gray-box Certificates for Graph Neural Networks

Yan Scholten[1], Jan Schuchardt[1], Simon Geisler[1], Aleksandar Bojchevski[2], Stephan Günnemann[1]

[1]Technical University of Munich [2]CISPA Helmholtz Center for Information Security

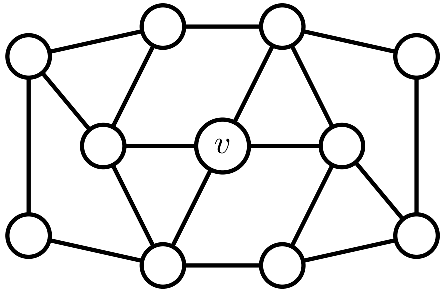
Motivation: Gray-box Certificates for GNNs

Context

- GNNs are susceptible to adversarial examples
- Certificates provide provable robustness guarantees

Problem

- White-box certificates: Only certify specific models
- Black-box certificates: Ignore properties of the classifier



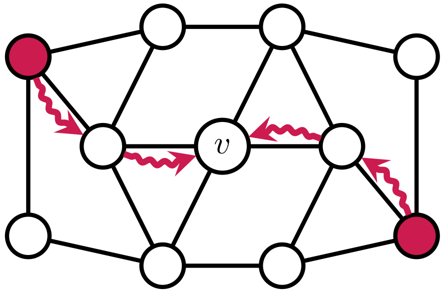
Solution

- Gray-box certificates: Exploit message-passing principles
- Robustness certificates against much stronger adversaries

Threat Model

Adversaries control multiple nodes & manipulate features

● Adversarial node ⬅ Adversarial message

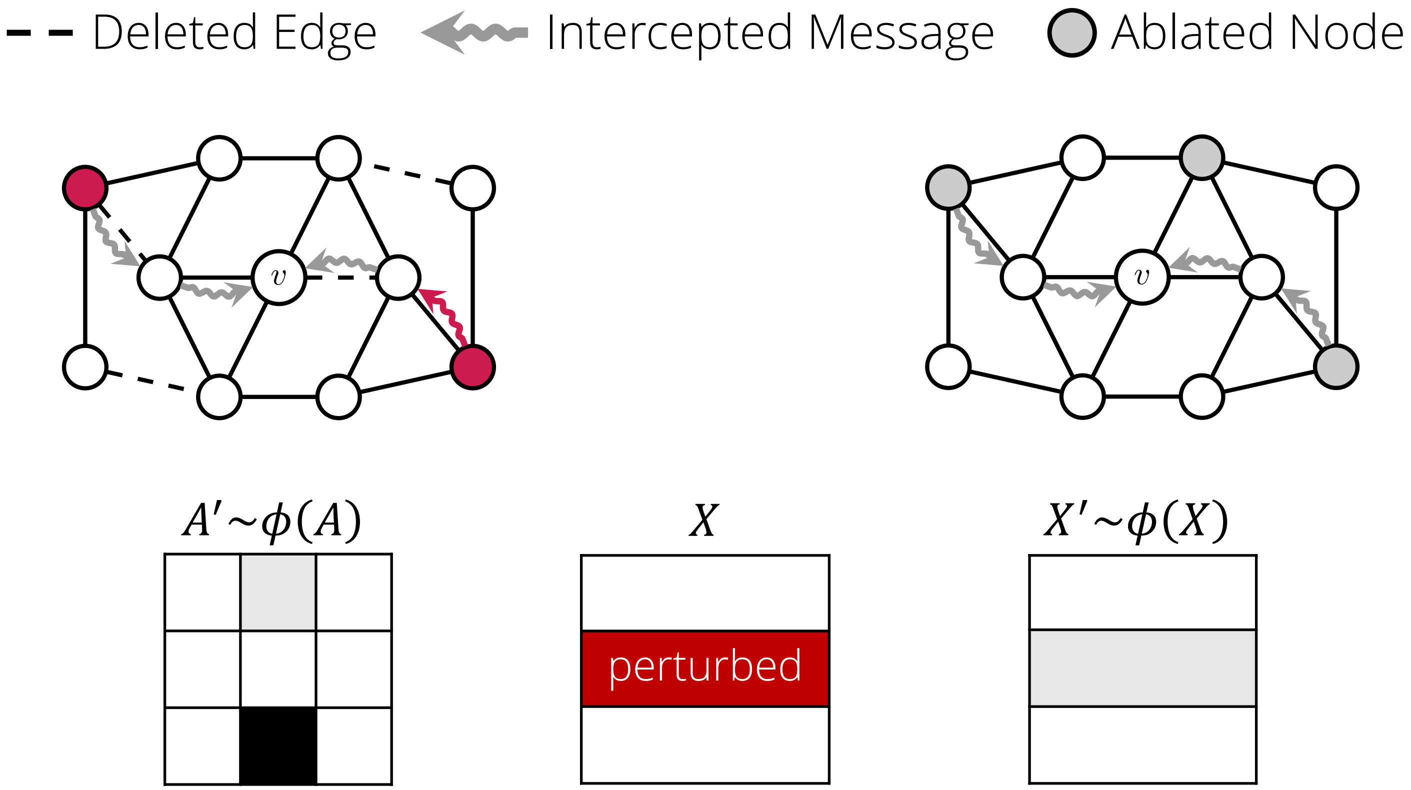


Class A → Class B

How can we limit the propagation of adversarial messages?

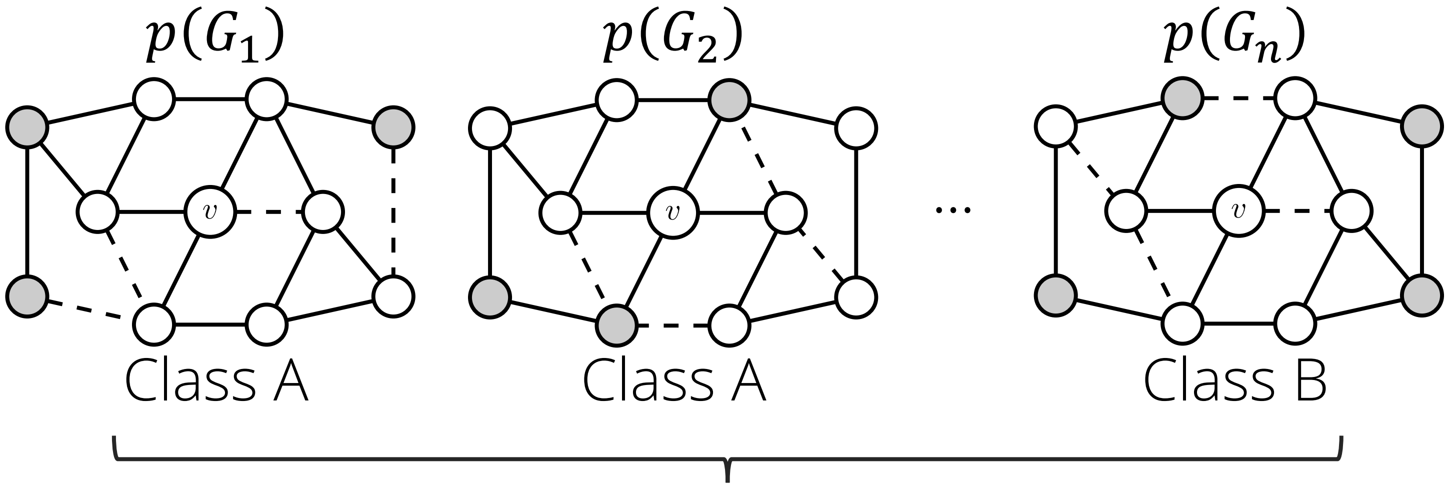
Gray-box Certificates for Graph Neural Networks

Exploit message-passing principles: Intercept messages

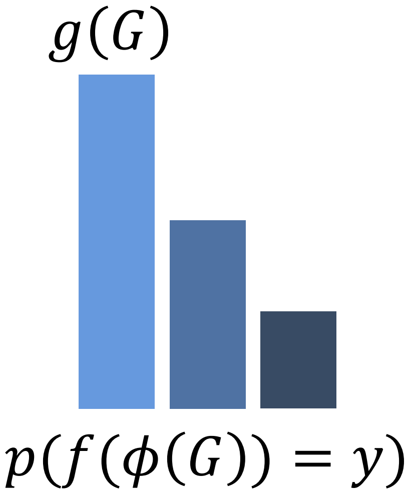


Randomized Message-Interception Smoothing

Majority vote under randomized message-interception



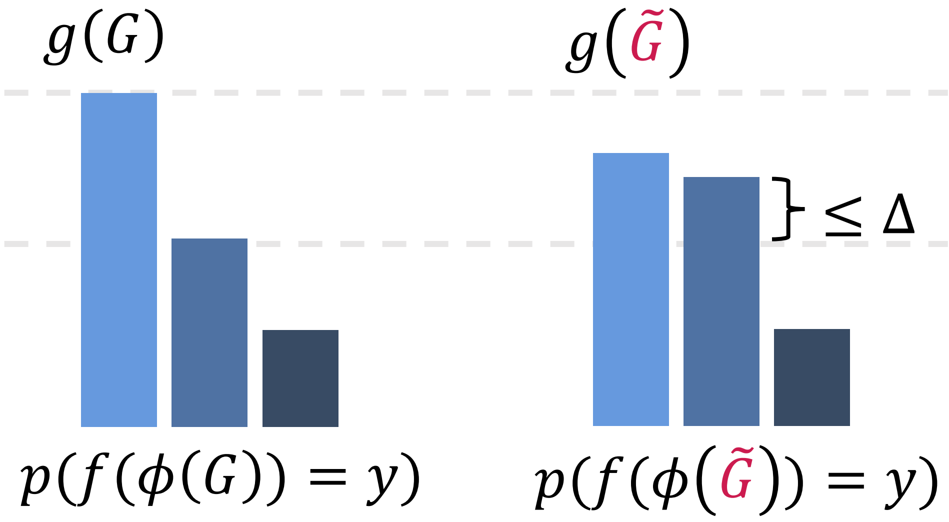
majority vote: $g(G)$ = class A



Interception Smoothing Certificates

Provable robustness certificates: Worst-case assumption

- One adversarial message is enough to change the prediction



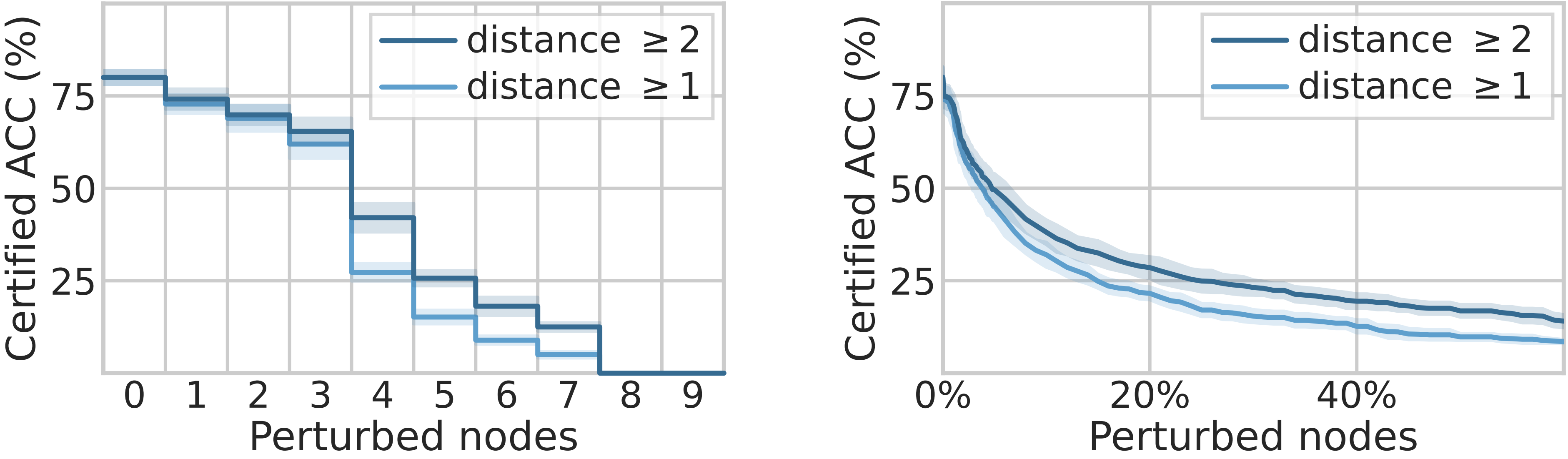
If adversary does not control enough probability mass

$\Rightarrow$ $g(G) = g(\tilde{G})$ for any graph $\tilde{G} \in \mathcal{B}_r(G)$

Certificates against Strong Adversaries

Robustness of Smoothed GAT on Cora-ML

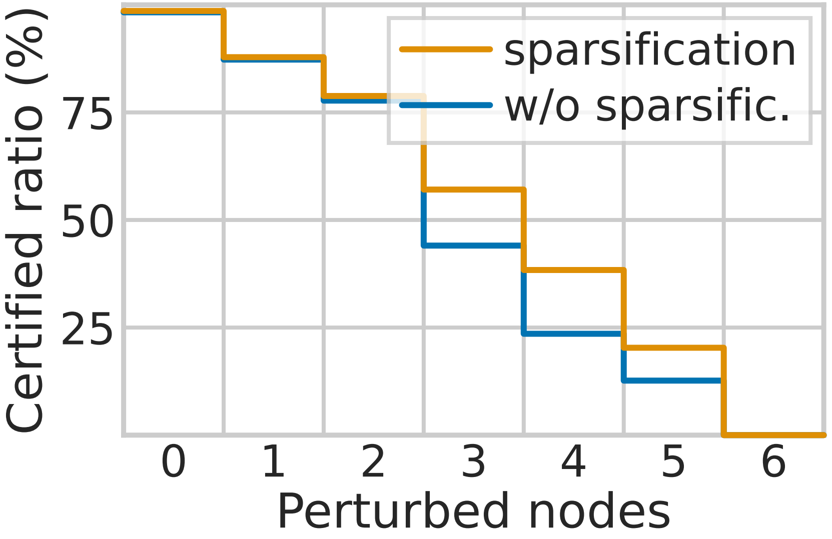
- Stronger certificates against more distant nodes

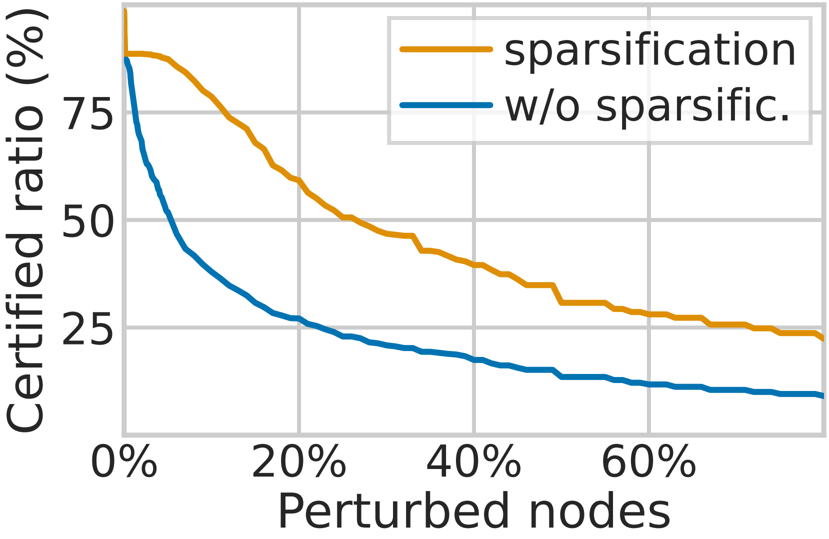


Stronger Certificates for Sparser Graphs

Sparsification

- Reduces messages to intercept
- Reduces nodes that send messages

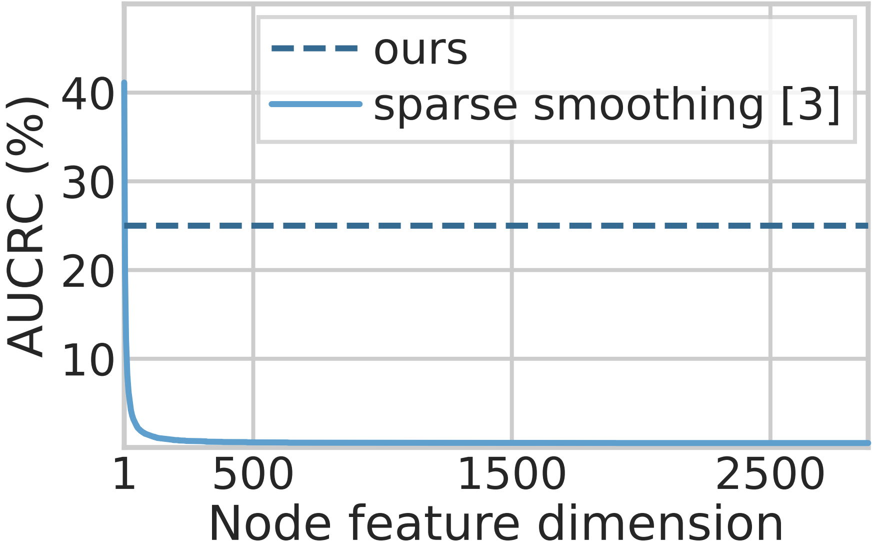




First Certificate against Stronger Adversaries

We certify robustness against features perturbations of arbitrary magnitude

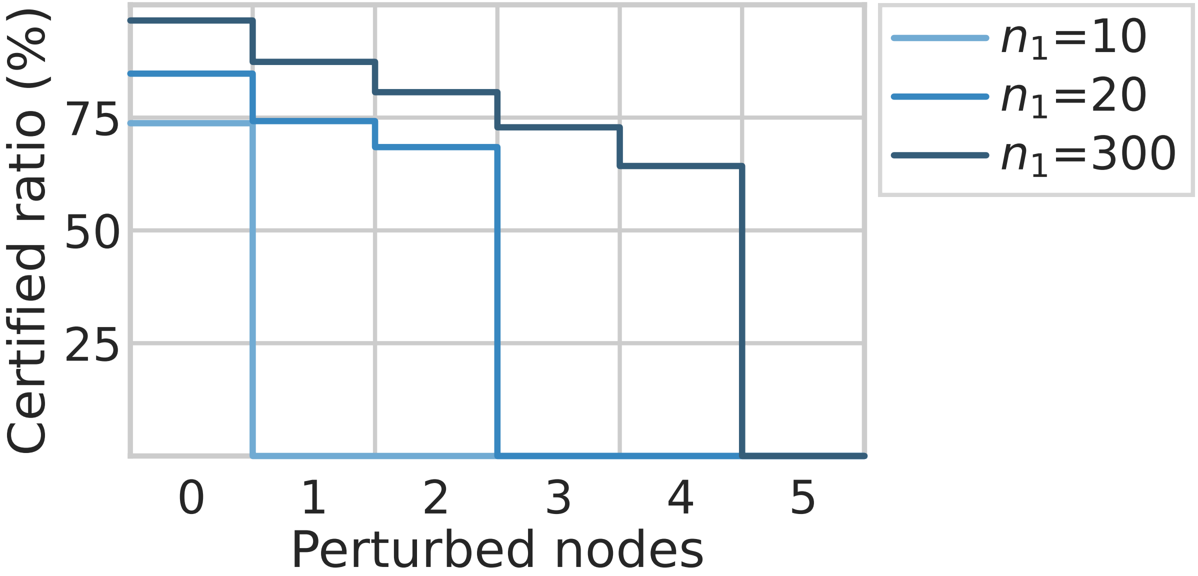
- Existing certificates certify only a few attributes in the graph



[3] Aleksandar Bojchevski, Johannes Gasteiger, and Stephan Günnemann. Efficient Robustness Certificates for Discrete Data: Sparsity-Aware Randomized Smoothing for Graphs, Images and More. ICML 2020.

Efficient Message-Interception Smoothing

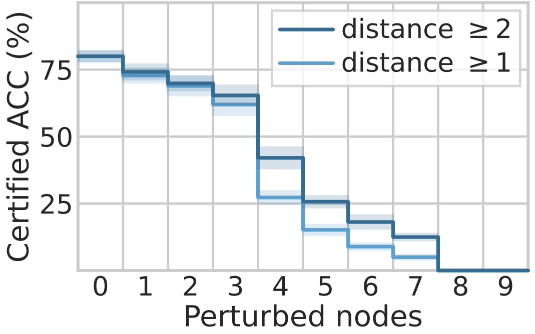
Certificates on Cora-ML: 17 seconds



tl;dr Gray-box Robustness Certificates for GNNs

Interception Smoothing: Gray-box Certificates for GNNs

- Exploit underlying message-passing principles of GNNs
- Certify robustness against strong adversaries
- Model-agnostic & efficient



www.cs.cit.tum.de/daml/interception-smoothing