Data Analytics and Machine Learning Group
Department of Computer Science
Technical University of Munich

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

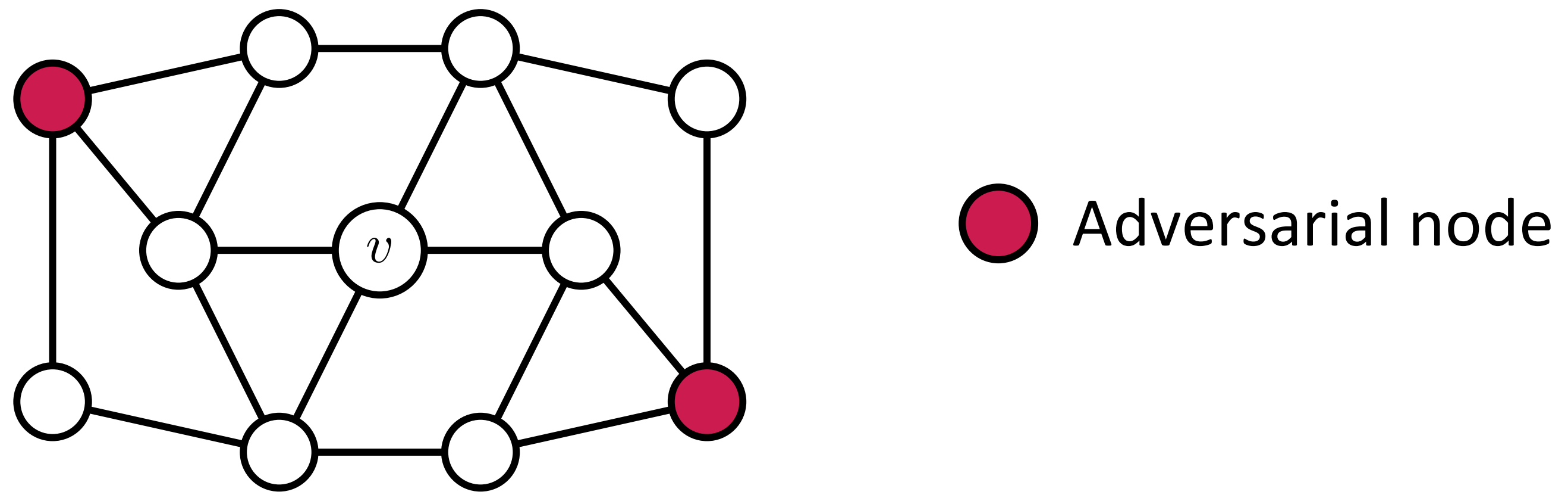
tl;dr: Gray-box Robustness Certificates for GNNs

- Exploit underlying message-passing principles
- Adversaries control multiple nodes in the graph and perturb node features arbitrarily
- Model-agnostic & efficient

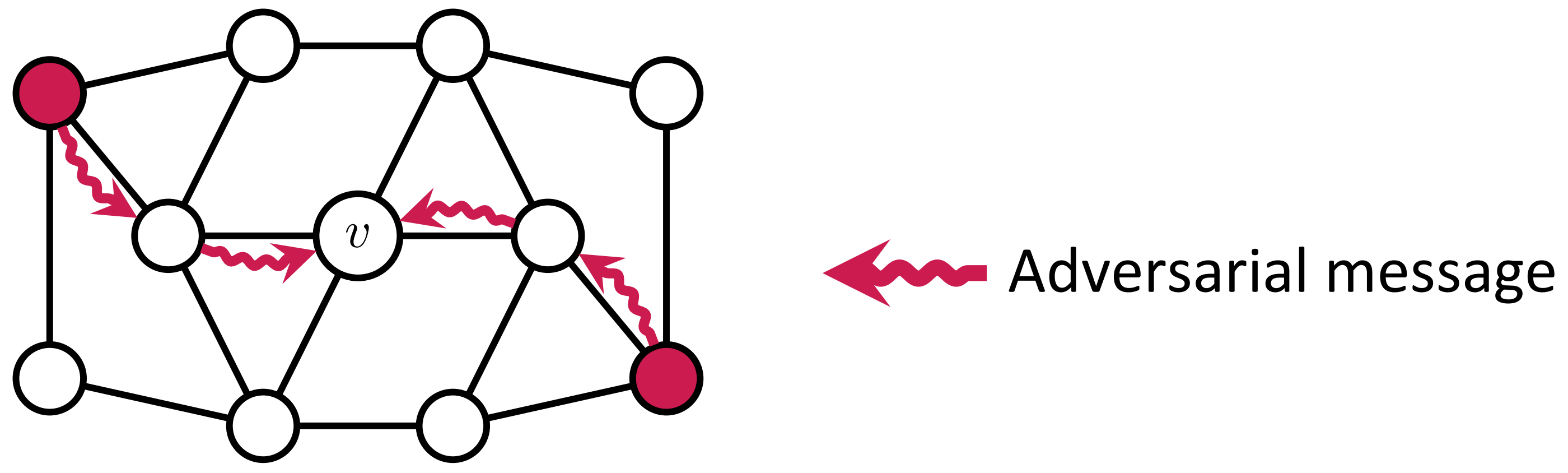
Motivation

GNNs are susceptible to adversarial examples

If adversaries control multiple nodes & perturb features…



…GNNs will propagate adversarial information through the graph…



…allowing adversaries to alter the prediction for target nodes v:

Class A ⇒ Class B

Robustness certificates: Provable guarantees for stable predictions

Exiting robustness certificates are inadequate

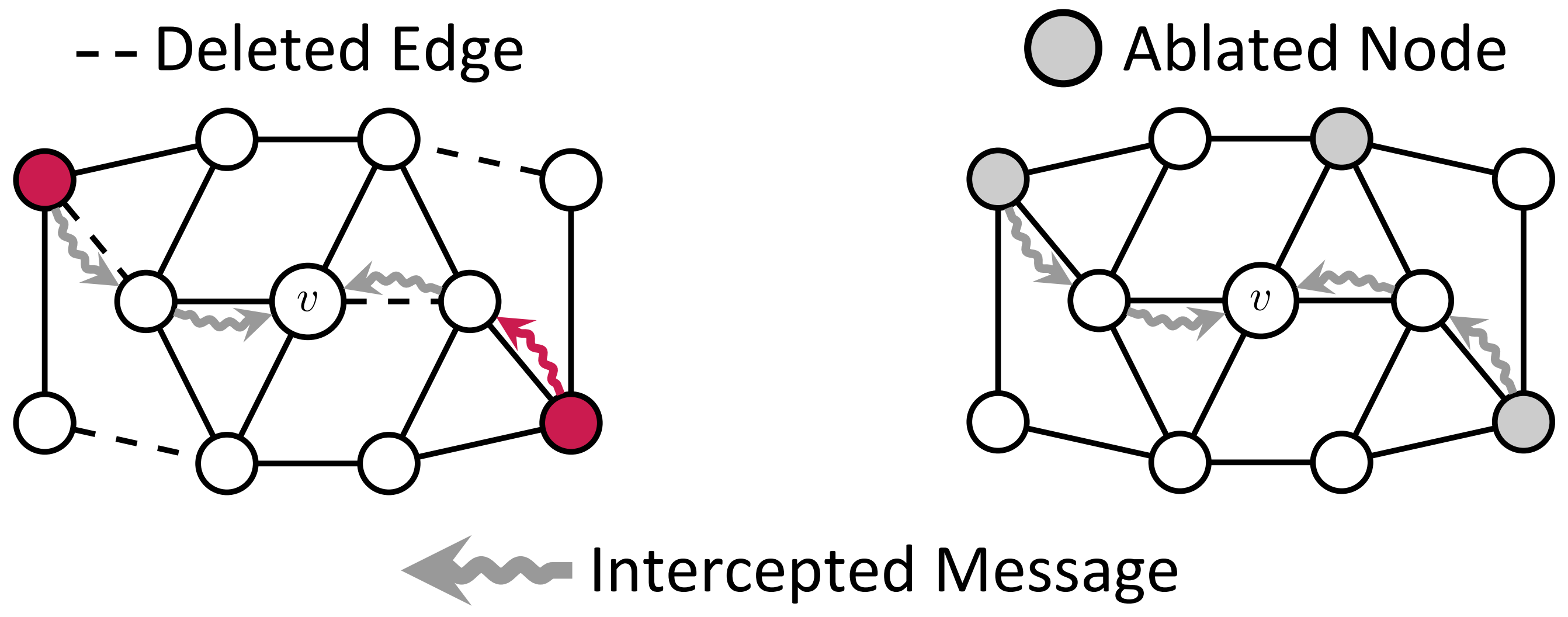
- White-box certificates only certify specific models
- Black-box certificates ignore properties of the classifier

We enhance model-agnostic black-box certificates by exploiting message-passing principles

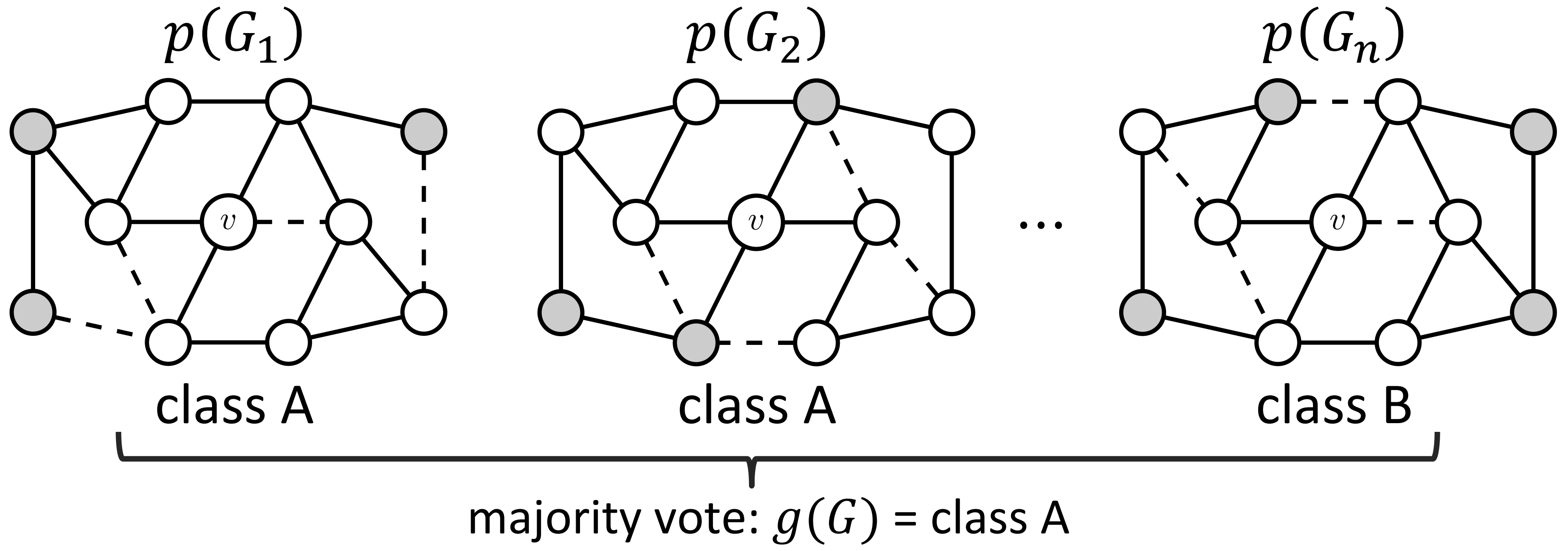
Interception Smoothing

Exploit message-passing principles & intercept messages

Intercept messages using edge deletion and node feature ablation

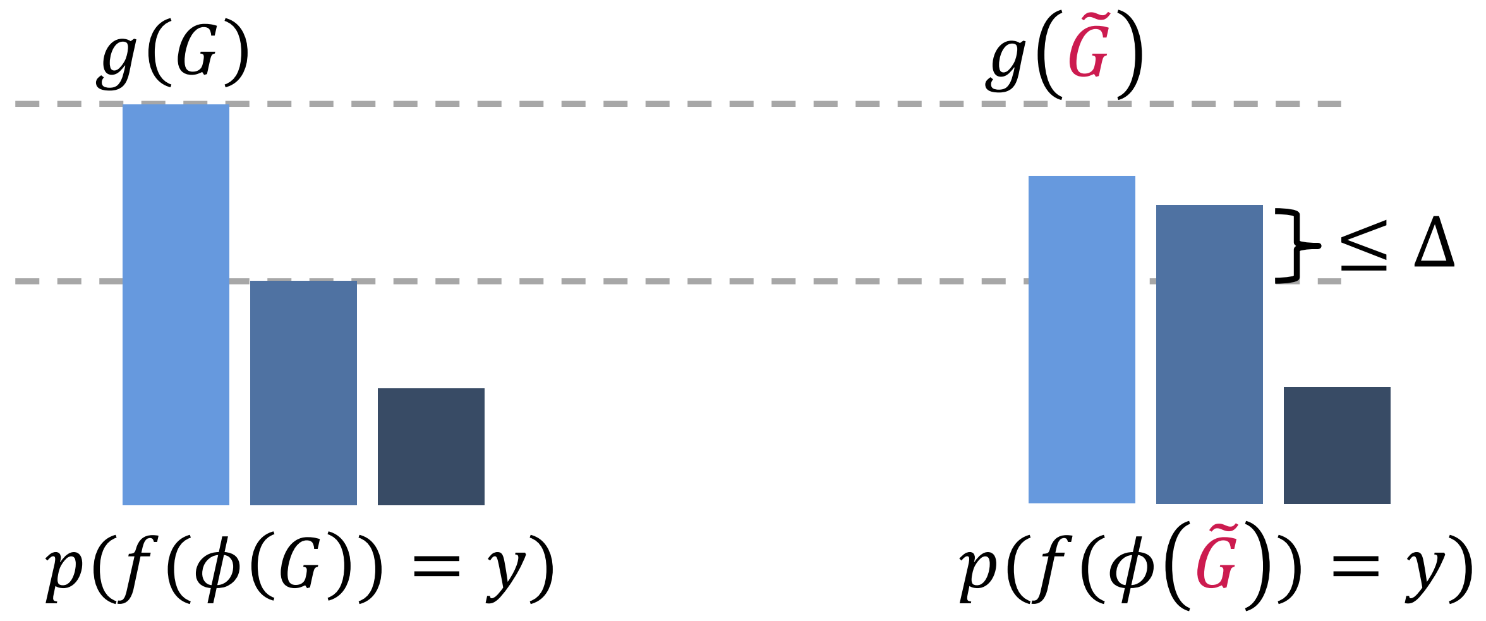


Constructing a smoothed classifier that intercepts messages



Provable robustness certificates for interception smoothing

Δ bounds probability to receive adversarial messages



If adversary does not control enough probability mass to change majority vote

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

Practical challenge: How to compute Δ for arbitrary graphs?

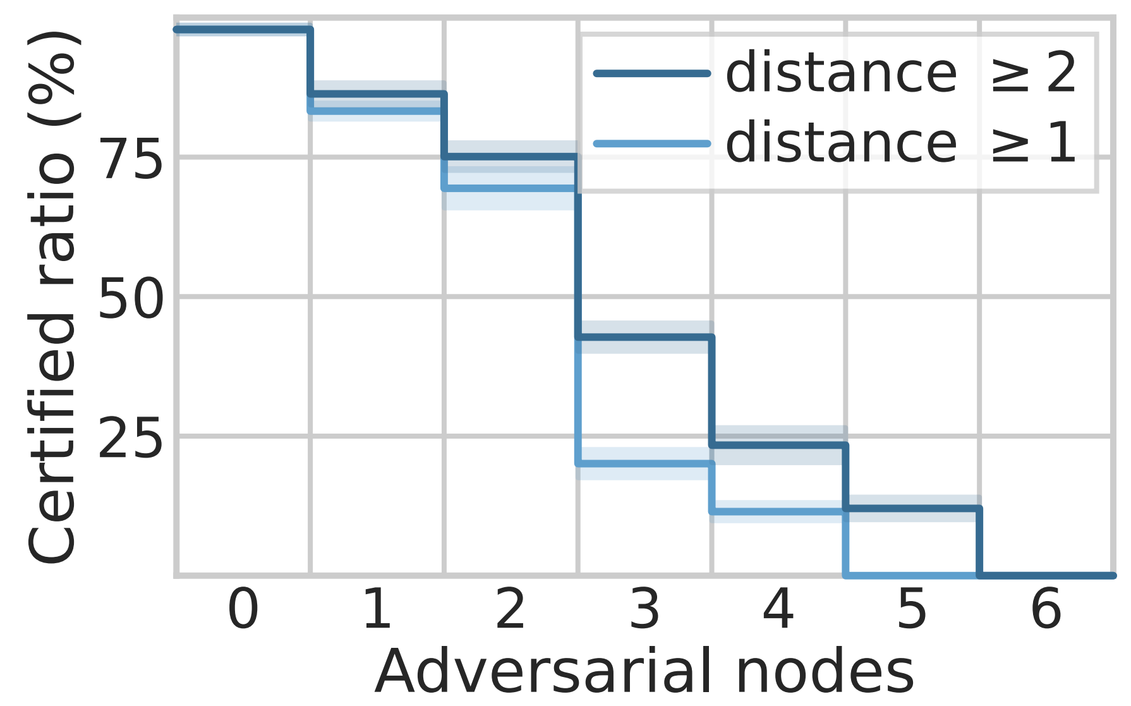
$$\Delta = \max_{|W|=r} p_\phi(\text{v receives any message from nodes in W})$$

⇒ Lower bound on certifiable robustness by relaxing to independent paths

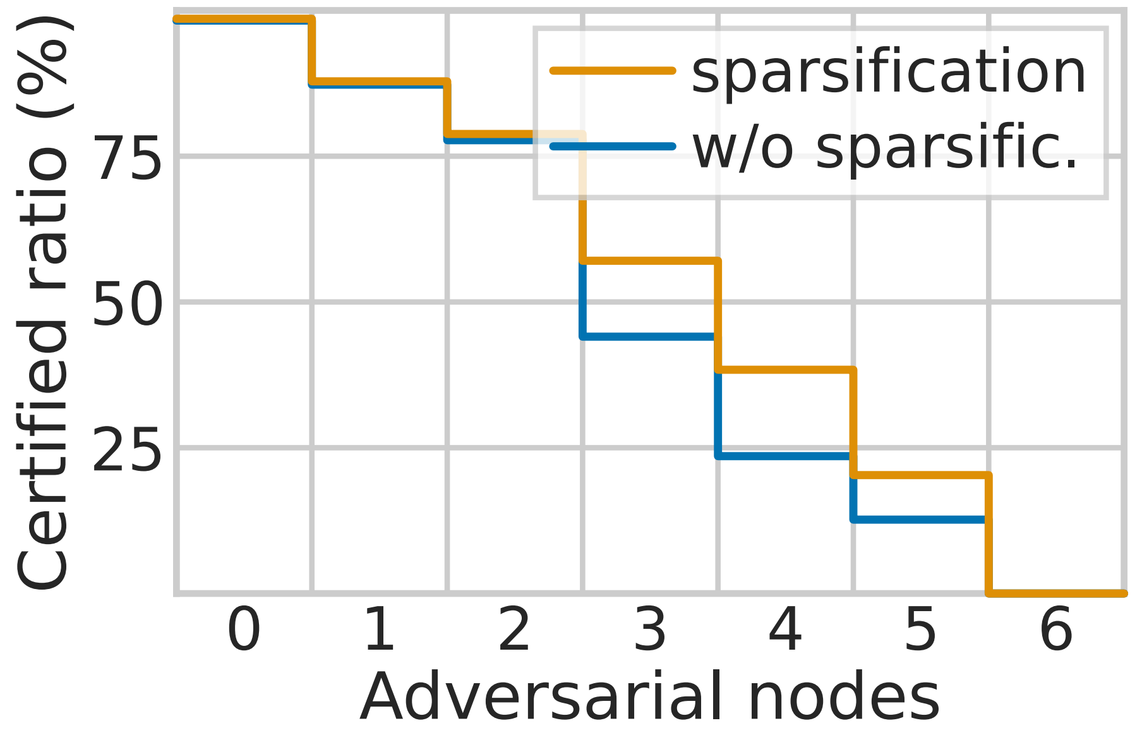
Experimental Evaluation

Robustness certificates against strong adversaries

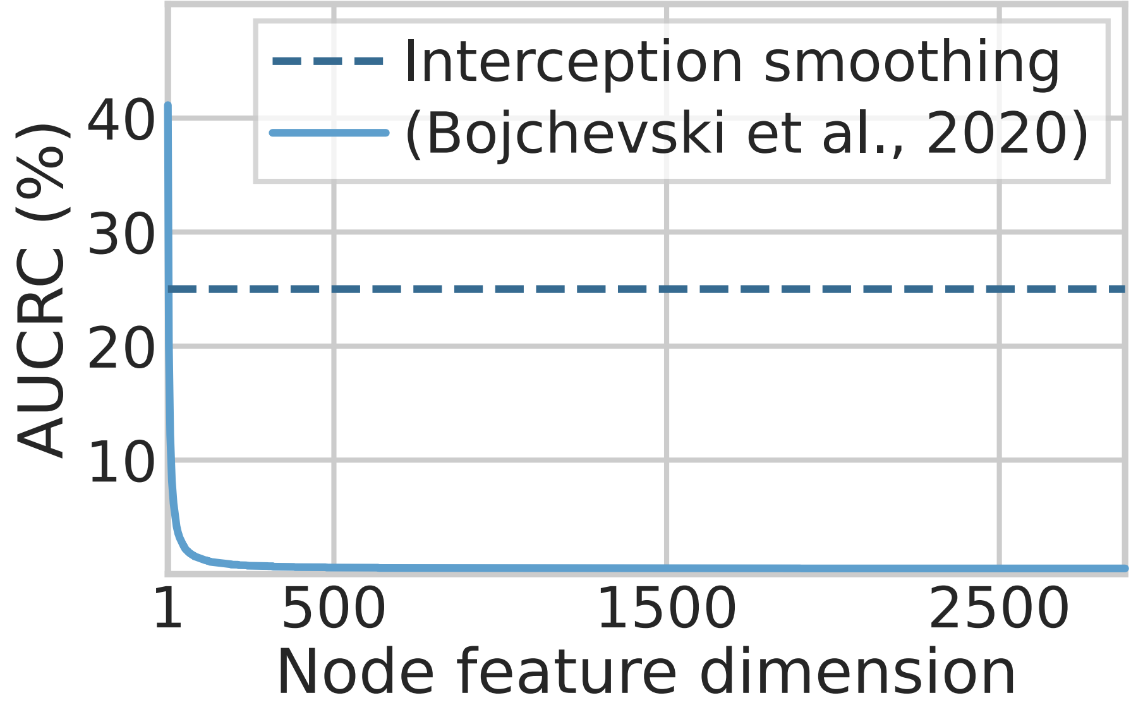
Adversaries control multiple nodes & perturb features arbitrarily



Stronger certificates for sparser graphs



Certificates independent of node feature dimensionality



Efficient certificates: 100x faster than previous methods