Efficient Robustness Certificates for Discrete Data

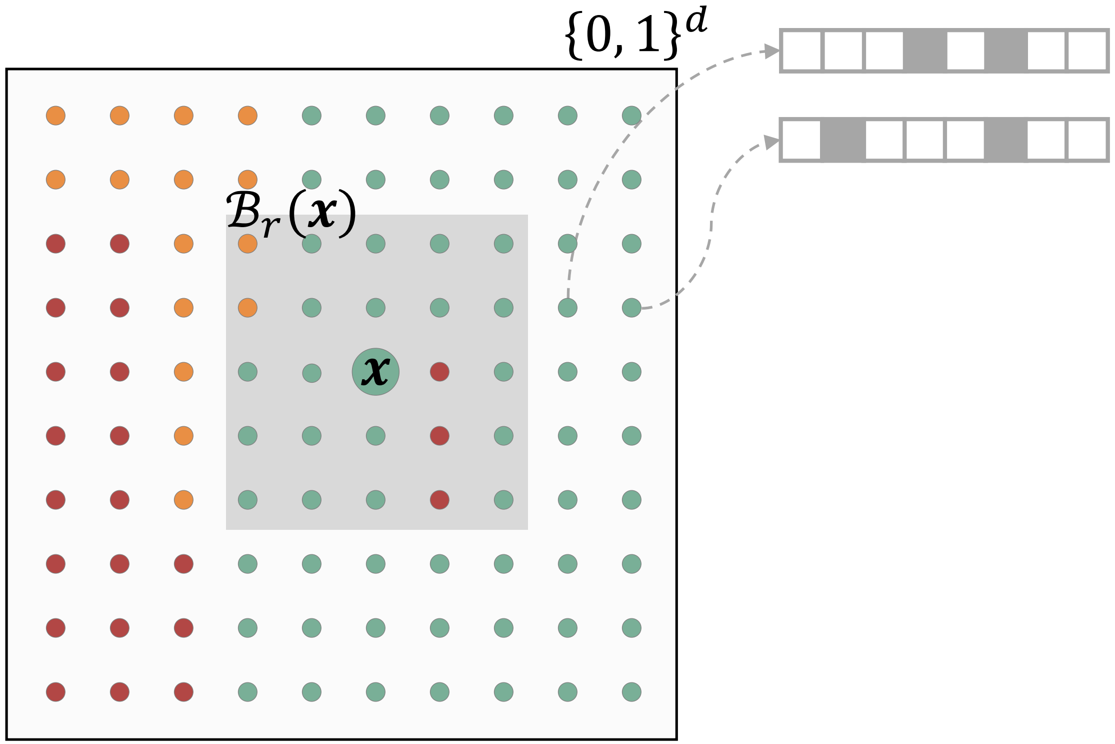
Sparsity-Aware Randomized Smoothing for Graphs, Images and More

Aleksandar Bojchevski, Johannes Klicpera, Stephan Günnemann

tl;dr Robustness Certificate

Guarantee that the prediction does not change for all $\tilde{\boldsymbol{x}}$ in a ball $\mathcal{B}_r(\boldsymbol{x})$ around the input $\boldsymbol{x}$

Here $\mathcal{B}_r(\boldsymbol{x})$ is the L_0 ball: the attacker can change up to r bits



tl;dr Robustness Certificate

Node-level Classification
Graph-level Classification

Graph Neural Network

Given any base classifier for discrete data

ResNet
Transformer
DNN
…

Discretized Images
Text
Molecules (SMILES)
…

Certify a smoothed classifier w.r.t. an L_0 adversary

tl;dr Tight, Efficient, & Sparsity-Aware

Sparsity-aware smoothing improves guarantees

Reduced complexity: $O(d^3)$ to $O(\mathrm{r})$

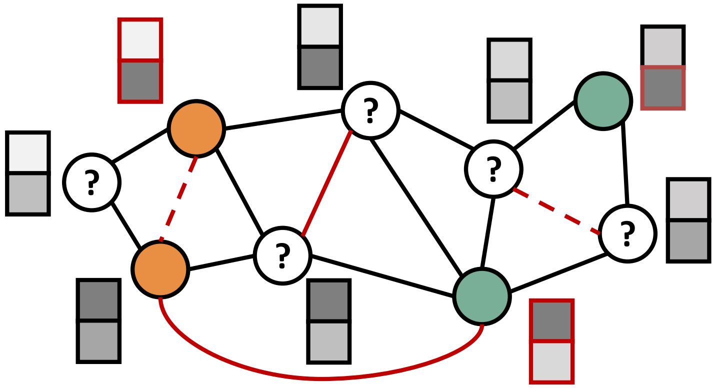
Results on Graphs, MNIST, ImageNet, …

tl;dr Certifying Graph Neural Networks

Any GNN: GCN, GAT, PPNP, GIN, …

Perturbing both graph and node attributes

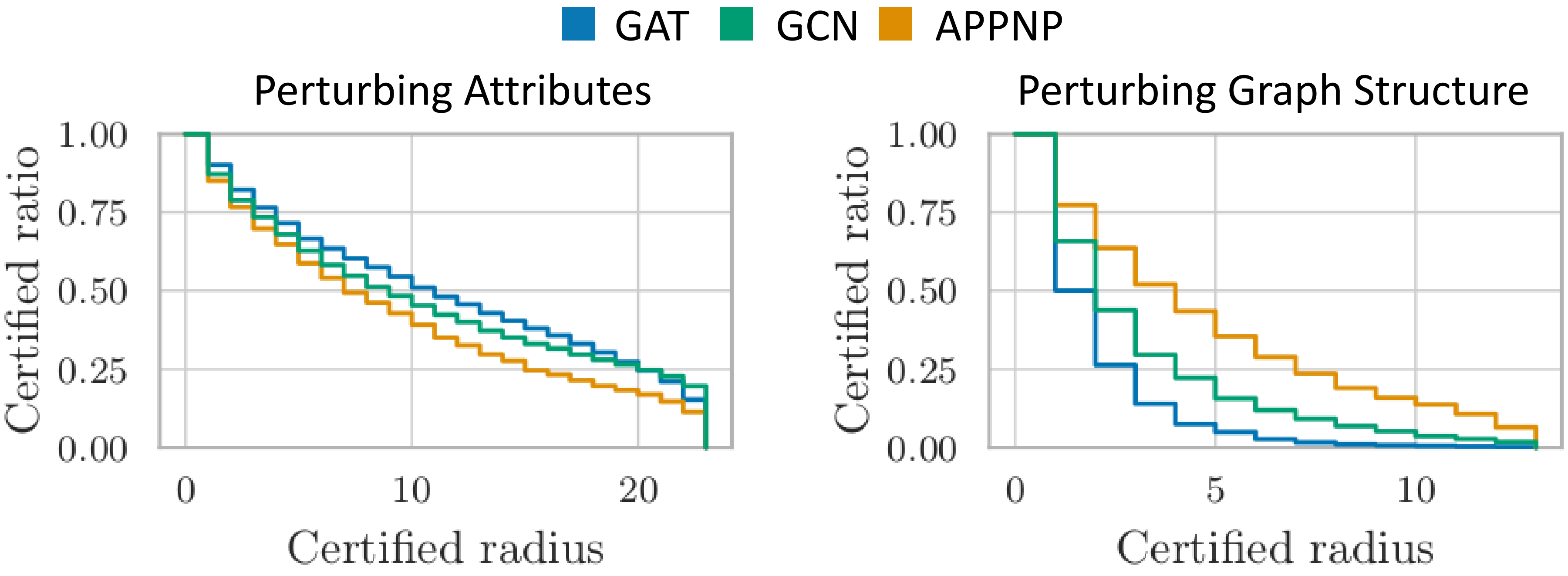
First certificate for graph-level classification



Perturbations:
- inserted edge
- deleted edge
- perturbed attribute

tl;dr Certifying Graph Neural Networks

Different GNNs have different robustness trade-offs



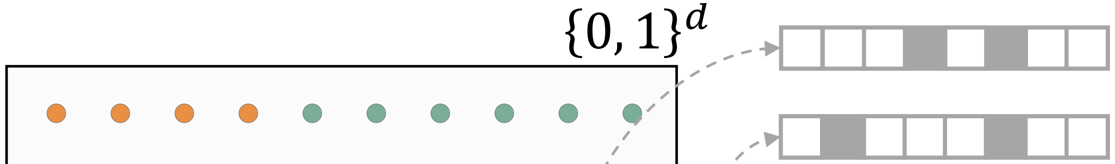
Randomly Smoothed Classifiers

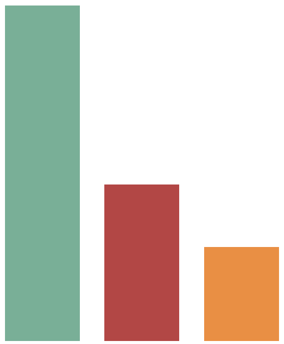
Given:

- Any base classifier $f\colon \mathcal{X} \to \mathcal{Y}$
- Any randomization scheme $\phi(\boldsymbol{x})$

Certify a smoothed classifier g

$$g(\boldsymbol{x}) = \underbrace{\underset{y \in \mathcal{Y}}{\operatorname{argmax}} \Pr\big(f\big(\phi(\boldsymbol{x})\big) = y\big)}_{\text{majority vote } y^*}$$

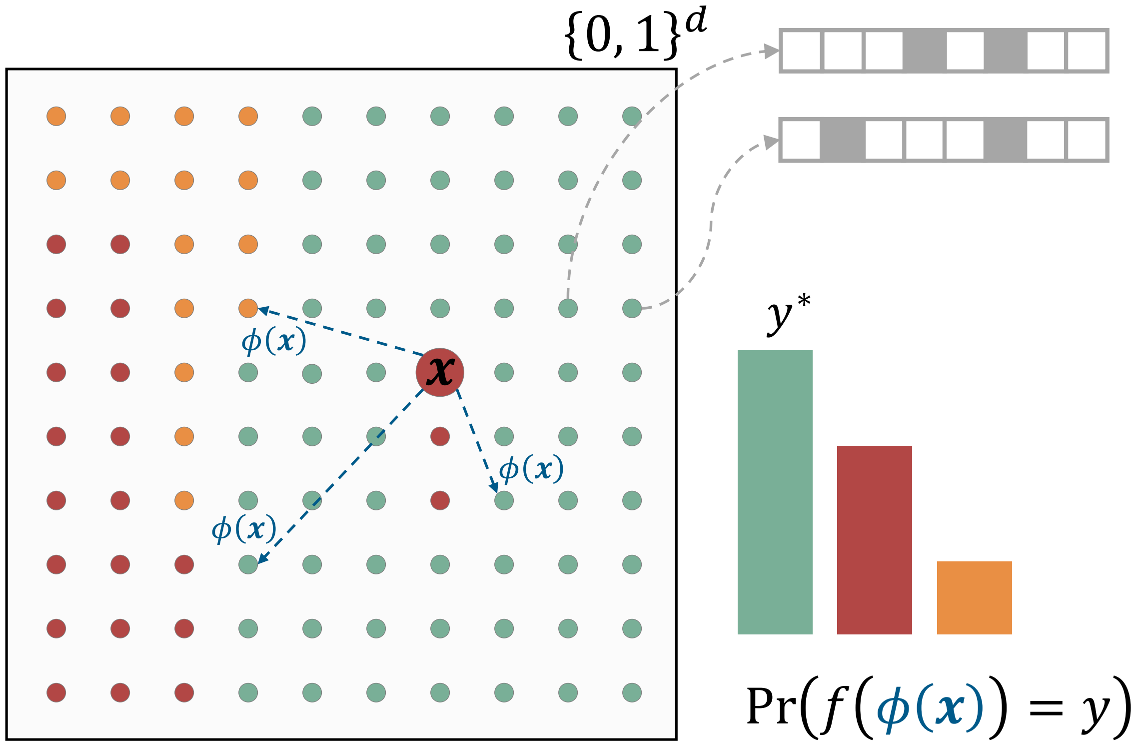




Certifying the Smoothed Classifiers

Majority vote $g(\boldsymbol{x})$ changes slowly

Example: $f(\boldsymbol{x}) =$ ●, but $\mathrm{g}(\boldsymbol{x}) =$ ●

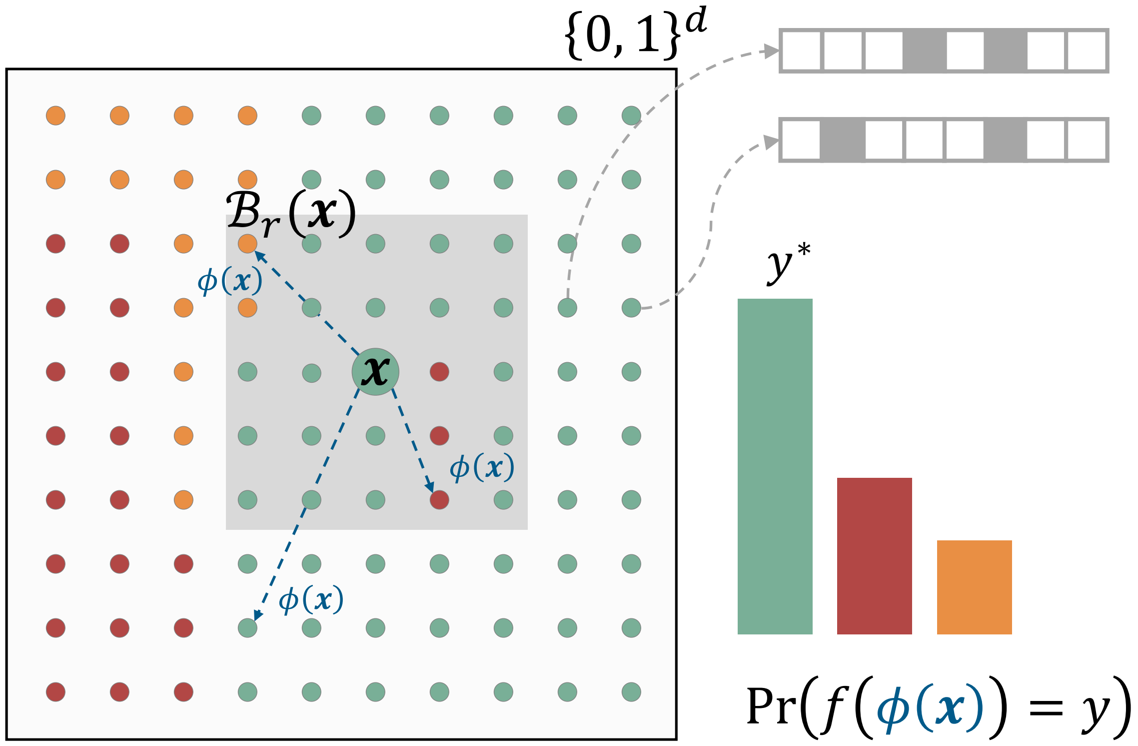


Randomly Smoothed Classifiers

Goal:

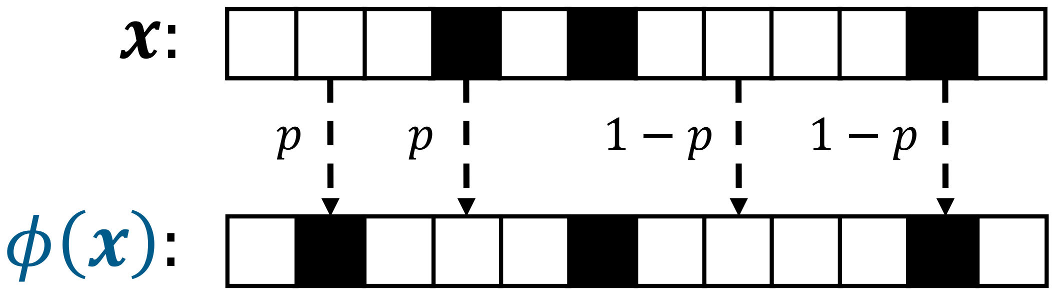
Guarantee that the majority votes does not change for all $\tilde{\boldsymbol{x}}$ in a ball $\mathcal{B}_r(\boldsymbol{x})$ around the input $\boldsymbol{x}$

For all $\tilde{\boldsymbol{x}}$, $\Pr\left(f\left(\phi(\tilde{\boldsymbol{x}})\right) = \bullet\right) \overset{?}{>} 0.5$



Choosing the Randomization Scheme $\phi(\boldsymbol{x})$

First idea: Randomly flip bits with probability p

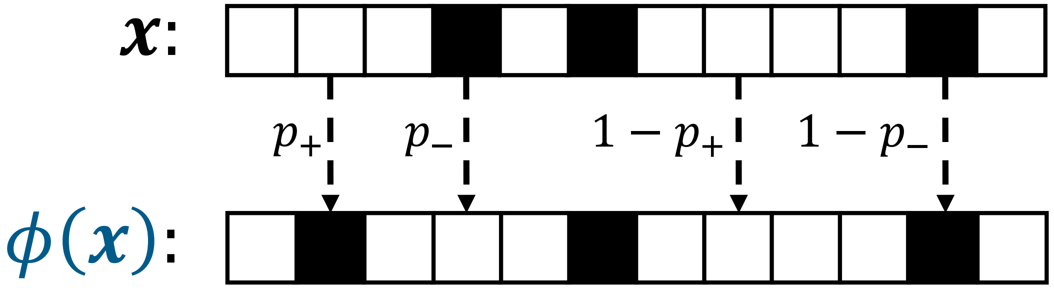


Higher p leads to better guarantees

Problem: For sparse data even moderately small p destroys the data

Choosing the Randomization Scheme $\phi(\boldsymbol{x})$

Sparsity aware: Treat zeros separately



Graphs: Insert edges with p_+, delete edges with p_-

We can afford to set p_- relatively high and p_+ relatively low without introducing too much noise in the data

Deriving the Certificate

The smoothed classifier is certifiably robust if

$$\min \Pr(f(\phi(\tilde{\boldsymbol{x}})) = y^*) \overset{?}{>} 0.5$$

subject to:

$$\tilde{\boldsymbol{x}} \in \mathcal{B}_r(\boldsymbol{x})$$

Find the $\tilde{\boldsymbol{x}}$ that minimizes the probability of the majority vote y^*

Constant Likelihood Ratio Regions

The smoothed classifier is certifiably robust if

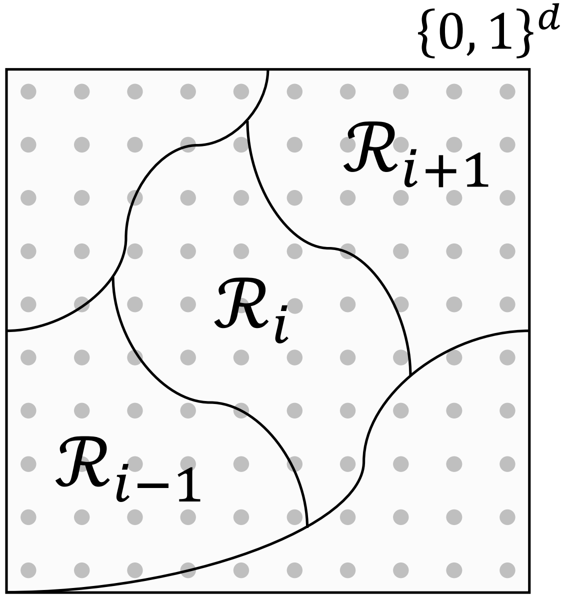
$$\min \textstyle\sum_i \Pr(\phi(\tilde{\boldsymbol{x}}) \in \mathcal{R}_i)\, h_i \overset{?}{>} 0.5$$

subject to:

$$\tilde{\boldsymbol{x}} \in \mathcal{B}_r(\boldsymbol{x})$$

$$h_i \in [0, 1]$$

$$\textstyle\sum_i \Pr(\phi(\boldsymbol{x}) \in \mathcal{R}_i)\, h_i = \underline{p_{y^*}}$$



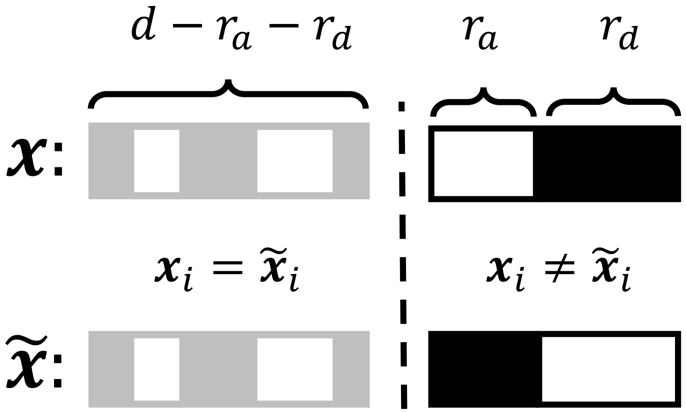
$$\frac{\Pr(\phi(\boldsymbol{x}) \in \mathcal{R}_i)}{\Pr(\phi(\tilde{\boldsymbol{x}}) \in \mathcal{R}_i)} = c_i$$

constant

Constant Likelihood Ratio Regions

Observation 1: We consider w.l.o.g. only dimensions where $\boldsymbol{x}_i \neq \tilde{\boldsymbol{x}}_i$

Observation 2: Number of regions is independent of d



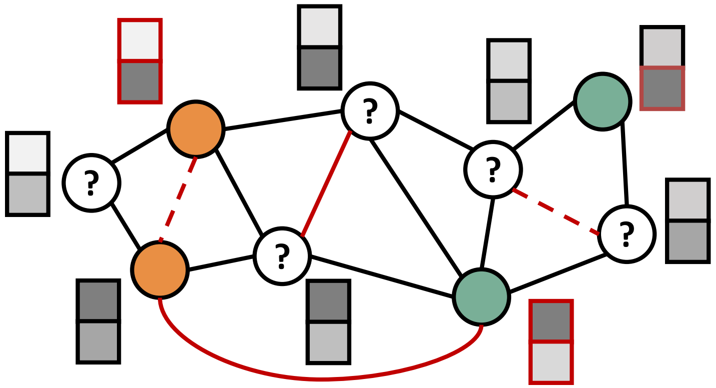
$\Pr(\phi(\boldsymbol{x}_i) = z_i) = \Pr(\phi(\tilde{\boldsymbol{x}}_i) = z_i)$
where $\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i$

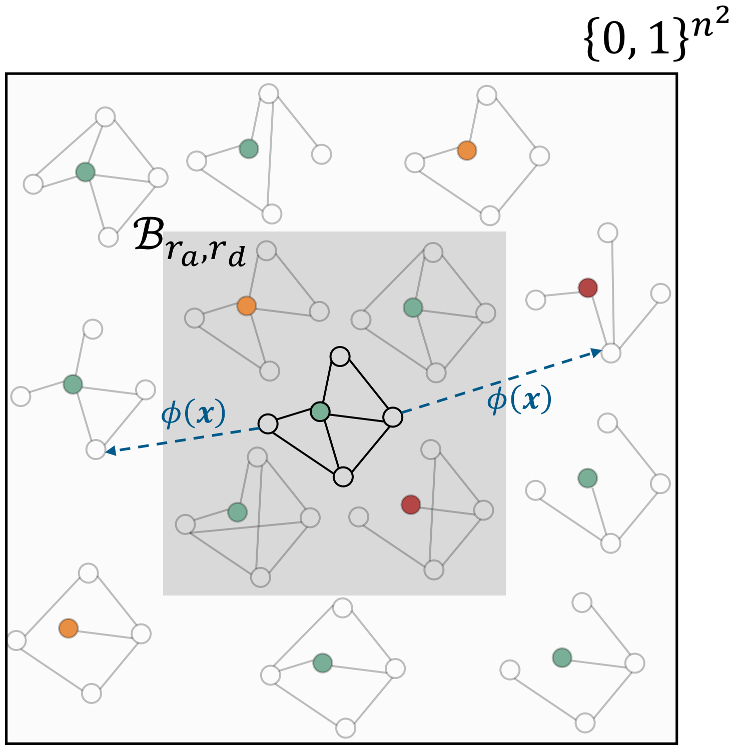
Threat model: $\mathcal{B}_{r_a, r_d} = \{\tilde{\boldsymbol{x}} : \text{added} \leq r_a \text{ bits, deleted} \leq r_d \text{ bits}\}$

GNNs: Setup

Threat model: Perturb either graph structure or attributes

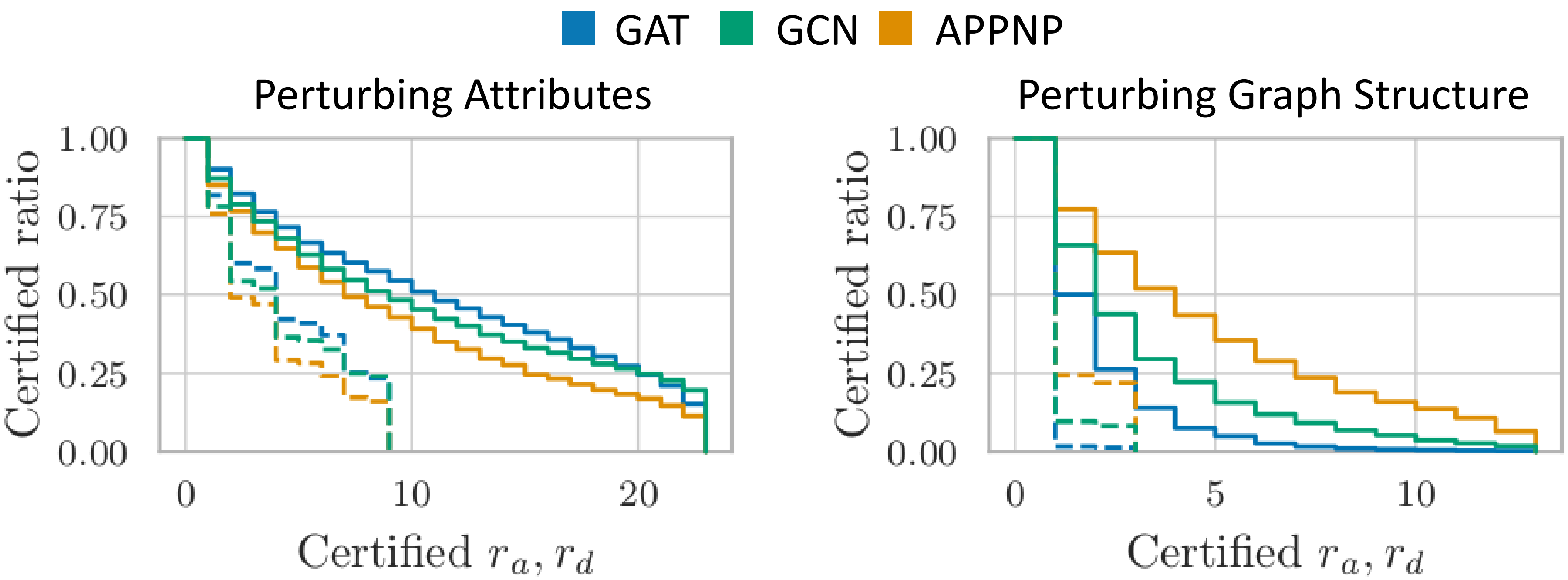
Task: Semi-supervised node classification





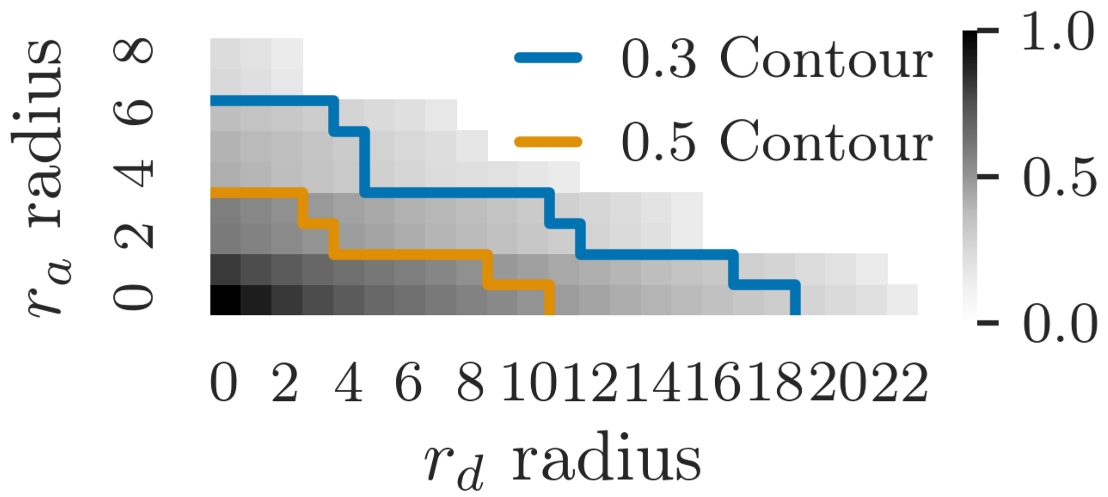
Results on Node Classification

GNNs are more robust to edge deletion than edge addition



Results on Node Classification

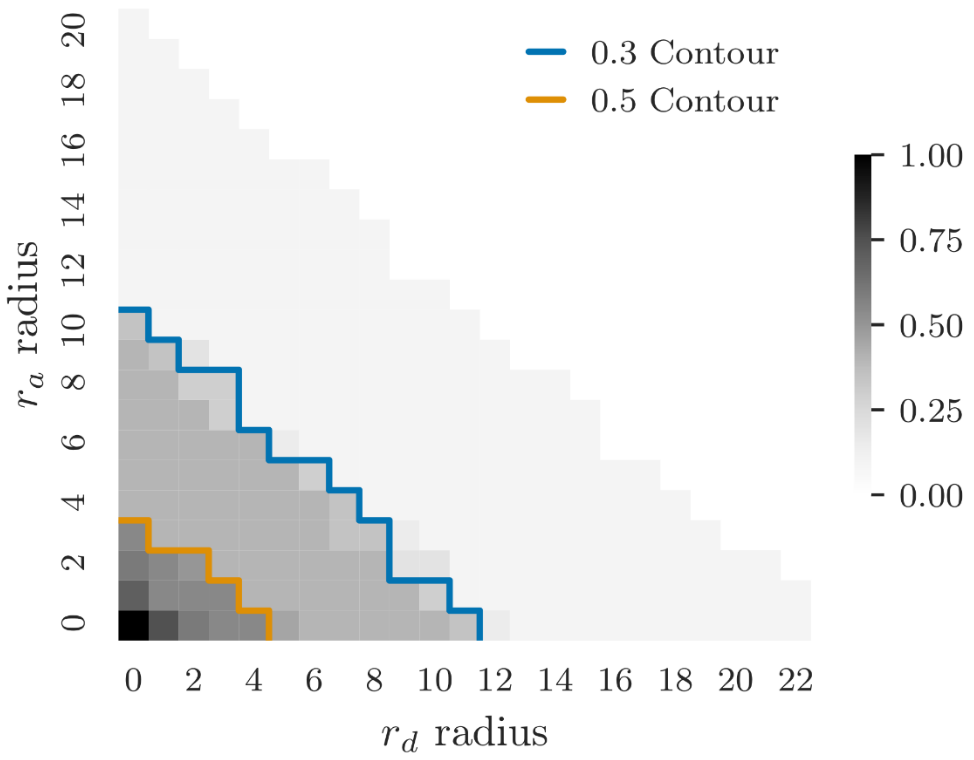
Models are more robust to edge deletion than edge addition



Average max r_d radius is 6.47 with sparse smoothing and 1.75 without

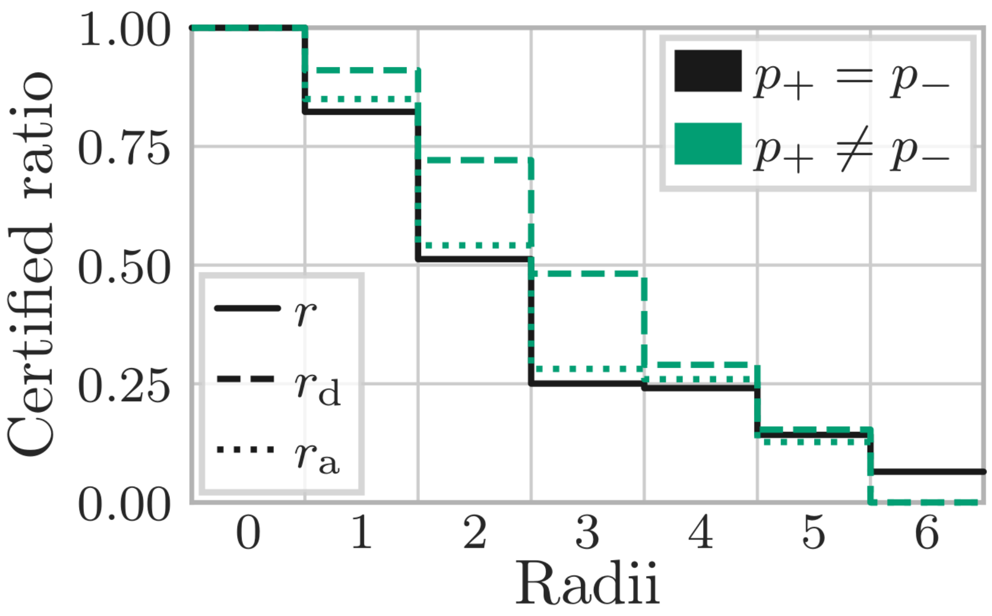
Results on Graph Classification

First certificate for the graph-level classification task



Results on MNIST

Sparsity-aware smoothing improves the certified ratio



Other results: ImageNet

Dramatically improved runtime for the exact same (tight) certificate

Certificate	**Type**	**Time**	$r = 1$	$r = 3$	$r = 5$	$r = 7$
Cohen et al. (2019)	Continuous	< 1 sec.	0.372	0.226	0.170	0.138
Dvijotham et al. (2020)	Discrete	< 1 sec.	0.362	0.224	0.136	0
Lee et al. (2019)	Discrete	4 days	**0.538**	**0.338**	**0.244**	**0.176**
Ours	Discrete	< 1 sec.	**0.538**	**0.338**	**0.244**	**0.176**

Model-agnostic, Tight, Efficient, & Sparsity-Aware Robustness Certificate

Code & Project Page: https://www.daml.in.tum.de/sparse_smoothing/

Twitter: @abojchevski