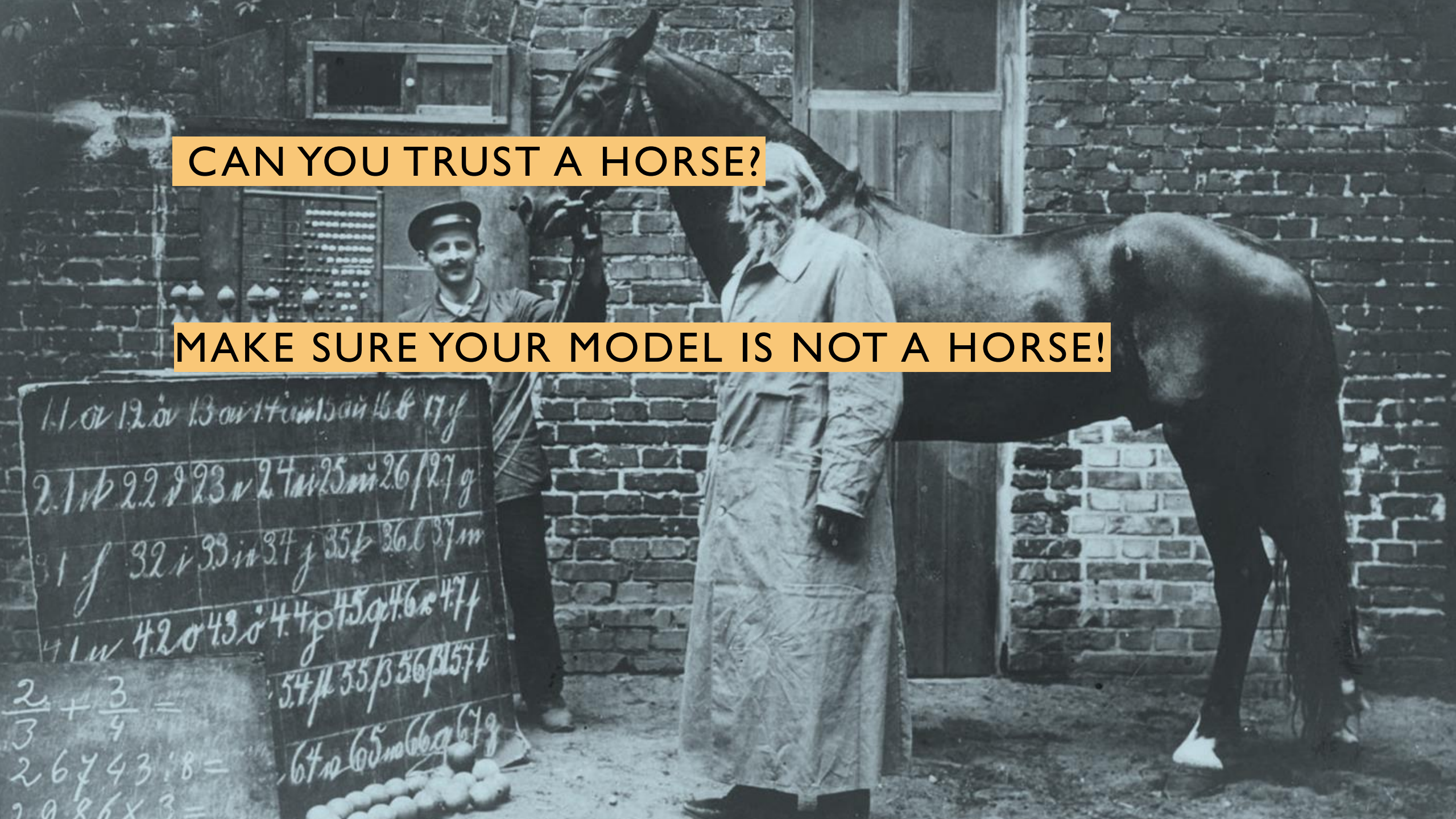
PROVABLY ROBUST MACHINE LEARNING ON GRAPHS

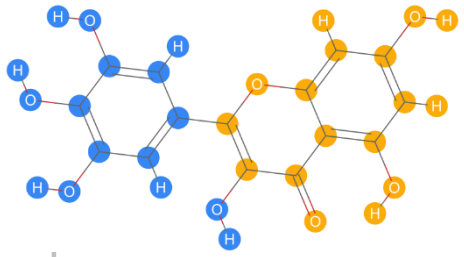
Aleksandar Bojchevski

22.03.2021

NEC Labs Europe

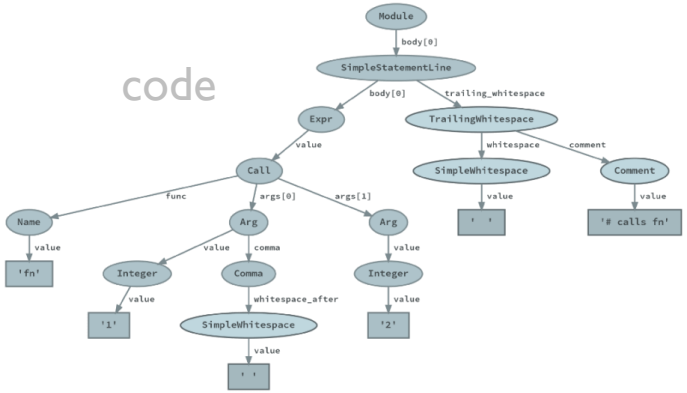

CAN YOU TRUST A HORSE?
MAKE SURE YOUR MODEL IS NOT A HORSE!

GRAPHS ARE EVERYWHERE

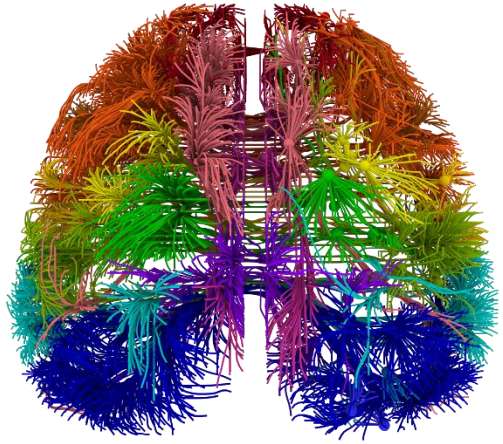


molecules

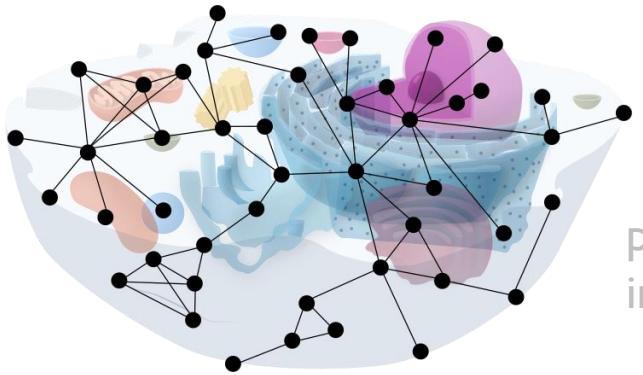
code



social networks



the brain



protein interactions

THE GRAPH ABSTRACTION

GRAPH-BASED MODELS ARE USED TO

Classify breast cancer

Detect fraud

Recommend items

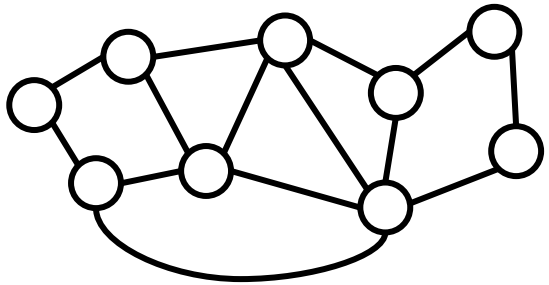
Classify proteins

Filter spam

Predict disease-gene associations

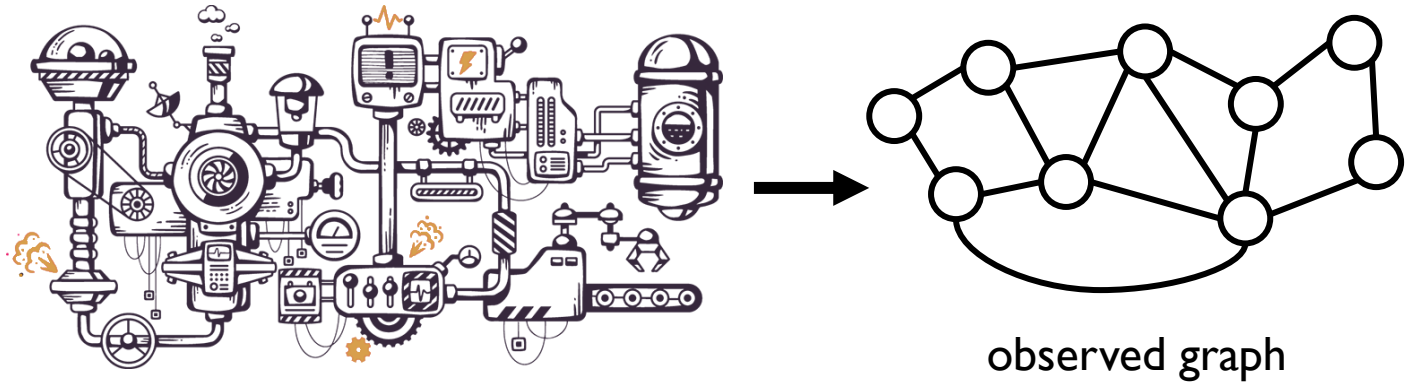
Induce object-oriented perception and reasoning

COMMON ASSUMPTIONS



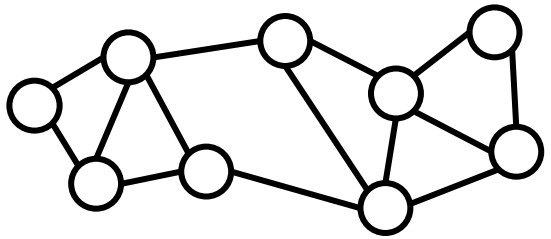
observed graph

NOISE

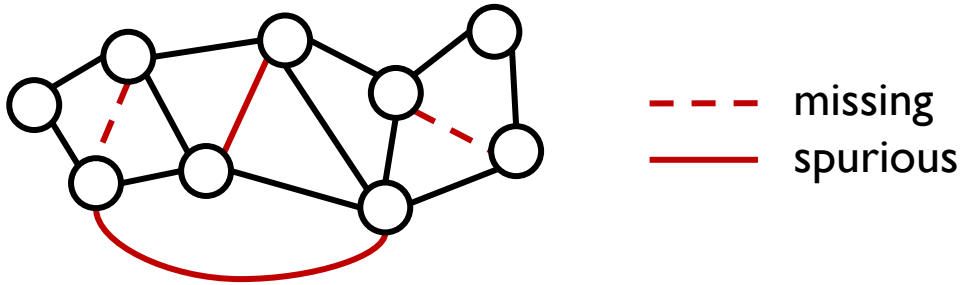


The graph is an output of a complex pipeline

NOISE AND ADVERSARIES

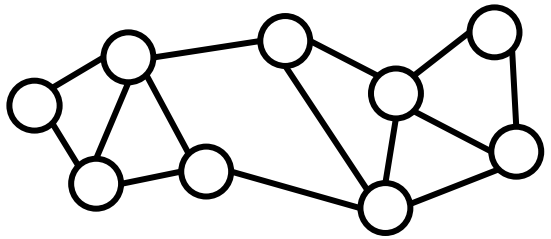


latent clean graph

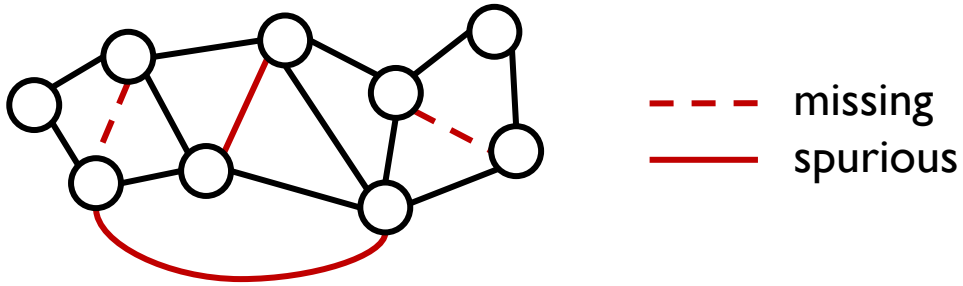


observed perturbed graph

NOISE AND ADVERSARIES



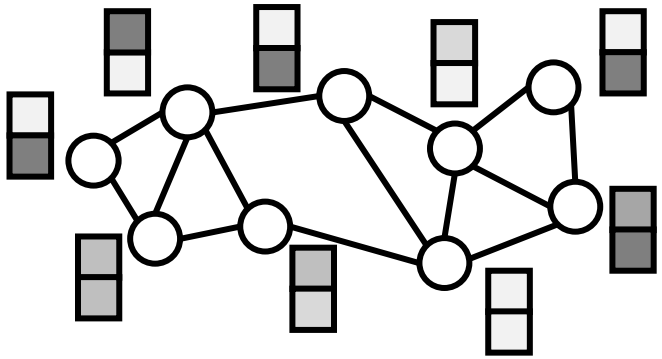
latent clean graph



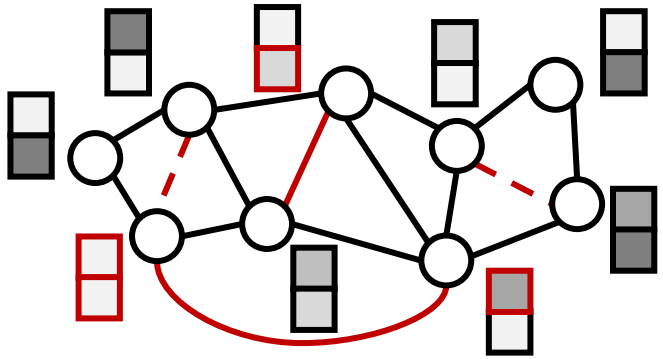
observed perturbed graph

Machine learning on graphs in real-world settings

FEATURES CAN BE PERTURBED TOO



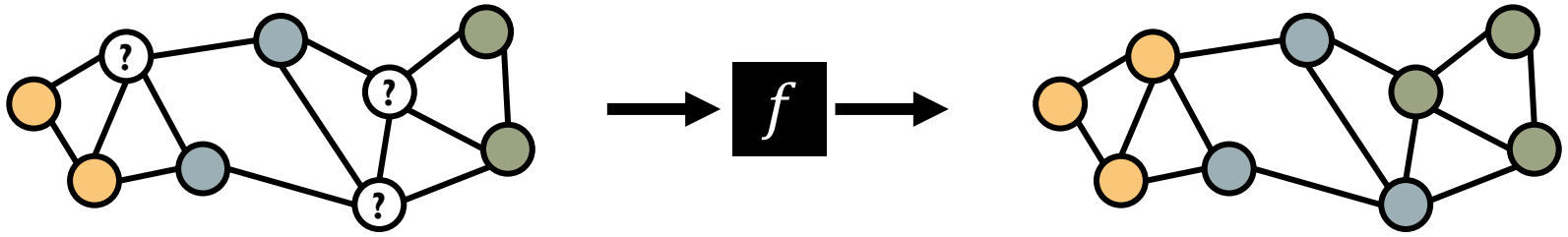
latent clean graph



observed perturbed graph

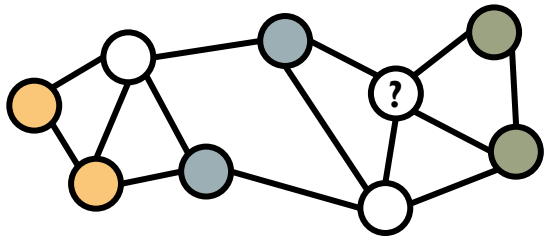
PROBLEM SETUP

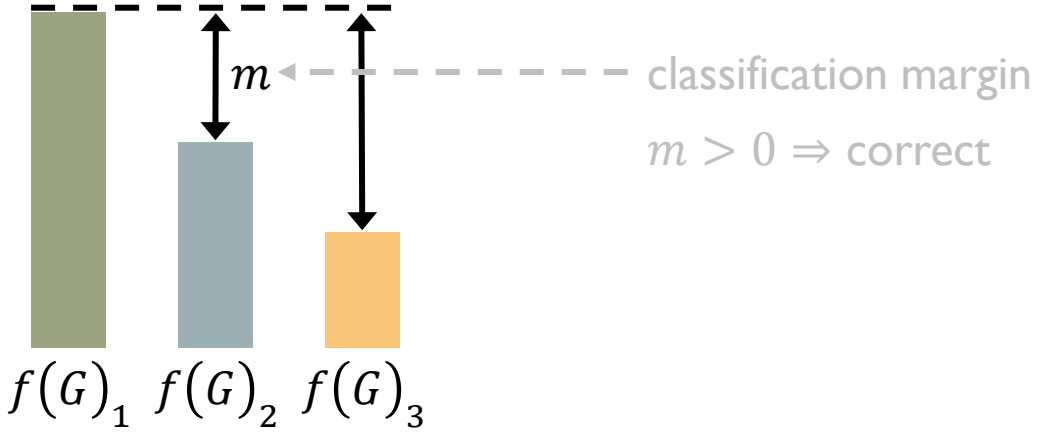
Semi-supervised node classification



PROBLEM SETUP

Semi-supervised node classification

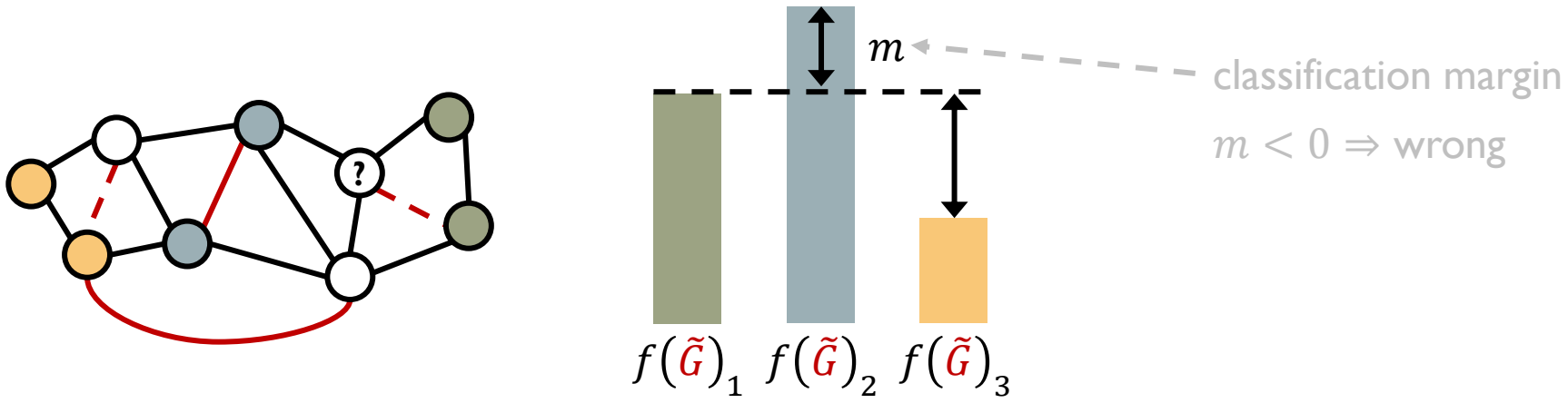




The prediction is ?

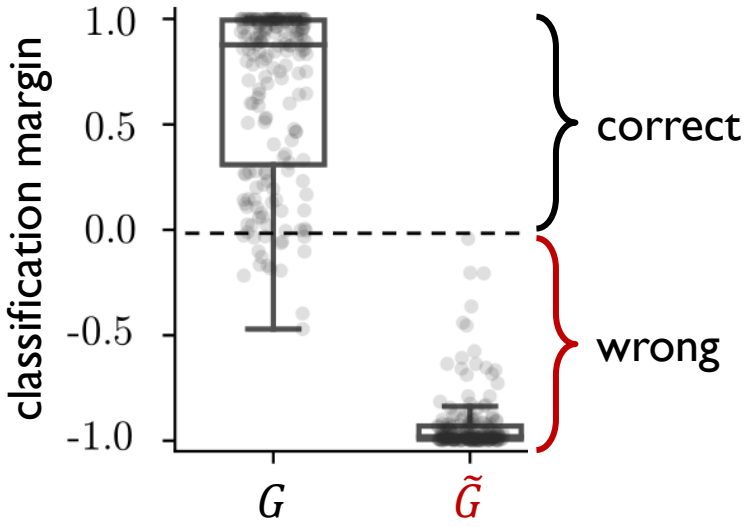
THE ADVERSARY WANTS TO

Change the prediction of a target node



The prediction is ? after perturbation

HOW EASY IS TO MANIPULATE THE PREDICTION?



Attackers can misclassify most nodes by perturbing just a few edges

ROBUSTNESS CERTIFICATE

Provable guarantee that the prediction does not change

e.g. all graphs that differ in up to 5 edges

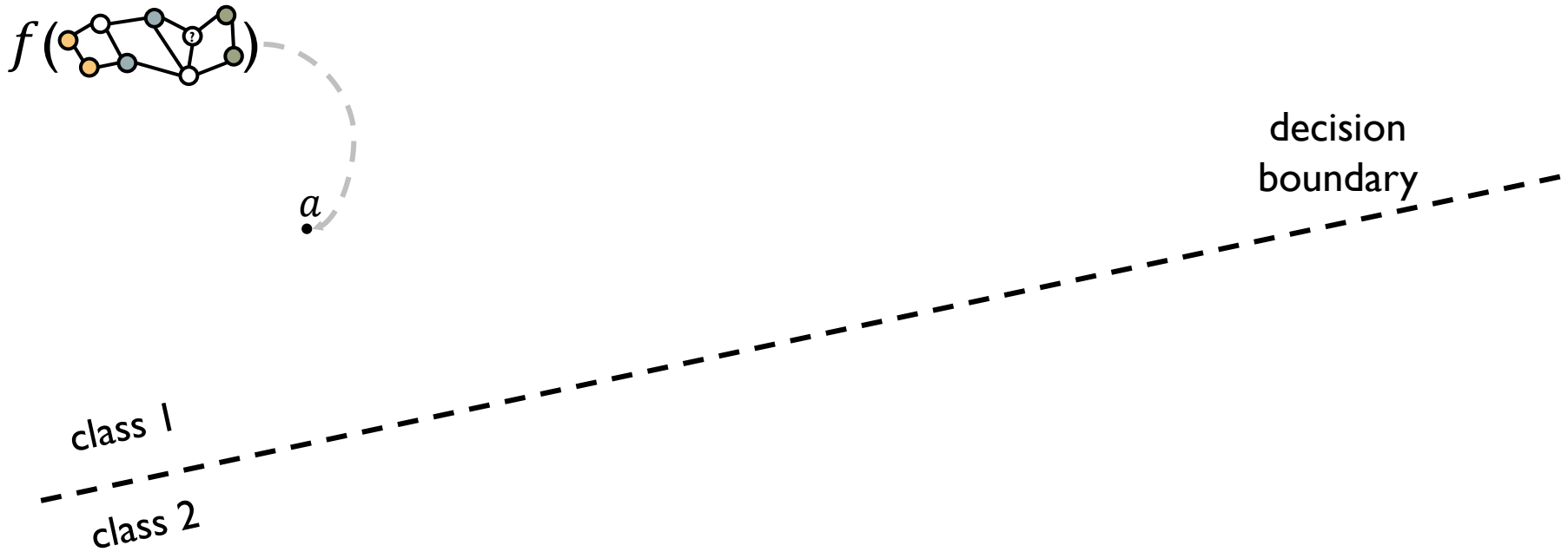
Verify whether for all admissible perturbed graphs $\tilde{G}$:

$$\underset{\text{class } y}{\operatorname{argmax}} f(\tilde{G})_y \overset{?}{=} \underset{\text{class } y}{\operatorname{argmax}} f(G)_y$$

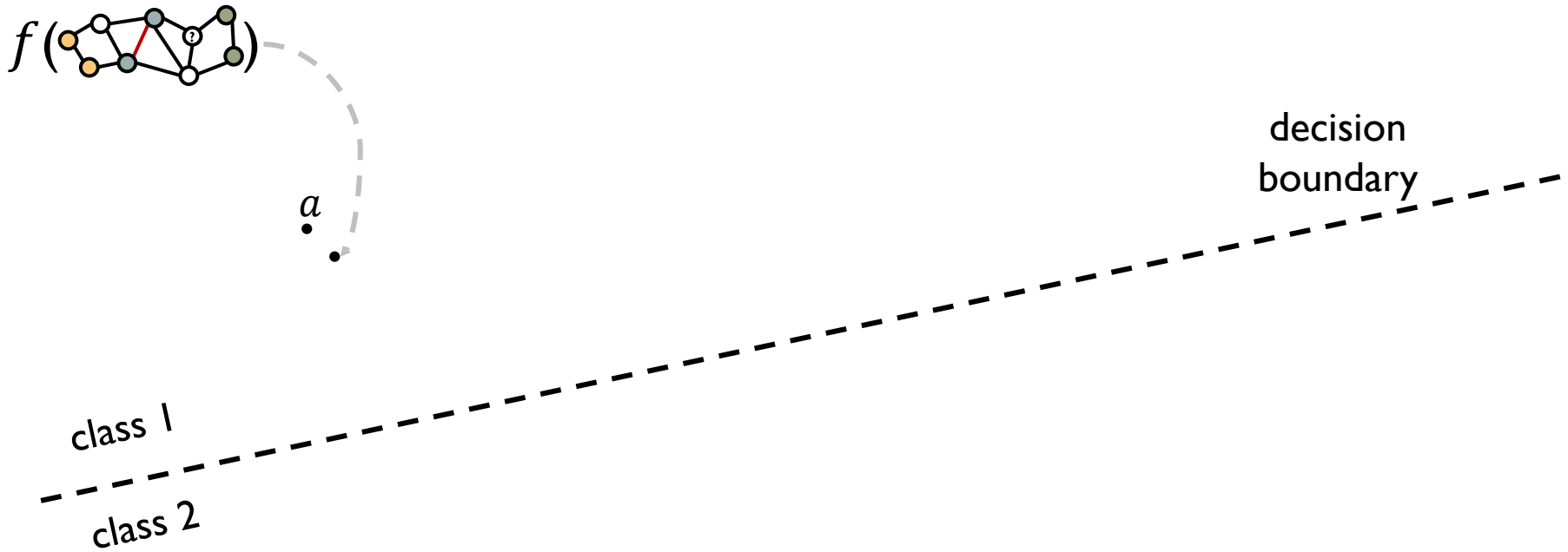
perturbed graph

clean graph

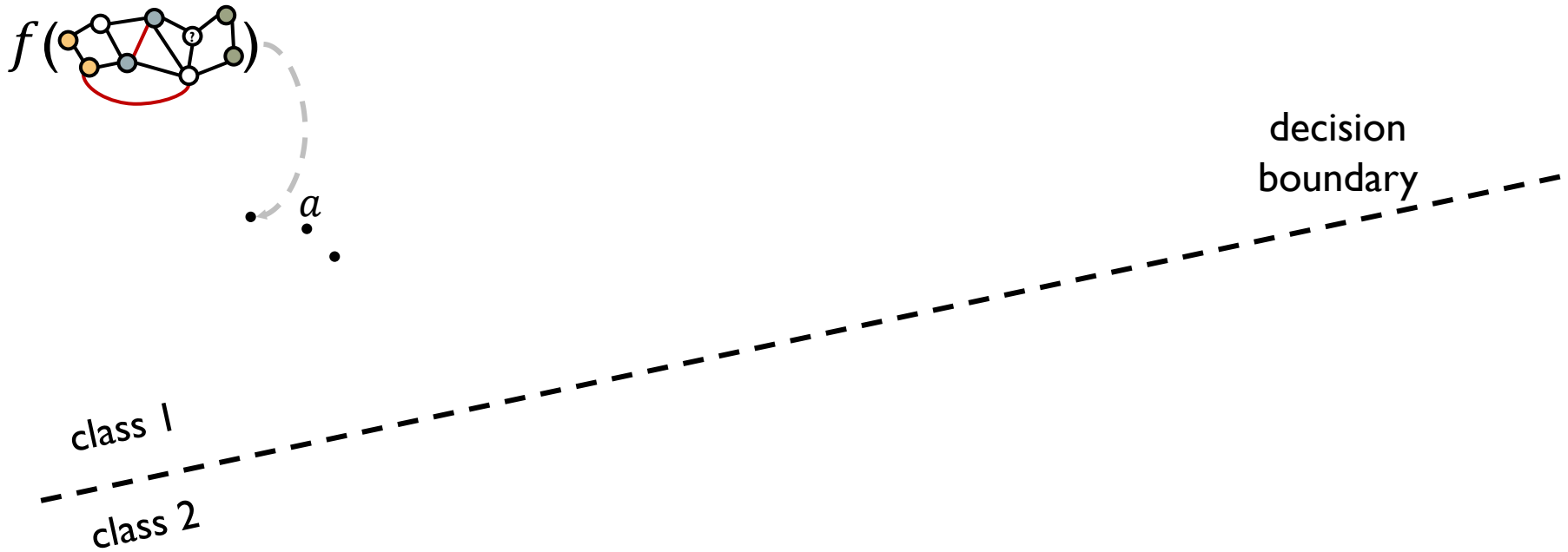
CERTIFICATE ILLUSTRATION



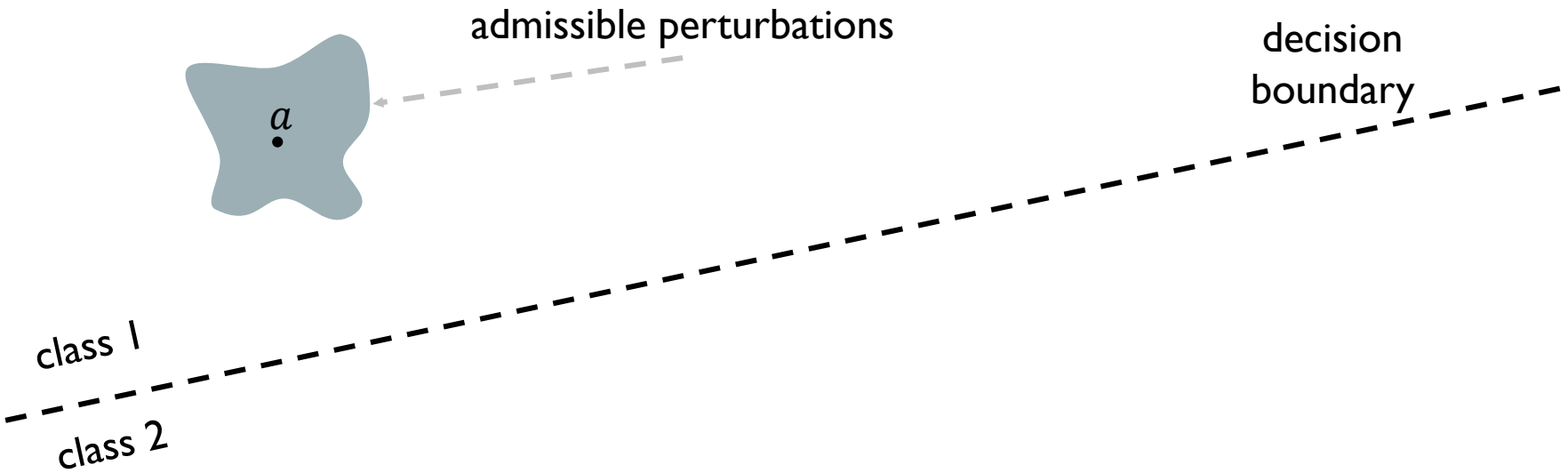
CERTIFICATE ILLUSTRATION



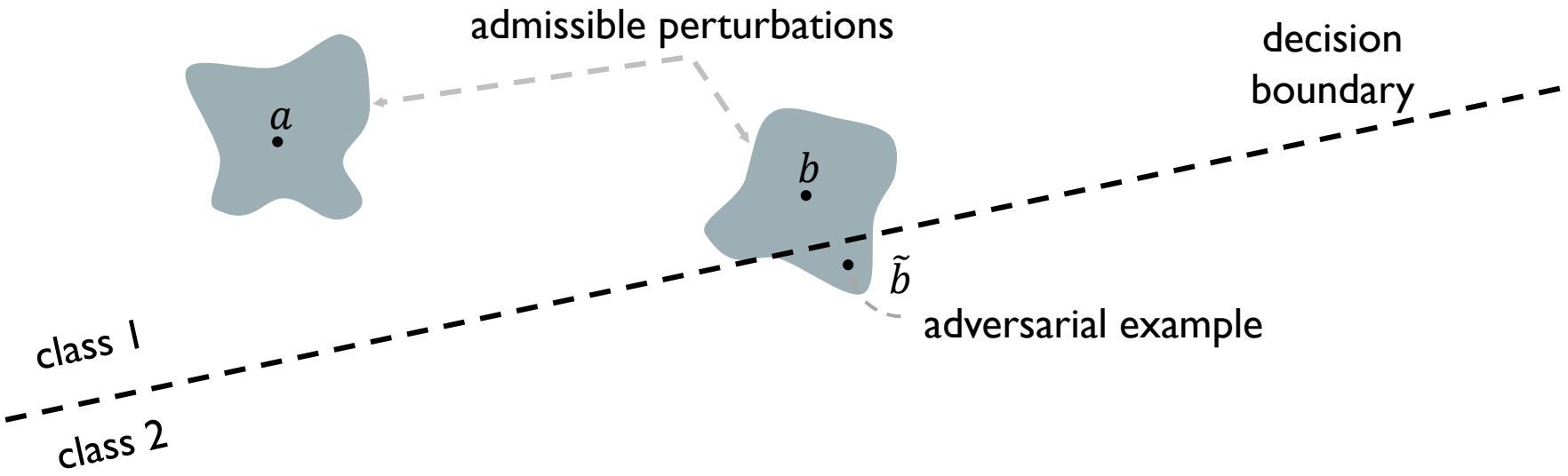
CERTIFICATE ILLUSTRATION



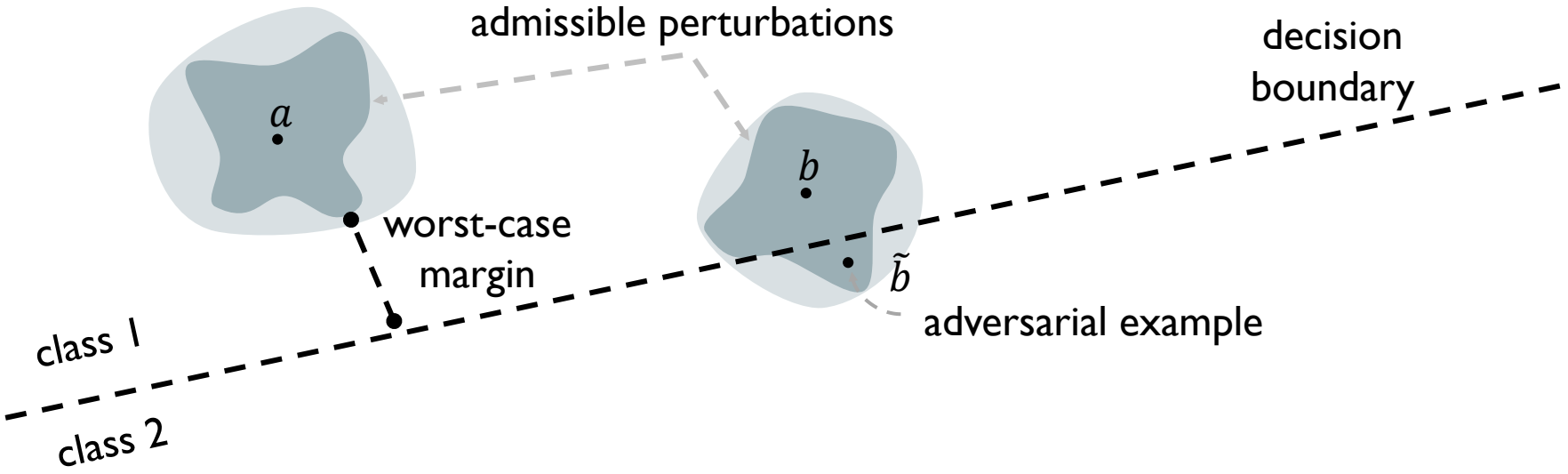
CERTIFICATE ILLUSTRATION



CERTIFICATE ILLUSTRATION



(LOWER BOUND ON) THE WORST-CASE MARGIN



COMPUTING THE WORST-CASE MARGIN

Model-specific Certificates

Model-agnostic Certificates

PAGERANK-BASED MODELS

Predictions are a linear function of personalized PageRank $\boldsymbol{\pi}_G$

$$f(G)_y = \boldsymbol{\pi}_G^T \boldsymbol{H}_{:,y}$$

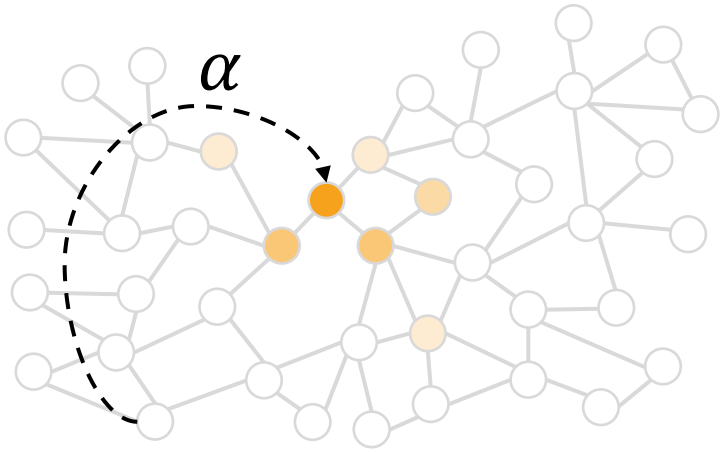
personalized PageRank vector for a given node

"logits" for class y

Models in this family: PPNP, PushNet, Label/Feature Propagation

BACKGROUND: PERSONALIZED PAGERANK (PPR)

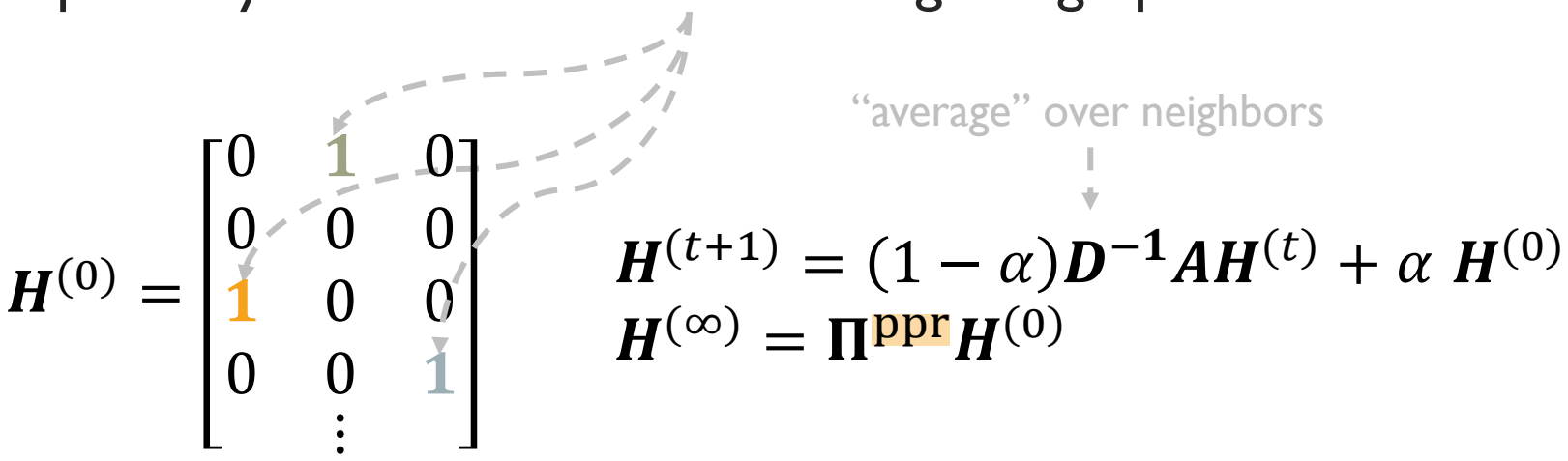
Stationary distribution of a random walk with teleport



The teleport probability α controls the effective neighborhood size

BACKGROUND: PPR AND LABEL PROPAGATION

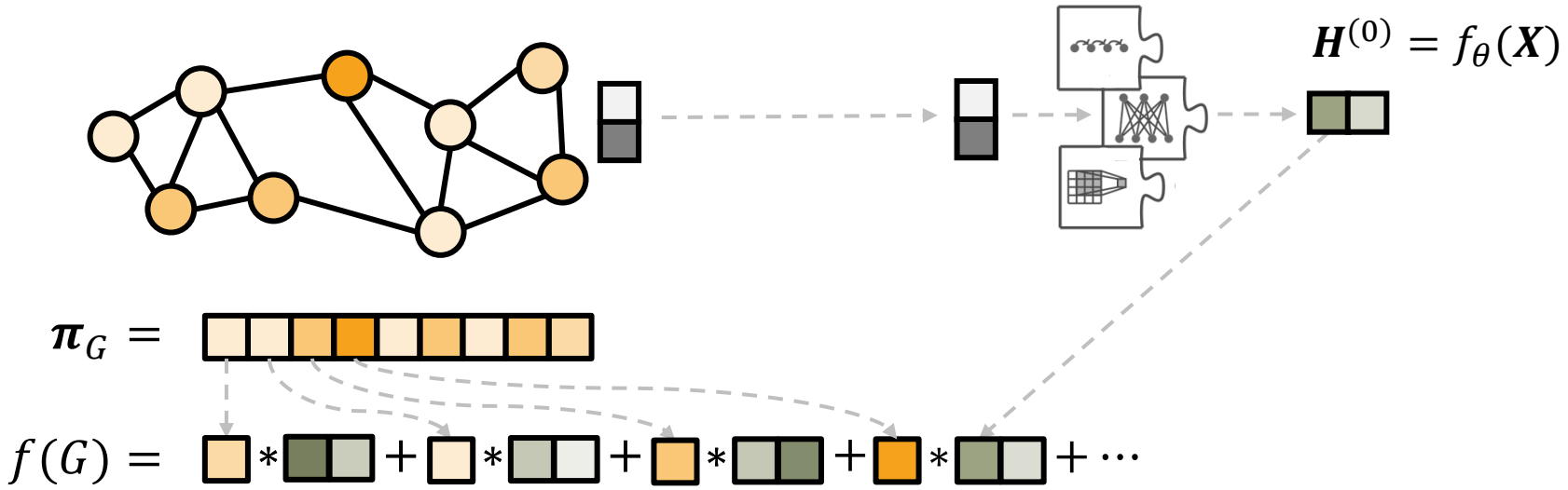
Repeatedly diffuse intial "beliefs" using the graph



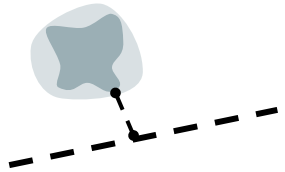
Probability that some node belongs to class $y \propto \boldsymbol{\pi}_G^T\, \boldsymbol{H}_{:,y}$

(A)PPNP: PREDICT THEN PROPAGATE

First map node features to initial beliefs then diffuse them with PPR



PAGERANK OPTIMIZATION

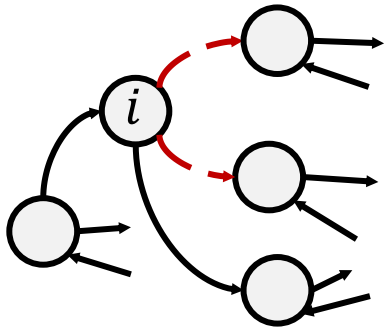
Computing  ⇔ PageRank optimization

$$m^* = \min_{\substack{\tilde{G} \in \text{admisible} \\ \text{perturbed graphs}}} \min_{\text{class } y \neq y^*} \boldsymbol{\pi}_{\tilde{G}}^T \left(\boldsymbol{H}_{:,y^*} - \boldsymbol{H}_{:,y} \right)$$

Minimize a linear function of PageRank over the set of graphs

THREAT MODEL – WHAT IS ADMISSIBLE?

An attacker has control over a set $\mathcal{F}$ of fragile edges



General and flexible:

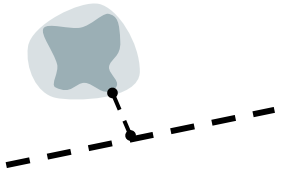
Scenario 1: $\mathcal{F} = \mathcal{E}$ # remove edges
Scenario 2: $\mathcal{F} = (\mathcal{V} \times \mathcal{V}) \backslash \mathcal{E}$ # add edges
Scenario 3: $\mathcal{F} =$ ...

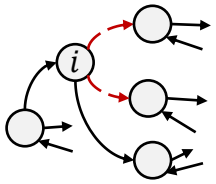
Global budget: perturb up to B edges in total

Local budget: perturb up to b_v edges for node v

Bojchevski, Günnemann. "Certifiable Robustness to Graph Perturbations". NeurIPS 2019.

PAGERANK OPTIMIZATION

Computing  ⇔ PageRank optimization ⇔ MDP

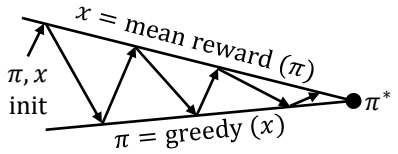


input graph

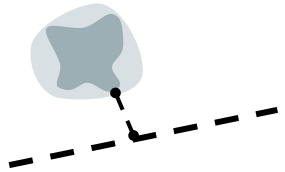
$\mathbb{E}(\boldsymbol{r})$ MDP

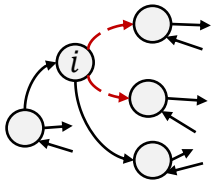
exponential action set

Policy Iter. poly time



PAGERANK OPTIMIZATION

Computing  ⇔ PageRank optimization ⇔ MDP

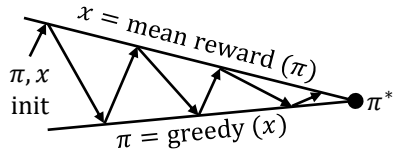


input graph

$\mathbb{E}(\boldsymbol{r})$ MDP

exponential
action set

Policy Iter.
poly time



auxiliary graph

$\sum(\boldsymbol{r})$ MDP

binary
action set

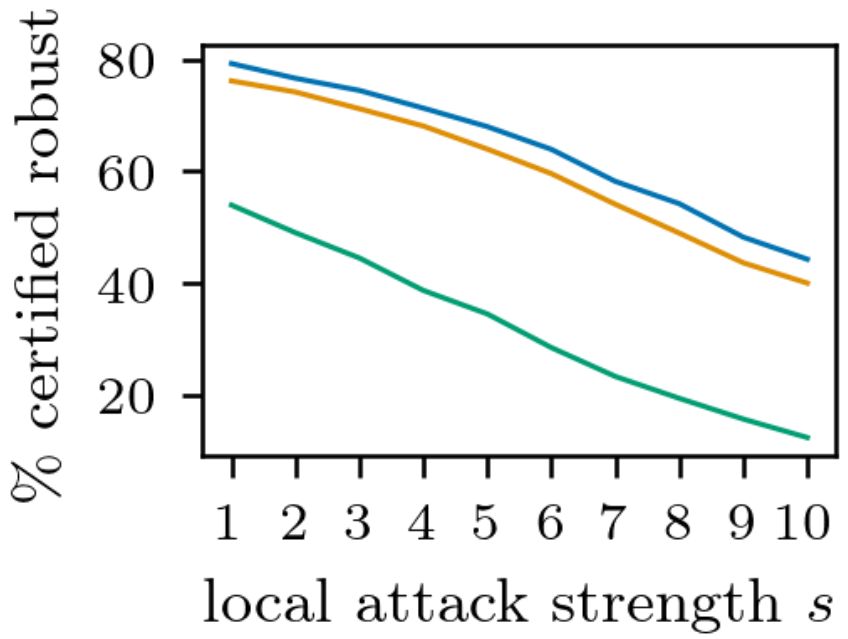
QCLP

$\min \boldsymbol{c}^T \boldsymbol{x}$
$s.t.\ \boldsymbol{Ax} = \boldsymbol{b}$
$\boldsymbol{x}^T \boldsymbol{Qx} = \boldsymbol{0}$

RLT

Bojchevski, Günnemann. "Certifiable Robustness to Graph Perturbations". NeurIPS 2019.

EXACT CERTIFICATES FOR PAGERANK-BASED MODELS



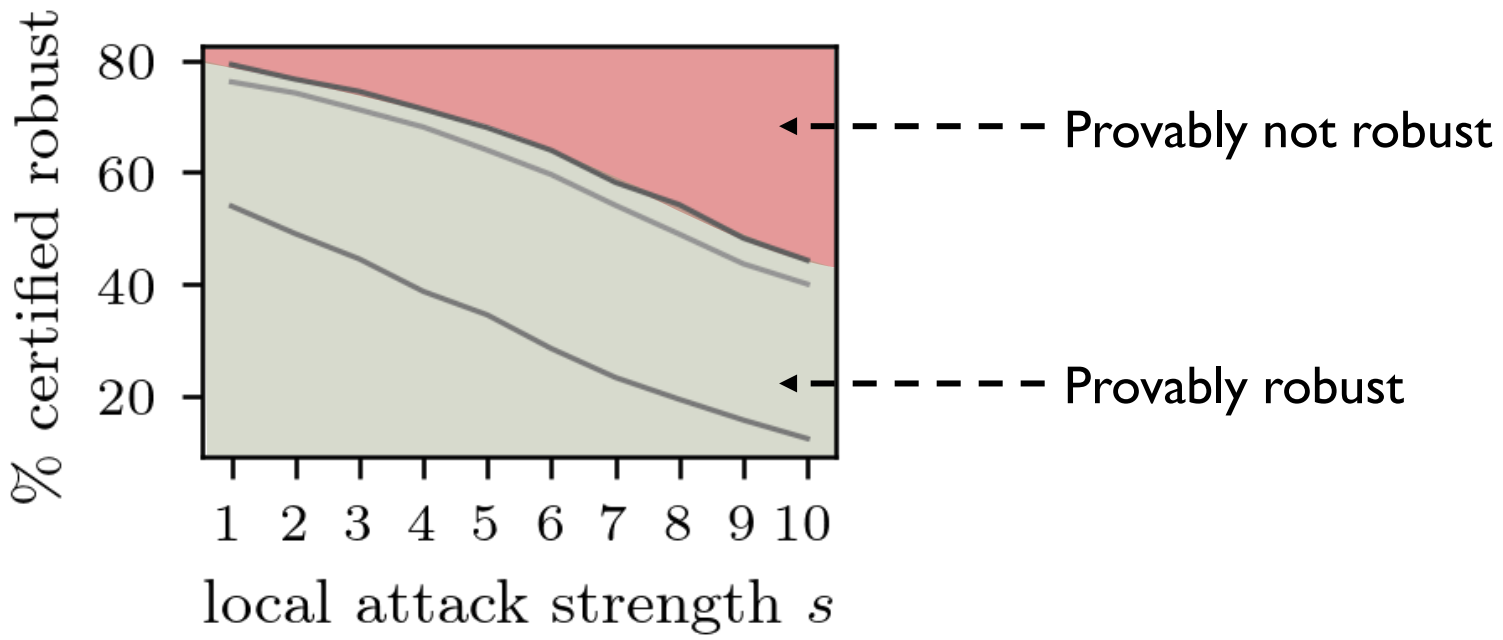
PPNP

Feature Propagation

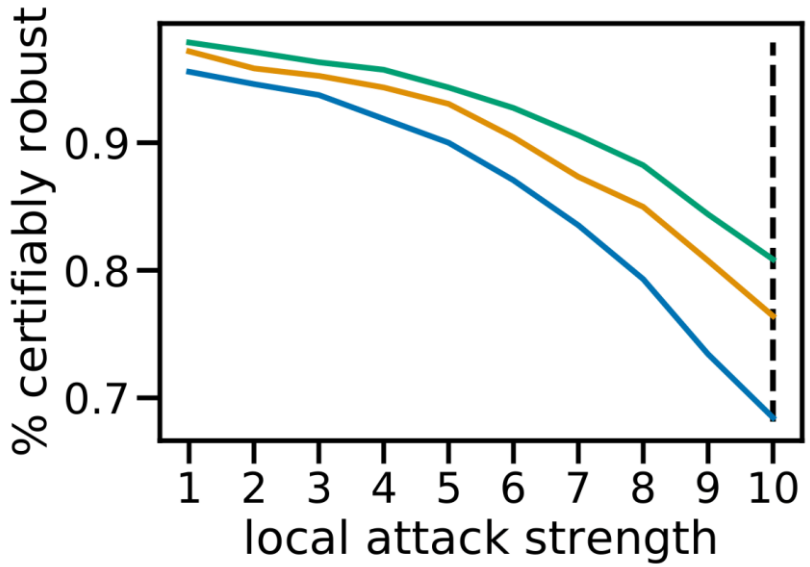
Label Propagation

Bojchevski, Günnemann. "Certifiable Robustness to Graph Perturbations". NeurIPS 2019.

EXACT CERTIFICATES FOR PAGERANK-BASED MODELS



ROBUST TRAINING



Training scheme:

(70%) Standard cross-entropy

(73%) Hinge-loss penalty

(72%) Robust cross-entropy

Trained with strength

COMPUTING THE WORST-CASE MARGIN

Model-specific Certificates

+ Takes advantage of problem structure (exact guarantees)

– Limited applicability

Model-agnostic Certificates

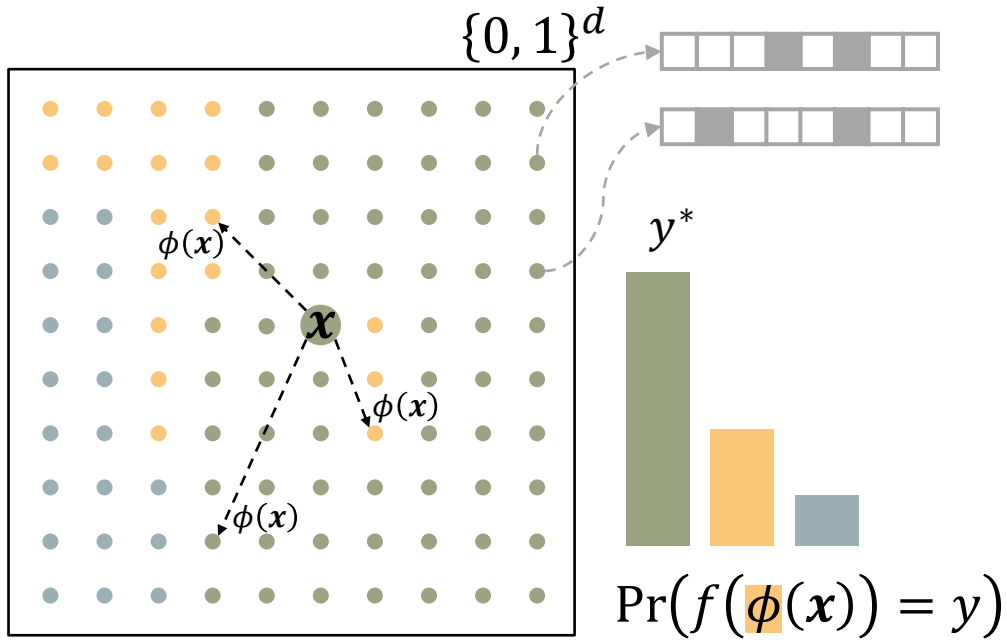
RANDOMLY SMOOTHED CLASSIFIERS

Any base classifier $f: \mathcal{X} \to \mathcal{Y}$

Randomization scheme $\phi(\boldsymbol{x})$

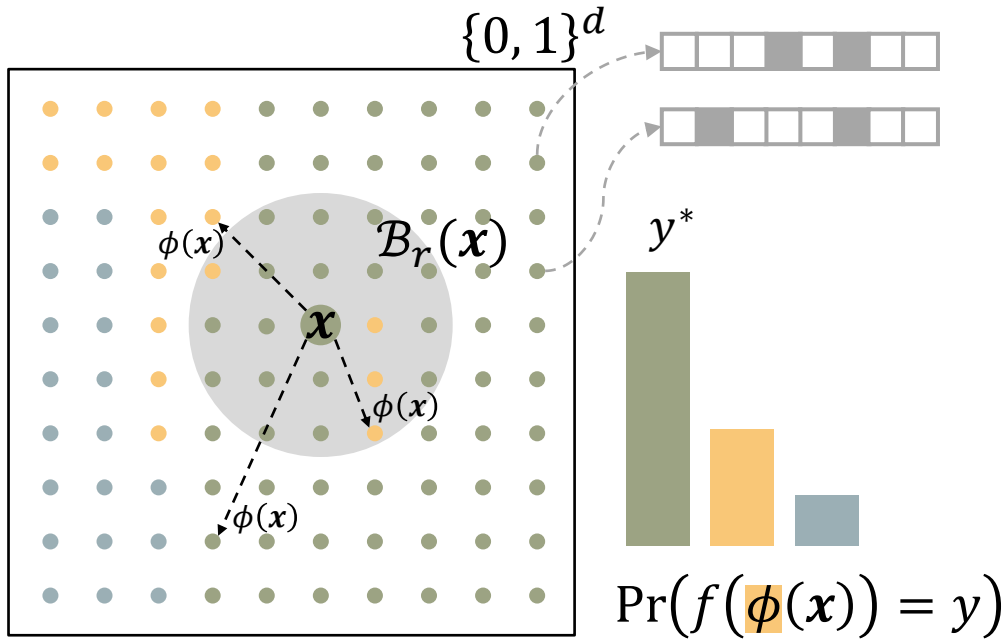
Certify a smoothed classifier g

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



RANDOMLY SMOOTHED CLASSIFIERS

Guarantee that the majority does not change in a ball $\mathcal{B}_r$ around $\boldsymbol{x}$



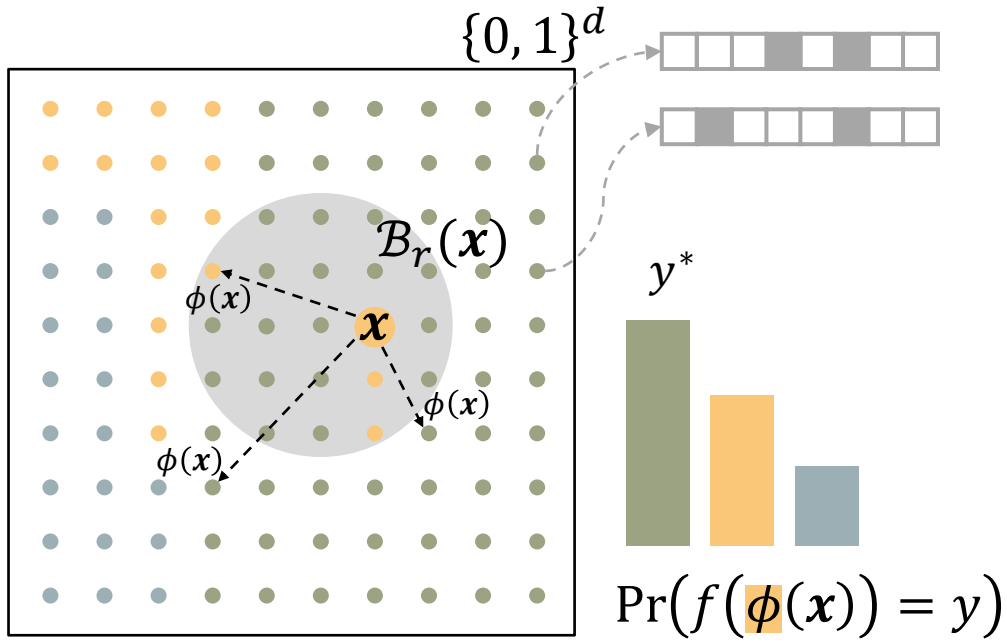
RANDOMLY SMOOTHED CLASSIFIERS

Guarantee that the majority does not change in a ball $\mathcal{B}_r$ around $\boldsymbol{x}$

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

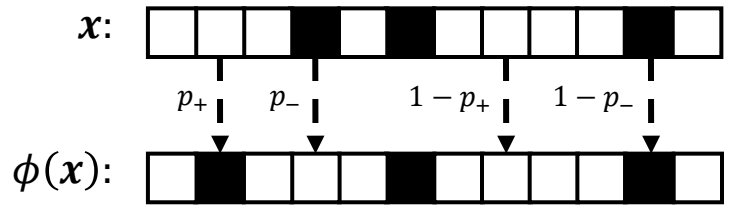
Verify whether for all $\tilde{\boldsymbol{x}} \in \mathcal{B}_r(\boldsymbol{x})$

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



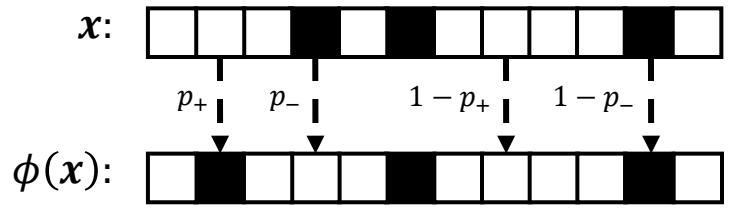
TWO NECESSARY IMPROVEMENTS

1. Sparsity-aware smoothing



TWO NECESSARY IMPROVEMENTS

1. Sparsity-aware smoothing



2. Dramatically more efficient: $O(n^4) \rightarrow O(r)$ ◂ - - - certified radius r
   - Independent of input size
   - 4 days → 2.5 milliseconds

number of nodes n

DERIVING THE CERTIFICATE

The smoothed classifier is certifiably robust if

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

subject to:

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

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

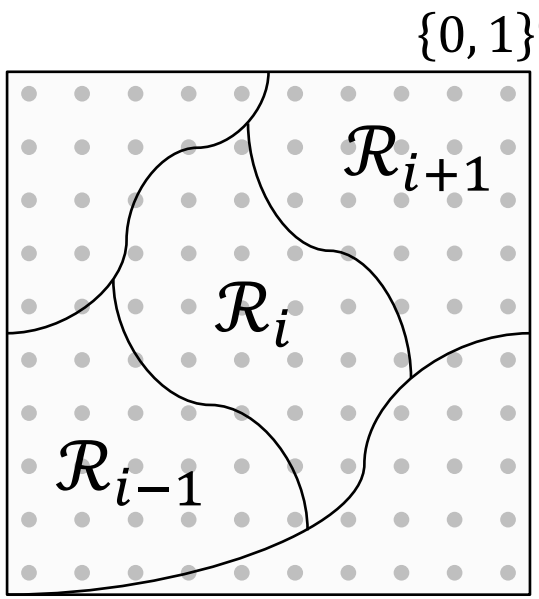
$$\min \sum_i \Pr(\phi(\tilde{x}) \in \mathcal{R}_i)\, h_i \overset{?}{>} 0.5$$

subject to:

$\tilde{x} \in \mathcal{B}_r(x)$ ← – – admissible

$h_i \in [0, 1]$

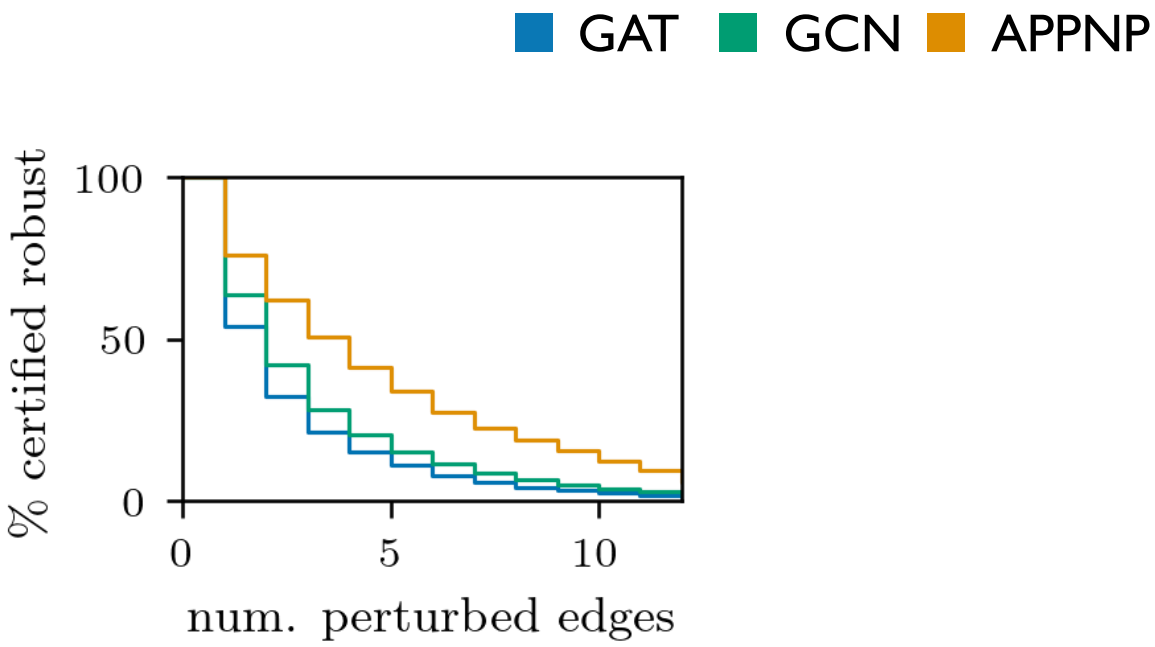
$\sum_i \Pr(\phi(x) \in \mathcal{R}_i)\, h_i = p_{y^*}$



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

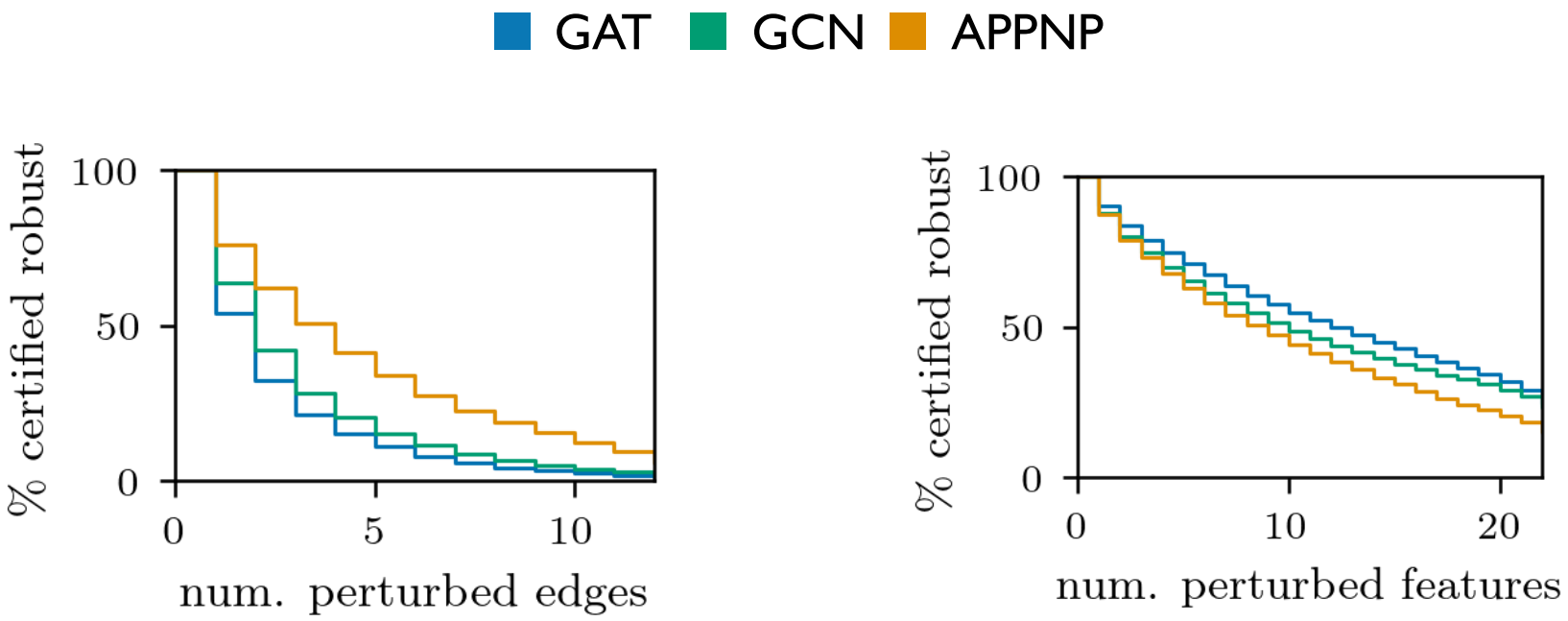
constant

GNNS HAVE DIFFERENT ROBUSTNESS TRADE-OFFS

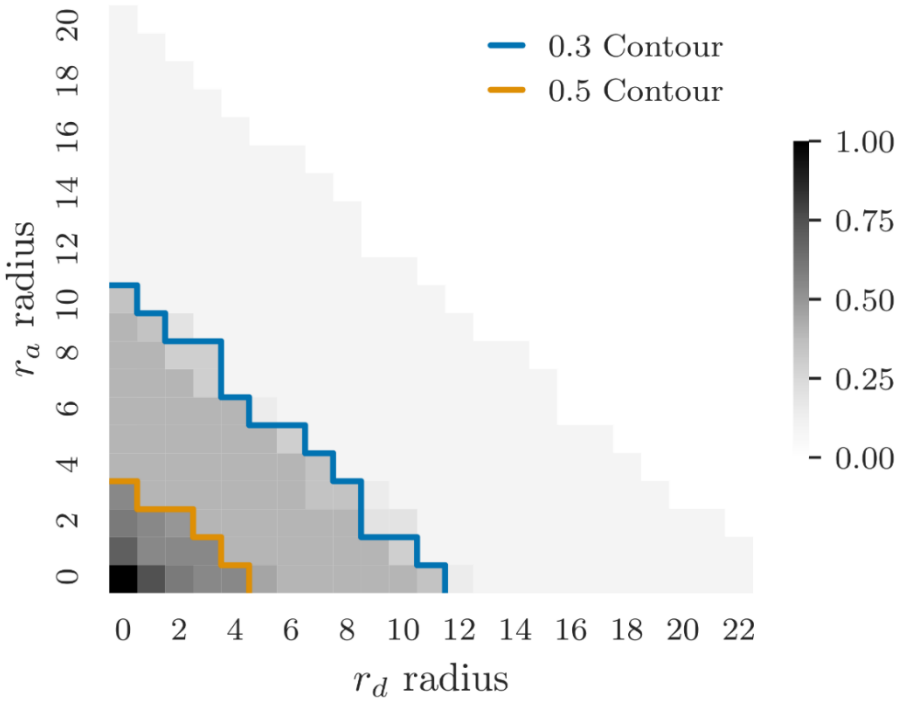


Bojchevski, Klicpera, Günnemann. "Efficient Robustness Certificates for Discrete Data: Sparsity-Aware Randomized Smoothing". ICML 2020

GNNS HAVE DIFFERENT ROBUSTNESS TRADE-OFFS



GRAPH-LEVEL CLASSIFICATION



CERTIFYING IMAGENET

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**

COMPUTING THE WORST-CASE MARGIN

Model-specific Certificate

+ Takes advantage of problem structure (exact guarantees)

– Limited applicability

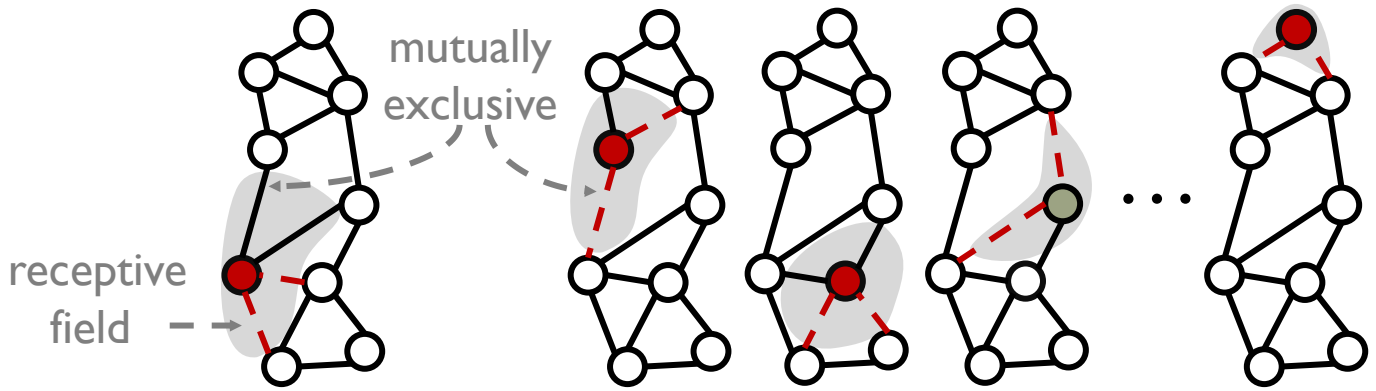
Model-agnostic Certificate

+ Can be applied to any classifier for discrete data (all GNNs)

– Does not capture all properties of the classifier

WHY ARE THE CERTIFICATES SO PESSIMISTIC?

They treat each node independently

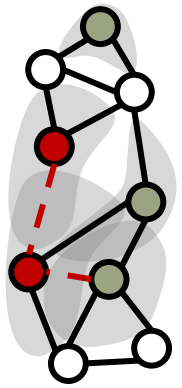


$$\sum_{\text{nodes}} \min_{\tilde{G}} \mathbb{I}\left[f_n(G) = f_n(\tilde{G})\right]$$

COLLECTIVE ROBUSTNESS CERTIFICATES

Verify how many nodes are simultaneously robust

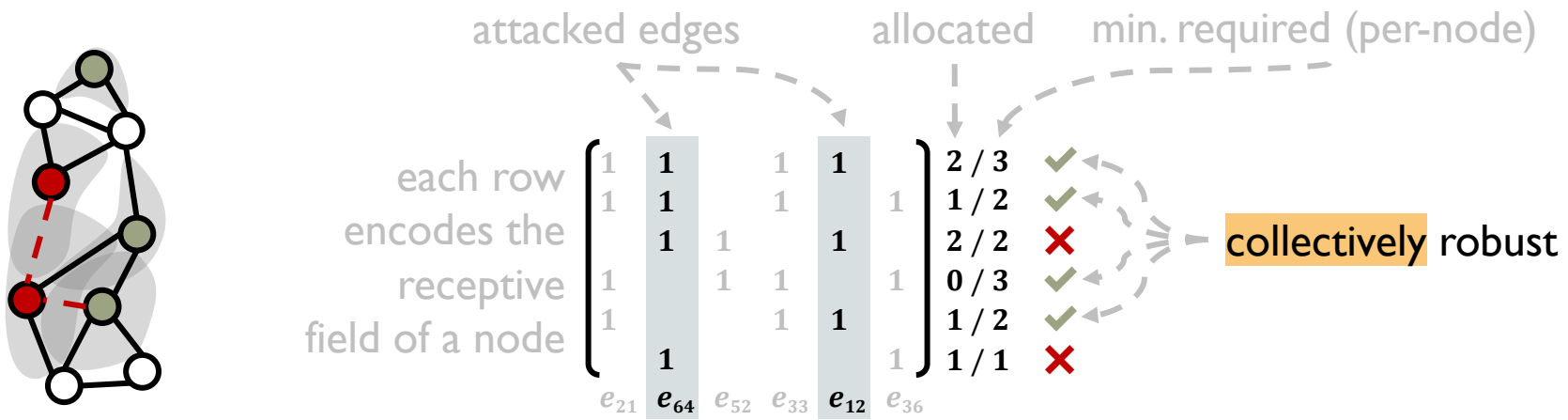
$$\sum_{\text{nodes}} \min_{\tilde{G}} \mathbb{I}\left[f_n(G) = f_n(\tilde{G})\right]$$



$$\min_{\tilde{G}} \sum_{\text{nodes}} \mathbb{I}\left[f_n(G) = f_n(\tilde{G})\right]$$

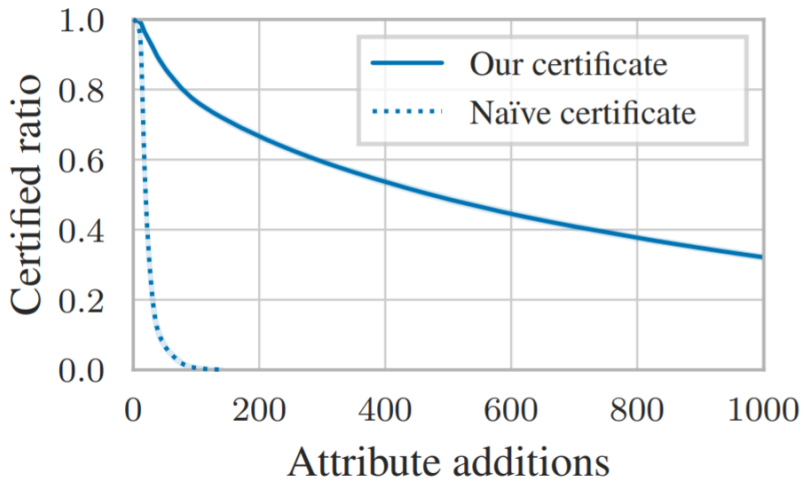
FUSE MULTIPLE SINGLE-NODE CERTIFICATES

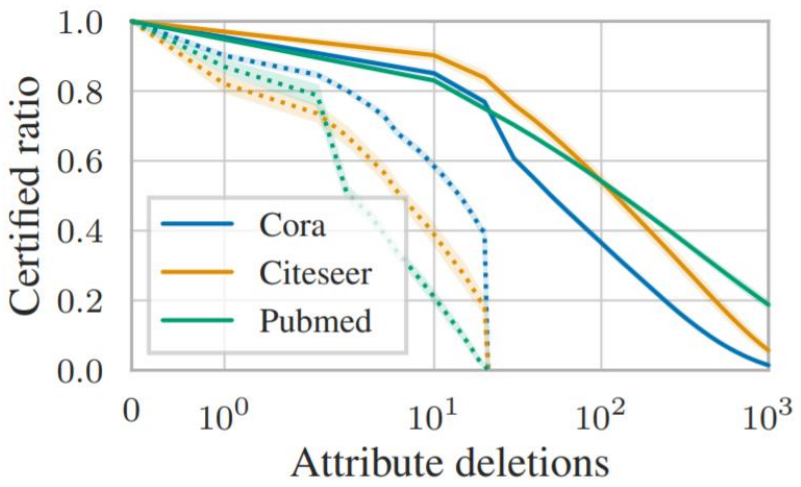
Into a provably stronger certificate by exploiting interdependence



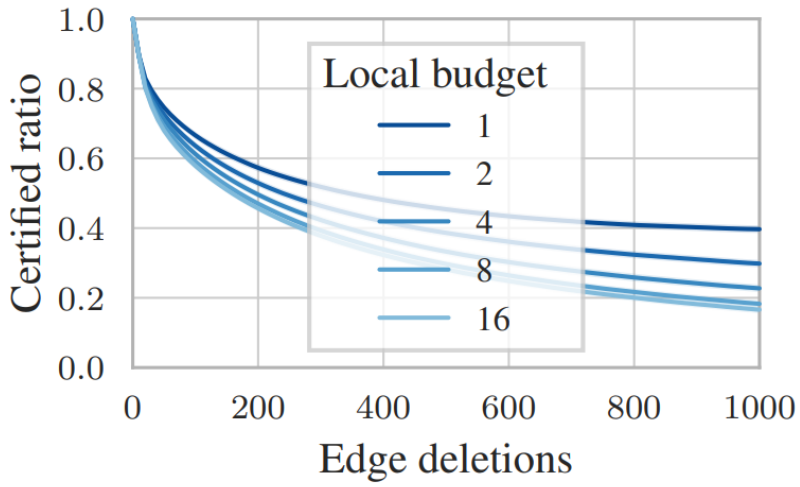
Budget allocation problem: pick k edges to misclassify most nodes

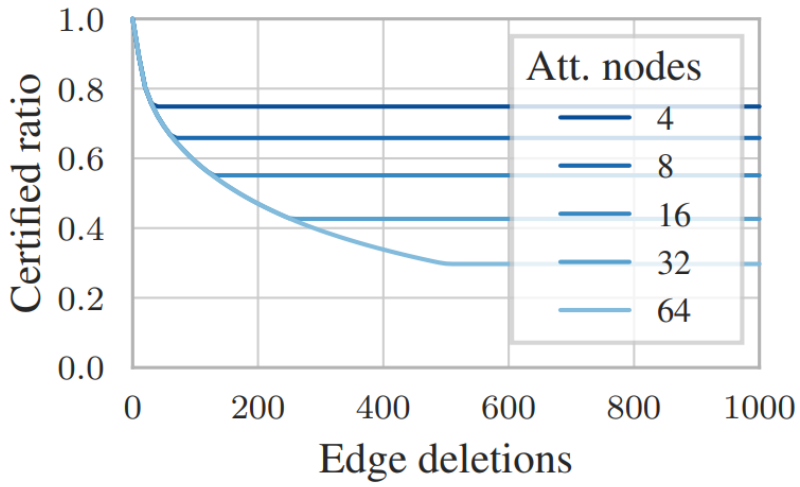
COLLECTIVE = DRAMATIC IMPROVEMENTS





MORE REALISTIC THREAT MODELS

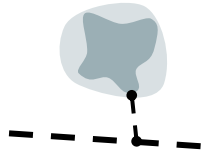




SUMMARY

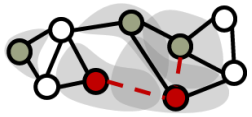
GNNs are vulnerable to small perturbations ⇒ Verify robustness

Model-specific: Exact guarantees for a family of PageRank-based models



Model-agnostic: Turn any classifier into a certifiable smoothed classifier

Collective: Certify simultaneously robust nodes (dramatic improvement)



Open: knowledge graphs, code, graph-level, representations, poisoning, …

www.daml.in.tum.de/graph-cert
www.daml.in.tum.de/sparse-smoothing
www.daml.in.tum.de/collective-robustness

@abojchevski