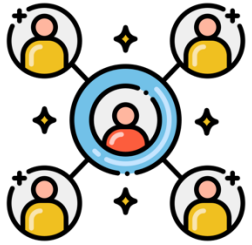
# Adversarial Defenses on Graph:

## Heuristic and certified

06.21

# Background

Real-world relationships are often abstracted into graphs.



Social network

Transaction network

Recommendation network

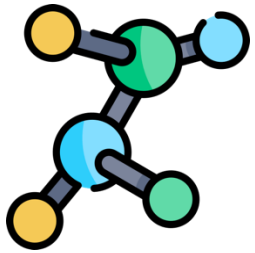
TASK: node classification with GNN

Threats: Evasion attack against node classifier

Unnoticeable perturbations on graph structure and node attributes could lead to misclassification of target node by high-accuracy GNN models.

# Background

Real-world relationships are often abstracted into graphs.



Molecular graph

Traffic network

Healthcare graph

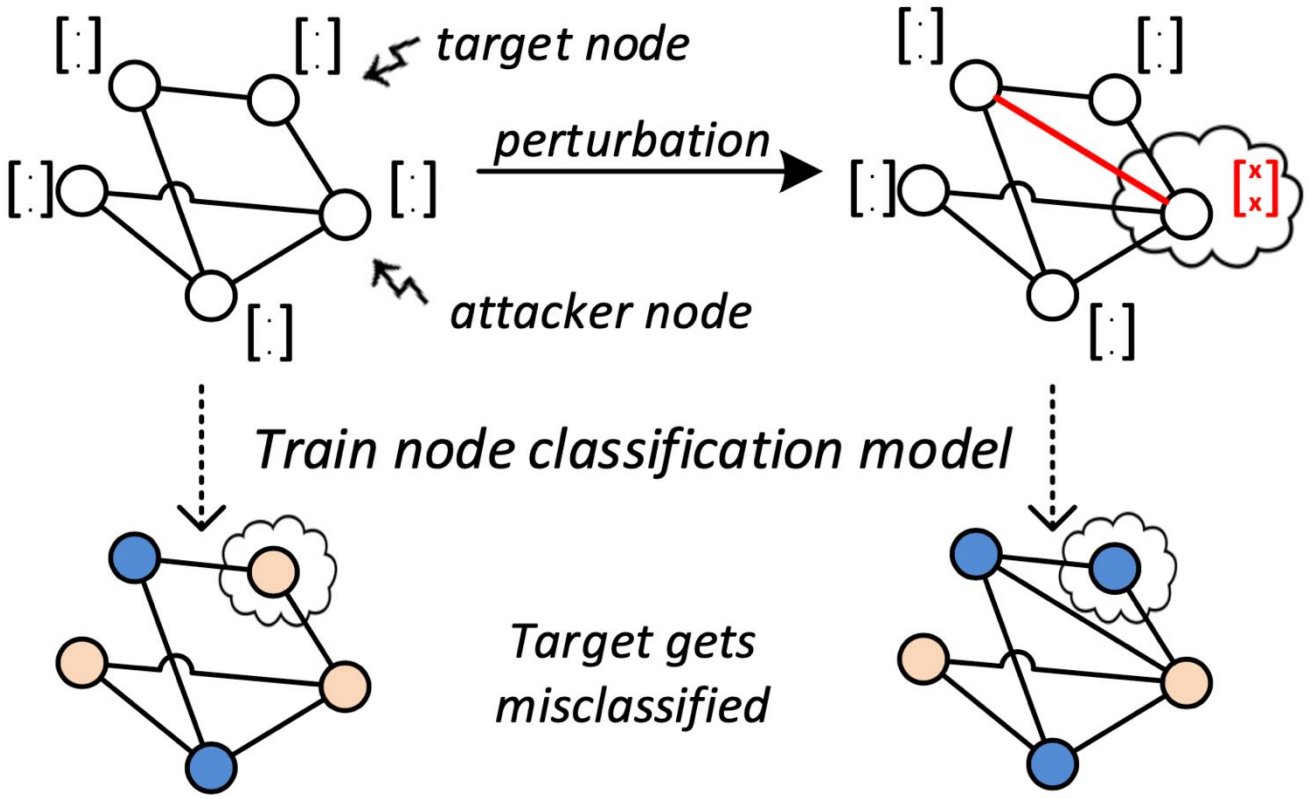
TASK: graph classification with GNN

Threats: Evasion attack against graph classifier

Unnoticeable perturbations on graph structure and node attributes could lead to misclassification of graph by high-accuracy GNN models.

# Background

Threats: Evasion attack against node classifier




**Notations**

$G = (X, A)$: graph with $N$ vertices

$X \in \{0,1\}^{N \times D}$: attribute matrix

$A \in \{0,1\}^{N \times N}$: adjacent matrix

$v \in \mathcal{V}$ : target node

$f_\theta : \mathcal{V} \rightarrow \{1,2 \ldots K\}$ : node classifier

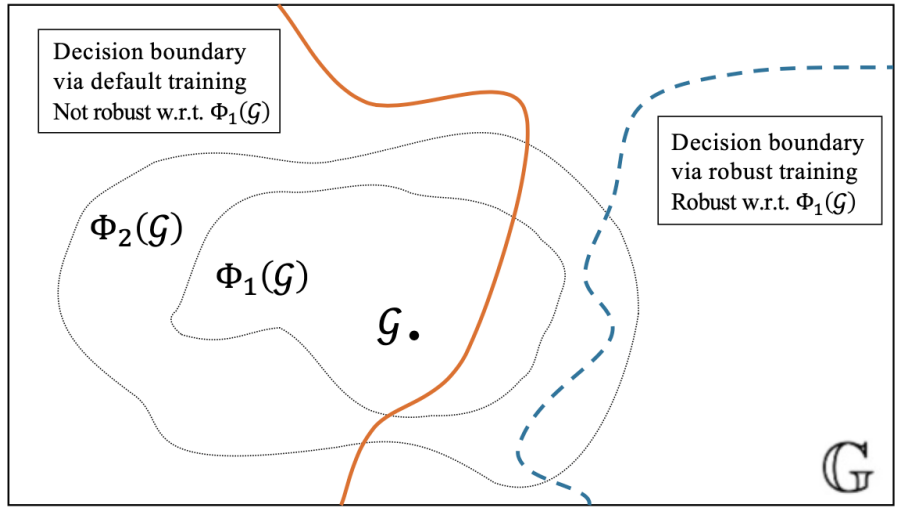
$g_\theta : G \rightarrow \{1,2 \ldots K\}$ : graph classifier

# Defenses: heuristic and certified

Heuristic defenses

Adversarial training: improves model robustness by including adversarial examples, which are intentionally perturbed inputs designed to deceive the model, during the training process.

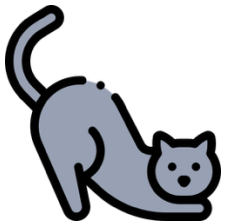
- PGD adversarial training




Attack detection: remove perturbations from the data, i.e., to revert the performed malicious changes and obtain a 'cleaner' graph

- the predicted class distribution changes for attacked nodes
- Jaccard similarly between the nodes' attributes
- affect mainly the high-rank (low-valued) singular components of the graph's adjacency matrix

# Certified defenses



Existing heuristic defense methods are only effective for certain attacks and are vulnerable to newly designed attacks.

Attacks and defenses are forming a cat-and-mouse chasing game.

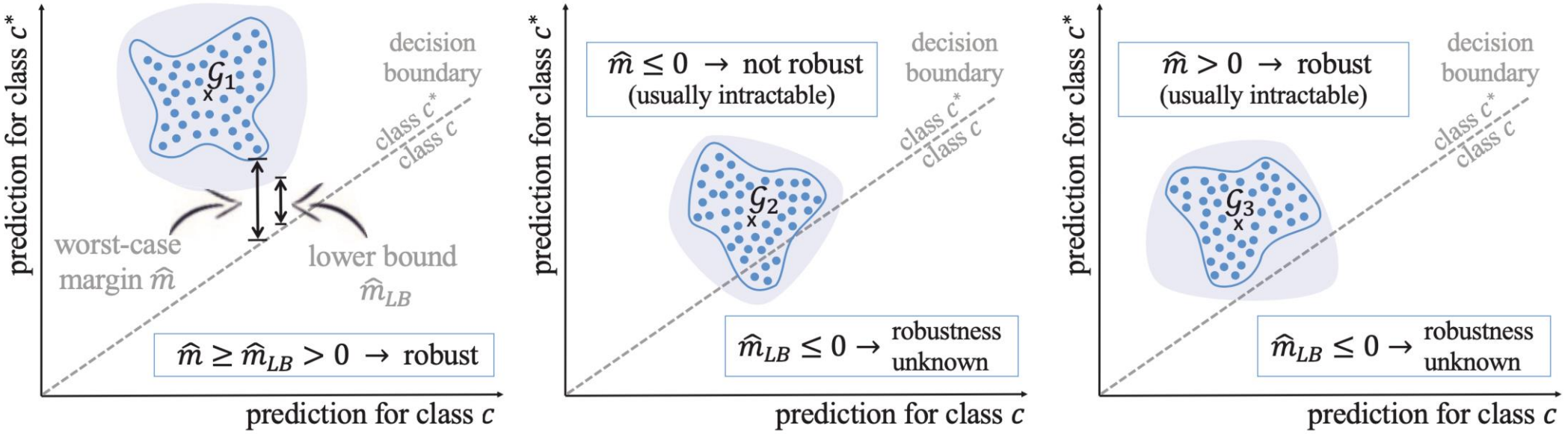
Certified defense provide provable robustness. These methods provide so called *robustness certificates*, giving formal guarantees that no perturbation regarding a specific perturbation model will change the prediction.

# Certified defenses

## Model specific certificate

Model-specific certificates are designed for a specific class of GNN models (e.g., 2-layer GCNs) and a specific task such as node-level classification.

A common theme is to phrase certification as a constrained optimization problem:



# Certified defenses

## Model specific certificate

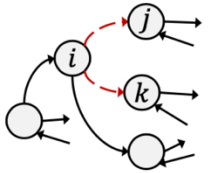
Optimization perspective:

- Node attribute perturbation: obtain lower bounds via a relaxation to a linear program (GCN / node)
- Edge deletion perturbation: reduces the problem to a jointly constrained bilinear program (GCN / node)
- Edge perturbation: Lagrange dualization and convex envelope (GCN / graph)

PageRank perspective (label propagation, GNN / node)

### Certificate ⇔ PageRank Optimization ⇔ MDP

Computing certificates amounts to finding optimal PageRank which can be done efficiently via a Markov Decision Process




node = state

transition law

$(\mathcal{S}, (\mathcal{A}_i)_{i\in\mathcal{S}}, p, r)$

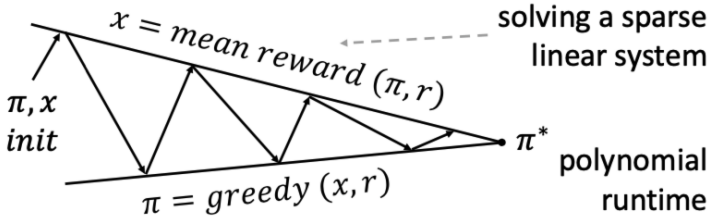
reward

$\mathcal{A}_i = \{\emptyset, \{j\}, \{k\}, \{j,k\}\}$ actions = subsets of fragile edges

**Local budget: find optimal fragile edges with policy iteration**

$$r = H_{c*}^{(0)} - H_c^{(0)}$$

set reward to the logit difference




**Local + Global budget: NP-Hard, augment graph & solve a QP**

# Certified defenses

## Model agnostic certificate

However, the white-box access to model architecture of these certificates is simultaneously their limitation: The proposed certificates capture only a subset of the existing GNN models and any GNN yet to be developed likely requires a new certification technique as well.

Model-agnostic certificates can be applied to any GNN, i.e., model architecture agnostic.

Method: **randomized smoothing**

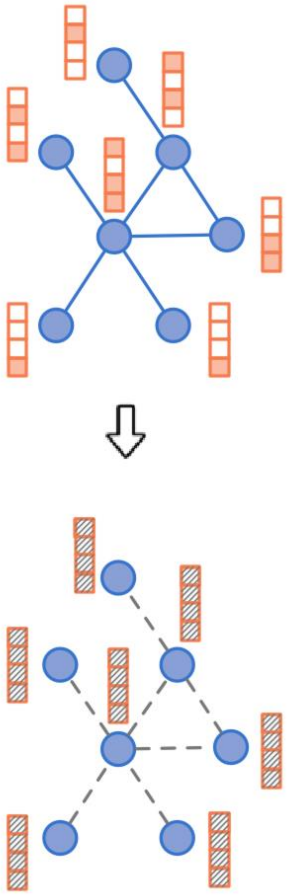
Given an input graph, RS obtain categorical prediction with a smooth version of base classifier $f$:

$$g_\theta(v|\tilde{G}) := \arg_y \max \Pr_{\tilde{G}\sim\phi(G)}[f_{\theta*}(v|\tilde{G}) = y],$$

# Randomized smoothing

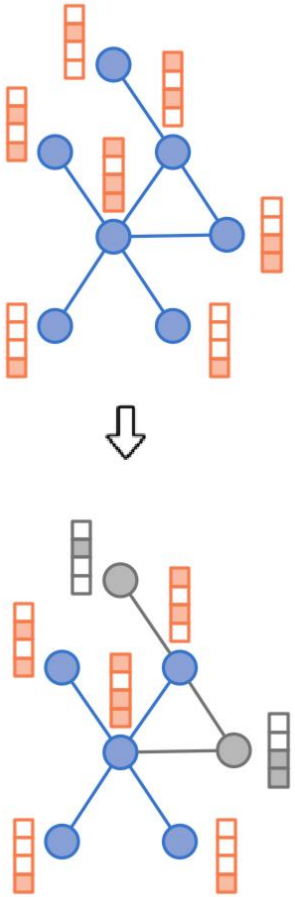
## Sparse smoothing

Add Bernoulli noise on all node attributes & edges



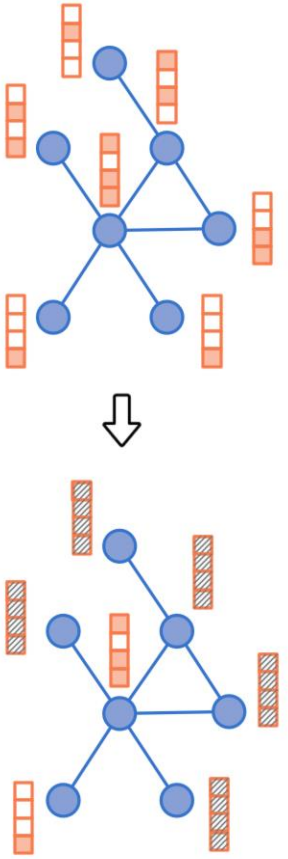
## Interception smoothing

Ablate randomly selected nodes



## Hierarchical smoothing

Add Bernoulli noise on randomly selected nodes' attributes



# Comparison

| type | method | pros | cons |
|---|---|---|---|
| Heuristic defense | Adversarial training | • High accuracy<br>• Applicable to different GNN | • Vulnerable to new attacks<br>• Need retraining or fine-tuning on white-box GNN |
| | Attack detection | • Simple<br>• Applicable to different GNN | • Vulnerable to new attacks |
| Certified defense | Optimization | • Give robustness certificate | • Specific to GNN architecture |
| | Randomized smoothing | • Give robustness certificate<br>• Applicable to different GNN | • Need retraining or fine-tuning on white-box GNN<br>• Accuracy degradation |

🤔 **Need provable robustness for arbitrary GNN without white-box access.**

# CiDer: Certified robustness via Diffusion model on graph

CiDer provides robustness certificate for arbitrary black-box pretrained GNN model.

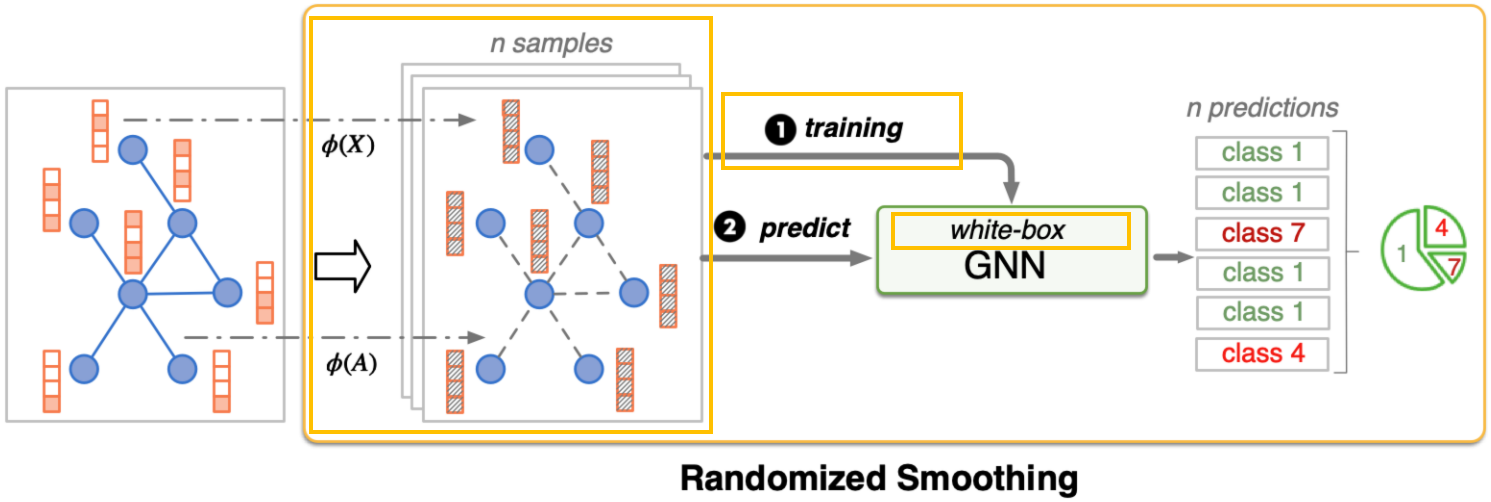
CiDer instantiates the randomized smoothing framework by prepending a discrete diffusion model before a high-accuracy classifier. The denoise process can effectively remove added randomized noise and retain high accuracy on clean samples.

$$g_\theta(v|\tilde{G}) := \arg_y \max \Pr_{\tilde{G}\sim\phi(G)}[f_{\theta*}(v|\tilde{G}) = y],$$

$$f_{\theta*}(v|G+\delta) := f_\theta(v|\mathcal{D}(G+\delta)).$$

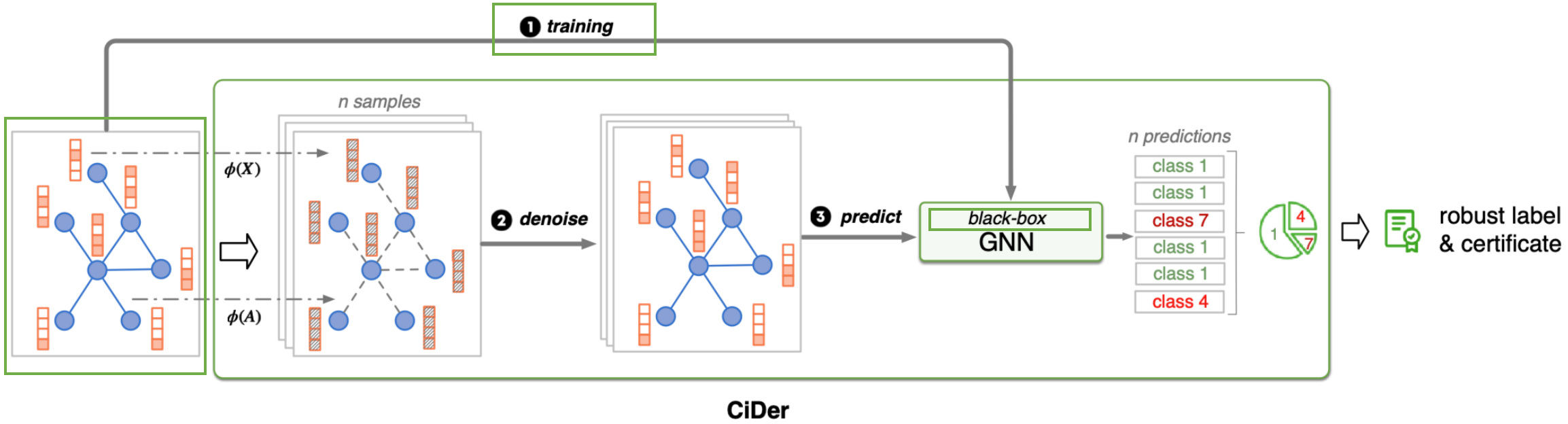
D3PM: Discrete Denoising Diffusion Probabilistic Model

# CiDer vs RS



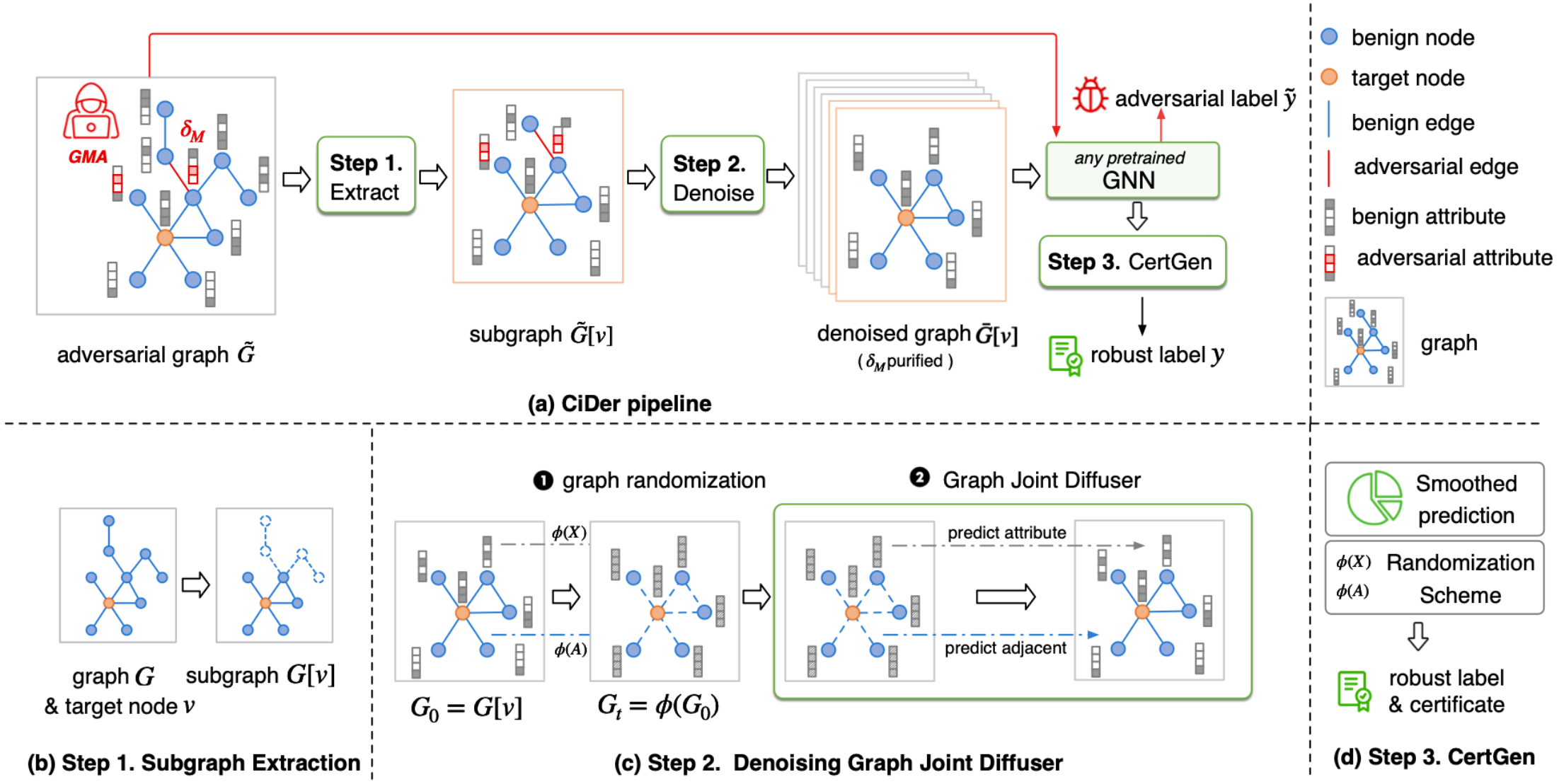

**Randomized Smoothing**

- *Need data augmentation training on GNN*
- *Accuracy drop on clean samples*
- *Retraining on different $\phi(\cdot)$ parameter (noise scale)*




**CiDer**

# CiDer pipeline



# Conclusion

- **Contribution**
  - CiDer provide robustness certificate for arbitrary black-box pretrained GNN model.
  - We propose a novel diffusion model named Graph Joint Diffuser.
    - Subgraph extraction (improve the computation efficiency).
    - Joint diffusion (model the complicated relations between nodes and edges).
  - Experiments on various data sets demonstrate the effectiveness and efficiency of the method.

- **Advantages over randomized smoothing**
  - No more retraining or fine-tuning on GNN (black-box oracle).
  - One denoiser fits all (all GNN, all noise scale, all noise combination over attr and adj).
  - Improved clean and certified accuracy.