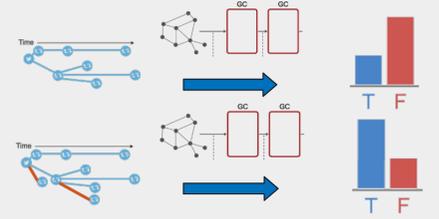


## Motivation

Graph Neural Networks (GNNs) are effective in many graph-related tasks, but are **vulnerable to targeted adversarial attacks**. Under adversarial attack, the victimized samples are perturbed in such a way that they are not easily noticeable, but they lead to **wrong predictions**. This limitation of GNNs has arisen concerns on adopting them in **safety-critical applications**. Our main objective is to **introduce robustness certificates** in our graph classification model.



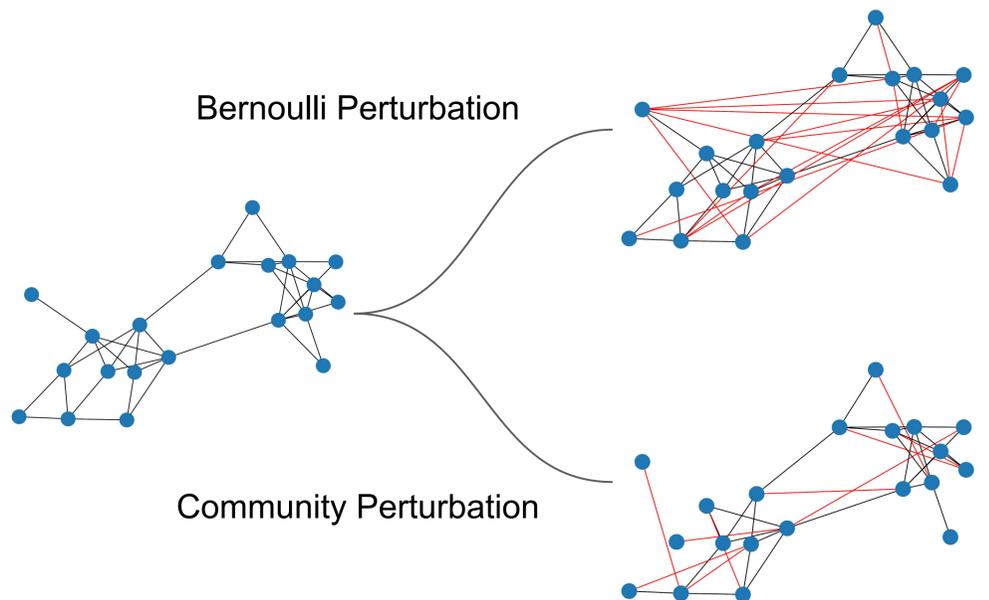
## Framework

- Let  $f : \mathcal{X} \rightarrow \mathcal{Y}$  be a trained classifier, we define a **certified radius**  $R$  as:  

$$f(x) = c \text{ for } c \in \mathcal{Y} \implies \forall \tilde{x} \in \mathcal{B}_{\|\cdot\|_0}(x, R), f(\tilde{x}) = c$$
- We transform our classifier  $f$  into a **smoothed classifier**  $g$  defined as:  

$$g(x) = \arg \max_{c \in \mathcal{Y}} \mathbb{P}_{X \sim \phi(x)}(f(X) = c)$$
- The current classical perturbation  $\phi(x)$  for graph is a **Bernoulli noise** over the edges  

$$\phi(x) = x \oplus \epsilon \text{ with } \forall i \in [N], \epsilon_i \sim \text{Bern}(p)$$
- Given a perturbation  $\phi(x)$  a certified radius  $R$  can be computed for every graph for the smoothed model  $g(x)$



## Method

- We introduce a new perturbation method with edge dependency relying on the **community structure** of the graph:

$$\forall e \in C_{i,j}, P(\phi(x)_e \neq x_e) = p_{i,j}$$

## Result

- We test our new method on a synthetic dataset. Our dataset consists of graphs generated by a Stochastic Block Model (SBM) and Erdős–Rényi model (ER). A GNN is trained on a task consisting in distinguishing the type of graphs.
- We achieve a higher average  $\ell_0$  robust radius for a fixed model accuracy.

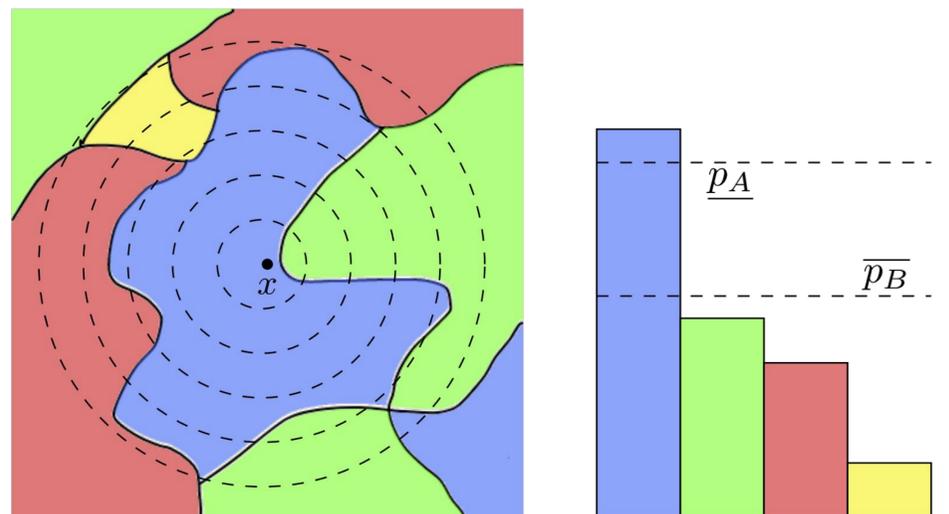
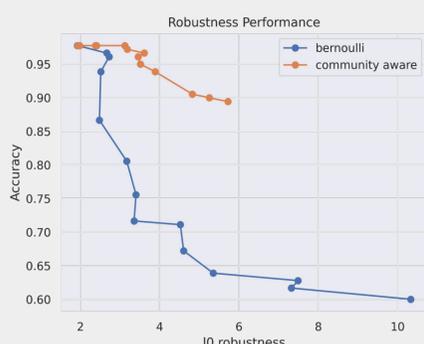
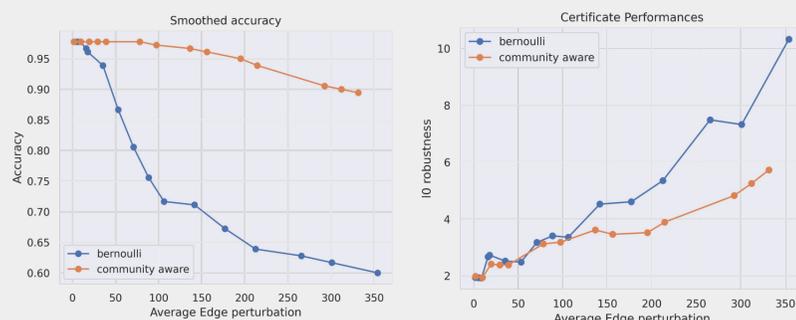


Illustration of a smoothed classifier: at every point  $x$  a neighborhood vote is performed according to a distribution  $\phi(x)$  centered on  $x$ .

## References

- [1] Gao et al, "Certified Robustness of Graph Classification against Topology Attack with Randomized Smoothing." *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020.
- [2] Cohen et al, "Certified adversarial robustness via randomized smoothing." *International Conference on Machine Learning*. PMLR, 2019.
- [3] Bojchevski et al, "Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more." *International Conference on Machine Learning*. PMLR, 2020.