Structure-Aware Robustness Certificates for Graph Classification

Pierre Osselin[*1] **Henry Kenlay**[*1] **Xiaowen Dong**[1]

[1]Department of Engineering Science, University of Oxford, Oxford, UK

Abstract

Certifying the robustness of a graph-based machine learning model poses a critical challenge for safety. Current robustness certificates for graph classifiers guarantee output invariance with respect to the total number of node pair flips (edge addition or edge deletion), which amounts to an l_0 ball centred on the adjacency matrix. Although theoretically attractive, this type of isotropic structural noise can be too restrictive in practical scenarios where some node pairs are more critical than others in determining the classifier's output. The certificate, in this case, gives a pessimistic depiction of the robustness of the graph model. To tackle this issue, we develop a randomised smoothing method based on adding an anisotropic noise distribution to the input graph structure. We show that our process generates structural-aware certificates for our classifiers, whereby the magnitude of robustness certificates can vary across different pre-defined structures of the graph. We demonstrate the benefits of these certificates in both synthetic and real-world experiments.

1 INTRODUCTION

Graph-based machine learning models have made considerable strides in the last couple of years, with applications ranging from NLP [Wu et al., 2023], combinatorial optimization [Drori et al., 2020] and protein function prediction [Gligorijević et al., 2021]. As these tools become more common, studying their vulnerability to potential adversarial examples becomes paramount for safety.

Robustness certification is an active field of research whose goal is to develop certificates guaranteeing invariance of the model prediction with respect to some input perturbations. Diverse methods have been used to achieve this goal, from interval bound propagation [Gowal et al., 2019], convex relaxation [Raghunathan et al., 2018], Lipschitz bounds computation [Huang et al., 2021] or randomised smoothing [Wang et al., 2021]. Given a data point $\mathbf{x}$ and a set of perturbed inputs $\mathcal{B}(\mathbf{x})$, a robustness certificate verifies that a model's prediction $f(\mathbf{x})$ remains unchanged for all other inputs in the perturbation set. That is, for all $\mathbf{x}' \in \mathcal{B}(\mathbf{x})$ it holds that $f(\mathbf{x}) = f(\mathbf{x}')$. Often the set of perturbed inputs $\mathcal{B}(\mathbf{x})$ is parameterised, for example by a closed-ball $\mathcal{B}_r(\mathbf{x}) = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') \leq r\}$ under some distance function d and radius r. In this case, we are interested in knowing the largest r that we can certify for, where r is called the certified radius.

In the context of robustness certification of graph classifiers against structural perturbation, a common choice of perturbation set is the set of all graphs reachable from an input graph $\mathbf{x}$ by up to r node pair flips (edge additions and deletions)[1]. This corresponds to a closed ball on the upper triangle entries of the adjacency matrix where the distance is induced by the ℓ_1 norm and the bottom triangle entries are determined by the constraint that the adjacency matrix is symmetric (assuming for simplicity the graph is unweighted and undirected). In some cases, however, different node pairs of the graph can be more predictive of the ground truth label than others. A real-world example is classification of molecular structures, where the edges that constitute key substructure (e.g., a ring) are more critical in determining the class label than the rest. A synthetic example is further presented in Fig. 1. In such situations, certifying according to a total number of edge additions or deletions might gives a pessimistic certified radius, because the set of perturbed inputs may include perturbations which consist of flipping many critical node pairs (in terms of determining the graph

[*] Equal contribution.

[1] We use the terminology of node pair flip instead of edge flip to emphasise that we are considering the addition of edges that do not exist in the original graph as well as the deletion of existing edges.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

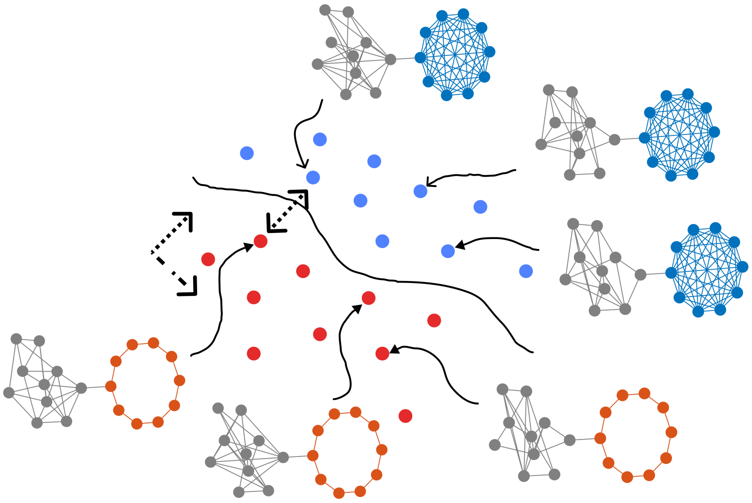


Figure 1: An example of graph classification where the class label is determined by the blue or red part of the input graph, but not the grey part. In this case, perturbations to the former will be more likely to affect classification outcome than the latter. In other words, the classification decision boundary is sensitive to some perturbations to the input graph structure (which move the point towards the boundary), but robust to others (which keep the point away from the boundary).

label).

In this work, we address this problem by defining disjoint regions of node pairs and proposing robustness certificates that verify that the prediction of the classifier will not change for a potentially different number of node pair flips for each region. Such disjoint regions may be either obtained via domain knowledge (as in the molecule example mentioned above) or assigned based on the level of importance of node pairs to the classification task. Our approach then relies on randomised smoothing, which is a powerful framework to produce robustness certificates which hold with high probability. Given some noise distribution over the input, randomised smoothing transforms a base model f into a smoothed model g for which we can provide probabilistic robustness guarantees.

Existing randomised smoothing approaches for certifying graph classification mostly consider an isotropic noise distribution that flips each node pair with a fixed probability [Jia et al., 2020, Gao et al., 2020, Wang et al., 2021]. The certificate in this case corresponds to the total number of node pair flips. Instead, we propose using an anisotropic noise distribution based on the predefined regions whereby the probability of flipping a node pair depends on to which region (if any) the node pair belongs. We show that smoothed classifiers constructed using this anisotropic noise distribution naturally lead to structure-aware robustness certificates whereby different numbers of node flips are certified for each of the regions. We demonstrate the benefits of our approach on both synthetic and real-world experiments [2]. To the best of our knowledge, our method is one of the first of its kind that allows for flexible and structure-aware graph certification in the input graph domain.

2 PRELIMINARIES

Let $\mathcal{X}$ be a data space and $f : \mathcal{X} \to \mathcal{Y}$ be a classifier which maps each point $\mathbf{x} \in \mathcal{X}$ to a label $y \in \mathcal{Y}$. Let $\phi : \mathcal{X} \to \mathcal{P}(\mathcal{X})$ be a noise distribution over our data, that is, $\phi(\mathbf{x})$ returns a distribution over $\mathcal{X}$ for every point $\mathbf{x} \in \mathcal{X}$. We write $f(\phi(\mathbf{x}))$ to denote the random variable $\mathbb{P}_{\mathbf{z}\sim\phi(\mathbf{x})}(f(\mathbf{z}))$. This represents the distribution of outputs of the base classifier given the randomisation scheme applied to the input. We define g to be the smoothed classifier of our base classifier f as

$$g(\mathbf{x}) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}(f(\phi(\mathbf{x})) = y). \qquad (1)$$

The smoothed classifier can be interpreted as a neighbourhood vote, where the output is the mode of the output of the classifier when inputs are sampled from the distribution $\phi(\mathbf{x})$. An illustration of a smoothed classifier is given in Fig. 2.

2.1 CERTIFYING A SMOOTHED CLASSIFIER

We can construct a lower and upper bound $\underline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p,y)$ and $\overline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p,y)$ on $\mathbb{P}(f(\phi(\tilde{\mathbf{x}})) = y)$, which is the probability of class y under our classifier f smoothed by ϕ and evaluated on an arbitrary point $\tilde{\mathbf{x}} \in \mathcal{X}$:

$$\underline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p,y) = \min_{\substack{h\in\mathcal{H}:\\ \mathbb{P}(h(\phi(\mathbf{x}))=y)=p}} \mathbb{P}(h(\phi(\tilde{\mathbf{x}})) = y) \qquad (2)$$

[2] Code to reproduce our numerical experiments can be found at https://github.com/pierreosselin/structureaware.git

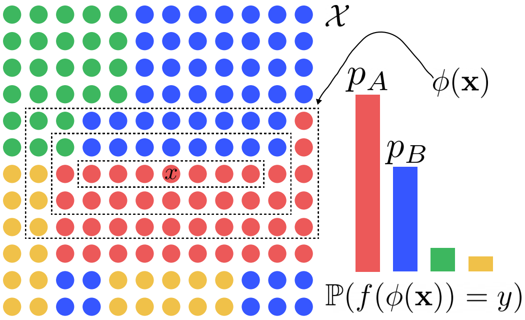


Figure 2: Illustration of a smoothed classifier: at every point $\mathbf{x}$ a neighborhood vote is performed according to a distribution $\phi(\mathbf{x})$ centered on $\mathbf{x}$. The figure was inspired by Günnemann [2020].

$$\overline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p,y) = \max_{\substack{h\in\mathcal{H}:\\ \mathbb{P}(h(\phi(\mathbf{x}))=y)=p}} \mathbb{P}(h(\phi(\tilde{\mathbf{x}}))=y) \tag{3}$$

In this definition, $\mathcal{H}$ is the set of measurable classifiers with respect to ϕ. Because the optimisation constraints include the base classifier $f \in \mathcal{H}$ it follows that

$$\underline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p,y) \leq \mathbb{P}(f(\phi(\tilde{\mathbf{x}}))=y) \leq \overline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p,y). \tag{4}$$

We define a perturbation set $\mathcal{B}_r(\mathbf{x})$ as a family of sets $\mathcal{B}_r(\mathbf{x}) \subseteq \mathcal{X}$ parameterised by some $r \geq 0$ such that $\mathcal{B}_r(\mathbf{x}) \subseteq \mathcal{B}_{r'}(\mathbf{x})$ if and only if $r \leq r'$. This includes open or closed balls with respect to a metric over $\mathcal{X}$. We say that the smoothed classifier g is certified at $\mathbf{x}$ in some perturbation set $\mathcal{B}_r(\mathbf{x})$ if the output of g is the same for all neighbouring points $\tilde{\mathbf{x}} \in \mathcal{B}_r(\mathbf{x})$ in the perturbation set.

Following Cohen et al. [2019], we will write c_A to be the output class of $g(\mathbf{x})$ which is returned with probability p_A, and c_B to be the "runner-up" class, i.e., the class distinct from c_A with the next highest probability p_B. From the framework of Lee et al. [2019], we define the notion of a point-wise certificate around a point $\mathbf{x}$ by verifying if the following holds:

$$\min_{\tilde{\mathbf{x}}\in\mathcal{B}_r(\mathbf{x})} \Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A) > 0, \tag{5}$$

where

$$\Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A) \triangleq \underline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_A, c_A) - \overline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_B, c_B). \tag{6}$$

Eq. 6 can be thought of as a margin, which gives the difference between a lower bound on p_A and an upper bound on p_B for some point $\tilde{\mathbf{x}}$ in the perturbation set. Eq. 5 then indicates if this property holds for all $\tilde{\mathbf{x}}$ in the perturbation set.

2.2 CERTIFIED RADIUS

We can define the certified radius to be the largest value of r so that we can certify with respect to the predefined perturbation set $\mathcal{B}_r(\mathbf{x})$. Formally this can be defined as:

$$R(\mathbf{x}) = \sup r, \text{ s.t. } \min_{\tilde{\mathbf{x}}\in\mathcal{B}_r(\mathbf{x})} \Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A) > 0. \tag{7}$$

If the perturbation set is an open or closed ball, $R(\mathbf{x})$ is the radius of the largest ball we can certify over.

2.3 COMPUTING THE CERTIFICATE

2.3.1 Computing a point-wise certificate

For a fixed $\mathbf{x}$ and neighbouring point $\tilde{\mathbf{x}}$, we can compute Eq. 7 following the method described by Lee et al. [2019]. First, we partition the space $\mathcal{X} = \bigcup_i \mathcal{R}_i$ into disjoint regions of equal likelihood ratios $\mathcal{R}_k = \{\mathbf{z} \in \mathcal{X} : \frac{\mathbb{P}(\phi(\tilde{\mathbf{x}})=\mathbf{z})}{\mathbb{P}(\phi(\mathbf{x})=\mathbf{z})} = \eta_k\}$ where without loss of generality we can assume $\eta_k \in \mathbb{R}$ are in an ascending order. The quantity $\Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A)$ can then be computed by solving the following linear programming (LP) problems [Lee et al., 2019, Lemma 2]:

$$\underline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_A, c_A) = \min_{\mathbf{h}} \mathbf{h}^T\tilde{\mathbf{r}} \quad \text{s.t.} \quad \mathbf{h}^T\mathbf{r} = p_A, \quad 0 \leq \mathbf{h} \leq 1 \tag{8}$$

and similarly,

$$\overline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_B, c_B) = \min_{\mathbf{h}} \mathbf{h}^T\tilde{\mathbf{r}} \quad \text{s.t.} \quad \mathbf{h}^T\mathbf{r} = p_B, \quad 0 \leq \mathbf{h} \leq 1. \tag{9}$$

The variable $\mathbf{h}$ corresponds to optimising over the classifiers that are optimised over in Eq. 2 and Eq. 3. The vectors $\mathbf{r}$ and $\tilde{\mathbf{r}}$ are $\mathbf{r}_i = \mathbb{P}(\phi(\mathbf{x}) \in \mathcal{R}_i)$ and $\tilde{\mathbf{r}}_i = \mathbb{P}(\phi(\tilde{\mathbf{x}}) \in \mathcal{R}_i)$ respectively. This LP problem can be solved via a greedy approach [Lee et al., 2019]. Given the ratios η_k are ordered in an ascending manner, starting with $\mathbf{h} = \mathbf{0}$ we can assign $\mathbf{h}_i = 1$ for the indices $i = 1, \ldots, k-1$ such that we choose the largest k with $\mathbf{h}^T\mathbf{r} \leq p_A$, and set the value of $\mathbf{h}_k$ such that $\mathbf{h}^T\mathbf{r} = p_A$. We can solve Eq. 9 in a similar way.

Given this efficient way to compute a certificate, there remain some quantities that must be calculated. For our certificate, we introduce the methods to compute them in Section 3. The first is partitioning the space $\mathcal{X}$ into disjoint unions of equal likelihood ratios. We provide a method to do so in Proposition 1. Next, the values η_k must be computed, or at least given in closed form, so the regions can be sorted in ascending order. Proposition 2 gives an analytic closed-form. Finally, a closed form for $\mathbb{P}(\phi(\mathbf{x}) = \mathbf{z})$ allows us to compute $\mathbf{r}$. We can compute $\tilde{\mathbf{r}}$, required to solve the linear programs, by noticing that $\mathbb{P}(\phi(\tilde{\mathbf{x}}) = \mathbf{z}) = \eta_k\mathbb{P}(\phi(\mathbf{x}) = \mathbf{z})$. We compute $\mathbb{P}(\phi(\mathbf{x}) = \mathbf{z})$ in Proposition 3.

2.3.2 Computing a regional certificate

To certify over some allowed set $\mathcal{B}_r(\mathbf{x})$ we need to find the worst case $\tilde{\mathbf{x}} \in \mathcal{B}_r(\mathbf{x})$ to solve Eq. 5. Through particular choices of parameterising $\tilde{\mathbf{x}}$, one can find equal likelihood regions that give rise to equal likelihood ratios values that are independent of the exact value of $\tilde{\mathbf{x}}$, turning the pointwise certificate into a regional certificate, where the region is given by specific parameterisation. We give such an example in Proposition 2.

2.4 RANDOMISED SMOOTHING FOR GRAPH CLASSIFICATION

Although our method is applicable to any setting with discrete data, we are interested in the robustness of graph classification models to structural perturbation of undirected, unweighted graphs[3]. In this case, the input domain is the set of finite graphs $\mathcal{X} = \cup_{i=1}^{\infty} \mathcal{X}_n$ where $\mathcal{X}_n$ is the space of graphs with n nodes. We can represent a graph with n nodes as a binary vector of size $\binom{n}{2}$ where each entry indicates the presence or absence of an edge. Without loss of generality we will treat graphs as binary vectors $\mathbf{x}$ from here on in[4].

We are interested in robustness with respect to node pair flips which can represent an edge addition (change a zero to a one in $\mathbf{x}$) or edge deletion (change a one to a zero). This may be interpreted as adding "structural noise" to the input graph. Two candidates for the distribution of such noise have been proposed in the literature. The first one applies an independent Bernoulli distribution to every node pair. That is, $\phi(\mathbf{x})_i = \mathbf{x}_i \oplus \epsilon_i$ with $\epsilon_i \sim \mathrm{Bern}(p)$ for a fixed p, where $\oplus$ is the bitwise XOR operator. We refer to this noise distribution as isotropic, as it flips each node pair with equal probability. This approach is used by Jia et al. [2020], Wang et al. [2021], Gao et al. [2020]. The second approach, a sparsity-aware noise distribution [Bojchevski et al., 2020], gives a different probability of edge flipping depending on whether the edge is present in the graph. If an edge exists between a node pair then it is flipped with probability p_-, whereas if a node pair does not exist between two nodes it is added with probability p_+. This distribution can be written as $P(\phi(\mathbf{x})_i \neq \mathbf{x}_i) = p_-^{\mathbf{x}_i} p_+^{1-\mathbf{x}_i}$. The proposed framework is conceptually similar to this latter case; however, we consider a generic partition of node pairs into one of many node pair sets and these pairs are perturbed according to their membership of the sets. This leads to the derivation of robustness certificates that are aware of the structural characteristics of the input graph with respect to the classification labels.

3 RANDOMISED SMOOTHING WITH ANISOTROPIC NOISE

Given $\mathbf{x} \in \mathcal{X}_n$, suppose we divide our input space of node pairs up into disjoint regions $\bigsqcup_{i \in I} \mathcal{C}_i$ such that each node pair belongs to exactly one region and there are a total of C regions. We define a noise distribution where independent Bernoulli distributions are applied to every node pair, and where the parameter of the Bernoulli distribution is shared within every set $\mathcal{C}_i$:

$$\phi(\mathbf{x})_k = \mathbf{x}_k \oplus \epsilon_k, \text{ where } \epsilon_k \sim \mathrm{Bern}(p_i) \text{ and } k \in \mathcal{C}_i, \quad (10)$$

Let $\mathbf{R} \in \mathbb{Z}^C$ be a tuple of integers such that $0 \leq \mathbf{R}_i \leq |\mathcal{C}_i|$ and let $\mathcal{B}_{\mathbf{R}}(x) = \{\mathbf{z} \in \mathcal{X}_n : \|\mathbf{z}_{\mathcal{C}_i} - \mathbf{x}_{\mathcal{C}_i}\|_0 \leq \mathbf{R}_i\}$ be a perturbation set. Let $\tilde{\mathbf{x}} \in \mathcal{B}_{\mathbf{R}}(x)$ and $\mathcal{J} = \{i : \mathbf{x}_i \neq \tilde{\mathbf{x}}_i\}$ be indices of $\mathbf{x}$ which are perturbed to give $\tilde{\mathbf{x}}$. Furthermore, let $\mathcal{J}_i = \mathcal{J} \cap \mathcal{C}_i$ be indices where $\mathbf{x}$ is perturbed in collection $\mathcal{C}_i$. In our case a maximum radius means maximizing the radius on every regions $\mathcal{C}_i$.

First, we develop a set decomposition on which likelihood ratios can be computed.

Proposition 1 *We define regions $\mathcal{R}_{\mathbf{Q}} = \{\mathbf{z} \in \mathcal{X}_n : \|\mathbf{z}_{\mathcal{J}_i} - \mathbf{x}_{\mathcal{J}_i}\|_0 = Q_i : i \in I\}$ that represent points $\mathbf{z}$ which agree with $\mathbf{x}$ by exactly Q_i bits in sub-regions $\mathcal{J}_i$. Then $\mathcal{X}_n$ can be represented by the following disjoint union*

$$\bigcup_{\mathbf{0} \leq \mathbf{Q} \leq \mathbf{R}} \mathcal{R}_{\mathbf{Q}}, \quad (11)$$

where vector inequalities are element-wise.

Next, the likelihood ratio is fixed for elements $\mathbf{z}$ in any one region. This likelihood ratio has the following closed form.

Proposition 2 *Consider a region $\mathcal{R}_{\mathbf{Q}} = \{\mathbf{z} \in \mathcal{X}_n : \|\mathbf{z}_{\mathcal{J}_i} - \mathbf{x}_{\mathcal{J}_i}\|_0 = Q_i\}$ then for all $\mathbf{z} \in \mathcal{R}_{\mathbf{Q}}$ the following holds*

$$\eta_{\mathbf{Q}}^{\mathcal{R}} = \frac{\mathbb{P}(\phi(\tilde{\mathbf{x}}) = \mathbf{z})}{\mathbb{P}(\phi(\mathbf{x}) = \mathbf{z})} = \prod_{i=1}^{C} \left(\frac{1-p_i}{p_i}\right)^{R_i - 2Q_i}. \quad (12)$$

Finally, we can compute the likelihood of a smoothed input belonging to these regions:

Proposition 3 *The probability of the output of a smoothed input $\phi(\mathbf{x})$ belonging to a region $\mathcal{R}_{\mathbf{Q}}$ is given by*

$$\mathbb{P}(\phi(\mathbf{x}) \in \mathcal{R}_{\mathbf{Q}}) = \prod_{i=1}^{C} \mathrm{Bin}(R_i - Q_i | R_i, p_i), \quad (13)$$

where $\mathrm{Bin}(R_i - Q_i|R_i, p_i)$ is the probability mass function of the binomial distribution giving probability of $R_i - Q_i$ successes from R_i trials each with success probability p_i.

All proofs are provided in Appendix 1. Using these results we can provide robustness certificates of the smoothed classifier. Given $\mathbf{x} \in \mathcal{X}$ and our noise distribution, we can compute the values p_A and p_B. In practice, these quantities are not available in closed form and are estimated via

[3] This work can be extended to the setting of directed graphs as well as the task of node classification.

[4] Note that due to isomorphism multiple different binary vectors can represent the same graph.

sampling, as in Bojchevski et al. [2020]. A more detailed description is given in Appendix 2, which gives probabilistic certificates according to some confidence intervals. Without loss of generality we order the regions as $\mathcal{R}_1, \ldots \mathcal{R}_T$ (where $T = \prod_i (R_i + 1)$). The corresponding ratios $\eta_{\mathbf{Q}}^{\mathcal{R}}$ as given by Proposition 2 are ordered as $\eta_1 \leq \ldots \leq \eta_T$. From these elements, $\Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A)$ can be computed through Eq. 6 and the optimisation problem Eq. 7 can be solved. In practice, the optimisation of Eq. 7 can be solved efficiently by leveraging symmetries displayed by $\Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A)$ when $\tilde{\mathbf{x}}$ varies. This property is more thoroughly described in Appendix 2. A complete description of the algorithm to compute the proposed certificate is provided in Appendix 3, along with its computational complexity.

4 RELATED WORK

Robustness certificate. The earliest work proposing a robustness certificate for graph-based models is Zügner and Günnemann [2019], where the authors consider semi-supervised node classification over graphs with binary attributes using graph neural networks. The admissible perturbations are those bounded under the ℓ_0 semi-norm. Following the approach of Wong and Kolter [2018], a convex relaxation is considered and a dual problem is solved via linear programming. Other early works that consider certificates under topological change include Bojchevski and Günnemann [2019] who propose a certificate for a class of linear graph neural networks based on PageRank diffusion. Jin et al. [2020] propose the first certificate for graph classifiers which consist of a single graph convolutional layer followed by a pooling layer and a final linear layer. A convex relaxation of the adversarial polytope based on Lagrange duality is considered. Although the computed certificates are exact, the framework is specific to this simple graph classifier model and is expensive to compute. Recently, Jin et al. [2022] consider certifying graph classifiers under a different admissible perturbation measure, i.e., the orthogonal Gromov-Wasserstein discrepancy.

Randomised smoothing based approaches. Randomised smoothing was originally proposed by Lecuyer et al. [2019] and Singla and Feizi [2020] in the context of image classification. Cohen et al. [2019] study this approach further, giving a tight certificate for perturbations bounded by an ℓ_2 norm and improving the scalability of the approach. They also outline some disadvantages of randomised smoothing, such as the need for high levels of noise or a substantial number of samples to certify to a large certified radius. Another consideration of randomised smoothing is their accuracy-robustness trade-off, controlled by the level of noise.

Existing randomised smoothing approaches for certifying graph-based classification models consider certificates for the number of edge flips. This is achieved by flipping an edge between each node pair independently with the same probability p. Jia et al. [2020] apply this to community detection algorithms, whereas Gao et al. [2020] and Wang et al. [2021] apply the same principles to graph classification. The closest method from our work is the sparsity-aware certificate proposed in Bojchevski et al. [2020]. The authors note that due to the sparse nature of many real-world graphs, adding and deleting edges between node pairs with the same probability leads to many more edges being added than deleted. To account for this, the authors propose using different probabilities for adding and deleting edges. This method is different from ours as the Bernoulli probabilities change with the input graph whereas ours depend on predefined sets of node pairs. These sets in our case account for varying levels of importance of the node pairs in predicting a graph label.

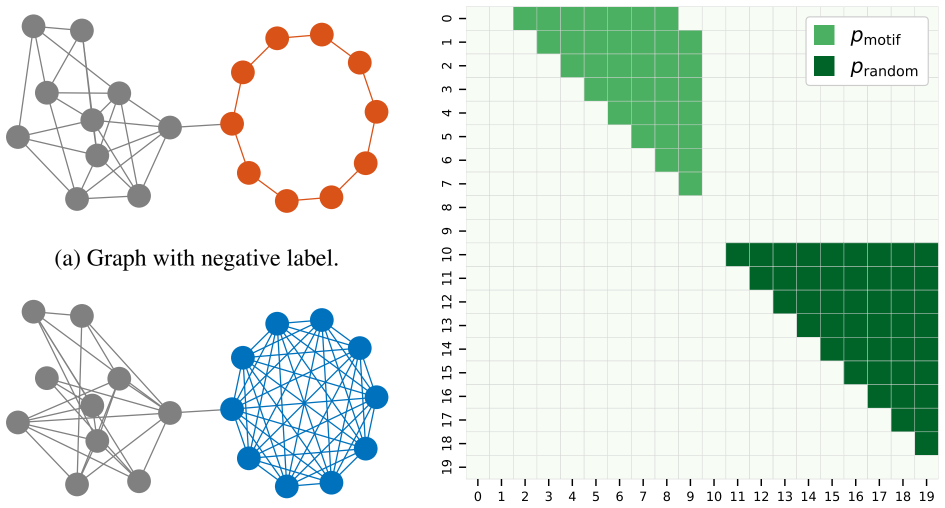
5 EXPERIMENTS

5.1 SYNTHETIC EXPERIMENT

We motivate the use of anisotropic noise by considering inputs $\mathbf{x}$ that are an element of some space $\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2$ where $\oplus$ is the direct sum. Consider a point that is close to the decision boundary in the $\mathcal{X}_1$ subspace but far from the decision boundary in the $\mathcal{X}_2$ subspace as illustrated in Fig. 1. An isotropic certificate cannot certify beyond the small distance to the decision boundary in $\mathcal{X}_1$. However, by design our certificate can certify the distances of $\mathcal{X}_1$ and $\mathcal{X}_2$ jointly allowing us to certify a small distance in the $\mathcal{X}_1$ subspace but a large distance in the $\mathcal{X}_2$ subspace.

We design a synthetic graph classification data set whereby the graphs are constructed using a motif which determines the class label (corresponding to the important subspace $\mathcal{X}_1$) connected to a randomly generated graph (corresponding to the unimportant subspace $\mathcal{X}_2$) which is independent of the class label. We can consider edges in the motif part to be in one node pair set $\mathcal{C}_{\text{motif}}$ and edges from the random part in $\mathcal{C}_{\text{random}}$. Given a model that solves this task, we would expect changes in the motif to move the input closer or further away from a decision boundary but changes in the random part of the graph to move the point parallel to the direction of the decision boundary. In other words, we should be able to certify a large number of node pairs in $\mathcal{C}_{\text{random}}$ by applying a large value of noise p_{random} without hurting the accuracy of the smoothed classifier. We cannot certify a large number of node pairs in $\mathcal{C}_{\text{motif}}$ without a drop in accuracy, so we choose a small value of noise p_{motif} to retain high accuracy.

More concretely, we generate a binary classification problem where each graph has a motif part of $n_{\text{motif}} = 10$ nodes where a cycle determines a negative label and a complete graph determines a positive label. We then generate a random part using a connected Erdős-Rényi graph [Gilbert, 1959] with $n_{\text{random}} = 10$ nodes and parameter $p = 0.5$. We join these graphs using a single edge. See Fig. 3a for an



(a) Graph with negative label.

(b) Graph with positive label.

(c) Noise matrix.

Figure 3: Example graphs with a positive and negative label. Blue nodes and edges denote a motif part of a positive label and red nodes and edges denote a motif part of a negative label. The noise matrix show how edges are perturbed with p_{motif} being the noise parameter for node pairs in the motif part and p_{random} being the noise applied to node pairs in the random part. Notice that only the internal edges of the motif are perturbed, and the bridge edge is not perturbed. Only the upper triangle of the noise is shown only; in practice this is sampled and applied to the upper and lower triangle of the adjacency matrix so the graph remains undirected.

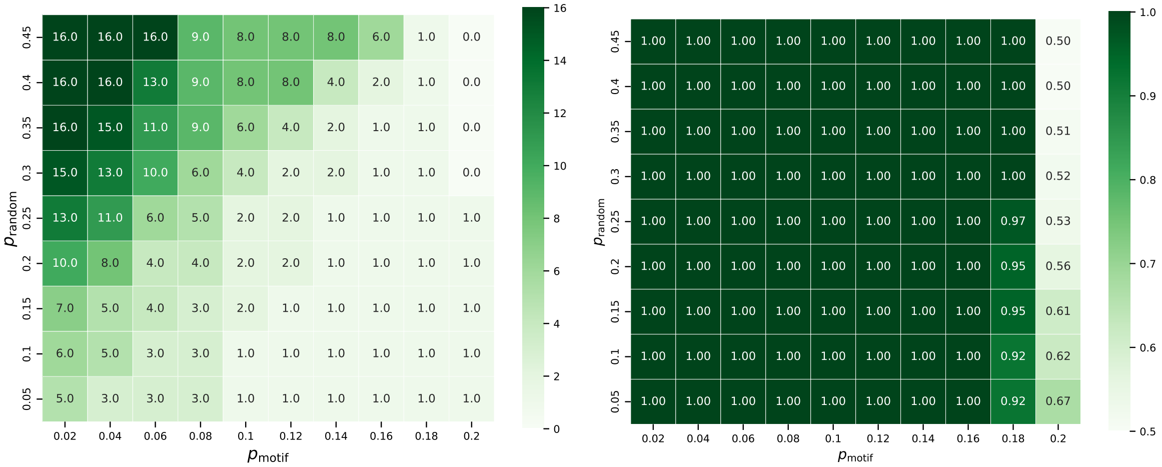
example of the negative class and Fig. 3b for an example of the positive class.

We generate balanced train, validation and test sets of size 1000, 1000, and 100 respectively. The test set is smaller as the randomised smoothing procedure is computationally expensive. This is because to estimate p_A a large number of random inputs need to be generated and classified using the model. We train as a base classifier an SVM using a node label histogram kernel [Sugiyama and Borgwardt, 2015] where the node label corresponds to the node degree. Let $c(\mathcal{G}, d)$ be a function that counts the number of nodes in a graph $\mathcal{G}$ with degree d. Then the kernel applied to graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ can be written as $\kappa(\mathcal{G}_1, \mathcal{G}_2) = \sum_{d=0}^{\infty} c(\mathcal{G}_1, d) \cdot c(\mathcal{G}_2, d)$ which is well defined for finite graphs. We use this model as we expect it to be sensitive to the motif structure that determines the label; the negative label gives a large value in the $c(\cdot, 2)$ component whereas the positive label gives a large value in the $c(\cdot, n_{\mathrm{motif}} - 1)$ component. Indeed, the base classifier gets 100% accuracy on the train, validation and test data sets. We present further results for different choices of kernel in Appendix 4.

For the certification procedure, we apply noise separately for node pairs in the motif part and node pairs in the random part. We apply noise to the internal edges of the motif part only (and not to those of the red cycle, see Fig. 3). We do not apply noise to node pairs where one node lies in the motif and one does not. The noise matrix is shown in Fig. 3c. We generate $100,000$ perturbations per test sample and use these to estimate the output of the smoothed classifier and generate a certificate. We experiment with varying the number of perturbations in Appendix 4. We use a confidence level of $\alpha = 0.99$ to estimate p_A. We also compute certificates using isotropic noise for comparison.

For our certificate with anisotropic noise we consider $\mathbf{p} = (p_{\mathrm{motif}}, p_{\mathrm{random}}) \in \mathbf{P}$ where $\mathbf{P} = \{0.02, 0.04, \ldots, 0.2\} \times \{0.05, 0.1, \ldots, 0.45\}$. Recall that p_{motif} is the noise parameter for the motif part and p_{random} is the noise parameter for the random part. For the isotropic certificate we consider $p \in \{0.02, 0.04, \ldots, 0.2\}$. For the anisotropic certificate, we certify over perturbation pairs $\mathbf{r} = (r_{\mathrm{motif}}, r_{\mathrm{random}})$ which means that with high probability r_{motif} edge flips in the motif part and r_{random} edge flips in the random part will not change the label of the smoothed classifier. This is different from the isotropic certificate which guarantees the label does not change for r edge flips anywhere in the graph.

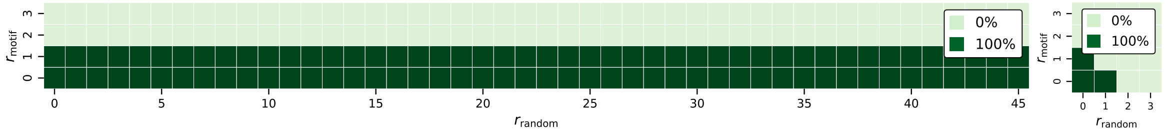
We first analyze how the noise vector $\mathbf{p}$ influences the certificates. We introduce a score to evaluate the noise parameters. For each $\mathbf{r}$ if we can certify strictly more than half of the test samples in this perturbation space then we add 1 to the score. To motivate the utility of this score, consider a smoothed model with large values of noise. In the limit, a perfectly smooth classifier will be constant everywhere. In other words, it will predict the same label for all inputs and give a certified accuracy of 50% for a balanced binary classification task. This classifier could be certified for arbitrary



(a) Score.

(b) Test-set accuracy.

Figure 4: The test-set accuracy of the smoothed classifier and the certified volume for various values of $\mathbf{p} = (p_{\text{motif}}, p_{\text{random}})$.



(a) Anisotropic.

(b) Isotropic.

Figure 5: Comparison of an anisotropic certificates with $\mathbf{p} = (0.02, 0.45)$ and isotropic certificates for various levels of p. We omit some values of p for the isotropic certificate for readability.

numbers of edge flips. For this reason, metrics such as total certified area averaged over samples do not necessarily tell us if a noise parameter is useful.

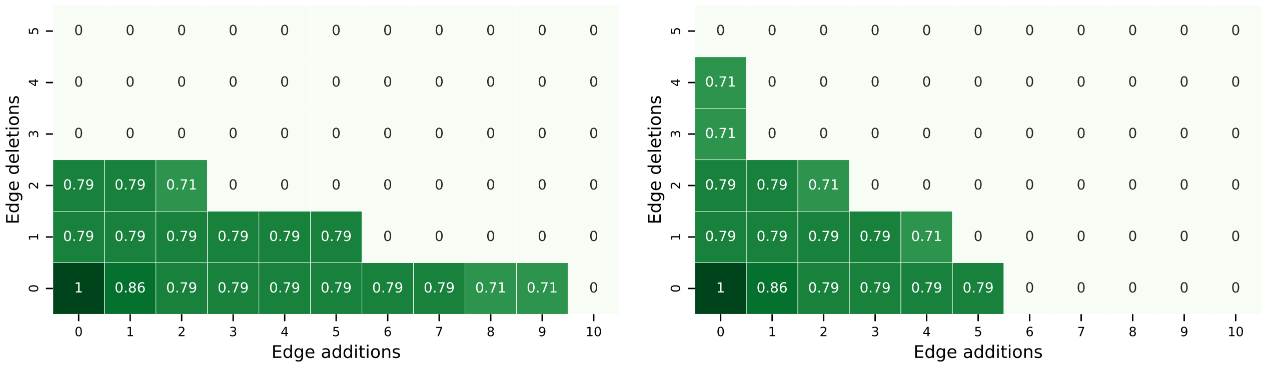
Fig. 4a shows our smoothed classifier has the highest score when p_{motif} is small and p_{random} is large. Fig. 4b shows large values of p_{random} does not effect the accuracy of the smoothed classifier, but if p_{motif} becomes too large the accuracy begins to drop. These results are expected: p_{motif} cannot be too large as the motif part is more important to determining the label. This motivates to fix p_{motif} to be small and increase p_{random}, retaining high test accuracy whilst increasing the number of edge flips we can certify in $\mathcal{C}_2$.

We take a closer look at the smoothed classifier given by $\mathbf{p} = (0.02, 0.45)$, one of the smoothed classifiers with the highest observed score. We are interested in a smooth model with high test set accuracy that can certify for many values of $\mathbf{r}$. Our model has 100% certified accuracy. The proportion of the test set that can be certified for varying values of $\mathbf{r}$ is shown in Fig. 5a. As the figure demonstrates we can certify 100% of the test-set samples to 0 or 1 edge flips in the motif part of the graph and up to 45 edge flips in the random part of the graph. This is the maximum number of possible node pairs in the random part, so we can certify any perturbation in this part of the graph. The smoothed classifier using isotropic noise can also achieve 100% test set accuracy for all values of noise we tested. We show the certification results for when $p = 0.02$, as this is the only value that allows us to certify the entire test-set for one edge flip. We plot the proportion of the test set that can be certified at using this value of isotropic noise in Fig. 5b. Using larger values of noise for the isotropic certificate allows for some of the test-set to be certified at a radius of 2, but it can no longer certify the entire test set at radius 1. By using anisotropic noise and being specific about where edges are being certified, we can certify 46 edge flips with 100% accuracy compared to 1 edge flip with 100% accuracy.

5.2 REAL-WORLD EXPERIMENT

We also experiment using the real-world data set MUTAG [Debnath et al., 1991]. In this data set each graph represents a molecule and the goal is to predict the molecules mutagenicity on a specific bacterium, which is encoded into a binary label. Each node has one of 7 discrete node labels corresponding to atomic number which is one-hot encoded. The data set contains a total 188 molecular graphs.

We train a base classifier as a graph neural network which has a single GCN layer [Welling and Kipf, 2017] with 64 hidden units, followed by a max pooling layer and a linear



(a) Anisotropic.

(b) Sparsity-aware.

Figure 6: A comparison between the anisotropic certificate and the sparsity-aware certificate. Each entry represents the ratio of correctly classified test-set samples that could be certified at a specified number of edge deletions and additions.

layer. We use a $80\%/10\%/10\%$ train/validation/test split on which to train and optimise the model hyperparameters. We train for a maximum of 500 epochs using the AdamW optimiser [Loshchilov and Hutter, 2019] with weight decay of 10^{-3}. The initial learning rate is 10^{-3} and it is decayed by 0.5 every 50 epochs.

We compare our certificate to Bojchevski et al. [2020], referred here as a sparsity-aware certificate, as this is the only non-isotropic certificate used for graph classification that we are aware of. We consider node pairs where there is an edge in the original graph as $\mathcal{C}_1$ and all other node pairs as $\mathcal{C}_2$. In this scenario, we can certify edge deletions and additions in a comparable way. Following the setup described in Bojchevski et al. [2020] we consider $p_1 = 0.4$ which corresponds to the probability of deleting an edge and $p_2 = 0.2$ which corresponds to the probability of adding an edge. We apply noise during training to make the model more robust.

The values computed in Proposition 2 differ between the two approaches as $\mathbb{P}(\phi(\tilde{\mathbf{x}}) = \mathbf{z})$ is computed differently. Furthermore, the probability $\phi(\mathbf{x})$ belonging to a region in the anisotropic approach is a product of Binomial distributions (Proposition 3) whereas for the sparsity-aware certificate this probability follows a Poison-Binomial distribution. If the assignment of node pairs was dependent on the individual sample, this would generalise our approach further, and would also generalise the sparsity-aware certificate.

Our model has a test-set accuracy of 84%. In Fig. 6 we plot the ratio of correctly predicted test points that are certified for varying numbers of edge deletions and additions. We make a few observations from these results. The first is that for values of $\mathbf{r}$ where both methods can certify test points, our method certifies the same quantity of points and in some cases more. The second is that there are two values of $\mathbf{r}$ where the sparsity-aware certificate can certify test samples but the anisotropic certificate cannot. However, there are five values of $\mathbf{r}$ where the anisotropic can certify but the sparsity-aware certificate cannot. Finally, we note that in this experiment, as well as the synthetic experiments, we find our certificates tend to be oblong, i.e. if p_i is larger than we tend to certify for larger values in the r_i direction. This is advantageous in the case where some node pairs are considered more important to the classification label (as demonstrated in the synthetic experiment).

6 CONCLUSIONS

We propose in this work, to the best of our knowledge, one of the first methods that introduces structure-aware robustness certificates in the context of classifying undirected, unweighted graphs. To achieve this, we leverage a flexible, anisotropic noise distribution in the framework of randomised smoothing and develope an efficient algorithm to compute certificates. We apply these certificates to a synthetic experiment and demonstrate a clearly improved robustness of graph classifiers that cannot be achieved with an isotropic certificates. We also validate our certificate on real-world experiments and show superior results compared to an existing approach.

Our approach requires defining a priori which edges a user believes to be more or less important to determining the graph label. Such knowledge may come from domain expertise (which we simulate in the synthetic experiment), or we may treat edge deletions and additions differently as in the sparsity-aware approach of Bojchevski et al. [2020]. We may also consider approaches that have been used to identify edges that may be vulnerable to attack. For example, a previous work found edges vulnerable to adversarial attack are those not captured by a low-rank approximation of the adjacency [Entezari et al., 2020]. Another line of work propose that edges where the end-point node features have low Jaccard index are potentially vulnerable [Wu et al., 2019]. Beyond this, one could learn the importance of the node pairs in a data-driven fashion. One such example is the recent work of Sui et al. [2022] where the authors propose to learn the causal relations between node pairs and model outcome. We leave these directions for future work. Finally, even though we have applied our method in the context of graph classification, it can also used for any type of task based on a discrete domain such as binary image classification.

Acknowledgements

P.O and H.K acknowledge support from the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (EP/L015897/1). X.D. acknowledges support from the Oxford-Man Institute of Quantitative Finance and the EPSRC (EP/T023333/1).

References

Aleksandar Bojchevski and Stephan Günnemann. Certifiable robustness to graph perturbations. In *Advances in Neural Information Processing Systems*, pages 8319–8330, 2019.

Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. *International Conference on Machine Learning*, pages 1003–1013, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

Iddo Drori, Anant Kharkar, William R Sickinger, Brandon Kates, Qiang Ma, Suwen Ge, Eden Dolev, Brenda Dietrich, David P Williamson, and Madeleine Udell. Learning to solve combinatorial optimization problems on real-world graphs in linear time. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 19–24, 2020.

Negin Entezari, Saba A Al-Sayouri, Amirali Darvishzadeh, and Evangelos E Papalexakis. All you need is low (rank) defending against adversarial attacks on graphs. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 169–177, 2020.

Zhidong Gao, Rui Hu, and Yanmin Gong. Certified robustness of graph classification against topology attack with randomized smoothing. *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6, 2020.

Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):1–14, 2021.

Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4842–4851, 2019.

Stephan Günnemann. ECML-PKDD 2020 Keynote: Certifiable robustness of ml models for graphs. `https://www.youtube.com/watch?v=HkISG9bdAl0&t=2571s`, 2020.

Yujia Huang, Huan Zhang, Yuanyuan Shi, J Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local lipschitz bounds. *Advances in Neural Information Processing Systems*, 34:22745–22757, 2021.

Jinyuan Jia, Binghui Wang, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of community detection against adversarial structural perturbation via randomized smoothing. *Proceedings of The Web Conference 2020*, pages 2718–2724, 2020.

Hongwei Jin, Zhan Shi, Venkata Jaya Shankar Ashish Peruri, and Xinhua Zhang. Certified robustness of graph convolution networks for graph classification under topological attacks. *Advances in neural information processing systems*, 33:8463–8474, 2020.

Hongwei Jin, Zishun Yu, and Xinhua Zhang. Certifying robust graph classification under orthogonal gromov-wasserstein threats. In *Advances in Neural Information Processing Systems*, 2022.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. *IEEE Symposium on Security and Privacy (SP)*, pages 656–672, 2019.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations, ICLR*, 2019.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. *Advances in Neural Information Processing Systems*, 31, 2018.

Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks. In *International conference on machine learning*, pages 8981–8991. PMLR, 2020.

Mahito Sugiyama and Karsten Borgwardt. Halting in random walk kernels. *Advances in neural information processing systems*, 28, 2015.

Y. Sui, X. Wang, J. Wu, M. Lin, X. He, and T.-S. Chua. Causal attention for interpretable and generalizable graph classification. In *ACM SIGKDD Conference on Knowledge Discovery and Data MiningAugust*, page 1696–1705, 2022.

Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of graph neural networks against adversarial structural perturbation. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1645–1653, 2021.

Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations, ICLR*, 2017.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. Adversarial examples on graph data: Deep insights into attack and defense. *arXiv preprint arXiv:1903.01610*, 2019.

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, Bo Long, et al. Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning*, 16(2): 119–328, 2023.

D. Zügner and S. Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 246–256, 2019.