Structure-Aware Robustness Certificates for Graph Classification (Supplementary Material)

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

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

1 PROOFS OF PROPOSITIONS

1.1 PROOF OF PROPOSITION 1

Disjoint Unions. Let $\mathbf{z} \in \mathcal{R}_{\mathbf{Q}}$ and $\tilde{\mathbf{z}} \in \mathcal{R}_{\mathbf{Q}'}$ such that for some $i \in I$ we have $Q_i \neq Q_i^{'}$. If $\mathbf{z} = \tilde{\mathbf{z}}$, it implies that $||\mathbf{z}_{\mathcal{J}_i} - \mathbf{x}_{\mathcal{J}_i}|| = Q_i$ and $||\tilde{z}_{\mathcal{J}_i} - \mathbf{x}_{\mathcal{J}_i}|| = Q_i^{'}$ which is a contradiction.

Partition. $|\mathcal{J}_i| \leq R_i$, and $||\mathbf{z}_{\mathcal{J}_i} - \mathbf{x}_{\mathcal{J}_i}|| \leq Q_i$ hence $\mathcal{X} = \cup_{Q \leq R} \mathcal{R}_Q^R$.

1.2 PROOF OF PROPOSITION 2

As the noise for each entry is independent we can decompose the probabilities as so

$$\frac{P(\phi(\tilde{\mathbf{x}}) = \mathbf{z})}{P(\phi(\mathbf{x}) = \mathbf{z})} = \prod_{k \in [N]} \frac{P(\phi(\tilde{\mathbf{x}})_k = \mathbf{z}_k)}{P(\phi(\mathbf{x})_k = \mathbf{z}_k)}. \tag{1}$$

Furthermore, as each components belongs to exactly one edge community.

$$\prod_{k \in [N]} \frac{P(\phi(\tilde{\mathbf{x}})_k = \mathbf{z}_k)}{P(\phi(\mathbf{x})_k = \mathbf{z}_k)} = \prod_{i=1}^{I} \prod_{k \in \mathcal{C}_i} \frac{P(\phi(\tilde{\mathbf{x}})_k = \mathbf{z}_k)}{P(\phi(\mathbf{x})_k = \mathbf{z}_k)}. \tag{2}$$

We note that for k where $\tilde{\mathbf{x}}_k = \mathbf{x}_k$ this fraction is one, so we can focus on terms when $\tilde{\mathbf{x}}_k \neq \mathbf{x}_k$. In equations this can be written as

$$\prod_{i=1}^{I} \prod_{k \in \mathcal{C}_i} \frac{P(\phi(\tilde{\mathbf{x}})_k = \mathbf{z}_k)}{P(\phi(\mathbf{x})_k = \mathbf{z}_k)} = \prod_{i=1}^{I} \prod_{k \in \mathcal{J}_i} \frac{P(\phi(\tilde{\mathbf{x}})_k = \mathbf{z}_k)}{P(\phi(\mathbf{x})_k = \mathbf{z}_k)} \tag{3}$$

We can consider what the terms are equal to when $\mathbf{x}_k = \mathbf{z}_k$ and when $\mathbf{x}_k \neq \mathbf{z}_k$ (assuming that $\mathbf{x}_k \neq \tilde{\mathbf{x}}_k$). We get

$$\frac{P(\phi(\tilde{\mathbf{x}})_k = \mathbf{z}_k)}{P(\phi(\mathbf{x})_k = \mathbf{z}_k)} = \begin{cases} \frac{p_i}{1-p_i} & \text{if } \mathbf{x}_k = \mathbf{z}_k \text{ and } \mathbf{x}_k \neq \tilde{\mathbf{x}}_k \\ \frac{1-p_i}{p_i} & \text{if } \mathbf{x}_k \neq \mathbf{z}_k \text{ and } \mathbf{x}_k \neq \tilde{\mathbf{x}}_k \end{cases}. \tag{4}$$

In total there are R_i terms in each product, of which Q_i are the first case and $R_i - Q_i$ are in case two. Thus

[*]Equal contribution.

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

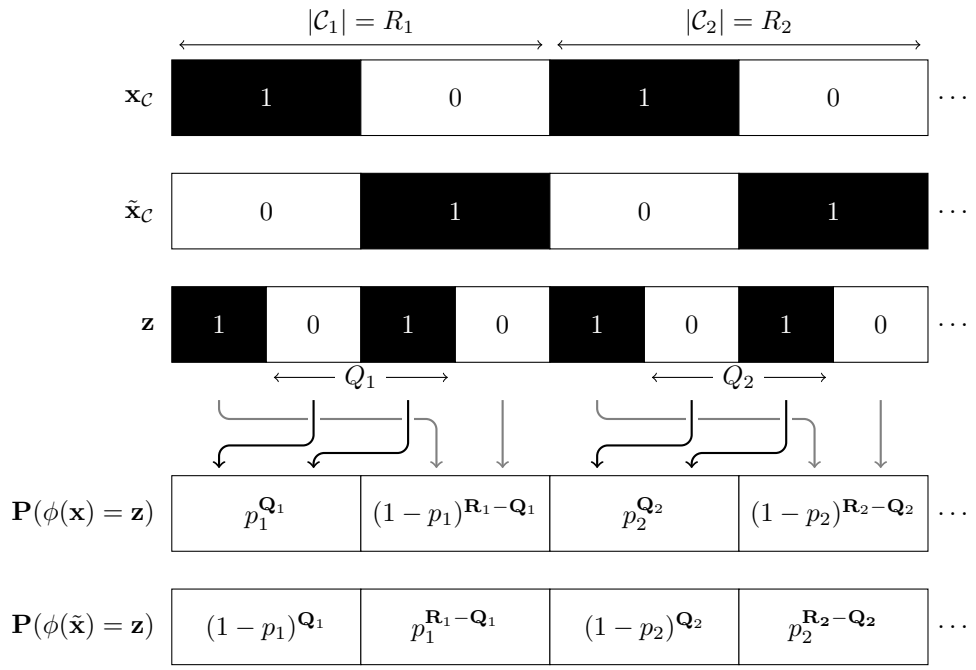

Figure 1: Pictorial representation of where the terms in Proposition 2 come from.

$$\prod_{i=1}^{I} \prod_{k \in \mathcal{J}_i} \frac{P(\phi(\tilde{\mathbf{x}})_k = \mathbf{z}_k)}{P(\phi(\mathbf{x})_k = \mathbf{z}_k)} = \prod_{i=1}^{I} \left(\frac{p_i}{1-p_i}\right)^{Q_i} \left(\frac{1-p_i}{p_i}\right)^{R_i - Q_i} \tag{5}$$

$$= \prod_{i=1}^{C} \left(\frac{p_i}{1-p_i}\right)^{2Q_i - R_i} \tag{6}$$

$$= \prod_{i=1}^{C} \left(\frac{1-p_i}{p_i}\right)^{R_i - 2Q_i} \tag{7}$$

as required. We provide Fig. 1 as a visual aid to the proof.

1.3 PROOF OF PROPOSITION 3

We have $\mathcal{R}_\mathbf{Q} = \{\mathbf{z} \in \mathcal{X} : \|\mathbf{z}_{\mathcal{J}_i} - \mathbf{x}_{\mathcal{J}_i}\|_0 = Q_i\}$. The probability $\mathbb{P}(\phi(\mathbf{x}) \in \mathcal{R}_\mathbf{Q})$ corresponds to each set R_i having Q_i entries not being flipped or equivalently $R_i - Q_i$ entries being flipped. Each node pair is flipped with a probability of p_i. Since all flips are independent we can express the probability as $\mathbb{P}(\phi(\mathbf{x}) \in \mathcal{R}_\mathbf{Q}) = \prod_{i=1}^{C} \mathrm{Bin}(R_i - Q_i | R_i, p_i)$.

2 IMPLEMENTATION

2.1 NOISE SAMPLING

In order to sample from the anisotropic noise defined in eq. (11), we propose an illustration in Fig. 2. Given disjoint regions of node pairs $\mathcal{C}_i$, new graphs are sampled by adding independent Bernoulli samples with parameters given by the regions to the appropriate part of the graph.

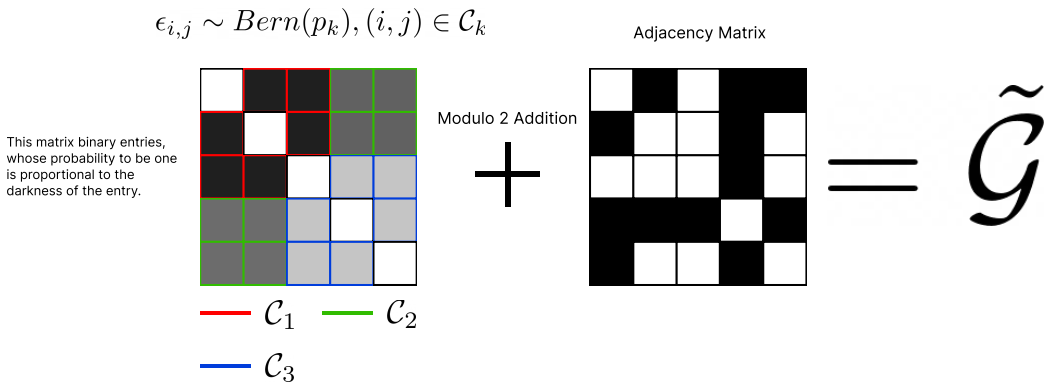

Figure 2: A comparison between the anisotropic certificate and the sparsity-aware certificate. Each entry represents the ratio of correctly classified test-set samples that could be certified at a specified number of edge deletions and additions.

2.2 ESTIMATIONS OF PROBABILITIES

The quantities $p_y(\mathbf{x})$ cannot be computed in closed form for general f. Hence, we resolve to lower bound p_A and upper bound $p_y(\mathbf{x}), y \neq c_A$ via sampling. To achieve this, we use the Clopper-Pearson interval. Cai [2005].

2.3 SYMMETRIES CERTIFICATION

Solving the optimization problem defined in Eq. (8) is difficult as certificates have to be computed for every $\tilde{\mathbf{x}}$ in the ball around $\mathbf{x}$: $\mathcal{B}_r(\mathbf{x})$. However, in practice, $\Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A)$ displays some symmetries depending on the noise distribution $\phi(\mathbf{x})$.

In the case of isotropic noise, the regions $\mathcal{H}_k$ and values η_k only depends on $\|\mathbf{x} - \tilde{\mathbf{x}}\|_0$. This implies $\Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A) = \Phi_{\mathbf{x},\tilde{\mathbf{x}}'}(p_A, c_A)$ for all $\tilde{\mathbf{x}}, \tilde{\mathbf{x}}' \in \mathcal{S}_r(\mathbf{x})$ which reduce the search on every spheres.

In the case of anisotropic noise, the regions $\mathcal{H}_k$ and values η_k only depends on $\|\mathbf{x}_{\mathcal{C}_i} - \tilde{\mathbf{x}}_{\mathcal{C}_i}\|_0$. This implies $\Phi_{\mathbf{x},\tilde{\mathbf{x}}}(p_A, c_A) = \Phi_{\mathbf{x},\tilde{\mathbf{x}}'}(p_A, c_A)$ for all $\tilde{\mathbf{x}}, \tilde{\mathbf{x}}' \in \mathcal{S}_{\mathbf{R}}(\mathbf{x})$.

3 ALGORITHM

The full algorithm of our method is given in Alg. 1 and its complexity is analyzed below.

3.1 ALGORITHMIC COMPLEXITY: CERTIFICATION

Let $\mathbf{x}$ be a graph, N the number of samples to perform the Clopper-Pearson statistical test, n the number of nodes in the graph, and $\mathbf{R} \in \prod_i [|\mathcal{C}_i|]$ a given radius to certify. The certification algorithm proceeds as follow:

1. Sample N graphs from the noise distribution with complexity $\mathcal{O}(Nn^2)$, this step is very easily parallelizable.
2. Forward the N sampled graphs through the model. Given a model forward complexity of $\mathcal{O}(m(n))$ (we omit potential depency on node or edge feature dimension), the total complexity is $\mathcal{O}(Nm(n))$, this step is very easily parallelizable.
3. From estimates (p_A, p_B) and noise distribution ϵ find optimal radius $\mathbf{R}$. and $T_{\mathbf{R}} = \prod_i (R_i + 1)$:
   (a) Compute the vectors $\eta_{\mathbf{Q}}^{\mathcal{R}}$ and sort them, with respective complexity $\mathcal{O}(CT_{\mathbf{R}})$ and $\mathcal{O}(T_{\mathbf{R}} \log(T_{\mathbf{R}}))$.
   (b) Solve the linear programs of eq. (9) and (10) and verify $\underline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_A, c_A) - \overline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_B, c_B) > 0$, with complexity $\mathcal{O}(T)$

   The total complexity becomes $\mathcal{O}(Nn^2 + Nm(n) + CT_{\mathbf{R}} + T_{\mathbf{R}} \log(T_{\mathbf{R}}))$

Regarding the model complexity, some example complexity are the following:

1. Graph Neural Network: the complexity is quadratic in the number of nodes due to matrix multiplication: $m(n) = \mathcal{O}(n^2)$
2. Label kernel: the complexity is linear in the number of edges $\mathcal{O}(\mathcal{E}) = \mathcal{O}(n^2)$

Algorithm 1 Structure aware randomized smoothing

inputs: Graph to certify $\mathbf{x}$, noise perturbation $\boldsymbol{\epsilon}$, anisotropic structure $(\mathcal{C}_i)_{i\in I} \subset [n]^2$, graph classification model $m : \mathcal{X} \to \mathcal{Y}$, number of samples N and upper bounds on certificate radii $(\mathbf{R}_{max,i})_{i\in I}$.
initialize: Train model m on classification data $\mathcal{D}$ or load model parameters.
voting
  for $i = 1, ..., N$ **do**
    Sample random graph $\tilde{\mathbf{x}}_i \sim \mathbf{x} \oplus \boldsymbol{\epsilon}$
    Compute model prediction $y_i \in \mathcal{Y}$
  end for
  Compute distribution label frequency from $(y_i)_{i\in[N]}$, denoted $(p_y, y)_{y\in\mathcal{Y}}$, and identify the most frequent and runner-up (second most frequent) class (p_A, c_A) and (p_B, c_B)
certification
  for $\mathbf{R} \in \prod_i [|\mathbf{R}_{max,i}|]$ **do**
    Compute $\eta^{\mathcal{R}}_{\mathbf{Q}}$ according to the formula (13) and sort them.
    Compute $\mathbb{P}(\phi(\mathbf{x})) \in \mathcal{R}_{\mathbf{Q}}$ with formula (14).
    Solve the linear programs described in eq. (9) and (10) greedily
    Verify $\underline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_A, c_A) - \overline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_B, c_B) > 0$
  end for
return Grid of certification for $\mathbf{R} \in \prod_i [|\mathbf{R}_{max,i}|]$

3.2 ALGORITHMIC COMPLEXITY: OPTIMAL RADIUS

Let $\mathbf{x}$ be a graph, N the number of samples to perform the Clopper-Pearson statistical test, n the number of nodes in the graph. The algorithm to find the optimal radius proceeds as follow:

1. Sample N graphs from the noise distribution with complexity $\mathcal{O}(Nn^2)$, this step is very easily parallelizable.
2. Forward the N sampled graphs through the model. Given a model forward complexity of $\mathcal{O}(m(n))$ (we omit potential depency on node or edge feature dimension), the total complexity is $\mathcal{O}(Nm(n))$, this step is very easily parallelizable.
3. From estimates (p_A, p_B) and noise distribution ϵ find optimal radius $\mathbf{R}$. Select a vector, $\mathbf{R} \in \prod_i [|\mathcal{C}_i|]$, let $T_{\mathbf{R}} = \prod_i (R_i + 1)$ and $T = \prod_i (R_{i,max} + 1)$:
   (a) Compute the vectors $\eta^{\mathcal{R}}_{\mathbf{Q}}$ and sort them, with respective complexity $\mathcal{O}(CT_{\mathbf{R}})$ and $\mathcal{O}(T_{\mathbf{R}} \log(T_{\mathbf{R}}))$.
   (b) Solve the linear programs of eq. (9) and (10) and verify $\underline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_A, c_A) - \overline{\rho_{\mathbf{x},\tilde{\mathbf{x}}}}(p_B, c_B) > 0$, with complexity $\mathcal{O}(T)$

   We output the pareto front $\mathbf{R}$ according to the partial ordering $\mathbf{R} \preceq \mathbf{R}' \iff \forall i, \mathbf{R}_i \leq \mathbf{R}'_i$

The total naive complexity is $\mathcal{O}(Nn^2 + Nm(n) + CT^2 + T^2 \log(T))$. However, we want to point out there are multiple places the complexity could drastically improve.

1. First, the last point is problem agnostic, meaning that, given the estimates (p_A, p_B) (first and second highest label probabilities) and the noise distribution ϵ, the corresponding optimal radii $\mathbf{R}$ can be computed. Given specific scenario, this opens the possibility to precompute tables $\mathbf{R}(p_A, p_B, \epsilon)$. This can be used to directly output $\mathbf{R}$ or use it to find the optimal $\mathbf{R}$ quicker.
2. Second, the linear program described in equation (9) and (10) can be efficiently solved greedily. Given we know the closed-form formula for μ_k, making the ordering explicitly dependant on $\mathbf{Q}$, one can compute them only when necessary.
3. Finally, the partial ordering defined previously is, in practice indicative of the robustness certification, i.e. if we cannot certify a certain radius, a larger radius won't be certified either. Although we don't propose a formal proof of this property, it holds true in practice, as one can see on the experiment results, and could be exploited for more efficient search, similar to a multidimensional binary search.

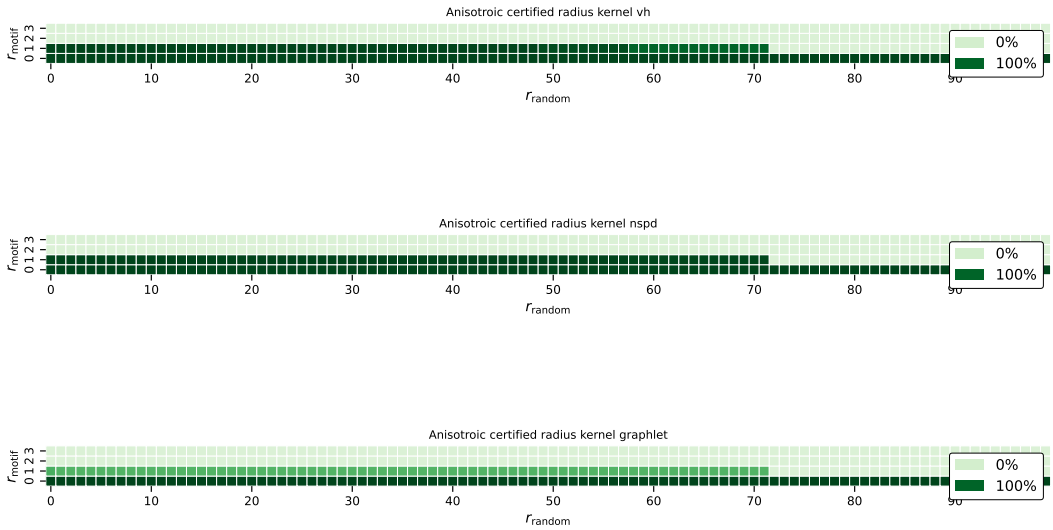

Figure 3: Influence of the underlying classifier on the anisotropic certificate radius.

4 ADDITIONAL RESULTS

Varying the base classifier In Figure 3, we compare our anisotropic certification performance across three kernels, the graphlet Sampling kernel Shervashidze et al. [2009], the neighbourhood subgraph pairwise distance kernel Costa and De Grave [2010] and vertex Histogram kernels Sugiyama and Borgwardt [2015] for a sample size of $N = 10,000$. In general, a model that is robust to noise will lead to certificates with large radii.

Number of sampled perturbations In Figure 4, we analysed the impact of sample size when computing the anisotropic certification radius in our synthetic experiments. The certificate performs poorly for a small number of samples. This is because the lower bound on p_A becomes very loose.

References

T Tony Cai. One-sided confidence intervals in discrete distributions. *Journal of Statistical planning and inference*, 131(1): 63–88, 2005.

Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 26th International Conference on Machine Learning*, pages 255–262. Omnipress; Madison, WI, USA, 2010.

Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *Artificial intelligence and statistics*, pages 488–495. PMLR, 2009.

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

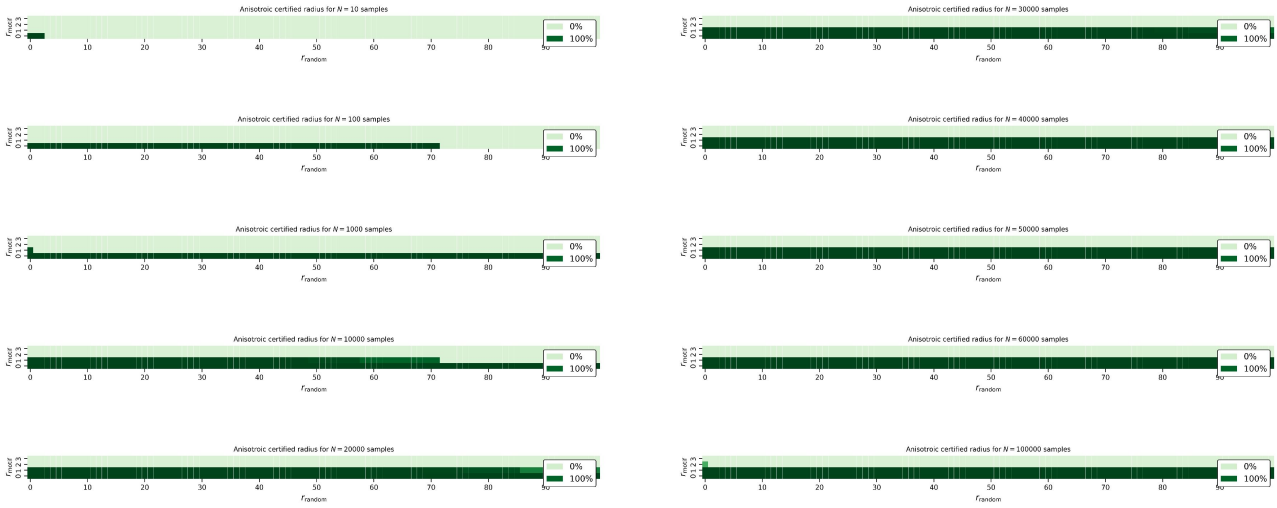

Figure 4: Influence of sample size on anisotropic certificate radius.