Certified Robustness on Structural Graph Matching

Anonymous authors
Paper under double-blind review

Abstract

The vulnerability of graph matching (GM) to adversarial attacks has received increasing attention from emerging empirical studies, while the certified robustness of GM has not been explored. Inspired by the technique of randomized smoothing, in this paper, for the first time to our best knowledge, the certified robustness on GM is defined and a new certification strategy is designed called Structure-based Certified Robustness of Graph Matching (SCR-GM). Structural prior information of nodes is used to construct a joint smoothing distribution matrix with physical significance, which certifies a wider range than those obtained by previous iterative optimization methods. Furthermore, we propose a certified space that can be used to derive a strictly certified radius and two extra radii for evaluation. Experimental results on GM datasets reveal that our strategy achieves state-of-the-art ℓ_2 certified accuracy and regions. Source code will be made publicly available.

1 Introduction

As a well-known NP-hard problem in its general form (Yan et al., 2016) with wide applications e.g. in computer vision and pattern recognition, graph matching (GM) refers to establishing correspondences among two (Cho et al., 2010) or multiple graphs (Jiang et al., 2021). Given two input graphs $\mathcal{G}_1 = \{\mathbf{V}_1, \mathbf{E}_1\}$ and $\mathcal{G}_2 = \{\mathbf{V}_2, \mathbf{E}_2\}$ with two sets of annotated nodes $\mathbf{z}^1 \in \mathbb{R}^{n_1 \times 2}$ and $\mathbf{z}^2 \in \mathbb{R}^{n_2 \times 2}$ (assumed in Euclidean space in this paper). Here, $\mathbf{V}_1 \in \mathbb{R}^{d_v \times n_1}$ and $\mathbf{E}_1 \in \mathbb{R}^{d_e \times m_1}$ represent the feature matrix of n_1 nodes and m_1 edges (likewise for $\mathbf{V}_2$ and $\mathbf{E}_2$). The similarities between nodes and edges are formulated into a global affinity matrix $\mathbf{K} \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$, whose diagonal and off-diagonal elements store the node-to-node and edge to-edge affinities. It aims to maximize the overall affinity score J of the matching nodes and the edges (Leordeanu & Hebert, 2005) in the form of quadratic assignment problem (QAP) (Loiola et al., 2007):

$$\begin{aligned} &\max_{\mathbf{X}} J(\mathbf{X}) = \mathrm{vec}(\mathbf{X})^\top \mathbf{K}\, \mathrm{vec}(\mathbf{X}), \\ &\text{s.t. } \mathbf{X} \in \{0,1\}^{n_1 \times n_2}, \mathbf{X}\mathbf{1}_{n_1} = \mathbf{1}_{n_1}, \mathbf{X}^\top \mathbf{1}_{n_2} \leq \mathbf{1}_{n_2}, \end{aligned} \tag{1}$$

where $\mathrm{vec}(\mathbf{X})$ denotes the column-wise vector of the matching solution $\mathbf{X} \in \{0,1\}^{n_1 \times n_2}$ which can be a partial permutation matrix when $n_1 < n_2$. One common approach is to relax $\mathbf{X}$'s raw binary constraint into a continuous one (between [0,1]), especially in the form of (partial) doubly-stochastic matrix $\mathbf{S} \in [0,1]^{n_1 \times n_2}$ of which the sum of rows/columns is 1 (or zero for partial case). The final $\mathbf{X}$ can be obtained by the Hungarian algorithm (Burkard & Dell'Amico, 2009): $\mathbf{X} = \mathrm{Hung}(\mathbf{S})$.

Eq. 1 can also directly incorporate deep nets to obtain the learned affinity matrix $\mathbf{K}$ by learning the raw attributes of the graphs e.g. CNNs for images from which the visual graphs are extracted, as well as learning the structure via graph neural networks (GNNs) (Wang et al., 2019): $\mathbf{K}$=NN$(\mathcal{G}_1, \mathcal{G}_2)$.

Studies on robustness of machine learning models have attracted wide attention, while the robustness of combinatorial solvers is an emergning and unmatured topic (Geisler et al., 2021; Lu et al., 2021). Under the deep GM paradigm, Ren et al. (2022) reveal that the combinatorial GM algorithms can also be sensitive to (additive) noise perturbations not only in appearance but also for structure, similar to the node classification models (Dai et al., 2018; Sun et al., 2018), and an empirical defense algorithm via an appearance-aware regularizer is proposed. So far, there still lacks principled certified defense to provide theoretical robustness guarantees for GM (let alone other combinatorial problems). In fact, existing certified robustness mechanisms (including randomized smoothing) in the graph domain (Rong et al., 2019; Bojchevski et al., 2020; Zügner & Günnemann, 2020; Jia et al., 2020) are confined to unconstrained node or graph-level classification/prediction within a single graph, which cannot be readily adopted for solving the cross-graph and combinatorial problems with structured output like the permutation matrix in GM.

Certifiable robustness studies solvers whose prediction at any point x is verifiably constant within some set around x (Wong & Kolter, 2018). As a recent promising approach to achieve certified defense, randomized smoothing (RS) (Lecuyer et al., 2019; Cohen et al., 2019) provides a general robust guarantee applicable to large-scale neural networks against arbitrary attacks. Given an input x and a base classifier, randomized smoothing constructs a 'smoothed classifier' which is certifiable within the region characterized by x and the smoothing distribution $\mathcal{D}$. RS has been used in certifying different models, e.g., image classification (Yang et al., 2020) and object detection in vision (Chiang et al., 2020).

As an initiative for applying RS to GM[1], in this paper we mainly consider two major challenges to solve. **C1: varying-size of input graphs.** It is not suitable to certify graphs with different sizes by using an identical smoothing distribution. **C2: dependency of nodes in graph.** The graph structure as a whole carries important information for certification. For the first challenge, we could refer to data-dependent certified robustness methods on image classification task. Some data-dependent methods (Alfarra et al., 2022; Eiras et al., 2021; Hong & Hong, 2022; Labarbarie et al., 2022) are proposed recently to vary and optimize the smoothing distributions $\mathcal{D}$ for larger certification region. Therefore, these methods can also be used to construct varying smoothing distributions for graphs with varying sizes. For the second challenge, we expect smoothing distributions constructing correlations between nodes in a graph, which is lacking for current randomized smoothing. Data-dependent methods consider little on the heterogeneity and structure of inputs. For example, Alfarra et al. (2022) treat all pixels in one image equally, Eiras et al. (2021) treat pixels differently but cannot reveal their correlation. Thus none of them can overcome the second challenge.

In this paper, we aim to solve certified robustness of GM, by analyzing the individual matching robustness of each node, instead of the whole variation of the output matching matrix $\mathbf{X}$ in Eq. 1. In particular, we study the node classification task when converting the relaxed solution $\mathbf{S}$ into the final matching $\mathbf{X}$ (see Eq. 1 and the discussion therein), as such the RS-type certification phase can be naturally introduced during the classification stage.

Specifically, we propose the **S**tructure-based **C**ertified **R**obustness of **G**raph **M**atching (SCR-GM) which adopts joint Gaussian distribution instead of independent homogeneous distribution to construct the smoothing solvers. As adversarial attacks tend to perturb the strongly correlated nodes at the same time, the additive noise sampled from joint distribution with structural information and physical meaning can reveal this correlation. According to our theoretical analysis, we obtain the robustness guarantee on GM which describes a certified ℓ_2-norm space ant its lower bound radius. In addition, we propose another two radii to help evaluate the robustness more comprehensively. We evaluate our strategy on Pascal VOC dataset (Everingham et al., 2010) with Berkeley annotations (Bourdev & Malik, 2009) and simulation dataset with random node sets. Experimental results reveal that our strategy outperforms the previous works (Cohen et al., 2019; Alfarra et al., 2022; Eiras et al., 2021) on structural GM for ℓ_2 certified accuracy and regions. Our contributions are as follows:

1) We propose a general framework for incorporating existing RS-based techniques for certifying graph matching solvers, as long as (which is often the case for both learning-based and classic solvers) it involves a post-binarization step that converts the relaxed matching $\mathbf{S}$ (by an arbitrary relaxed GM solver) to node matching.

2) Based our proposed framework, we present the first definition, to our best knowledge (see Eq. 5) of certified robustness for a graph matching solver.

3) We propose a certification method dubbed structure-based certified robustness of GM (SCR-GM) (see Sec. 4.3). It uses jointly distributed noise to model dependent node matching certification.

4) A certified space and lower bound radius are derived to guarantee robustness of graph matching. Two radii are also devised for more complete evaluation of robustness, which complements potentially safe regions and largest feasible perturbations.

[1] Another challenge is how to better handle the constraints of $\mathbf{X}$, which is related to how to extend the certification of the specific GM problem to other combinatorial solvers, which we leave for future work.

2 Related work

We discuss works on certified robustness related to randomized smoothing and robustness of GM. **Certified Robustness related to Randomized Smoothing** Lecuyer et al. (2019) propose randomized smoothing firstly as a certified adversarial defense, and use it to train the first certifiably robust classifier for ImageNet. However, its guarantees are loose, then Cohen et al. (2019) shows that adding Gaussian noise to classifiers enjoys a strict ℓ_2 certification radius, with follow-ups presenting new RS-type techniques, such as optimal perturbations at different norms, and certified robustness definitions for different tasks. Alfarra et al. (2022) show that the variance of the Gaussian distributions can be optimized at each input so as to maximize the certification region. Meanwhile, Eiras et al. (2021) extend isotropic smoothing distributions to generalized anisotropic counterparts. Hong & Hong (2022) adopt the same anisotropic defination and further design a noise generator to efficiently fine-tune the distributions. A recent work (Labarbarie et al., 2022) that relies on information geometry techniques manages to prove larger regions than previous methods.

However, all previous smoothing distributions $\mathcal{D}$ deprive the favorable prior knowledge which mainly refers to the node location and graph structure in GM. Moreover, all of them at most certify a single image or graph but do not consider the combinatorial nature of the prediction as in GM.

Robustness of Graph Matching Approximate GM solvers have been developed over the decades from traditional learning-free methods (Emmert-Streib et al., 2016) to learning-based ones (Yan et al., 2020). The seminal work (Zanfir & Sminchisescu, 2018) proposes a deep neural network based pipeline for visual GM, in which the visual appearance features are learned via CNN, with subsequent variants (Wang et al., 2019; Rolínek et al., 2020), among which a major improvement is to explore the structural information using different techniques e.g. GNN, rather than only appearance features for node/edge attributes as done in (Zanfir & Sminchisescu, 2018). Our work treats the GM solver as blackbox regardless it is learning-based or not, as long as it involves a continuous relaxation to obtain the intermediate double-stochastic matrix.

There is also an emerging line of research on adversarial attack and defense on (deep) GM. The earlier work (Yu et al., 2019b) proposes a robust graph matching (RGM) model to improve the robustness against perturbations e.g. distortion, rotation, outliers and noise. Zhang et al. (2020) devise an adversarial attack model for deep GM networks, which uses kernel density estimation to construct dense regions such that the neighboring nodes are indistinguishable. Ren et al. (2021) devise two specific topology attacks in GM: inter-graph dispersion and intra-graph combination attacks, and propose a resilient defense model. Ren et al. (2022) design an attack perturbing input images and their hidden graphs together for deep (visual) GM, and further propose appearance-aware regularizers to enlarge the disparity among similar keypoints for defense. However, the above defense methods are all heuristic and lacks robustness certification in face of other unseen attacks.

3 Preliminaries on Randomized Smoothing

The original RS (Cohen et al., 2019) can transform an arbitrary base classifier f into a smoothed classifier g that is certifiably robust under ℓ_2 norm. For any input x, the smoothed classifier g returns the most probable prediction of f for the random variable $\mathcal{N}(x;\sigma^2 I)$, which is defined by:

$$g(x) = \underset{c\in\mathcal{Y}}{\arg\max}\mathbb{P}(f(x+\varepsilon)=c), \tag{2}$$

where $\varepsilon \sim \mathcal{N}\left(0,\sigma^2 I\right)$ is isotropic Gaussian noise perturbing the input x. Then the certified radius within which the output is unchanged for $g(x+\delta)=c_A$ that measures the certified robustness is:

$$R = \|\delta\|_2 < \frac{\sigma}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \tag{3}$$

where the most probable class c_A is returned with probability p_A and the 'runner-up' class is returned with probability p_B. $\underline{p_A}$ and $\overline{p_B}$ are lower bound and upper bound of p_A and p_B respectively, and Φ^{-1} is the inverse of the standard Gaussian cumulative distribution function. The smoothed classifier g is robust around x within the ℓ_2 radius in Eq. 3.

To enhance the certification, Alfarra et al. (2022) and Eiras et al. (2021) propose isotropic and anisotropic distributions to maximize the certified region respectively. However, none of them can explicitly encode the prior information of the inputs (e.g. the graph topology in GM) which means their distributions are randomly initialized. Differently, we propose a correlation matrix to reveal the structural information in graphs, and in turn construct a joint Gaussian distribution to replace the single Gaussian distribution, which not only makes the initial distribution physically meaningful, but also eliminates the optimization process of finding the largest certified region.

4 Methodology

We first define the smoothed GM solver (be either neural network or traditional solver) and propose a robustness guarantee. We then devise a new certification strategy SCR-GM using a physically meaningful joint smoothing distribution. We also give two new radii to aid evaluating robustness.

4.1 Problem Formulation

For pairwise GM with input $\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)$, we mainly focus on the effect of perturbing two sets of annotated nodes $\mathbf{z}^1 \in \mathbb{R}^{n_1 \times 2}$ and $\mathbf{z}^2 \in \mathbb{R}^{n_2 \times 2}$. For visual GM (Zanfir & Sminchisescu, 2018; Ren et al., 2022) as widely considered in literature, $\mathbf{z}^1$ and $\mathbf{z}^2$ are node coordinates obtained by human annotation or keypoint detectors. During the certification for perturbing nodes, here we consider the node coordinates as the input while keep the node/edge attributes as unchanged. The robustness guarantees of perturbing features are given in Appendix B.

As discussed in Sec. 3, randomized smoothing (RS) is a technique for constructing a smoothed function g from an arbitrary base function f. In this paper, we technically convert a whole matching problem into a set $\mathbf{F}$ regarding with binary classification based on the intermediate matrix $\mathbf{S}$. The set $\mathbf{F}$ can be expressed as: $\mathbf{F} = \{f_i | f_i : \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right) \rightarrow r_j, i \in n_1, j \in n_2\}$, where f_i represents that the i-th node in $\mathbf{z}^1$ matches the j-th node in $\mathbf{z}^2$ and r_j represents the j-th node r_j in $\mathbf{z}^2$. Such a conversion allows us to certify the matching robustness for a single node, avoiding an imprecise certification for the entire matching matrix. The smoothed network g_i returns whichever node in $\mathbf{z}^2$ is most likely to match the node in $\mathbf{z}^1$ when the input is perturbed by joint smoothing noise:

$$\begin{gathered} g_i = \underset{r_j \in \mathbf{z}^2}{\arg\max} \mathbb{P}(f_i\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon, \mathbf{z}^2\right) = r_j), \\ \text{where } \varepsilon \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}\right), i \in n_1, j \in n_2. \end{gathered} \tag{4}$$

For convenience, we simplify $f_i\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)$ to $f_i(\mathbf{z}^1)$ and derive the results by perturbing $\mathbf{z}^1$ only, as it is equivalent to robustness certification under joint perturbation to $\mathbf{z}^1$ and $\mathbf{z}^2$. Furthermore, we propose a method which defines the smoothed function for certifying whole $\mathbf{X}$ as introduced in Appendix. E. The distribution of noise ε is a joint Gaussian distribution matrix whose variance represents the correlation between nodes. In addition, $\boldsymbol{\Sigma}$ is a hyperparameter for certified function which controls a robustness/accuracy trade-off and will be detailed in Sec. 4.3. Note that for robustness certification, we only consider those nodes that can obtain a unique solution $\arg\max$ in Eq. 4.

4.2 Robustness Guarantee

Suppose that when the base function f_i solves for the optimal matching of node i in $\mathbf{z}^1$, the most probable node r_A in $\mathbf{z}^2$ is returned with probability $p_A = \max_{s_i \in \mathbf{S}_i} s_i$, where $\mathbf{S}_i$ is the i-th row of $\mathbf{S}$. Similarly, the probability of "runner-up" node r_B in $\mathbf{z}^2$ is denoted as p_B, $p_B = \max_{s_i \in \mathbf{S}_i, r_B \neq r_A} s_i$. We adopt an ℓ_2 certified space to guarantee robustness of graph matching.

Theorem 1 (**ℓ_2 certified space**) *Let $f_i(\mathbf{z}^1)$ be node matching function, g_i be defined as in Eq. 4, and $\varepsilon \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}\right)$. If $\underline{p_A} \in [0,1]$ and $\overline{p_B} \in [0,1]$ satisfy:*

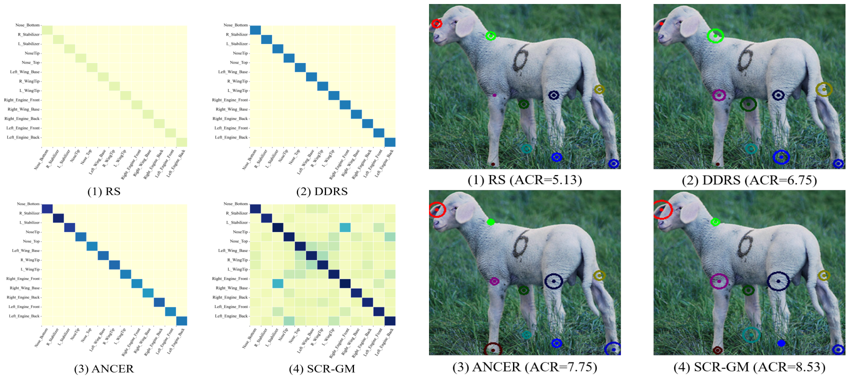
$$\mathbb{P}\left(f_i(\mathbf{z}^1 + \varepsilon) = r_A\right) \geq \underline{p_A} \geq \overline{p_B} \geq \mathbb{P}(f_i(\mathbf{z}^1 + \varepsilon) = r_B), \tag{5}$$

then for $g_i(\mathbf{z}_1 + \delta) = r_A$, we can get the certified ℓ_2 space for the addictive noise δ:

$$\|\delta^\top \mathbf{B}^{-1}\| < \frac{1}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \tag{6}$$

where $\mathbf{B}^\top \mathbf{B} = \boldsymbol{\Sigma}$, and $\mathbf{B} \in \mathbb{R}^{n_1 \times n_1}$ is a full rank and real symmetric matrix based on the node correlation in node matrix $\mathbf{z}^1$, and $\underline{p_A}$ and $\overline{p_B}$ are the lower bound of p_A and upper bound of p_B.

The detail settings and properties of $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are described and illustrated in Section 4.3. The complete proof of Theorem 1 is presented in Appendix A.



(1) RS (2) DDRS (3) ANCER (4) SCR-GM

(1) RS (ACR=5.13) (2) DDRS (ACR=6.75) (3) ANCER (ACR=7.75) (4) SCR-GM (ACR=8.53)

(a) Smoothing distributions (b) Certified radius and ACR of a sample.

Figure 1: The smoothing distribution of the four RS methods in Fig. 1(a) (including RS (Cohen et al., 2019), data-dependent randomized smoothing (DDRS) (Alfarra et al., 2022), anisotropic certification RS (ANCER) (Eiras et al., 2021) and SCR-GM, whose details are in experiment part), on Pascal VOC. The inter-node correlation gradually increases over the four methods from (1) to (4) with darker color. Fig. 1(b) shows certified radius $\|\delta\|_{\text{lower}}$ and their ACR (Zhai et al., 2020). The center of the circle in Fig. 1(b) represents the position of the node in $\mathbf{z}^{\mathbf{1}}$, while the radius of the circle represents its corresponding certified radius $\|\delta\|_{\text{lower}}$ calculated by Eq. 9. ACR shows the overall certified robustness: the higher ACR, the better overall certified robustness.

Lemma 1 (Eigenvalue Comparison) *For a real symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, with $\lambda_{\max}$ and $\lambda_{\min}$ as its maximum and minimum of eigenvalues, then $\forall \mathbf{X} \in \mathbb{R}^n$, $\lambda_{\min}\mathbf{X}^\top\mathbf{X} \leq \mathbf{X}^\top\mathbf{A}\mathbf{X} \leq \lambda_{\max}\mathbf{X}^\top\mathbf{X}$.*

Based on Lemma 1 and the certified space in Eq. 6, we can further obtain a certified ℓ_2 norm radius:

$$\|\delta^\top \mathbf{B}^{-1}\|^2 = \delta^\top \mathbf{\Sigma}^{-1}\delta, \tag{7}$$

$$\delta^\top \mathbf{\Sigma}^{-1}\delta \leq \lambda_{\max}\delta^\top\delta, \tag{8}$$

$$\|\delta\|_{\text{lower}} < \frac{1}{2\sqrt{\lambda_{\max}}}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \tag{9}$$

where $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{\Sigma}^{-1}$. We let the upper bound of $\delta^\top\mathbf{\Sigma}^{-1}\delta$ satisfy the constraint of Eq. 6, therefore a lower bound on $\|\delta\|$ can be obtained as $\|\delta\|_{\text{lower}}$. Eq. 6 is an exact constraint on the perturbation space which is a hyperellipsoid, while Eq. 9 describes minor axis of the hyperellipsoid. Both of them are general expressions for arbitrary GM solvers and joint Gaussian smoothing distributions which will be shown in Sec. 4.3.

4.3 Joint Smoothing Distribution

In contrast to isotropic (Alfarra et al., 2022) and anisotropic (Eiras et al., 2021) distributions, SCR-GM reflects the structure of graph while achieving efficiency by avoiding gradient optimization.

We first construct the correlation matrix $\mathbf{B}$ based on the similarity between nodes in matrix $\mathbf{z}^1$. $\mathbf{B}$ is a full rank and real symmetric matrix whose element b_{mn} denotes the correlation between m-th and n-th node in $\mathbf{z}_1$. We define a similarity using Euclidean distance as follows:

$$b_{mn} = 1/(1 + \frac{d_{mn}}{\gamma}), \tag{10}$$

where d_{mn} is the Euclidean distance between the m-th and n-th nodes, and γ is the normalization coefficient which controls the degree of correlation. We also uses other three similarity measures to construct $\mathbf{B}$ including cosine similarity, pearson similarity and dice similarity as in Appendix C. Nodes in close proximity are more susceptible to perturbations with similar intensity, while perturbations added to nodes at larger distances are almost independent.

The diagonal elements in $\mathbf{B}$ indicate the intensity of perturbation at nodes, while the non-diagonal elements reveal the correlation between nodes. Then by $\mathbf{B}^\top\mathbf{B} = \mathbf{\Sigma}$, we can get the smoothing distribution $\mathbf{\Sigma}$ to sample the additive noise for the input. $\mathbf{\Sigma}$ is a positive definite matrix, which determines the feasibility of radii derived in this work.

In contrast, the distribution in (Eiras et al., 2021) is a diagonal matrix with different diagonal elements, which cannot represent the correlation between nodes; and the distribution in (Alfarra et al., 2022) is a diagonal matrix with the same diagonal elements, which directly treats all nodes without difference. In fact when inter-node correlations and the differences of noise intensity are neglected, $\mathbf{\Sigma}$ can degenerate into the above two distributions. Therefore, $\mathbf{\Sigma}$ is a generalized setting that allows all distributions to be compared in the same framework.

For comparison, we need to keep $\mathbf{\Sigma}$ at the same order of magnitude as the previous three distributions (Cohen et al., 2019; Eiras et al., 2021; Alfarra et al., 2022). We take a similar strategy as that in (Eiras et al., 2021) to ensure that:

$$\min_i \frac{1}{\lambda_i^x} r\left(x, \mathbf{\Sigma}^x\right) \geq \min_i \theta_i^x r\left(x, \mathbf{\Theta}^x\right), \tag{11}$$

where λ_i^x is the eigenvalue of $(\mathbf{\Sigma}^x)^{-1}$, $\mathbf{\Theta}^x$ is the distribution in (Eiras et al., 2021), θ_i^x is the diagonal element of $\mathbf{\Theta}^x$ and $r = \frac{1}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right)$. Therefore, the four distributions mentioned above can be calculated and analyzed incrementally. The visualization of the four distributions calculated from a same original σ (Cohen et al., 2019) are shown in Fig. 1(a). Moreover, $\mathbf{\Sigma}^x$ can trade-off the certified accuracy and radius, the eigenvalue λ_i^x is positively correlated with the certified accuracy and negatively correlated with the certified radius.

4.4 Evaluating Certificates

In Sec. 4.2, Eq. 6 reveals the certified space which is however difficult to quantify and compare. Though Eq. 9 represents a certified and quantifiable form, it actually ignores a large portion of the certified space. We therefore propose two more effective radii to help evaluate the robustness. Eq. 9 is the certification for the worst case of the input, Eq. 13 is the certification for all cases and Eq. 12 reveals the maximum potential for immunity to perturbations. Combining the three radii allows a complete evaluation of the robustness for solvers.

By Lemma 1 and Eq. 7 we define a maximum radius of the certified space:

$$\|\delta\|_{\max} = \frac{1}{2\sqrt{\lambda_{\min}}}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right)), \tag{12}$$

where $\lambda_{\min}$ is the minimum eigenvalue of $\mathbf{\Sigma}^{-1}$, and $\delta^\top \mathbf{\Sigma}^{-1} \delta \geq \lambda_{\min} \delta^\top \delta$. $\|\delta\|_{\max}$ denotes the ℓ_2-norm maximum value for all possible perturbations. Inspired by (Eiras et al., 2021), we can also measure the certified space in terms of ellipsoidal volume. By using the formula for the volume of the ellipsoid: $\mathcal{V}\left(\mathcal{R}\right) = r^n \sqrt{\pi^n} / \Gamma(n/2+1) \prod_{i=1}^n \xi_i$ (Kendall, 2004) where ξ_i is the i-th radius of the ellipsoid, we can get a proxy radius $\|\delta\|_{\text{volume}}$ as:

$$\|\delta\|_{\text{volume}} = r\sqrt{\pi} / \left(\sqrt[n]{\Gamma(n/2+1)} \sqrt[2n]{1/\prod_i^n \lambda_i} \right), \tag{13}$$

where $r = \frac{1}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right)$ and λ_i is the eigenvalue of $\mathbf{\Sigma}^{-1}$. When all λ_i are the same, the certification result will be the same as the traditional method (Cohen et al., 2019). As described in section 4.2, the certified space is a hyperellipsoid geometrically, $\|\delta\|_{\text{lower}}$ represents the minor axis, $\|\delta\|_{\max}$ represents the major axis, $\|\delta\|_{\text{volumn}}$ is a proxy radius of a hypersphere with the same volume as the hyperellipsoid. The whole certification process is shown in Algorithm 1.

5 Experiments

We evaluate our strategy in three aspects: i) For deep graph matching, we compare three radii in Eq. 9, Eq. 12 and Eq. 13 obtained by different certified methods on four GM networks; ii) For non-learning GM methods, we perform synthetic experiments on the widely-used solver RRWM (Cho et al., 2010); iii) We reveal the impact of $\mathbf{\Sigma}$ on the certification results by ablation study.

5.1 Evaluation Settings

Following the GM literature (Wang et al., 2021), we mainly evaluate our method on Pascal VOC dataset (Everingham et al., 2010) with Berkeley annotations (Bourdev & Malik, 2009). All the experiments are conducted on CPU (Intel(R) Core(TM) i7-7820X CPU @ 3.60GHz) and GPU

Algorithm 1 Graph Matching Robustness Certification with SCR-GM.

Input: Graph pair $(\mathcal{G}_1, \mathcal{G}_2)$ of size $\mathbf{z}_1$ and $\mathbf{z}_2$; set of base classifier $\mathbf{F}$; DDRS (Alfarra et al., 2022) and ANCER (Eiras et al., 2021); original σ; normalization coefficient γ; sampling times k_0.
Output: Matching set $\mathbf{M}$ and radius set $\mathbf{\Delta}$.

```
Obtain data-dependent $\sigma_x^*$ by adapting (see details in Appendix C) an off-the-shelf DDRS method (Alfarra et al., 2022) to the graph setting;
Obtain Anisotropic $\mathbf{\Theta}^x$ by adapting (see details in Appendix C)) an off-the-shelf ANCER method (Eiras et al., 2021);
Obtain $\mathbf{B}$ and regularized $\mathbf{\Sigma}$ described in Sec. 4.3 according to Eq. 10 and 11;
Sample $k_0$ noisy samples for left node matrix:$\mathbf{z}_1^{1\prime}, \ldots, \mathbf{z}_{k_0}^{1\prime} \sim \mathcal{N}\left(\mathbf{z}^1, \mathbf{\Sigma}\right)$.
Compute the matching result for nodes in $\mathbf{z}^1$:
    $\mathbf{M} = \{m_i | \arg\max_{r_j \in \mathbf{z}^2} \sum_{k=1}^{k_0} \mathbb{I}\left\{f_i\left(\mathbf{z}_k^{1\prime}\right) = r_j\right\}\}$.
Sample $k (k = 10k_0)$ noisy samples for $\mathcal{G}_1$'s node matrix:$\mathbf{z}_1^{1\prime}, \ldots, \mathbf{z}_k^{1\prime} \sim \mathcal{N}\left(\mathbf{z}^1, \mathbf{\Sigma}\right)$.
Calculate one-sided confidence lower bound $\underline{p_A}$ and $\overline{p_B}$ using $\mathbf{M}$ as described in (Cohen et al., 2019) for every node in $\mathbf{z}^{\mathbf{1}}$, get set $\underline{\mathbf{P}_A}$ and $\overline{\mathbf{P}_B}$.
for $\underline{p_A}$ and $\overline{p_B}$ in $\underline{\mathbf{P}_A}$ and $\overline{\mathbf{P}_B}$ do
    if $\underline{p_A} < \frac{1}{2}$ then
        $m_i$ ABSTAIN; set $||\delta_i||_{\text{lower}}$=$||\delta_i||_{\max}$=$||\delta_i||_{\text{volume}}$=0, append $\mathbf{\Delta}$;
        // Discard nodes with low matching confidence.
    else
        Compute radius $||\delta_i||_{\text{lower}}$, $||\delta_i||_{\max}$ and $||\delta_i||_{\text{volume}}$ described in Sec. 4.4, append $\mathbf{\Delta}$.
    end if
end for
return M, $\mathbf{\Delta}$
```

(GTX 2080 Ti GPU). We validate the certified robustness on four representative deep GM models: GMN (Zanfir & Sminchisescu, 2018), PCA-GM (Wang et al., 2019), CIE-H (Yu et al., 2019a), NGMv2 (Wang et al., 2021) and also a non-deep method RRWM (Cho et al., 2010). In this work, data processing and parameter settings are the same as the original papers unless otherwise specified.

The compared methods include RS (Cohen et al., 2019), DDRS (Alfarra et al., 2022) and ANCER (Eiras et al., 2021). Since the anisotropic method in (Hong & Hong, 2022) is the same as in (Eiras et al., 2021) and (Hong & Hong, 2022) does not provide any code, we choose to compare with (Eiras et al., 2021). We follow the procedure as much similar as possible to that in (Cohen et al., 2019). In (Cohen et al., 2019), the certified accuracy (CA) is defined as: $CA(R) = \mathbb{E}_{x,y}\left[\mathbb{1}(\|\delta\| \geq R)\mathbb{1}\{g(x) = y\}\right]$. In our method, g represents the smoothed function defined in Eq. 4, x denotes the input node in test set, and y is its ground truth matching node. $\|\delta\|$ denotes the certified radius calculated by Eq. 9, Eq. 12, Eq. 13, R is the scale of x-axis, $\mathbb{1}$ is an indicator function. To quantify the improvement, we use Average Certified Radius (ACR) in (Zhai et al., 2020): $\mathbb{E}_{x,y}\left[\|\delta\|\mathbb{1}\{g(x) = y\}\right]$. We use ℓ_2^{lower}, $\ell_2^{\max}$ and ℓ_2^{Σ} to express $||\delta||_{\text{lower}}$, $||\delta||_{\max}$ and $||\delta||_{\text{volume}}$ in the experiments.

5.2 Experiments on Deep Graph Matching

We first set the initial σ of RS to $\sigma \in \{1, 5, 10, 15, 20\}$, and calculate the smoothing distribution of σ_x^* in DDRS and $\mathbf{\Theta}^x$ in ANCER, where iteration number in DDRS and ANCER is equal to 100. Then we set normalization coefficient $\gamma = 5$ and compute the joint distribution matrix $\mathbf{\Sigma}$ of SCR-GM. Fig. 1(b) shows the certified radius $||\delta||_{\text{lower}}$ and ACR on a sample of our method and baselines which indicates that the overall certified robustness of our methods is superior to the baselines. Then we evaluate our strategy on four deep GM methods, the relationship of top-1 certified accuracy and three radii (ℓ_2^{lower}, $\ell_2^{\max}$ and ℓ_2^{Σ}) are plotted in Fig. 2, which only shows the case of $\sigma = 5$. When the radius on x-axis is the same, the higher the certified accuracy on y-axis, the better the certified robustness. The certified accuracy of our method is slightly lower sometimes than baselines when $||\delta||_{\text{lower}}$ is small. However, when $||\delta||_{\text{lower}}$ is large, the accuracy of baseline decreases significantly or even fails completely while our method maintains a more respectable accuracy. When evaluating using $||\delta||_{\max}$ and $||\delta||_{\text{volume}}$, the advantages of our method are more obvious.

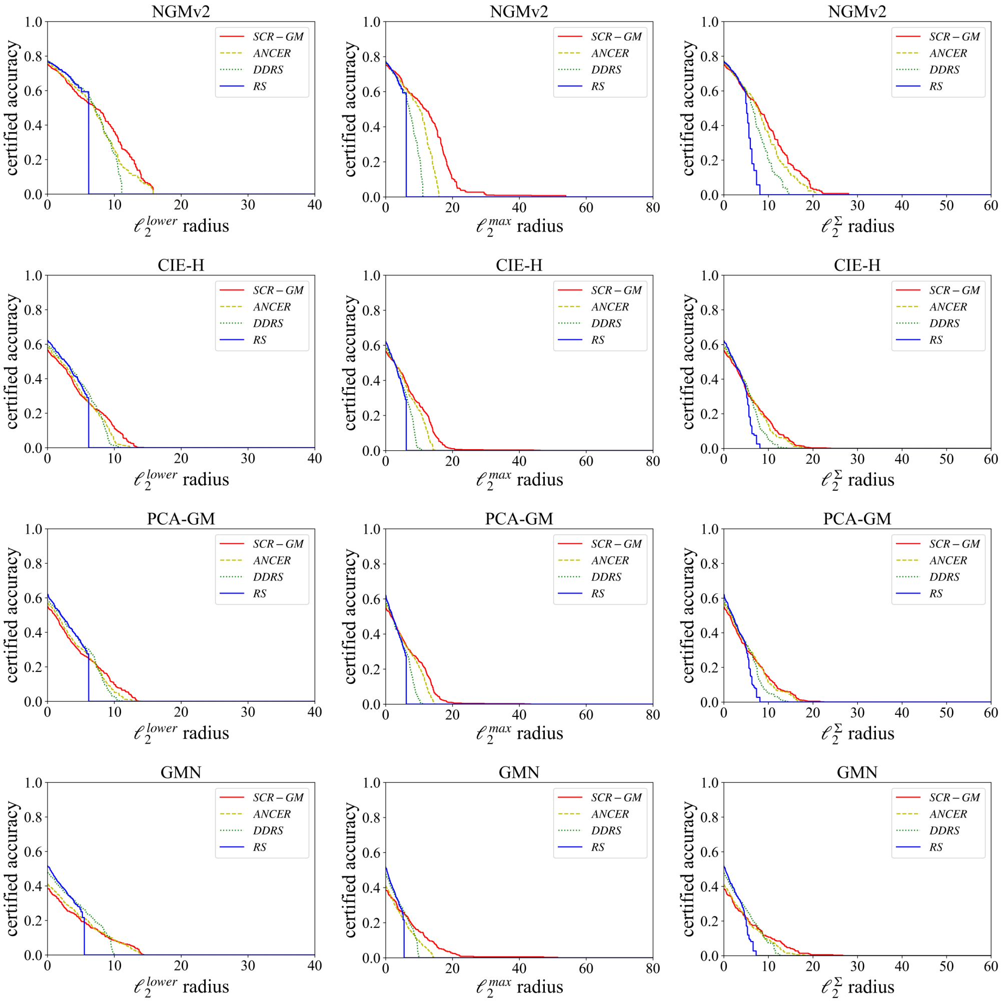
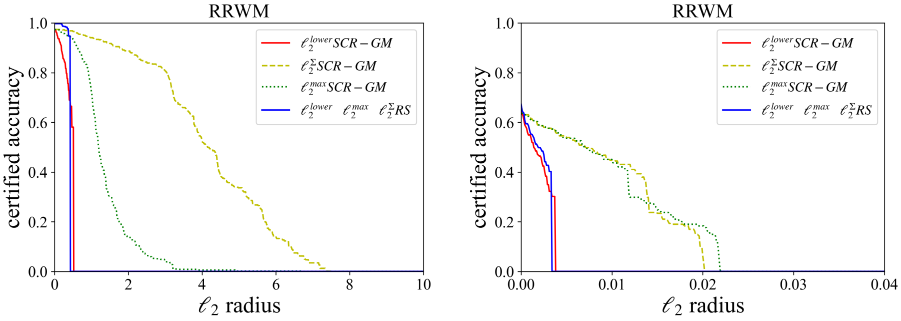


Figure 2: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certification by different RS-type methods ($\sigma = 5$) on four GM methods. SCR-GM almost always achieves a larger certification radius while maintaining the similar certified accuracy.

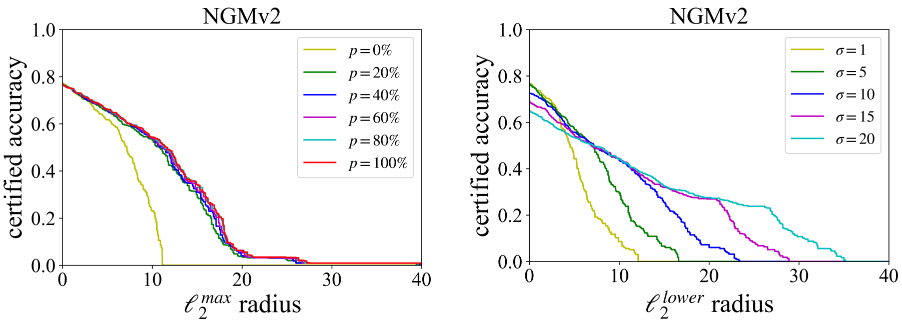


(a) Perturbing node location

(b) Perturbing feature

Figure 3: Certified robustness on non-learning method RRWM by perturbing node location in Fig. 3(a) and perturbing features in Fig. 3(b) on a simulation dataset.

We calculate the ACR $\|\delta\|_{\text{lower}}$ of four different RS-type methods ($\sigma = 5$) and four GM methods as shown in Tab. 1, which indicates that our method shows a better certified robustness performance over the whole dataset. To show the impact of certified robustness on the accuracy of the solvers,



(a) Vary correlation ratio

(b) The trade-off of $\mathbf{\Sigma}$

Figure 4: (a) The effect of joint smoothing distribution $\mathbf{\Sigma}$ on Pascal VOC using SCR-GM for NGMv2. As $\mathbf{\Sigma}$ embodies more correlations between nodes, the wider certified radius is achieved. (b) $\mathbf{\Sigma}$ can trade-off the certified accuracy and radius using SCR-GM for NGMv2 on Pascal VOC.

we use Tab. 2 to show the accuracy of base function, the standard accuracy and certified accuracy of different certified radius $||\delta||_{\text{lower}}$ using NGMv2 algorithm on Pascal VOC dataset. More results are detailed in Appendix D.1.

5.3 Experiments on Non-Learning GM Methods

For non-learning GM, we certify the effectiveness of SCR-GM using simulation experiments on classic non-learning solver RRWM. First we randomly generate two sets of node matrices and calculate their affinity matrix $\mathbf{K}$ using Gaussian kernel affinity function. Then we obtain the robustness results by perturbing node locations and edge features respectively using RS and SCR-GM smoothing distributions. We set $\sigma = 0.5$ and $\sigma = 0.004$ respectively in Fig. 3(a) and 3(b). Our method has similar performance corresponding to the same $||\delta||_{\text{lower}}$ as the baseline. Moreover, it performs better on the other two cases which indicates that the guarantee space certified by our method is wider and its overall robustness is better. We only compare the results using RS and SCR-GM in this experiment, because DDRS and ANCER require the gradient optimization of networks, and they are not applicable to non-learning GM solvers.

5.4 The Effect of Joint Smoothing Distribution

First, we simplify $\mathbf{B}$ by retaining only the higher correlation values in the matrix according to the correlation radio p and setting other values to 0. The radio is set to $p \in \{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$ where 100% represents SCR-GM retaining all the correlation coefficients and 0% represents ANCER without correlation coefficients. Results in Fig. 4(a) demonstrate the effectiveness of the Σ which can be used to get a better certified robustness properties. Then, we verify the impact of initial σ for $\mathbf{\Sigma}$ and the results are plotted in Fig. 4(b). Hyperparameter σ determines the scale of $\mathbf{\Sigma}$ which controls a trade-off between certified robustness and accuracy.

6 Conclusion and Outlook

We have proposed a definition of certified robustness on structural graph matching and design a method SCR-GM that utilizes the correlation between nodes to construct a joint smooth distribution. We obtain ℓ_2 norm certified space and radius for certification. For evaluation, we propose two additional radii by eigenvalue properties. Experiments on deep GM networks and classic solvers show that our method achieves a state-of-art robustness guarantee.

Potential impact & limitations. The currently technique is confined with the graph in Euclidean space (and specifically 2D graphs for experiments), a more general formulation is QAP where the perturb may be directly added on the affinity matrix $\mathbf{K}$. A significant direction is enabling robustness certification on the combinatorial solvers whereby GM is one of such cases. We hope this work can inspire subsequent research in this promising area where theoretical results are welcomed given the recent intensive empirical studies (Bengio et al., 2021; Yan et al., 2020).

REFERENCES

Motasem Alfarra, Adel Bibi, Philip Torr, and Bernard Ghanem. Data dependent randomized smoothing. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 290(2): 405–421, 2021.

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

Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 1365–1372. IEEE, 2009.

RM Burkard and S Dell'Amico. Martello assignment problems. *SIAM, Society for Industrial and Applied Mathematics: Philadelphia, PA, USA*, 2009.

Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems*, 33:1275–1286, 2020.

Minsu Cho, Jungmin Lee, and Kyoung Mu Lee. Reweighted random walks for graph matching. In *European conference on Computer vision*, pp. 492–505. Springer, 2010.

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

Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In *International conference on machine learning*, pp. 1115–1124. PMLR, 2018.

Francisco Eiras, Motasem Alfarra, M Pawan Kumar, Philip HS Torr, Puneet K Dokania, Bernard Ghanem, and Adel Bibi. Ancer: Anisotropic certification via sample-wise volume maximization. *arXiv preprint arXiv:2107.04570*, 2021.

Frank Emmert-Streib, Matthias Dehmer, and Yongtang Shi. Fifty years of graph matching, network alignment and network comparison. *Information sciences*, 346:180–197, 2016.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Simon Geisler, Johanna Sommer, Jan Schuchardt, Aleksandar Bojchevski, and Stephan Günnemann. Generalization of neural combinatorial solvers through the lens of adversarial robustness. *arXiv preprint arXiv:2110.10942*, 2021.

Hanbin Hong and Yuan Hong. Certified adversarial robustness via anisotropic randomized smoothing. *arXiv preprint arXiv:2207.05327*, 2022.

Kenneth Hung and William Fithian. Rank verification for exponential families. *The Annals of Statistics*, 47(2):758–782, 2019.

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

Zetian Jiang, Tianzhe Wang, and Junchi Yan. Unifying offline and online multi-graph matching via finding shortest paths on supergraph. *TPAMI*, 43(10):3648–3663, 2021.

Maurice G Kendall. *A Course in the Geometry of n Dimensions*. Courier Corporation, 2004.

Pol Labarbarie, Hatem Hajri, and Marc Arnaudon. Riemannian data-dependent randomized smoothing for neural networks certification. *arXiv preprint arXiv:2206.10235*, 2022.

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

M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Tenth IEEE International Conference on Computer Vision*, 2005.

Eliane Maria Loiola, Nair Maria Maia de Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. A survey for the quadratic assignment problem. *European journal of operational research*, 176(2):657–690, 2007.

Han Lu, Zenan Li, Runzhong Wang, Qibing Ren, Junchi Yan, and Xiaokang Yang. Mind your solver! on adversarial attack and defense for combinatorial optimization. *arXiv preprint arXiv:2201.00402*, 2021.

Jiaxiang Ren, Zijie Zhang, Jiayin Jin, Xin Zhao, Sixing Wu, Yang Zhou, Yelong Shen, Tianshi Che, Ruoming Jin, and Dejing Dou. Integrated defense for resilient graph matching. In *International Conference on Machine Learning*, pp. 8982–8997. PMLR, 2021.

Qibing Ren, Qingquan Bao, Runzhong Wang, and Junchi Yan. Appearance and structure aware robust deep visual graph matching: Attack, defense and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15263–15272, 2022.

Michal Rolínek, Vít Musil, Anselm Paulus, Marin Vlastelica, Claudio Michaelis, and Georg Martius. Optimizing rank-based metrics with blackbox differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7620–7630, 2020.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Lichao Sun, Yingtong Dou, Carl Yang, Ji Wang, Philip S Yu, Lifang He, and Bo Li. Adversarial attack and defense on graph data: A survey. *arXiv preprint arXiv:1812.10528*, 2018.

Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3056–3065, 2019.

Runzhong Wang, Junchi Yan, and Xiaokang Yang. Neural graph matching network: Learning lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

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

Junchi Yan, Xu-Cheng Yin, Weiyao Lin, Cheng Deng, Hongyuan Zha, and Xiaokang Yang. A short survey of recent advances in graph matching. In *ICMR*, 2016.

Junchi Yan, Shuang Yang, and Edwin Hancock. Learning graph matching and related combinatorial optimization problems. In *IJCAI*, 2020.

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pp. 10693–10705. PMLR, 2020.

Tianshu Yu, Runzhong Wang, Junchi Yan, and Baoxin Li. Learning deep graph matching with channel-independent embedding and hungarian attention. In *International conference on learning representations*, 2019a.

Yu-Feng Yu, Guoxia Xu, Min Jiang, Hu Zhu, Dao-Qing Dai, and Hong Yan. Joint transformation learning via the l 2, 1-norm metric for robust graph matching. *IEEE transactions on cybernetics*, 51(2):521–533, 2019b.

Andrei Zanfir and Cristian Sminchisescu. Deep learning of graph matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2684–2693, 2018.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. *arXiv preprint arXiv:2001.02378*, 2020.

Zijie Zhang, Zeru Zhang, Yang Zhou, Yelong Shen, Ruoming Jin, and Dejing Dou. Adversarial attacks on deep graph matching. *Advances in Neural Information Processing Systems*, 33:20834–20851, 2020.

Daniel Zügner and Stephan Günnemann. Certifiable robustness of graph convolutional networks under structure perturbations. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1656–1665, 2020.

The appendix contains rich supplemental material to enrich the technical details of the paper, as well as our originally proposed techniques for certifying the overall solution $\mathbf{X}$.

A Proofs of Theorem 1

Here we provide the complete proof for Theorem 1. We first prove the following Lemma 2 which is inspired by the Neyman-Pearson for Gaussians lemma derived in (Cohen et al., 2019) and introduce Lemma 3 which makes random vector independent after linear transformation.

Lemma 2 (Neyman-Pearson for Joint Gaussian Noise) *Let $X \sim \mathcal{N}(x, \Sigma)$ and $Y \sim \mathcal{N}(x+\delta, \Sigma)$. Let $h : \mathbb{R}^d \to \{0,1\}$ be any deterministic or random function. Then:*

1. If $S = \left\{k \in \mathbb{R}^d : \delta^T \Sigma^{-1} k \le \beta\right\}$ for some β and $\mathbb{P}(h(X)=1) \ge \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y)=1) \ge \mathbb{P}(Y \in S)$.

2. If $S = \left\{k \in \mathbb{R}^d : \delta^T \Sigma^{-1} k \ge \beta\right\}$ for some β and $\mathbb{P}(h(X)=1) \le \mathbb{P}(X \in S)$, then $\mathbb{P}(h(Y)=1) \le \mathbb{P}(Y \in S)$.

Proof. This lemma is the special case of Neyman-Pearson when X and Y are joint Gaussian noises with means x and $x+\delta$. It suffices to simply show that for any β, there is some $t > 0$ for which:

$$\begin{aligned} \left\{k : \delta^T \Sigma^{-1} k \le \beta\right\} &= \left\{z : \frac{\mu_Y(k)}{\mu_X(k)} \le t\right\}, \\ \left\{k : \delta^T \Sigma^{-1} k \ge \beta\right\} &= \left\{z : \frac{\mu_Y(k)}{\mu_X(k)} \ge t\right\}. \end{aligned} \tag{14}$$

For ease of representation, we use $S \in \mathbb{R}^{d\times d}$ (with element s_{ij}) instead of Σ^{-1}. The likelihood ratio for this choice of X and Y turns out to be:

$$\begin{aligned} \frac{u_Y(k)}{u_X(k)} &= \frac{\exp\left(-\frac{1}{2}(k-(x+\delta))^T\Sigma^{-1}(k-(x+\delta))\right)}{\exp\left(-\frac{1}{2}(k-x)^T\Sigma^{-1}(k-x)\right)} \\ &= \frac{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(k_i-(x_i+\delta_i)\right)s_{ij}\left(k_j-(x_j+\delta_j)\right)\right)}{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(k_i-x_i\right)s_{ij}\left(k_j-x_j\right)\right)} \\ &= \exp\left(\delta^T\Sigma^{-1}k - \delta^T\Sigma^{-1}x - \frac{1}{2}\delta^T\Sigma^{-1}\delta\right) \\ &= \exp\left(\delta^T\Sigma^{-1}k + b\right) \le t, \end{aligned}$$

where b is a constant, specifically $b = -\delta^T\Sigma^{-1}x - \frac{1}{2}\delta^T\Sigma^{-1}\delta$. Therefore given any β, we may take $t = \exp(\beta + b)$ and get this correlation:

$$\begin{aligned} \delta^T\Sigma^{-1}k \le \beta &\Longleftrightarrow \exp\left(\beta+b\right) \le t, \\ \delta^T\Sigma^{-1}k \ge \beta &\Longleftrightarrow \exp\left(\beta+b\right) \ge t. \end{aligned} \tag{15}$$

Lemma 3 (Joint Gaussian Distribution) *If there is a random vector $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^n$ is the mean vector. A positive semi-definite real symmetric matrix $\Sigma \in \mathbb{S}_{++}^{n\times n}$ is the covariance matrix of X. There is a full rank matrix $B \in \mathbb{R}^{n\times n}$, which makes $X = BZ + \mu$, $Z \sim \mathcal{N}(\mathbf{0}, I)$ and $B^\top B = \Sigma$.*

Then we can prove Theorem 1, recall:

Theorem 1.

Let $f_i(\mathbf{z}^1)$ be node matching function, g_i be defined as in Eq. 4, and $\varepsilon \sim \mathcal{N}(0, \mathbf{\Sigma})$. If $\underline{p_A} \in [0,1]$ and $\overline{p_B} \in [0,1]$ satisfy:

$$\mathbb{P}\left(f_i(\mathbf{z}^1+\varepsilon) = r_A\right) \ge \underline{p_A} \ge \overline{p_B} \ge \mathbb{P}(f_i(\mathbf{z}^1+\varepsilon) = r_B). \tag{16}$$

Then for $g_i(\mathbf{z}_1 + \delta) = r_A$, we can get the certified ℓ_2 space for the addictive noise δ:

$$\|\delta^\top \mathbf{B}^{-1}\| < \frac{1}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \tag{17}$$

where $\mathbf{B}^\top\mathbf{B} = \mathbf{\Sigma}$, $\mathbf{B} \in \mathbb{R}^{n_1 \times n_1}$ is a full rank and real symmetric matrix based on the physical relationships in node matrix $\mathbf{z}^1$, and $\underline{p_A}$ and $\overline{p_B}$ are the lower bound of p_A and the upper bound of p_B, respectively.

To show that $g_i(\mathbf{z}_1 + \delta) = r_A$, it follows from the definition of g_i that we need to show that

$$\mathbb{P}\left(f_i(\mathbf{z}^1 + \delta + \varepsilon) = r_A\right) \geq \mathbb{P}(f_i(\mathbf{z}^1 + \delta + \varepsilon) = r_B).$$

In the derivation, r_B is actually not just "runner-up" node, but any node that is different from r_A. We define the random variables:

$$\begin{aligned} X &:= \mathbf{z}^1 + \varepsilon = \mathcal{N}\left(\mathbf{z}^1, \Sigma\right), \\ Y &:= \mathbf{z}^1 + \delta + \varepsilon = \mathcal{N}\left(\mathbf{z}^1 + \delta, \Sigma\right). \end{aligned}$$

We know that:

$$\begin{aligned} \mathbb{P}\left(f_i(X) = r_A\right) &\geq \underline{p_A}, \\ \mathbb{P}\left(f_i(X) = r_B\right) &\leq \overline{p_B}. \end{aligned} \tag{18}$$

Our goal is to show that

$$\mathbb{P}\left(f_i(Y) = r_A\right) > \mathbb{P}\left(f_i(Y) = r_B\right). \tag{19}$$

According to lemma 2, we can define the half-spaces:

$$\begin{aligned} A &= \left\{k : \delta^T \Sigma^{-1}(k - \mathbf{z}^1) \leq \|\delta^T \Sigma^{-1}\mathbf{B}\| \Phi^{-1}\left(\underline{p_A}\right)\right\}, \\ B &= \left\{k : \delta^T \Sigma^{-1}(k - \mathbf{z}^1) \geq \|\delta^T \Sigma^{-1}\mathbf{B}\| \Phi^{-1}\left(1 - \overline{p_B}\right)\right\}. \end{aligned}$$

Claim 1 shows that $\mathbb{P}(X \in A) = \underline{p_A}$, therefore we can get $\mathbb{P}\left(f_i(X) = r_A\right) \geq \mathbb{P}(X \in A)$. Hence we may apply Lemma 2 with $h(z) := \mathbf{1}\left[f_i(z) = r_A\right]$ to conclude:

$$\mathbb{P}\left(f_i(Y) = r_A\right) \geq \mathbb{P}(Y \in A). \tag{20}$$

Similarly, we obtain $\mathbb{P}\left(f_i(X) = r_B\right) \leq \mathbb{P}(X \in B)$. Hence we may apply Lemma 2 with $h(z) := \mathbf{1}\left[f_i(z) = r_B\right]$ to conclude:

$$\mathbb{P}\left(f_i(Y) = r_B\right) \leq \mathbb{P}(Y \in B). \tag{21}$$

Combining Eq. 20 and 21, we can get the conditions of Eq. 19:

$$\mathbb{P}\left(f(Y) = r_A\right) \geq \mathbb{P}(Y \in A) > \mathbb{P}(Y \in B) \geq \mathbb{P}\left(f(Y) = r_B\right). \tag{22}$$

According to Claim 3 and Claim 4, we can get $\mathbb{P}(Y \in A)$ and $\mathbb{P}(Y \in B)$ as:

$$\begin{aligned} \mathbb{P}(Y \in A) &= \Phi\left(\Phi^{-1}\left(\underline{p_A}\right) - \frac{\delta^T \Sigma^{-1}\delta}{\|\delta^T \Sigma^{-1}\mathbf{B}\|}\right), \\ \mathbb{P}(Y \in B) &= \Phi\left(\Phi^{-1}\left(\overline{p_B}\right) + \frac{\delta^T \Sigma^{-1}\delta}{\|\delta^T \Sigma^{-1}\mathbf{B}\|}\right). \end{aligned} \tag{23}$$

Finally, we obtain that $\mathbb{P}(Y \in A) > \mathbb{P}(Y \in B)$ if and only if:

$$\begin{aligned} \frac{\delta^T \Sigma^{-1}\delta}{\|\delta^T \Sigma^{-1}\mathbf{B}\|} &< \frac{1}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right), \\ \frac{\delta^T (\mathbf{B}^T\mathbf{B})^{-1}\delta}{\|\delta^T (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}\|} &< \frac{1}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right). \end{aligned}$$

Because $\mathbf{B}$ is a real symmetric matrix ($\mathbf{B}^T = \mathbf{B}$), we can finally get:

$$\|\delta^T \mathbf{B}^{-1}\| < \frac{1}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right),$$

which recovers the theorem statement.

A.1 Linear transformation and derivation

We obtain four equations based on linear transformation:

Claim 1. $\mathbb{P}(X \in A) = \underline{p_A}$

Proof. Recall that $A = \left\{k : \delta^T\Sigma^{-1}(k - \mathbf{z_1}) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right\}$ and $X \sim \mathcal{N}(\mathbf{z_1}, \Sigma)$, according to lemma 3, we can get:

$$\begin{aligned}
\mathbb{P}(X \in A) =& \ \mathbb{P}\left(\delta^T\Sigma^{-1}(X - \mathbf{z_1}) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}\mathcal{N}(0, \Sigma) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}\mathbf{B}\mathcal{N}(0, I) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \mathbb{P}\left(\|\delta^T\Sigma^{-1}\mathbf{B}\|\mathcal{N}(0, 1) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \Phi\left(\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \underline{p_A}.
\end{aligned}$$

Claim 2. $\mathbb{P}(X \in B) = \overline{p_B}$

Proof. Recall that $B = \left\{k : \delta^T\Sigma^{-1}(k - \mathbf{z_1}) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1 - \overline{p_B}\right)\right\}$ and $X \sim \mathcal{N}(\mathbf{z_1}, \Sigma)$, according to lemma 3, we can get:

$$\begin{aligned}
\mathbb{P}(X \in B) =& \ \mathbb{P}\left(\delta^T\Sigma^{-1}(X - \mathbf{z_1}) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1 - \overline{p_B}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}\mathcal{N}(0, \Sigma) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1 - \overline{p_B}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}\mathbf{B}\mathcal{N}(0, I) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1 - \overline{p_B}\right)\right) \\
=& \ \mathbb{P}\left(\|\delta^T\Sigma^{-1}\mathbf{B}\|\mathcal{N}(0, 1) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1 - \overline{p_B}\right)\right) \\
=& \ 1 - \Phi\left(\Phi^{-1}\left(1 - \overline{p_B}\right)\right) \\
=& \ \overline{p_B}.
\end{aligned}$$

Claim 3. $\mathbb{P}(Y \in A) = \Phi\left(\Phi^{-1}\left(\underline{p_A}\right) - \frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right)$

Proof. Recall that $A = \left\{k : \delta^T\Sigma^{-1}(k - \mathbf{z_1}) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right\}$ and $Y \sim \mathcal{N}(\mathbf{z_1} + \delta, \Sigma)$, according to lemma 3, we can get:

$$\begin{aligned}
\mathbb{P}(Y \in A) =& \ \mathbb{P}\left(\delta^T\Sigma^{-1}(Y - \mathbf{z_1}) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}\mathcal{N}(\delta, \Sigma) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}(\mathbf{B}\mathcal{N}(0, I) + \delta) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}\mathbf{B}\mathcal{N}(0, I) + \delta^T\Sigma^{-1}\delta \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right)\right) \\
=& \ \mathbb{P}\left(\|\delta^T\Sigma^{-1}\mathbf{B}\|\mathcal{N}(0, 1) \leq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p_A}\right) - \delta^T\Sigma^{-1}\delta\right) \\
=& \ \mathbb{P}\left(\mathcal{N}(0, 1) \leq \Phi^{-1}\left(\underline{p_A}\right) - \frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right) \\
=& \ \Phi\left(\Phi^{-1}\left(\underline{p_A}\right) - \frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right).
\end{aligned}$$

Claim 4. $\mathbb{P}(Y \in \mathbf{B}) = \Phi\left(\Phi^{-1}\left(\overline{p_B}\right) + \frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right)$

Proof. Recall that $B = \left\{k : \delta^T\Sigma^{-1}(k - \mathbf{z_1}) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1 - \overline{p_B}\right)\right\}$ and $Y \sim \mathcal{N}(\mathbf{z_1} + \delta, \Sigma)$, according to lemma 3, we can get:

$$\begin{aligned}
\mathbb{P}(Y \in B) = \;& \mathbb{P}\left(\delta^T\Sigma^{-1}(Y-\mathbf{z_1}) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\
=\;& \mathbb{P}\left(\delta^T\Sigma^{-1}\mathcal{N}\left(\delta,\Sigma\right) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\
=\;& \mathbb{P}\left(\delta^T\Sigma^{-1}(\mathbf{B}\mathcal{N}\left(0,I\right)+\delta) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\
=\;& \mathbb{P}\left(\delta^T\Sigma^{-1}\mathbf{B}\mathcal{N}\left(0,I\right)+\delta^T\Sigma^{-1}\delta \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1-\overline{p_B}\right)\right) \\
=\;& \mathbb{P}\left(\|\delta^T\Sigma^{-1}\mathbf{B}\|\mathcal{N}\left(0,1\right) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(1-\overline{p_B}\right)-\delta^T\Sigma^{-1}\delta\right) \\
=\;& \mathbb{P}\left(\mathcal{N}\left(0,1\right) \geq \Phi^{-1}\left(1-\overline{p_B}\right)-\frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right) \\
=\;& \mathbb{P}\left(\mathcal{N}\left(0,1\right) \leq \Phi^{-1}\left(\overline{p_B}\right)+\frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right) \\
=\;& \Phi\left(\Phi^{-1}\left(\overline{p_B}\right)+\frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right)
\end{aligned}$$

B Robustness guarantee when perturbing features

For GM with input $\left(\mathcal{G}_1,\mathcal{G}_2,\mathbf{z}^1,\mathbf{z}^2\right)$ for matching prediction $\mathbf{X}$, we now focus on the effect of perturbing node features. Recall that the set $\mathbf{F}$ can be expressed as: $\mathbf{F}=\{f_i|f_i:\left(\mathcal{G}_1,\mathcal{G}_2,\mathbf{z}^1,\mathbf{z}^2\right)\rightarrow r_j, i\in n_1, j\in n_2\}$ where $\mathcal{G}_1=\{\mathbf{V}_1,\mathbf{E}_1\}$ and $\mathcal{G}_2=\{\mathbf{V}_2,\mathbf{E}_2\}$, f_i represents that the i-th node in $\mathbf{z}^1$ matches the j-th node in $\mathbf{z}^2$, r_j is the j-th node in $\mathbf{z}^2$. Now we define a new smoothed network g_i that returns whichever node in $\mathbf{z}^2$ is most likely to match the node in $\mathbf{z}^1$ when perturbing node features $\mathbf{V}_1\in\mathbb{R}^{d_v\times n_1}$ by joint smoothing distribution noise:

$$\begin{gathered}
g_i=\underset{r_j\in\mathbf{z}^2}{\arg\max}\mathbb{P}(f_i\left(\mathbf{V}_1+\varepsilon,\mathbf{E}_1,\mathbf{V}_2,\mathbf{E}_2,\mathbf{z}^1,\mathbf{z}^2\right)=r_j), \\
\text{where } \varepsilon\sim\mathcal{N}\left(0,\mathbf{\Sigma}\right), i\in n_1, j\in n_2.
\end{gathered} \tag{24}$$

For notational convenience, we simplify $f_i\left(\mathbf{V}_1+\varepsilon,\mathbf{E}_1,\mathbf{V}_2,\mathbf{E}_2,\mathbf{z}^1,\mathbf{z}^2\right)$ to $f_i(\mathbf{V}_1)$. Suppose that when the base function f_i solves for the optimal matching of node i in $\mathbf{z}^1$, the most probable node r_A is returned with probability $p_A=\max_{s_i\in\mathbf{S}_i}s_i$, where $\mathbf{S}_i$ is the i-th row of $\mathbf{S}$. The probability of "runner-up" node r_B is denoted as p_B, $p_B=\max_{s_i\in\mathbf{S}_i, r_B\neq r_A}s_i$. Similarly, we obtain an ℓ_2 certified space to guarantee robustness of graph matching when perturbing features as follows.

Theorem 2 (ℓ_2 certified space when perturbing features) *Let $f_i(\mathbf{V}_1)$ be node matching function, g_i be defined as in Eq. 24, and $\varepsilon\sim\mathcal{N}\left(0,\mathbf{\Sigma}\right)$. If $\underline{p_A}\in[0,1]$ and $\overline{p_B}\in[0,1]$ satisfy:*

$$\mathbb{P}\left(f_i(\mathbf{V}_1+\varepsilon)=r_A\right)\geq\underline{p_A}\geq\overline{p_B}\geq\mathbb{P}(f_i(\mathbf{V}_1+\varepsilon)=r_B), \tag{25}$$

then for $g_i(\mathbf{V}_1+\delta)=r_A$, we can get the certified ℓ_2 space for the addictive noise δ:

$$\|\delta^\top\mathbf{B}^{-1}\|<\frac{1}{2}\left(\Phi^{-1}\left(\underline{p_A}\right)-\Phi^{-1}\left(\overline{p_B}\right)\right), \tag{26}$$

where $\underline{p_A}$ and $\overline{p_B}$ are the lower bound of p_A and the upper bound of p_B respectively. We set $\mathbf{B}^\top\mathbf{B}=\mathbf{\Sigma}$ where $\mathbf{B}\in\mathbb{R}^{(d_v\times n_1)\times(d_v\times n_1)}$ is a diagonal matrix.

Different from the correlation matrix in Eq. 10, $\mathbf{B}$ is a diagonal matrix similar as (Eiras et al., 2021). However, $\mathbf{B}$ is obtained by structure-based prior knowledge rather than the optimization process in (Eiras et al., 2021). We divide the node feature $\mathbf{V}_1$ into n_1 parts and add independent identically distributed noise of the same intensity (denoted by $b_m, m\in n_1$) to each part. The noise intensity of m-th part b_m is defined as $b_m=\frac{d_m}{d}\sigma$ where d is the whole distance between nodes in $\mathbf{z^1}$, d_m is the distance between the m-th node and other nodes, the original σ is the same as described in (Cohen et al., 2019). This setting indicates that outlier points are more resistant to perturbation. Finally we can derive the same radius forms as Eq. 9, 12 and 13.

C Experimental Setup

In this work, we evaluate our strategy on deep graph matching networks and a classic non-learning solver. The procedures to obtain the baseline networks and the evaluation methods are detailed as follows.

C.1 Baseline of Certification methods

In terms of certification, the baselines we considered are RS (Cohen et al., 2019), DDRS (Alfarra et al., 2022) and ANCER (Eiras et al., 2021). We adapt the off-the-shelf DDRS and ANCER to obtain the data-dependent distribution σ_x^* and anisotropic distribution $\mathbf{\Theta}^x$ for graph matching. We add noise to graphs and use $p_A = \max_{s_i \in \mathbf{S}_i} s_i$ and $p_B = \max_{s_i \in \mathbf{S}_i, r_B \neq r_A} s_i$ to calculate the gap value $\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})$ ($\mathbf{S}_i$ is the i-th row of $\mathbf{S}$). The optimization equations and parameters remain the same as the original algorithms.

Then we use SCR-GM to get our joint distribution $\mathbf{\Sigma}$. Finally, we use the Monte Carlo algorithms in (Cohen et al., 2019) to sample noises according to different distributions and output three radii derived in Sec. 4.2 and 4.4. The sample number n and n_0 are set to 1000 and 100 due to the efficiency of graph matching networks, and other parameters are the same as the original network settings (Cohen et al., 2019). We also use hypothesis test (Hung & Fithian, 2019) as in (Cohen et al., 2019) by using α to represent the probability of getting incorrect matching results. In this paper, we set $\alpha = 0.001$, so there is a high probability (99.9% in this paper) to ensure the certification. α can be set arbitrarily small hence in theory our method is highly reliable.

C.2 Evaluation on Deep Graph Matching

For deep graph matching, we mainly evaluate our method on Pascal VOC dataset (Everingham et al., 2010) with Berkeley annotations (Bourdev & Malik, 2009). We follow the protocol of (Wang et al., 2021) and filter out poorly annotated images. In the experiment, we use 100 inputs (containing approximately 650 nodes) of 20 categories to certify the matching robustness.

We check our strategy on four representative deep graph matching methods: GMN (Zanfir & Sminchisescu, 2018), PCA-GM (Wang et al., 2019), CIE-H (Yu et al., 2019a) and NGMv2 (Wang et al., 2021), while use the checkpoints of these GM models collected by ThinkMatch (`https://github.com/Thinklab-SJTU/ThinkMatch`).

We directly evaluate the certified robustness of these networks without fine-tune training.

C.3 Evaluation on non-learning method

For non-learning method, we mainly evaluate our method on simulation data which contains randomly generated node pairs. In the experiment, we use 100 inputs (each contains 5-10 nodes randomly) and evaluate the strategy on classic solver RRWM (Cho et al., 2010). For evaluation, we extract node features and calculate the affinity matrix $\mathbf{K}$ using Gaussian kernel affinity function. Then we perturb node locations and features separately and obtain the certified robustness results.

C.4 Similarity measures

In addition to Eq. 10, we also uses other three similarity measures to construct B including cosine similarity, pearson similarity and dice similarity as follows.

For two points in the Euclidean space $\mathbb{R}^n$: $A = (a_1, a_2, \cdots, a_n)$ and $B = (b_1, b_2, \cdots, b_n)$, cosine similarity is defined as follows:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\|_2 \cdot \|B\|_2} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}} \in [-1, 1]. \tag{27}$$

Table 1: The ACR $\|\delta\|_{\text{lower}}$ of four different RS-type methods (σ = 5) and four GM methods on Pascal VOC dataset.

	NGMv2	**CIE-H**	**PCA-GM**	**GMN**
RS	4.189	2.880	2.745	2.037
DDRS	5.936	3.505	3.307	2.741
ANCER	6.300	3.367	3.179	2.517
SCR-GM	7.107*	3.726*	3.455*	2.745*

Table 2: The accuracy of base function (BA) of NGMv2, standard accuracy (SA) and certified accuracy (CA) of different certified radius $\|\delta\|_{\text{lower}}$ using NGMv2 algorithm (σ = 5) on Pascal VOC dataset.

	BA (%)	**SA (%)**	**CA (%) R=3.5**	**CA (%) R=7.0**	**CA (%) R=10.5**
SCR-GM	77.3	75.6	63.7	51.5*	36.4*
ANCER	77.3	76.5	64.2	49.1	23.8
DDRS	77.3	77.4*	66.6	50.5	18.2
RS	77.3	76.7	66.9*	0.0	0.0

Pearson similarity is defined as follows:

$$\text{Pearson Similarity}(A, B) = \frac{\text{cov}(A, B)}{\sigma_A \cdot \sigma_B} = \frac{\sum_{i=1} (a_i - \bar{A}) \cdot (b_i - \bar{B})}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{A})^2} \cdot \sqrt{\sum_{i=1}^{n} (b_i - \bar{B})^2}} \in [-1, 1], \tag{28}$$

where $\bar{A} = \sum_{i=1}^{n} a_i/n, \bar{B} = \sum_{i=1}^{n} b_i/n$.

Dice similarity is defined as follows:

$$\text{Dice Similarity}(A, B) = \frac{2\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} (a_i^2 + b_i^2)}, \tag{29}$$

where A and B can not be zero point at the same time.

D Experimental Results

D.1 Certification Results of Deep Graph Matching

D.1.1 Perturbing node location

For perturbing node location, we report certified accuracy at ℓ_2^{lower}, $\ell_2^{\max}$ and ℓ_2^{Σ} radii, for each certified method RS (Cohen et al., 2019), DDRS (Alfarra et al., 2022), ANCER (Eiras et al., 2021) and SCR-GM, each network GMN (Zanfir & Sminchisescu, 2018), PCA-GM (Wang et al., 2019), CIE-H (Yu et al., 2019a) and NGMv2 (Wang et al., 2021), each original σ ($\sigma = 1,\ 5,\ 10,\ 15$ and 20). Figures 5, 6, 7 and 8 show certified results on different graph matching networks, respectively.

In addition, we certify the effect of the normalization parameter γ, and Fig. 12 shows the results on NGMv2 (Wang et al., 2021) algorithm and σ is set to 5. Tab. 3 shows the impact of different choices for constructing $\mathbf{B}$ on the certified robustness. $\mathbf{B}$ constructed by Euclidean distance and Dice similarity perform better on the certified robustness. The advantage of $\mathbf{B}$ constructed by Euclidean distance is more obvious when the radius is larger.

D.1.2 Perturbing features

For perturbing node features, we only compare our strategy with RS (Cohen et al., 2019) due to the excessive inefficiency of DDRS (Alfarra et al., 2022) and ANCER (Eiras et al., 2021). We set original σ as $\sigma = 0.25,\ 0.5,\ 1,\ 1.5$ and 2, other settings are the same as Appendix D.1.1. Fig. 9 shows certified results on different graph matching networks when perturbing node features.

Table 3: The impact of different similarity measures for constructing $\mathbf{B}$ on the certified robustness.

	SA (%)	**CA (%) R=3.0**	**CA (%) R=6.0**	**CA (%) R=9.0**	**CA (%) R=12.0**
Euclidean	75.2	64.3	53.3*	42.0*	24.1*
Dice	75.6*	65.5*	52.3	41.0	23.5
Cosine	75.6	65.1	52.0	41.0	23.5
Pearson	75.6	65.2	51.8	40.7	23.6

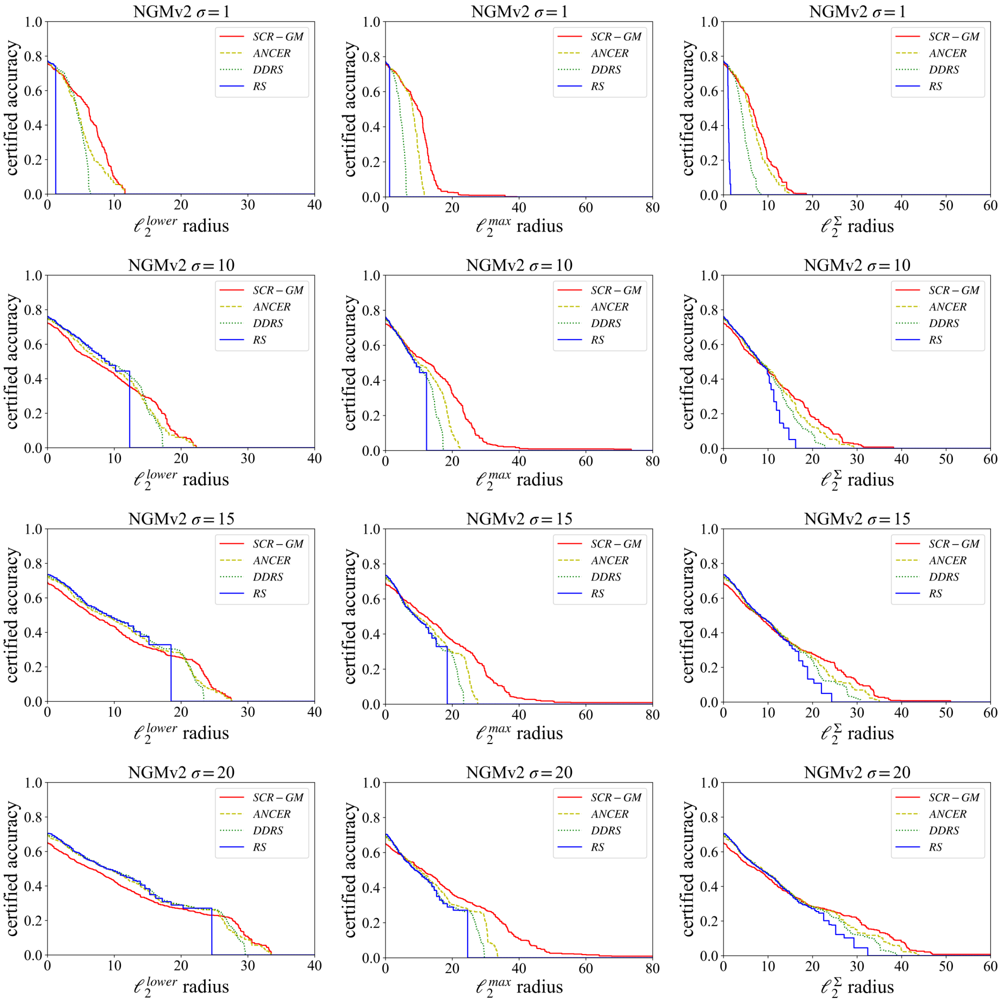


Figure 5: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certification by different RS-type methods on NGMv2 methods. Hyperparameter σ trade-off the certified accuracy and radii.

D.2 Certification Results of non-learning methods

In this section, we report certified accuracy at ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} radii, for certified method (Cohen et al., 2019) and SCR-GM on RRWM (Cho et al., 2010). We set original σ as $\sigma = 0.3,\ 0.4$ and 0.5 when perturbing node locations, while we set $\sigma = 0.001,\ 0.004$ and 0.006 when perturbing features. Fig. 13 and 14 show certified results on the classic solver.

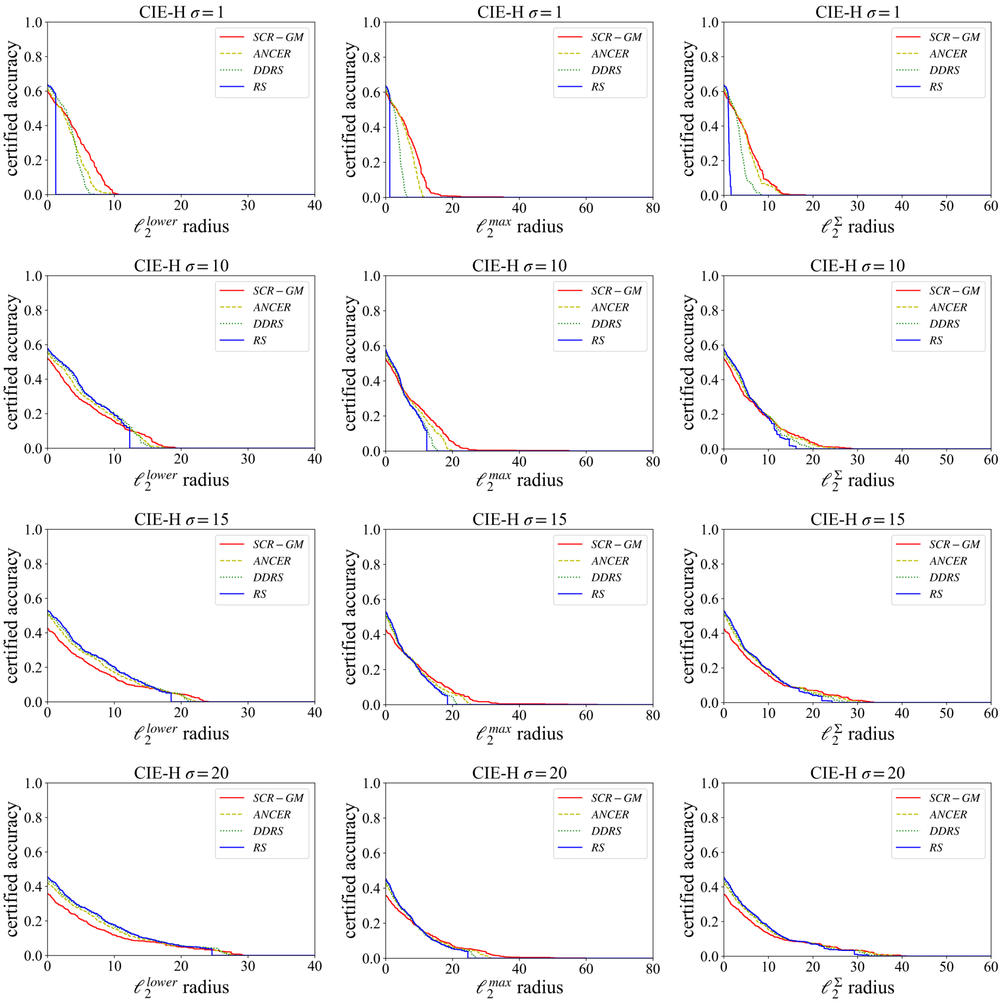


Figure 6: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certification by different RS-type methods on CIE-H methods.

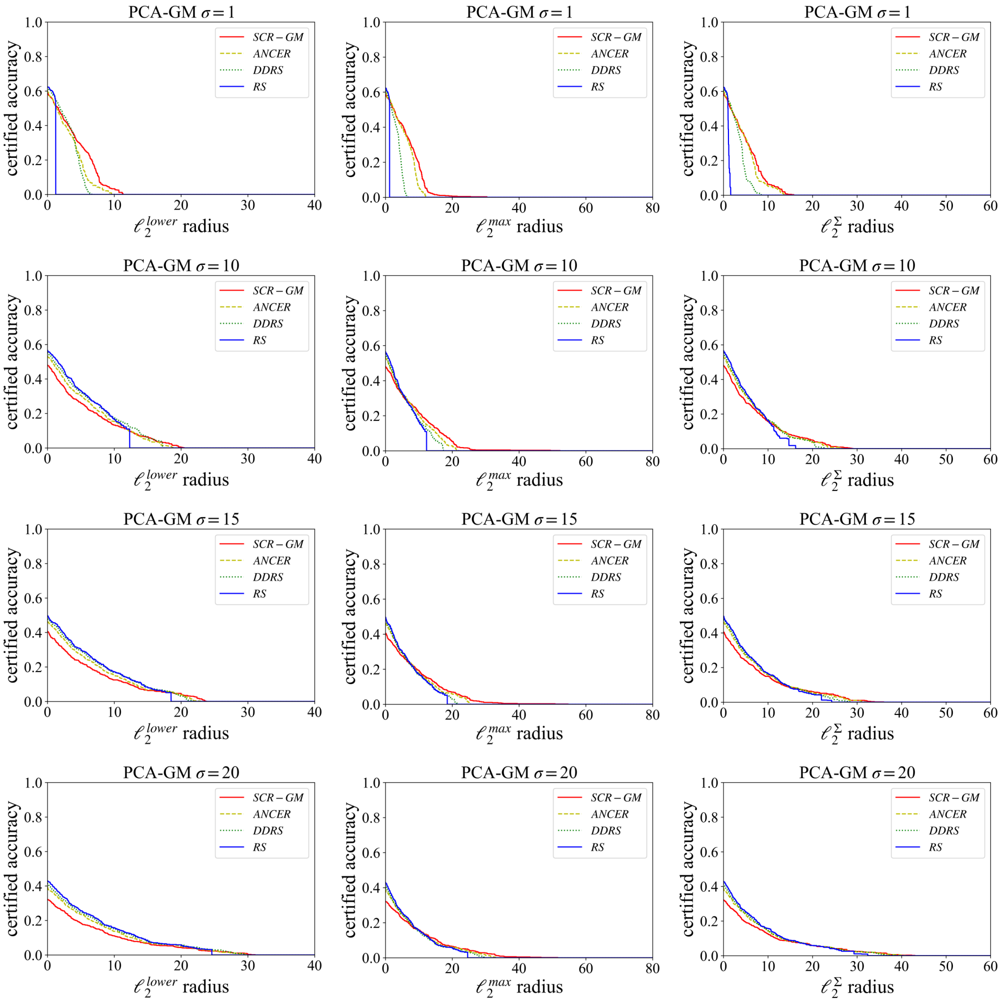


Figure 7: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certification by different RS-type methods on PCA-GM methods.

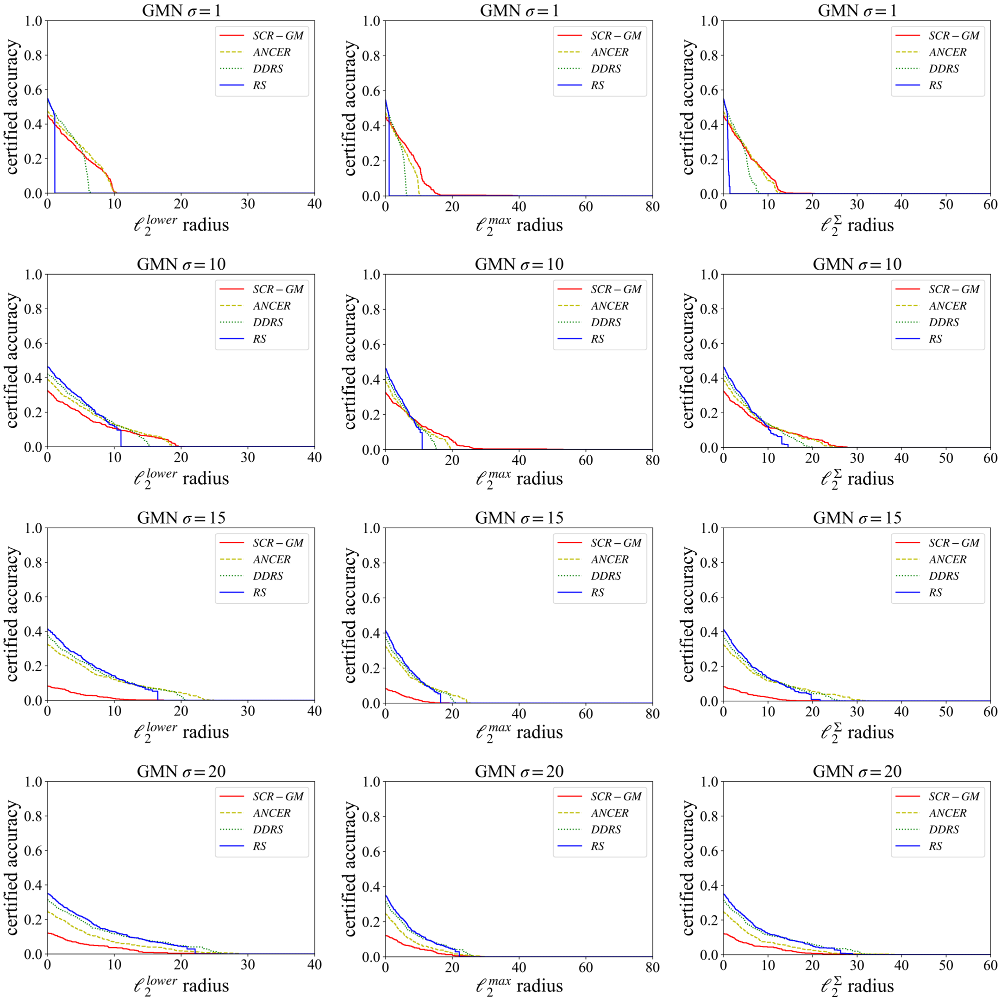


Figure 8: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certification by different RS-type methods on GMN methods.

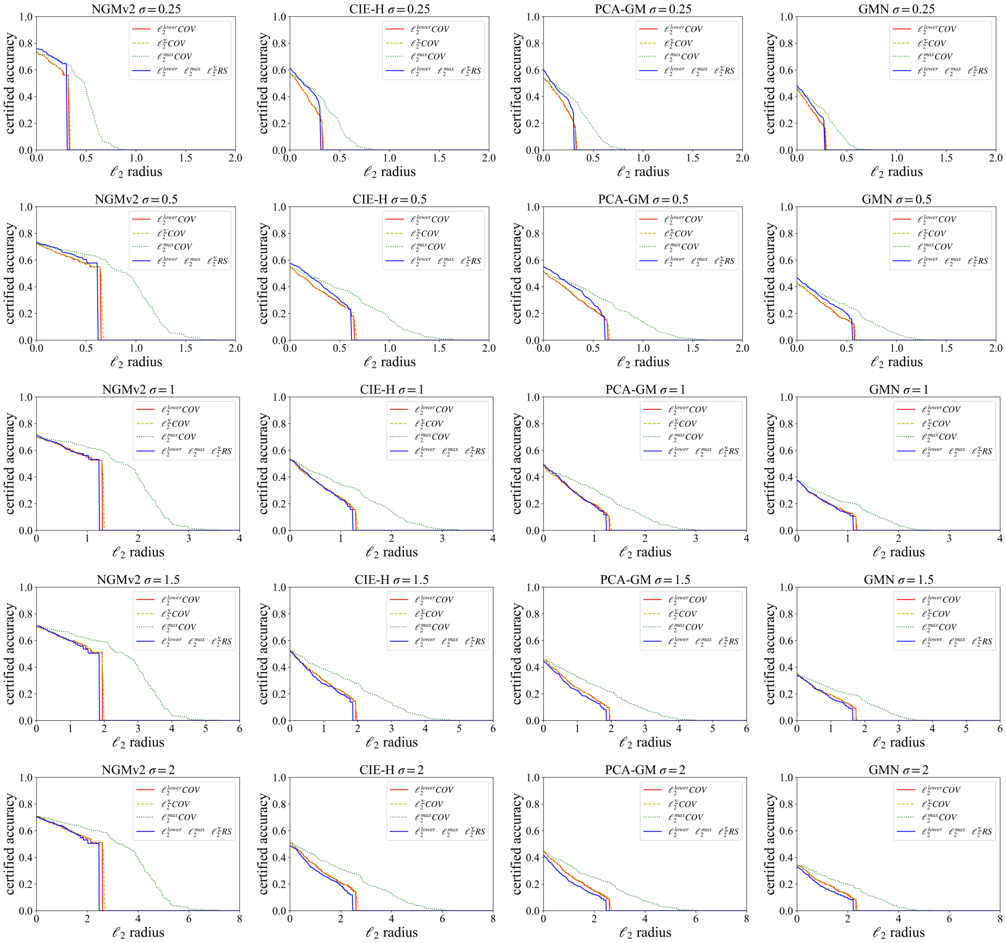


Figure 9: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certification by different RS-type methods on NGMv2, CIE-H, PCA-GM and GMN methods.

E CERTIFIED ROBUSTNESS OF THE SOLUTION X'S STRUCTURE

In Sec. 4, we focus on the certified robustness of node matching results in the graph rather than the whole graph matching result. Our work treats the GM solver as blackbox to get the relaxed matching $\mathbf{S}$, then uses a post-binarization step to to modify the output format X and get the node matching function set $\mathbf{F}$. Then we certify the robustness of $\mathbf{F}$.

However, we can also certify the robustness of the full matrix X which is able to utilize more graph structure information, as well as fully consider the constrains in Eq. 1.

E.1 DEFINITION

Consider a graph matching problem from input space to partial permutation matrices $\mathcal{X}$. As discussed above, randomized smoothing (RS) is a technique for constructing a smoothed function g from an arbitrary base function f. When queried at the input $\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)$, the smoothed function g returns whichever matrix X the base function f is most likely to return when $\mathbf{z}^1$ is perturbed by noise:

$$\begin{gathered} g = \underset{X \in \mathcal{X}}{\arg\max} \mathbb{P}(f\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon, \mathbf{z}^2\right) = X), \\ \text{where } \varepsilon \sim \mathcal{N}\left(0, \mathbf{\Sigma}\right). \end{gathered} \tag{30}$$

The distribution of additive noise ε is a joint Gaussian distribution matrix whose variance $\mathbf{\Sigma}$ represents the correlations between nodes. In addition, $\mathbf{\Sigma}$ is a hyperparameter for certified function which controls the robustness/accuracy trade-off.

E.2 ROBUSTNESS GUARANTEE FOR $\mathcal{X}$

We define a robustness guarantee with confidence $c \in [0, 1]$, which ensures that the similarity between the output matrix calculated by g and its ground truth matrix X_g is not less than a confidence c. Suppose that when the base function f solves $\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon, \mathbf{z}^2\right)$, its output matrices whose similarity to X_g is not less than c are returned with probability p:

$$\begin{gathered} \mathcal{X}' = \left\{ X_i \middle| \frac{X_i \cdot X_g}{X_g \cdot X_g} \geq c, X_i \in \mathcal{X} \right\}, \\ p = \mathbb{P}(X_i | X_i \in \mathcal{X}') \end{gathered} \tag{31}$$

Our main result is that smoothed function g is robust within a ℓ_2 certified space, which also holds if we replace p with a lower bound $\underline{p}$.

Theorem 3 (**ℓ_2 certified space for $\mathcal{X}$**) *Let f be a matching function, g be defined as in Eq. 30, and $\varepsilon \sim \mathcal{N}\left(0, \mathbf{\Sigma}\right)$. Suppose $X_A \in \mathcal{X}'$ and $\underline{p} \in (\frac{1}{2}, 1]$ satisfy:*

$$\mathbb{P}(f\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon, \mathbf{z}^2\right) = X_A, X_A \in \mathcal{X}') \geq \underline{p}. \tag{32}$$

Then we can get the certified ℓ_2 space for the addictive noise δ:

$$\|\delta^\top \mathbf{B}^{-1}\| < \Phi^{-1}\left(\underline{p}\right), \tag{33}$$

which guarantees $g\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \delta, \mathbf{z}^2\right) \in \mathcal{X}'$. In Eq. 6, $\mathbf{B}^\top \mathbf{B} = \mathbf{\Sigma}$ and $\mathbf{B} \in \mathbb{R}^{n_1 \times n_1}$ is a full rank and real symmetric matrix based on the node correlation in node matrix $\mathbf{z}^1$.

The detail settings and properties of $\mathbf{B}$ and $\mathbf{\Sigma}$ are the same as in Section 4.3.

Based on Lemma 1 and the certified space in Eq. 33, we can further obtain a certified ℓ_2 norm radius:

$$\|\delta\|_{\text{lower}} < \frac{1}{\sqrt{\lambda_{\max}}}\left(\Phi^{-1}\left(\underline{p}\right)\right), \tag{34}$$

where $\lambda_{\max}$ is the maximum eigenvalue of $\mathbf{\Sigma}^{-1}$.

We can define a maximum radius of the certified space:

$$\|\delta\|_{\max} = \frac{1}{\sqrt{\lambda_{\min}}}\left(\Phi^{-1}\left(\underline{p}\right)\right)), \tag{35}$$

Algorithm 2 Graph Matching Robustness Certification for X with SCR-GM.

Input: Graph pair $(\mathcal{G}_1, \mathcal{G}_2)$ of size $\mathbf{z}_1$ and $\mathbf{z}_2$; base function f of graph matching; DDRS (Alfarra et al., 2022) and ANCER (Eiras et al., 2021); original σ; normalization coefficient γ; sampling times k_0; matrix similarity confidence c.

Output: Matching result $\hat{X}_g$ and radius R.

```
Obtain data-dependent $\sigma_x^*$ by adapting (see details in Appendix C) an off-the-shelf DDRS method (Alfarra et al., 2022) to the graph setting;
Obtain Anisotropic $\mathbf{\Theta}^x$ by adapting (see details in Appendix C)) an off-the-shelf ANCER method (Eiras et al., 2021);
Obtain $\mathbf{B}$ and regularized $\mathbf{\Sigma}$ described in Sec. 4.3 according to Eq. 10 and 11;
Sample $k_0$ noisy samples for $\mathcal{G}_1$'s node matrix:$\mathbf{z}_1^{1\prime}, \dots, \mathbf{z}_{k_0}^{1}{}' \sim \mathcal{N}\left(\mathbf{z}^1, \mathbf{\Sigma}\right)$.
Compute the approximate ground truth matrix $\hat{X}_g$.
Sample $k (k = 10k_0)$ noisy samples for $\mathcal{G}_1$'s node matrix:$\mathbf{z}_1^{1\prime}, \dots, \mathbf{z}_k^{1\prime} \sim \mathcal{N}\left(\mathbf{z}^1, \mathbf{\Sigma}\right)$ and get an approximate output set $\hat{\mathcal{X}}$.
Calculate one-sided confidence lower bound $\underline{p}$ using set $\hat{\mathcal{X}}$ and Eq. 31.
if $\underline{p} < \frac{1}{2}$ then
    $X$ ABSTAIN; set $\|\delta_i\|_{\text{lower}}$=$\|\delta_i\|_{\max}$=$\|\delta_i\|_{\text{volume}}$=0, append $R$;
    //Discard matching result with low confidence.
else
    Compute radius $\|\delta_i\|_{\text{lower}}$, $\|\delta_i\|_{\max}$ and $\|\delta_i\|_{\text{volume}}$ described in Sec. 4.4, append $R$.
end if
return $\hat{X}_g$, $R$
```

where $\lambda_{\min}$ is the minimum eigenvalue of $\mathbf{\Sigma}^{-1}$. The proxy radius $\|\delta\|_{\text{volume}}$ is as follows:

$$\|\delta\|_{\text{volume}} = r\sqrt{\pi} / \left(\sqrt[n]{\Gamma(n/2+1)} \sqrt[2n]{1 / \prod_i^n \lambda_i} \right). \tag{36}$$

The whole robustness certification process is shown in Algorithm 2. In fact, we cannot get the real X_g and $\mathcal{X}$ during certification stage, so we use Monte Carlo sampling to estimate it. We first sample $f\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon, \mathbf{z}^2\right)$ with n_0 times and add all permutation matrices to get X_s, then we use Sinkhorn and Hungarian algorithm to approximate X_g. During certification, if the approximated $\hat{X}_g$ is not the same as the ground truth matrix X_g, we consider that the certification for this sample has failed. Then we sample $f\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon, \mathbf{z}^2\right)$ with n times and put all possible matrices into set $\hat{\mathcal{X}}$ to approximate $\mathcal{X}$. When n is large, $\hat{\mathcal{X}}$ and $\mathcal{X}$ are relatively close.

E.3 EXPERIMENTS

We evaluate our methods on deep graph matching networks and non-learning solvers. The evaluation settings are the same as in Sec. 5.1.

E.3.1 EXPERIMENTS ON DEEP GRAPH MATCHING

We focus on certifying the robustness of node locality and compare ℓ_2^{lower}, $\ell_2^{\max}$, ℓ_2^{Σ} certification using four certified methods on four deep GM algorithms.

We first set the initial σ of RS to $\sigma \in \{1, 5, 10, 15, 20\}$, the confidence $c = 0.9$ and calculate the smoothing distribution of σ_x^* in DDRS and $\mathbf{\Theta}^x$ in ANCER, where iteration number in DDRS and ANCER is equal to 100. Then we set normalization coefficient $\gamma = 5$ and compute the joint distribution matrix $\mathbf{\Sigma}$ of SCR-GM. Then we evaluate our strategy on four deep GM methods, the relationship of top-1 certified accuracy and three radii (ℓ_2^{lower}, $\ell_2^{\max}$ and ℓ_2^{Σ}) is plotted in Fig. 10. When the radius on x-axis is the same, the higher the certified accuracy on y-axis, the better the certified robustness.

Our method outperforms the baseline on NGMv2 algorithm, which means that the certified accuracy is higher when the radii (ℓ_2^{lower}, $\ell_2^{\max}$ and ℓ_2^{Σ}) is the same. On CIE-H and PCA-GM algorithms, the

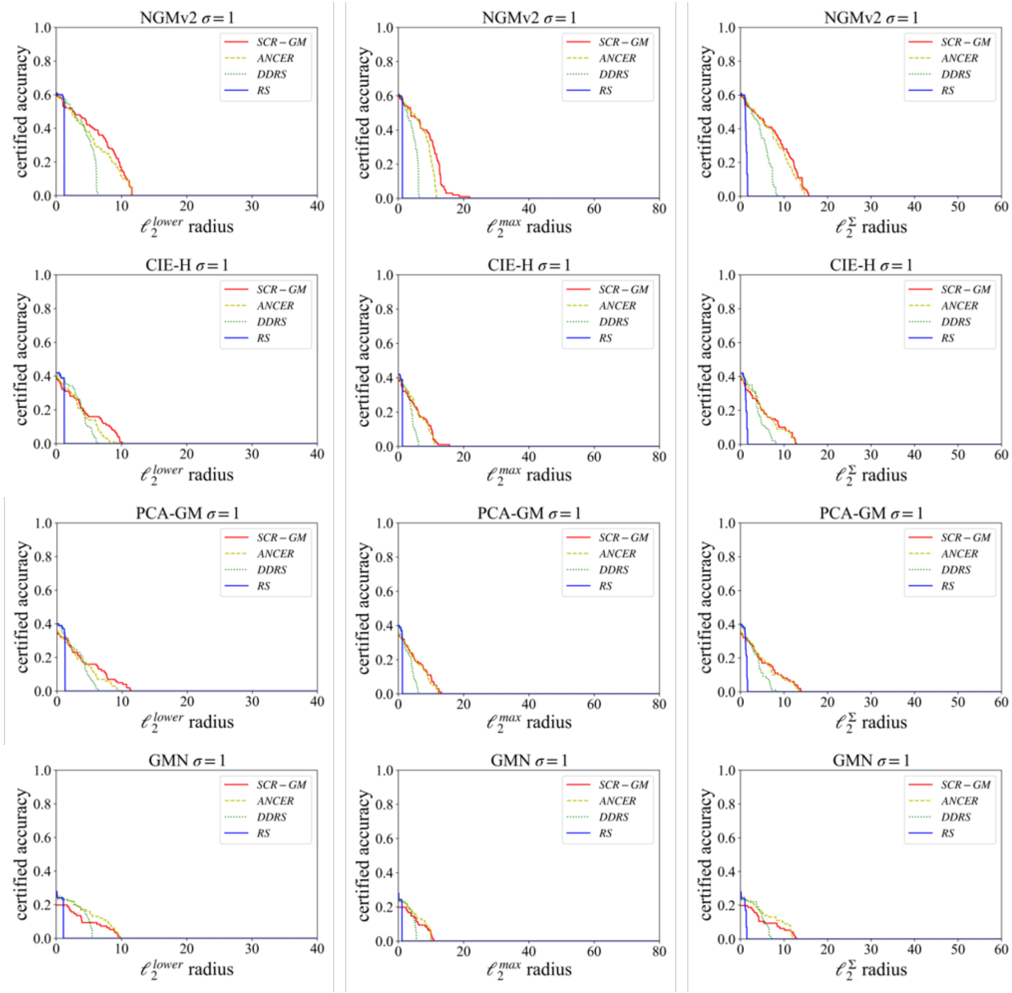
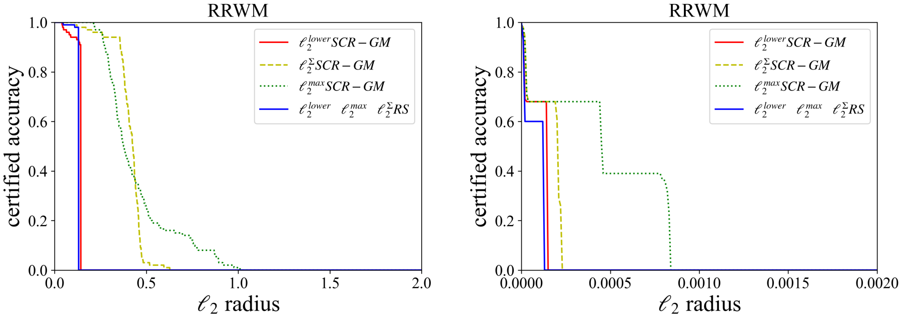


Figure 10: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certifying the full matrix X by different RS-type methods ($\sigma = 1$) on four GM methods. SCR-GM almost always achieves a larger certification radius while maintaining the similar certified accuracy.

certified accuracy of our method is slightly lower sometimes than baselines when ℓ_2^{lower} radius is small. However, when ℓ_2^{lower} radius is large, the accuracy of baselines decrease significantly or even fail completely while our method maintains a more respectable accuracy. When evaluating using $\ell_2^{\max}$ and ℓ_2^{Σ} radii, the certified results of our method are similar as the baselines. On GMN algorithm, our certification results are a bit worse than ANCER. In short, the certified robustness advantage of our method is more obvious on the algorithm with better matching accuracy itself.

E.3.2 Experiments on Non-Learning GM Methods

For non-learning GM, we certify the effectiveness of SCR-GM using simulation experiments on classic non-learning solver RRWM. First we randomly generate two sets of node matrices and calculate their affinity matrix $\mathbf{K}$ using Gaussian kernel affinity function. Then we obtain the robustness results by perturbing node locations and edge features respectively using RS and SCR-GM smoothing distributions. We set $\sigma = 0.1$ and $\sigma = 0.0001$ respectively in Fig. 11(a) and 11(b). Our method has similar performance of the certified accuracy corresponding to the same $\|\delta\|_{\text{lower}}$ with the baseline. However, our method performs better on $\|\delta\|_{\text{volume}}$ and $\|\delta\|_{\max}$ which indicates that the guarantee space certified by our method is wider and its overall robustness is better. We only compare the results using RS and SCR-GM in this experiment, because DDRS and ANCER require the gradient optimization of networks, and they are not applicable to non-learning GM solvers.



(a) Perturbing node location

(b) Perturbing feature

Figure 11: Certified robustness for the full matrix X on non-learning method RRWM by perturbing node location in Fig. 11(a) and perturbing features in Fig. 11(b) on a simulation dataset.

E.4 Proof

To show that $g\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \delta, \mathbf{z}^2\right) \in \mathcal{X}'$, it follows from the definition of g that we need to show that:

$$\mathbb{P}(f\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon + \delta, \mathbf{z}^2\right) = X_A, X_A \in \mathcal{X}') \geq \mathbb{P}(f\left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon + \delta, \mathbf{z}^2\right) = X_B, X_B \notin \mathcal{X}').$$

We define two random variables:

$$\begin{aligned} I &:= \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon, \mathbf{z}^2\right) = \left(\mathcal{G}_1, \mathcal{G}_2, \mathcal{N}\left(\mathbf{z}^1, \Sigma\right), \mathbf{z}^2\right) \\ O &:= \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1 + \varepsilon + \delta, \mathbf{z}^2\right) = \left(\mathcal{G}_1, \mathcal{G}_2, \mathcal{N}\left(\mathbf{z}^1 + \delta, \Sigma\right), \mathbf{z}^2\right). \end{aligned}$$

We know that:

$$\mathbb{P}(f(I) = X_A, X_A \in \mathcal{X}') \geq \underline{p}. \tag{37}$$

Our goal is to show that

$$\mathbb{P}(f(O) = X_A, X_A \in \mathcal{X}') > \mathbb{P}(f(O) = X_B, X_B \notin \mathcal{X}'). \tag{38}$$

According to lemma 2, we can define the half-spaces:

$$\begin{aligned} A &= \left\{k : \delta^T \Sigma^{-1}(k - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right\}, \\ B &= \left\{k : \delta^T \Sigma^{-1}(k - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \geq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right\}. \end{aligned}$$

Claim 1 shows that $\mathbb{P}(I \in A) = \underline{p}$, therefore we can get $\mathbb{P}(f(I) = X_A, X_A \in \mathcal{X}') \geq \mathbb{P}(I \in A)$. Hence we may apply Lemma 2 to conclude:

$$\mathbb{P}(f(O) = X_A, X_A \in \mathcal{X}') \geq \mathbb{P}(O \in A). \tag{39}$$

Similarly, we obtain $\mathbb{P}(f(I) = X_B, X_B \notin \mathcal{X}') \leq \mathbb{P}(I \in B)$. Hence we may apply Lemma 2 to conclude:

$$\mathbb{P}(f(O) = X_B, X_B \notin \mathcal{X}') \leq \mathbb{P}(O \in B). \tag{40}$$

Combining Eq. 39 and 40, we can get the conditions of Eq. 38:

$$\mathbb{P}(f(O) = X_A, X_A \in \mathcal{X}') \geq \mathbb{P}(O \in A) > \mathbb{P}(O \in B) \geq \mathbb{P}(f(O) = X_B, X_B \notin \mathcal{X}'). \tag{41}$$

According to Claim 3 and Claim 4, we can get $\mathbb{P}(O \in A)$ and $\mathbb{P}(O \in B)$ as:

$$\begin{aligned} \mathbb{P}(O \in A) &= \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta^T \Sigma^{-1} \delta}{\|\delta^T \Sigma^{-1} \mathbf{B}\|}\right), \\ \mathbb{P}(O \in B) &= \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta^T \Sigma^{-1} \delta}{\|\delta^T \Sigma^{-1} \mathbf{B}\|}\right). \end{aligned} \tag{42}$$

Finally, we obtain that $\mathbb{P}(O \in A) > \mathbb{P}(O \in B)$ if and only if:

$$\frac{\delta^T \Sigma^{-1} \delta}{\|\delta^T \Sigma^{-1} \mathbf{B}\|} < \Phi^{-1}\left(\underline{p}\right),$$
$$\frac{\delta^T (\mathbf{B}^T \mathbf{B})^{-1} \delta}{\|\delta^T (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}\|} < \Phi^{-1}\left(\underline{p}\right).$$

Since $\mathbf{B}$ is a real symmetric matrix ($\mathbf{B}^T = \mathbf{B}$), we can finally get:

$$\|\delta^T \mathbf{B}^{-1}\| < \Phi^{-1}\left(\underline{p}\right),$$

which recovers the theorem statement.

E.4.1 Linear transformation and derivation

We obtain four equations based on linear transformation: **Claim 1.** $\mathbb{P}(I \in A) = \underline{p}$

Proof. Recall that $A = \left\{k : \delta^T \Sigma^{-1}(k - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right\}$, according to lemma 3, we can get:

$$\begin{aligned}
\mathbb{P}(I \in A) &= \mathbb{P}\left(\delta^T \Sigma^{-1}(I - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\delta^T \Sigma^{-1} \mathcal{N}\left(0, \Sigma\right) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\delta^T \Sigma^{-1} \mathbf{B} \mathcal{N}\left(0, I\right) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\|\delta^T \Sigma^{-1} \mathbf{B}\| \mathcal{N}\left(0, 1\right) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \Phi\left(\Phi^{-1}\left(\underline{p}\right)\right) \\
&= \underline{p}.
\end{aligned}$$

Claim 2. $\mathbb{P}(I \in B) = 1 - \underline{p}$

Proof. Recall that $B = \left\{k : \delta^T \Sigma^{-1}(k - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \geq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right\}$, according to lemma 3, we can get:

$$\begin{aligned}
\mathbb{P}(I \in B) &= \mathbb{P}\left(\delta^T \Sigma^{-1}(I - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \geq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\delta^T \Sigma^{-1} \mathcal{N}\left(0, \Sigma\right) \geq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\delta^T \Sigma^{-1} \mathbf{B} \mathcal{N}\left(0, I\right) \geq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\|\delta^T \Sigma^{-1} \mathbf{B}\| \mathcal{N}\left(0, 1\right) \geq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= 1 - \Phi\left(\Phi^{-1}\left(\underline{p}\right)\right) \\
&= 1 - \underline{p}.
\end{aligned}$$

Claim 3. $\mathbb{P}(O \in A) = \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta^T \Sigma^{-1} \delta}{\|\delta^T \Sigma^{-1} \mathbf{B}\|}\right)$

Proof. Recall that $A = \left\{k : \delta^T \Sigma^{-1}(k - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right\}$ and $O \sim \left(\mathcal{G}_1, \mathcal{G}_2, \mathcal{N}\left(\mathbf{z}^1 + \delta, \Sigma\right), \mathbf{z}^2\right)$, according to lemma 3, we can get:

$$\begin{aligned}
\mathbb{P}(O \in A) &= \mathbb{P}\left(\delta^T \Sigma^{-1}(O - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\delta^T \Sigma^{-1} \mathcal{N}\left(\delta, \Sigma\right) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\delta^T \Sigma^{-1} (\mathbf{B} \mathcal{N}\left(0, I\right) + \delta) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\delta^T \Sigma^{-1} \mathbf{B} \mathcal{N}\left(0, I\right) + \delta^T \Sigma^{-1} \delta \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right)\right) \\
&= \mathbb{P}\left(\|\delta^T \Sigma^{-1} \mathbf{B}\| \mathcal{N}\left(0, 1\right) \leq \|\delta^T \Sigma^{-1} \mathbf{B}\| \Phi^{-1}\left(\underline{p}\right) - \delta^T \Sigma^{-1} \delta\right) \\
&= \mathbb{P}\left(\mathcal{N}\left(0, 1\right) \leq \Phi^{-1}\left(\underline{p}\right) - \frac{\delta^T \Sigma^{-1} \delta}{\|\delta^T \Sigma^{-1} \mathbf{B}\|}\right) \\
&= \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta^T \Sigma^{-1} \delta}{\|\delta^T \Sigma^{-1} \mathbf{B}\|}\right).
\end{aligned}$$

Claim 4. $\mathbb{P}(O \in \mathbf{B}) = \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right)$

Proof. Recall that $B = \left\{k : \delta^T\Sigma^{-1}(k - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p}\right)\right\}$ and $O \sim \left(\mathcal{G}_1, \mathcal{G}_2, \mathcal{N}\left(\mathbf{z}^1 + \delta, \Sigma\right), \mathbf{z}^2\right)$, according to lemma 3, we can get:

$$\begin{aligned}
\mathbb{P}(O \in B) =& \ \mathbb{P}\left(\delta^T\Sigma^{-1}((O - \left(\mathcal{G}_1, \mathcal{G}_2, \mathbf{z}^1, \mathbf{z}^2\right)) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}\mathcal{N}\left(\delta, \Sigma\right) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}(\mathbf{B}\mathcal{N}\left(0, I\right) + \delta) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p}\right)\right) \\
=& \ \mathbb{P}\left(\delta^T\Sigma^{-1}\mathbf{B}\mathcal{N}\left(0, I\right) + \delta^T\Sigma^{-1}\delta \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p}\right)\right) \\
=& \ \mathbb{P}\left(\|\delta^T\Sigma^{-1}\mathbf{B}\|\mathcal{N}\left(0, 1\right) \geq \|\delta^T\Sigma^{-1}\mathbf{B}\|\Phi^{-1}\left(\underline{p}\right) - \delta^T\Sigma^{-1}\delta\right) \\
=& \ \mathbb{P}\left(\mathcal{N}\left(0, 1\right) \geq \Phi^{-1}\left(\underline{p}\right) - \frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right) \\
=& \ \mathbb{P}\left(\mathcal{N}\left(0, 1\right) \leq -\Phi^{-1}\left(\underline{p}\right) + \frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right) \\
=& \ \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta^T\Sigma^{-1}\delta}{\|\delta^T\Sigma^{-1}\mathbf{B}\|}\right)
\end{aligned}$$

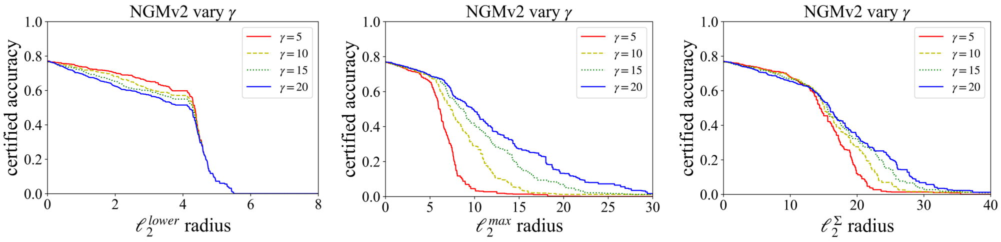


Figure 12: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} for certification by changing γ on NGMv2 methods. When γ is smaller, the correlation values are overall smaller, the hyperellipsoidal volume is smaller, and vice versa.

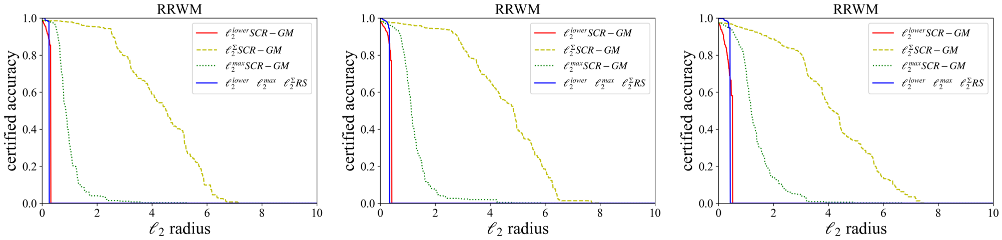


Figure 13: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certification on RRWM methods by perturbing node locations (from left to right $\sigma = 0.3,\ 0.4$ and 0.5).

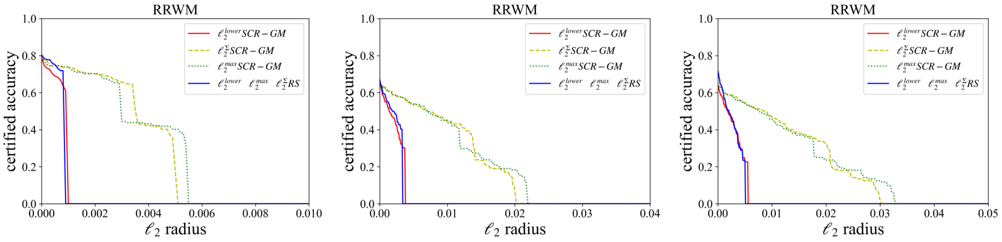


Figure 14: Top-1 certified accuracy on ℓ_2^{lower}, ℓ_2^{max} and ℓ_2^{Σ} certification on RRWM methods by perturbing features (from left to right $\sigma = 0.001,\ 0.004$ and 0.006).