COLLECTIVE ROBUSTNESS CERTIFICATES

Anonymous authors
Paper under double-blind review

ABSTRACT

In tasks like node classification, image segmentation, and named-entity recognition we have a classifier that simultaneously outputs multiple predictions (a vector of labels) based on a single input, i.e. a single graph, image, or document respectively. Existing adversarial robustness certificates consider each prediction independently and are thus overly pessimistic for such tasks. They implicitly assume that an adversary can use different perturbed inputs to attack different predictions, ignoring the fact that we have a *single* shared input. We propose the first collective robustness certificate which computes the number of predictions which are *simultaneously* guaranteed to remain stable under perturbation, i.e. cannot be attacked. We focus on Graph Neural Networks and leverage their locality property – perturbations only affect the predictions in a close neighborhood – to fuse multiple single-node certificates into a drastically stronger collective certificate. For example, on the Citeseer dataset our collective certificate for node classification increases the number of average certifiable feature perturbations from 7 to 351.

1 INTRODUCTION

Most classifiers are vulnerable to adversarial attacks (Akhtar & Mian, 2018; Hao-Chen et al., 2020). Slight perturbations of the data are often sufficient to manipulate their predictions. Even in scenarios where attackers are not present it is critical to ensure that models are robust since data can be noisy, incomplete, or anomalous. Here we study classifiers that collectively output many predictions based on a single input. This includes node classification, link prediction, molecular property prediction, image segmentation, part-of-speech tagging, named-entity recognition, and many other tasks.

Just as with single-output classifiers, one way of addressing the problem of adversarial attacks is to increase the robustness of models through various heuristic defenses. One example is adversarial training (Goodfellow et al., 2015), which has been applied to part-of-speech tagging (Han et al., 2020), semantic segmentation (Xu et al., 2020b) and node classification (Feng et al., 2019). Graph-related tasks in particular have spawned a rich assortment of defenses. These include Bayesian models (Feng et al., 2020), data-augmentation techniques (Entezari et al., 2020) and various robust network architectures (Zhu et al., 2019; Geisler et al., 2020). There are also robust loss functions which either explicitly model an adversary trying to cause misclassifications (Zhou & Vorobeychik, 2020) or use regularization terms derived from robustness certificates (Zügner & Günnemann, 2019). Other methods try to detect adversarially perturbed graphs (Zhang et al., 2019; Xu et al., 2020a) or directly correct perturbations using generative models (Zhang & Ma, 2020).

The problem with such heuristics is that they can only be evaluated based on their ability to defend against known adversarial attacks. Once heuristic defenses are established, they can be easily broken using novel attacks (Carlini & Wagner, 2017). We are therefore interested in deriving adversarial robustness certificates which provide provable guarantees and are by definition unbreakable. Existing certificates consider each prediction independently and are thus overly pessimistic for collective tasks.[1] They implicitly assume that an adversary can use different perturbed inputs to attack different predictions, ignoring the fact that we have a *single* shared input.

In this work we focus on node classification.[2] Here, the goal is assign a label to each node in a single (attributed) graph. By attacking a small part of the graph, i.e. perturbing the features or edges of just a few nodes, an adversary can affect a large number of nodes (Zügner & Günnemann, 2019).

[1] Chiang et al. (2020) certify multi-object detection, but they still treat each detected object independently.

[2] While we focus on node classification, our approach can easily be applied to other multi-output classifiers.

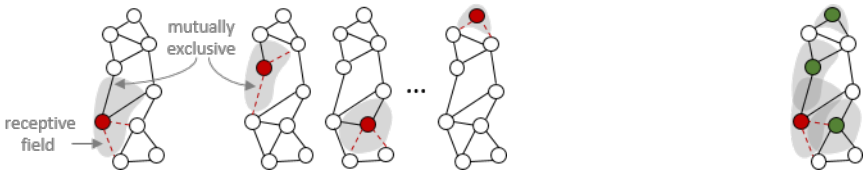


Figure 1: Previous certificates consider each node independently. Most nodes cannot be certified since the adversary is free to choose a different perturbed graph per node (left). This is impossible in practice due to mutually exclusive perturbations. The effective budget is much larger and leads to pessimistic robustness estimates. Our collective certificate enforces a single perturbed graph (right).

To successfully attack some nodes the attacker needs to insert certain edges in the graphs, while for another set of nodes the same edges must not be inserted. Since they are mutually exclusive, the attacker is forced to make a choice and can attack only one set of nodes (see Fig. 1). We argue that we should not consider nodes independently but rather all at once, i.e. to certify the overall accuracy.

We propose a collective certificate that computes the number of *simultaneously* certifiable nodes for which we can guarantee that their predictions will not change. Specifically, we fuse multiple single-node certificates, which we refer to as base certificates, into a drastically (and provably) stronger *collective certificate*. Our approach is independent of how the base certificates are derived, and any improvements to the base certificates directly translates to improvements in the collective certificate.

The key property which we exploit is *locality*. For example, in a k-layer message-passing graph neural network (Gilmer et al., 2017) the prediction for any given node depends only on the nodes in its k-hop neighborhood. Similarly, the predicted segment for any pixel depends only on the pixels in its receptive field, and the named-entity assigned to any word only depends on words in its surrounding.

How to obtain a single collective certificate instead of many individual certificates is so far unexplored. Assume we are given a clean graph $\mathcal{G}$, and a set of admissible perturbed graphs $\mathbb{B}_{\mathcal{G}}$. One straightforward way is to lower bound the number of individually certified nodes $\sum_n \min_{\mathcal{G}'_n \in \mathbb{B}_{\mathcal{G}}} \mathbf{1}_{f_n(\mathcal{G})=f_n(\mathcal{G}'_n)}$, where $f_n(\cdot)$ is the prediction for the n-th node. This approach is overly pessimistic since it allows n different perturbed graphs $\mathcal{G}'_n$. Instead, we lower bound the number of simultaneously certified nodes $\min_{\mathcal{G}' \in \mathbb{B}_{\mathcal{G}}} \sum_n \mathbf{1}_{f_n(\mathcal{G})=f_n(\mathcal{G}')}$, i.e. one perturbed graph $\mathcal{G}'$.

For classifiers that satisfy locality, perturbations to one part of the graph do not affect all nodes. Adversaries are thus faced with a budget allocation problem: It might be possible to attack different subsets of nodes via perturbations to different subgraphs, but performing all perturbations at once could violate the constraints imposed by $\mathbb{B}_{\mathcal{G}}$. The naïve approach discussed above ignores this, overestimating how many nodes can be attacked. We design a simple (mixed-integer) linear program (LP) that leverages locality by accounting for the receptive field of each node. In the LP, we explicitly represent the subgraph which can influence the prediction of $f(\cdot)_n$ for any $\mathcal{G}' \in \mathbb{B}_{\mathcal{G}}$.

We evaluate our approach on different datasets and with different base certificates. We show that incorporating locality alone is sufficient to obtain significantly better results. Our proposed certificate:

- Is the first *collective* certificate w.r.t. an adversary that simultaneously attacks multiple outputs.
- Fuses individual certificates into a provably stronger certificate by explicitly modeling *locality*.
- Is the first node classification certificate that can constrain both the global and local budget, and the number of adversary-controlled nodes, regardless of whether the base certificates support this.

2 Preliminaries

Data and models. We define our unperturbed data as an attributed graph $\mathcal{G} = (\boldsymbol{X}, \boldsymbol{A}) \in \mathbb{G}$ with $\mathbb{G} = \{0,1\}^{N \times D} \times \{0,1\}^{N \times N}$, consisting of N D-dimensional feature vectors and a directed $N \times N$ adjacency matrix. Each vertex is assigned one out of C classes by a multi-output classifier $f : \mathbb{G} \mapsto \{1, \dots, C\}^N$. Let $f_n(\mathcal{G}) = f_n(\boldsymbol{X}, \boldsymbol{A}) = f_n$ be the prediction for node n.

Collective threat model. Unlike previous certificates, we model an adversary that aims to change multiple predictions at once. Let $\mathbb{B}_{\boldsymbol{X}} \subseteq \{0,1\}^{N \times D}$ and $\mathbb{B}_{\boldsymbol{A}} \subseteq \{0,1\}^{N \times N}$ be sets of admissible

perturbed attribute and adjacency matrices respectively. Let $\mathbb{B}_{\mathcal{G}} = \mathbb{B}_{\boldsymbol{X}} \times \mathbb{B}_{\boldsymbol{A}}$ be the set of admissible graphs. Given a clean graph $\mathcal{G}$, the adversary tries to find a $\mathcal{G}' \in \mathbb{B}_{\mathcal{G}}$ that maximizes the number of misclassified nodes, i.e. $\sum_{n \in \mathbb{T}} \mathbf{1}_{f_n(\mathcal{G}) \neq f_n(\mathcal{G}')}$, for some set of target nodes $\mathbb{T} \subseteq \{1, \dots, N\}$.

Following prior work (Zügner & Günnemann, 2019), we constrain the set of admissible perturbed graphs $\mathbb{B}_{\mathcal{G}}$ through global and (optionally) local constraints on the number of changed bits. Our global constraints are parameterized by scalars $r_{\boldsymbol{X}_{\text{add}}}, r_{\boldsymbol{X}_{\text{del}}}, r_{\boldsymbol{A}_{\text{add}}}, r_{\boldsymbol{A}_{\text{del}}} \in \mathbb{N}_0$. They are an upper limit on how many bits can be added ($0 \to 1$) or deleted ($1 \to 0$) when perturbing $\boldsymbol{X}$ and $\boldsymbol{A}$. Our local constraints are parameterized by vectors $\boldsymbol{r}_{\boldsymbol{X}_{\text{add,loc}}}, \boldsymbol{r}_{\boldsymbol{X}_{\text{del,loc}}}, \boldsymbol{r}_{\boldsymbol{A}_{\text{add,loc}}}, \boldsymbol{r}_{\boldsymbol{A}_{\text{del,loc}}} \in \mathbb{N}_0^N$. They are an upper limit on how many bits can be added or deleted per row of $\boldsymbol{X}$ and $\boldsymbol{A}$, i.e. how much the attributes of a particular node can change and how many incident edges can be perturbed. We also introduce a function $\gamma : \mathbb{N}_0^4 \mapsto \mathcal{P}(\mathbb{G})$. Given budget parameters $\gamma(r_{\boldsymbol{X}_{\text{add}}}, r_{\boldsymbol{X}_{\text{del}}}, r_{\boldsymbol{A}_{\text{add}}}, r_{\boldsymbol{A}_{\text{del}}})$ returns the set of perturbed graphs that fulfil the corresponding global budget constraints (see Eq. 24).

Often, it is reasonable to assume that no adversary has direct control over the entire graph. Instead, a realistic attacker should only be able to perturb a small (adaptively chosen) subset of nodes. To model this, we introduce an additional parameter $\sigma \in \mathbb{N}$. For all $(\boldsymbol{X}', \boldsymbol{A}') \in \mathbb{B}_{\mathcal{G}}$, there must be a set of node indices $\mathbb{S} \subseteq \{1, \dots, N\}$ with $|\mathbb{S}| \leq \sigma$ such that for all $d \in \{1, \dots, D\}$ and $n, m \in \{1, \dots, N\}$:

$$\left(X'_{n,d} \neq X_{n,d} \implies n \in \mathbb{S}\right) \wedge \left(A'_{n,m} \neq A_{n,m} \implies n \in \mathbb{S} \vee m \in \mathbb{S}\right). \tag{1}$$

The set $\mathbb{S}$ is not fixed, but chosen by the adversary.[3] The resulting set $\mathbb{B}_{\mathcal{G}}$ is formally defined in § B.

Local predictions. Our certificate exploits the locality of predictions, i.e. the fact that predictions are only based on a subset of the input data. We characterize the receptive field of f_n via an indicator vector $\boldsymbol{\psi}^{(n)} \in \{0,1\}^N$ corresponding to rows in attribute matrix $\boldsymbol{X}$ and an indicator matrix $\boldsymbol{\Psi}^{(n)} \in \{0,1\}^{N \times N}$ corresponding to entries in adjacency matrix $\boldsymbol{A}$. For all $(\boldsymbol{X}', \boldsymbol{A}'), (\boldsymbol{X}'', \boldsymbol{A}'') \in \mathbb{B}_{\mathcal{G}}$:

$$\sum_{m=1}^{N} \sum_{d=1}^{D} \psi_m^{(n)} \mathbf{1}_{X'_{m,d} \neq X''_{m,d}} + \sum_{i=1}^{N} \sum_{j=1}^{N} \Psi_{i,j}^{(n)} \mathbf{1}_{A'_{i,j} \neq A''_{i,j}} = 0 \implies f_n(\boldsymbol{X}', \boldsymbol{A}') = f_n(\boldsymbol{X}'', \boldsymbol{A}''). \tag{2}$$

Eq. 2 enforces that as long as all nodes and edges for which $\boldsymbol{\psi}^{(n)} = 1$ or $\boldsymbol{\Psi}^{(n)} = 1$ remain unperturbed, the prediction f_n does not change. Put differently, changes to the rest of the data do not affect the prediction. Note that by adding or deleting edges the adversary can alter receptive fields. The receptive field indicators $\boldsymbol{\psi}^{(n)}, \boldsymbol{\Psi}^{(n)}$ correspond to the union of all receptive fields achievable under some threat model, i.e. all data points that influence f_n under some graph in $\mathbb{B}_{\mathcal{G}}$.

Base certificates. A base certificate is any procedure that can provably guarantee that the prediction f_n for a specific node n cannot be changed for any graph in the admissible set. We treat base certificates as black-box functions $\zeta_n(\cdot) \mapsto \{0,1\}$ that return 1 if node n is certifiably robust:

$$[\zeta_n(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}_{\mathcal{G}}) = 1] \implies [\forall (\boldsymbol{X}', \boldsymbol{A}') \in \mathbb{B}_{\mathcal{G}} : f_n(\boldsymbol{X}, \boldsymbol{A}) = f_n(\boldsymbol{X}', \boldsymbol{A}')] \tag{3}$$

The implication is one-directional, the functions $\zeta_n(\cdot)$ do not have to certify non-robustness.

3 COLLECTIVE CERTIFICATE

To improve clarity, in this section we only discuss the global budget constraints. All remaining constraints from the threat model can be easily expressed as linear constraints. You can find the certificate for the full threat model in § C. We first formalize the naïve collective certificate, which implicitly allows the adversary to use different graphs to attack different predictions. We then derive the proposed collective certificate, first focusing on attribute additions before extending it to arbitrary perturbations. We finally relax the certificate to a linear program to enable fast computation and discuss the certificate's tightness when using a randomized smoothing base certificate.

Naïve collective certificate. Given a clean input $(\boldsymbol{X}, \boldsymbol{A})$, a multi-output classifier f, a set of admissible perturbed graphs $\mathbb{B}_{\mathcal{G}}$, and a set $\mathbb{T}$ of target nodes, the naïve certificate evaluates $\sum_{n \in \mathbb{T}} \zeta_n(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}_{\mathcal{G}})$, i.e. we simply count the certifiable nodes. This is a lower bound on the

[3] Note that the adversary only needs to control one of the nodes incident to an edge in order to perturb it. We do not associate different costs for perturbing different nodes or edges, but such an extension is straightforward.

optimal value of $\sum_{n\in\mathbb{T}} \min_{(\boldsymbol{X}',\boldsymbol{A}')\in\mathbb{B}_\mathcal{G}} \mathbf{1}_{f_n(\boldsymbol{X},\boldsymbol{A})=f_n(\boldsymbol{X}',\boldsymbol{A}')}$, i.e. the number of predictions guaranteed to be stable under attack. We can directly see this from the base certificate definition in Eq. 3.

Collective certificate for attribute additions. To improve upon the naïve certificate, we want to determine the number of predictions that are not simultaneously attackable via a *single* graph:

$$\min_{(\boldsymbol{X}',\boldsymbol{A}')\in\mathbb{B}_\mathcal{G}} \sum_{n\in\mathbb{T}} \mathbf{1}_{f_n(\boldsymbol{X},\boldsymbol{A})=f_n(\boldsymbol{X}',\boldsymbol{A}')}. \tag{4}$$

Solving this problem is usually not tractable. However, as before, we can lower-bound the indicator functions via base certificates. For simplicity, let us assume that the adversary is only allowed to perform attribute additions. We later discuss how to generalize the certificate to arbitrary perturbations. Recall that $\gamma : \mathbb{N}_0^4 \mapsto \mathcal{P}(\mathbb{G})$ (see Eq. 24 in § B) returns the set of all graphs fulfilling the specified global budget constraints. We can combine γ with the base certificates to obtain

$$\min_{(\boldsymbol{X}',\boldsymbol{A}')\in\mathbb{B}_\mathcal{G}} \sum_{n\in\mathbb{T}} \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b,0,0,0\right)\right), \tag{5}$$

where $b = \sum_{(n,d):X_{n,d}=0} X'_{n,d}$ is the number of attribute additions for a given perturbed graph. Eq. 5 is a lower bound on Eq. 4. If $(\boldsymbol{X}', \boldsymbol{A}')$ is an adversarial example for f_n, then the set given by γ contains an adversarial example. Both $\mathbf{1}_{f_n(\boldsymbol{X},\boldsymbol{A})=f_n(\boldsymbol{X}',\boldsymbol{A}')}$ and ζ_n return 0. If $(\boldsymbol{X}', \boldsymbol{A}')$ is not an adversarial example, the indicator function returns 1, while ζ_n can return either 1 or 0. Note that γ is only dependent on the number of perturbations. It is thus sufficient to optimize over the number of additions per node, while enforcing the global budget constraint.

$$\min_{b\in\mathbb{N}_0} \sum_{n\in\mathbb{T}} \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b,0,0,0\right)\right) \text{ s.t. } b \leq r_{\boldsymbol{X}_\text{add}}. \tag{6}$$

There are two challenges: (1) The base certificate ζ_n does not account for locality, but simply considers the number of perturbations in the entire graph; (2) ζ_n might be a complex optimization problem, which is difficult to optimize through. We tackle (1) by constructing local base certificates.

Lemma 1 *Given base certificates ζ_n, multi-output classifier f, corresponding receptive field indicators $\boldsymbol{\psi}^{(n)} \in \{0,1\}^N$ and $\boldsymbol{\Psi}^{(n)} \in \{0,1\}^{N\times N}$, and a clean graph $(\boldsymbol{X}, \boldsymbol{A})$. Let $(\boldsymbol{X}', \boldsymbol{A}')$ be a perturbed graph. Define $\boldsymbol{X}'' \in \{0,1\}^{N\times D}$ and $\boldsymbol{A}'' \in \{0,1\}^{N\times N}$ as follows:*

$$\boldsymbol{X}''_{i,d} = \psi_i^{(n)} \boldsymbol{X}'_{i,d} + (1-\psi_i^{(n)})\boldsymbol{X}_{i,d}, \qquad \boldsymbol{A}''_{i,j} = \Psi_{i,j}^{(n)} \boldsymbol{A}'_{i,j} + (1-\Psi_{i,j}^{(n)})\boldsymbol{A}_{i,j}, \tag{7}$$

i.e. use values from the clean graph for bits that are not in f_n's receptive field. If there exists a set of graphs $\mathbb{B}'_\mathcal{G}$ with $(\boldsymbol{X}'', \boldsymbol{A}'') \in \mathbb{B}'_\mathcal{G}$ and $\zeta_n(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}'_\mathcal{G}) = 1$, then $f_n(\boldsymbol{X}, \boldsymbol{A}) = f_n(\boldsymbol{X}', \boldsymbol{A}')$.

See proof in § C. This lemma means that we can lower-bound $\mathbf{1}_{f_n(\boldsymbol{X},\boldsymbol{A})=f_n(\boldsymbol{X}',\boldsymbol{A}')}$ in Eq. 4 via $\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}'_\mathcal{G}\right)$, where $\mathbb{B}'_\mathcal{G} \ni (\boldsymbol{X}'', \boldsymbol{A}'')$ and $(\boldsymbol{X}'', \boldsymbol{A}'')$ is constructed by reverting all perturbations in $(\boldsymbol{X}', \boldsymbol{A}')$ that are outside f_n's receptive field. In particular, we can choose $\mathbb{B}'_\mathcal{G} = \gamma\left(\boldsymbol{b}^T\boldsymbol{\psi}^{(n)}, 0, 0, 0\right)$, where the vector $\boldsymbol{b} \in \mathbb{N}_0^N$ indicates the number of attribute additions for each node. We can then optimize over $\boldsymbol{b}$ to obtain a collective certificate that accounts for locality.

$$\min_{\boldsymbol{b}\in\mathbb{N}_0^N} \sum_{n\in\mathbb{T}} \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(\boldsymbol{b}^T\boldsymbol{\psi}^{(n)}, 0, 0, 0\right)\right) \text{ s.t. } ||\boldsymbol{b}||_1 \leq r_{\boldsymbol{X}_\text{add}}. \tag{8}$$

We now tackle issue (2) by noticing that each local base certificate can be characterized via a single scalar p_n. As we will see this also implies that we never explicitly need to compute $\gamma(\cdot)$. We define

$$\zeta'_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b,0,0,0\right)\right) = \max_{c\in\mathbb{N}_0} \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b+c,0,0,0\right)\right). \tag{9}$$

This function is a valid base certificate since $\gamma\left(b,0,0,0\right) \subseteq \gamma\left(b+c,0,0,0\right)$ for all $b, c \in \mathbb{N}_0$. If there is some c such that $\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b+c,0,0,0\right)\right) = 1$, we provably know that there is no adversarial example within $\gamma\left(b+c,0,0,0\right)$, so there cannot be an adversarial example within its subset $\gamma\left(b,0,0,0\right)$ either. By construction $\zeta'_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b,0,0,0\right)\right)$ is guaranteed to be monotonically decreasing in b (i.e. not oscillating between 1 and 0). We can exploit this to fully describe the certificate via a single scalar (see also § D): There must be some $p_n \in \mathbb{N}_0$ such that $\zeta'_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b,0,0,0\right)\right) = 0 \iff b \geq p_n$. Thus, we do not need to consider ζ_n and γ in Eq. 8;

instead, determining if the prediction of f_n is certifiably robust boils down to determining whether the number of perturbations in f_n's receptive field is smaller than p_n. We express this as a MILP:

$$\min_{\boldsymbol{b}\in\mathbb{N}_0^N, \boldsymbol{t}\in\{0,1\}^N} |\mathbb{T}| - \sum_{n\in\mathbb{T}} t_n \tag{10}$$

$$\text{s.t. } \psi^{(n)}\boldsymbol{b} \geq p_n t_n \; \forall n \in \{1,\dots,N\} \tag{11}$$

$$||\boldsymbol{b}||_1 \leq r_{\boldsymbol{X}_{\text{add}}}. \tag{12}$$

Eq. 12 ensures that the number of perturbations fulfills the global budget constraint. Eq. 11 ensures that the indicator t_n can only be set to 1, if the local perturbation on the l.h.s. exceeds or matches p_n, i.e. f_n is *not* robustly certified by ζ'_n. The adversary tries to minimize the number of robustly certified predictions in $\mathbb{T}$ (c.f. Eq. 6), which is equivalent to Eq. 10.

Collective certificate for arbitrary perturbations. Lemma 1 is defined for arbitrary perturbations. However, when multiple perturbation types are allowed we can no longer treat the base certificates as one-dimensional functions and we cannot just use scalars p_n. We generalize our approach. Define

$$\zeta'_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b_1, b_2, b_3, b_4\right)\right) = \max_{\boldsymbol{c}\in\mathbb{N}_0^4} \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b_1 + c_1, b_2 + c_2, b_3 + c_3, b_4 + c_4\right)\right). \tag{13}$$

This is again a valid certificate, as $\gamma\left(b_1, b_2, b_3, b_4\right) \subseteq \gamma\left(b_1 + c_1, b_2 + c_2, b_3 + c_3, b_4 + c_4\right)$ for all $\boldsymbol{c} \in \mathbb{N}_0^4$. If ζ_n returns 1 for some superset, then the subset $\gamma\left(b_1, b_2, b_3, b_4\right)$ cannot contain any adversarial examples either. By construction, it is also monotonically decreasing in all four elements of $\boldsymbol{b}$, i.e. $\forall \boldsymbol{b}, \boldsymbol{b}' \in \mathbb{N}_0^4 : \zeta'_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(\boldsymbol{b}\right)\right) \leq \zeta'_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(\boldsymbol{b}'\right)\right)$. Let $\mathbb{P}'^{(n)} = \left\{p \in \{1,\dots,N\}^4 | \zeta'_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(p_1, p_2, p_3, p_4\right)\right) = 0\right\}$ be the combinations of budgets for which the classifier f_n cannot be robustly certified by ζ'_n. The monotonicity of ζ'_n allows us to characterize $\mathbb{P}'^{(n)} \subseteq N^4$ by its pareto front (see also § E) w.r.t. the value in each component:

$$\mathbb{P}^{(n)} = \left\{\boldsymbol{p} \in \mathbb{P}'^{(n)} | \neg\exists \boldsymbol{p}' \in \mathbb{P}'^{(n)} : \boldsymbol{p}' \neq \boldsymbol{p} \wedge \forall d \in \{1,2,3,4\} : p'_d \leq p_d\right\}. \tag{14}$$

We encode the set $\mathbb{P}^{(n)}$ as a matrix $\boldsymbol{P}^{(n)} \in \mathbb{N}_0^{|\mathbb{P}^{(n)}|\times 4}$. To determine if f_n is robust we can then simply check if there is some pareto-optimal point $\boldsymbol{P}^{(n)}_{i,:}$ such that the amount of perturbation in f_n's receptive field matches or exceeds $\boldsymbol{P}^{(n)}_{i,:}$ in all four dimensions, which we can express as a MILP:

$$\min_{\left(\boldsymbol{Q}^{(n)}, \boldsymbol{s}^{(n)}\right)_{n=1}^N, \boldsymbol{b}_{\boldsymbol{X}_{\text{add}}}, \boldsymbol{b}_{\boldsymbol{X}_{\text{del}}}, \boldsymbol{B}_{\boldsymbol{A}}, \boldsymbol{t}} |\mathbb{T}| - \sum_{n\in\mathbb{T}} t_n \tag{15}$$

$$\text{s.t.} \qquad ||\boldsymbol{s}^{(n)}||_1 \geq t_n, \qquad Q^{(n)}_{p,d} \geq s^{(n)}_p, \tag{16}$$

$$(\boldsymbol{b}_{\boldsymbol{X}_{\text{add}}})^T \boldsymbol{\psi}^{(n)} \geq Q^{(n)}_{i,1} P^{(n)}_{p,1}, \qquad (\boldsymbol{b}_{\boldsymbol{X}_{\text{del}}})^T \boldsymbol{\psi}^{(n)} \geq Q^{(n)}_{p,2} P^{(n)}_{p,2}, \tag{17}$$

$$\sum_{m,m'\leq N} (1 - A_{m,m'})(\boldsymbol{\Psi}^N \odot B_{\boldsymbol{A}})_{m,m'} \geq Q^{(n)}_{p,3} P^{(n)}_{p,3}, \tag{18}$$

$$\sum_{m,m'\leq N} (1 - A_{m,m'})(\boldsymbol{\Psi}^N \odot B_{\boldsymbol{A}})_{m,m'} \geq Q^{(n)}_{p,4} P^{(n)}_{p,4}, \tag{19}$$

$$||b_{\boldsymbol{X}_{\text{add}}}||_1 \leq r_{\boldsymbol{X}_{\text{add}}}, \qquad ||b_{\boldsymbol{X}_{\text{add}}}||_1 \leq r_{\boldsymbol{X}_{\text{del}}}, \tag{20}$$

$$\sum_{(i,j):A_{i,j}=0} B_{\boldsymbol{A}i,j} \leq r_{\boldsymbol{A}_{\text{add}}}, \qquad \sum_{(i,j):A_{i,j}=1} B_{\boldsymbol{A}i,j} \leq r_{\boldsymbol{A}_{\text{del}}}, \tag{21}$$

$$\boldsymbol{s}^{(n)} \in \{0,1\}^{|\mathbb{P}^{(n)}|}, \qquad \boldsymbol{Q}^{(n)} \in \{0,1\}^{|\mathbb{P}^{(n)}|\times 4} \qquad \boldsymbol{t} \in \{0,1\}^N \tag{22}$$

$$\boldsymbol{b}_{\boldsymbol{X}_{\text{add}}}, \boldsymbol{b}_{\boldsymbol{X}_{\text{del}}} \in \mathbb{N}_0^N, \qquad \boldsymbol{B}_{\boldsymbol{A}} \in \{0,1\}^{N\times N}. \tag{23}$$

As before, we use a vector $\boldsymbol{t}$ with $t_n = 1$ indicating that f_n is not certified by ζ'_n. The adversary tries to find a budget allocation (parameterized by $\boldsymbol{b}_{\boldsymbol{X}_{\text{add}}}, \boldsymbol{b}_{\boldsymbol{X}_{\text{del}}}$ and $\boldsymbol{B}_{\boldsymbol{A}}$) that minimizes the number of robustly certified predictions in $\mathbb{T}$ (see Eq. 15). Eq. 20 and Eq. 21 ensure that the adversary's budget allocation is consistent with the global budget parameters characterizing $\mathbb{B}_{\mathcal{G}}$. The value of t_n is determined by the following constraints: First, Eq. 17 to Eq. 19 ensure that $Q^{(n)}_{p,d}$ is only set to 1 if the local perturbation matches or exceeds the pareto-optimal point corresponding to row p

of $\boldsymbol{P}^{(n)}$ in dimension d. The constraints in Eq. 16 implement logic operations on $\boldsymbol{Q}^{(n)}$: Indicator $s_p^{(n)}$ can only be set to 1 if $\forall d \in \{1,2,3,4\} : Q_{p,d}^{(n)} = 1$. Indicator t_n can only be set to 1 if $\exists p \in \{1,\dots,|\mathbb{P}^{(n)}|\} : s_p^{(n)} = 1$. Combined, these constraints enforce that if $t_n = 1$, there must be some point in $\mathbb{P}^{(n)}$ that is exceeded or matched by the amount of perturbations in all four dimensions.

LP-relaxation. For large graphs, finding an optimum to the mixed-integer problem is prohibitively expensive. In practice, we relax integer variables to reals and binary variables to $[0, 1]$. Semantically, the relaxation means that bits can be partially perturbed, nodes can be partially controlled by the attacker and classifiers can be partially certified. The resulting certificate is a linear program, which can be solved much faster. In particular, a prediction f_n can be (partially) uncertified (i.e. $1 > t_n > 0$), even if the entire available budget is not sufficient to match or exceed any point $p^{(n)} \in \mathbb{P}^{(n)}$ along all four dimensions. To resolve this issue, we remove all unreachable pareto-optimal points.

Tightness for randomized smoothing. One recent method for robustness certification is randomized smoothing (Cohen et al., 2019). In randomized smoothing, a base classifier $h : \mathbb{X} \mapsto \Delta_C$ that maps from some input space $\mathbb{X}$ to the C-dimensional simplex Δ_C is transformed into a smoothed classifier $g(x)$ with $g(x) = \operatorname{argmax}_{c\in\{1,\dots,C\}} \mathbb{E}\left[h(\phi(x))\right]_c$ where $\phi(x)$ is some randomization scheme. For the smoothed $g(x)$ we can then derive probabilistic robustness certificates. Randomized smoothing is a black-box method that only depends on h's expected output behavior under $\phi(x)$, and does not require any further assumptions. Building on prior randomized smoothing work for discrete data by Lee et al. (2019), Bojchevski et al. (2020) propose a smoothing distribution and corresponding certificate for graphs. Using their method as a base certificate to our collective certificate the resulting (non-relaxed) certificate is tight. That is, our mixed-integer collective certificate is the best certificate we can obtain for the specified threat model, if we do not use any information other than the classifiers' expected outputs and their locality. Detailed explanation and proof in § F.2.

Time complexity. For our method we need to construct monotonic base certificates and their pareto fronts. This has to be performed only once and the results can then be reused in evaluating the collective certificate with varying parameters. We discuss the algorithmic details of this pre-processing in § D and § E. The complexity of the collective certificate itself is based on the number of constraints and variables of the underlying (MI)LP. In total, we have $13\sum_{n=1}^{N}|\mathbb{P}^{(n)}| + 8N + 2e + 5$ constraints and $5\sum_{n=1}^{N}|\mathbb{P}^{(n)}| + 4N + e$ variables, where e are the number of edges in the unperturbed graph (we disallow edge additions). For single-type perturbations we have $\mathcal{O}(N + e)$ terms, linear in the number of nodes and edges. In practice, the relaxed LP takes at most a few seconds to certify robustness for single-type perturbations and a few minutes for multiple types of perturbations (see § 5).

4 LIMITATIONS

Locality. The proposed approach is designed to exploit locality. Without locality, it is equivalent to a naïve combination of base certificates that sums over perturbations in the entire graph. A non-obvious limitation is that our notion of locality breaks down if the receptive fields are data-dependent and can be arbitrarily extended by the adversary. Recall how we specified locality in Eq. 2: The indicators $\boldsymbol{\psi}^{(n)}$ and $\boldsymbol{\Psi}^{(n)}$ correspond to the union of all achievable receptive fields. Take for example a two-layer message-passing neural networks and an adversary that can add new edges. Each node is classified based on its 2-hop neighborhood. For any two nodes n, m, the adversary can construct a graph such that m is in f_n receptive field. We thus have to treat the f_n as global, even if for any single graph they might only process some subgraph. Nonetheless, our method still yields significant improvements for edge deletions and arbitrary attribute perturbations. As discussed in prior work (Zügner & Günnemann, 2020) edge addition in inherently harder and less relevant in practice.

5 EXPERIMENTAL EVALUATION

Experimental setup. We evaluate the proposed approach by certifying node classifiers on multiple graphs and with different base certificates. We use 20 nodes per class for train and validation set. We certify all remaining nodes. We repeat each experiment five times with different random initialization and data splits. Unless otherwise specified, we do not impose any local budget constraints or constraints on the number of attacker-controlled nodes. We compare the proposed method with the

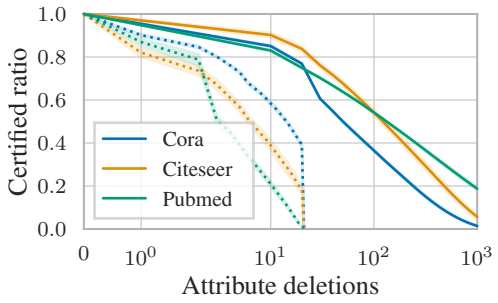


Figure 2: Certified ratios for smoothed GCN on Cora, Citeseer and PubMed, under varying $r_{\boldsymbol{X}_{\text{del}}}$. We compare the proposed certificate (solid lines) to the naïve certificate (dotted lines). Our method certifies orders of magnitude larger radii (note the logarithmic x-axis).

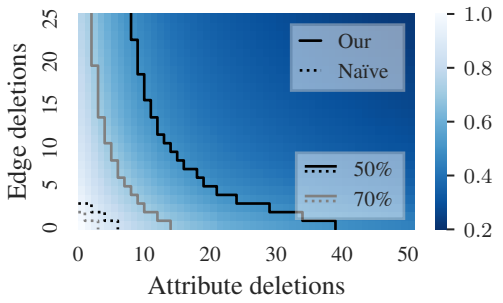


Figure 3: Two-dimensional colective certificate for smoothed GCN on Cora-ML under varying $r_{\boldsymbol{X}_{\text{del}}}$ and $r_{\boldsymbol{A}_{\text{del}}}$. The solid and dotted contour lines show ratios ≥ 0.5 and ≥ 0.7 for our vs. the naïve certificate respectively. Our method achieves much larger certified ratios and radii.

naïve collective certificate, which simply counts the number of predictions that are certified to be robust by the base certificate. All experiments are based on the relaxed linear programming version of the certificate. We assess the integrality gap to the mixed-integer version in § A. The code will be made publicly available. We uploaded a reference implementation as supplementary material.

Datasets, models and base certificates. We train and certify models on the following datasets: Cora-ML (McCallum et al. (2000); Bojchevski & Günnemann (2018); $N = 2810$, 7981 edges, 7 classes), Citeseer (Sen et al. (2008); $N = 2110$, 3668 edges, 6 classes), PubMed (Namata et al. (2012); $N = 19717$, 44324 edges, 3 classes), Reuters-21578 [4] ($N = 862$, 2586 edges, 4 classes) and WebKB (Craven et al. (1998); $N = 877$, 2631 edges, 5 classes). The graphs for the natural language corpora Reuters and WebKB are constructed using the procedure described in Zhou & Vorobeychik (2020), resulting in 3-regular graphs. We use four classifiers: Graph convolution networks (GCN) (Kipf & Welling, 2017), graph attention networks (GAT) (Veličković et al., 2018), APPNP (Klicpera et al., 2019), and robust graph convolution networks (RGCN) (Zhu et al., 2019). All classifiers are configured to have two layers, i.e. each node's classifier is dependent on its two-hop neighborhood. We use two types of base certificates: (Bojchevski et al., 2020) (randomized smoothing, arbitrary perturbations) and Zügner & Günnemann (2019) (convex relaxations of network nonlinearities, attribute perturbations). We provide a summary of all hyperparameters in § G.

Evaluation metrics. We report the *certified ratio* on the test set, i.e. the percentage of nodes that are certifiably robust under a given threat model, averaged over all data splits. We further calculate the standard sample deviation in certified ratio (visualized as shaded areas in plots) and the average wall-clock time per collective certificate. For experiments, in which only one global budget parameter is altered, we report the *average certifiable radius*, i.e. $\bar{r} = (\sum_{r=0}^{\infty} \phi(r) * r / \sum_{r=0}^{\infty} \phi(r))$, where $\phi(r)$ is the certified ratio for value r of the global budget parameter, averaged over all splits.

Attribute perturbations. We first evaluate the certificate for a single perturbation type. Using randomized smoothing as the base certificate, we evaluate the certified ratio of GCN classifiers for varying global attribute deletion budgets $r_{\boldsymbol{X}_{\text{del}}}$ on the citation graphs Cora, Citeseer and PubMed (for Reuters and WebKB, see § A). The remaining global budget parameters are set to 0. Fig. 2 shows that for all datasets, the proposed method yields significantly larger certified ratios than the naïve certificate and can certify robustness for much larger $r_{\boldsymbol{X}_{\text{del}}}$. The average certifiable radius $\bar{r}$ on Citeseer increases from 7.18 to 351.73. The results demonstrate the benefit of using the collective certificate, which explicitly models simultaneous attacks on all predictions. The average wall-clock time per certificate on Cora, Citeseer and PubMed is 2.0 s, 0.29 s and 336.41 s. Interestingly, the base certificate yields the highest certifiable ratios on Cora, while the collective certificate yields the highest certifiable ratios on PubMed. We attribute this to differences in graph structure, which are explicitly taken into account by the proposed certification procedure.

[4] Distribution 1.0, available from http://www.daviddlewis.com/resources/testcollections/reuters21578

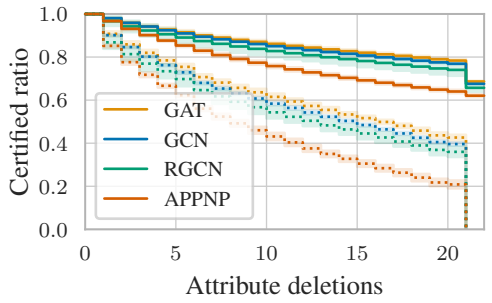


Figure 4: Comparison of certified ratios for GAT, GCN, RGCN and APPNP on Cora-ML under varying $r_{\boldsymbol{X}_{\text{del}}}$ for our (solid lines) and the naïve (dotted lines) collective certificate.

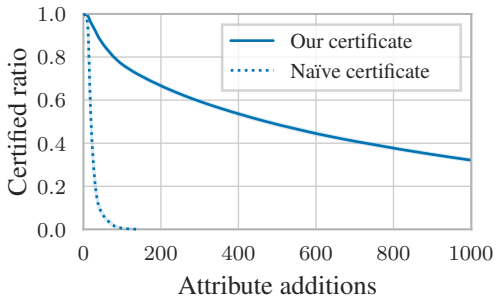


Figure 5: Certifying GCN on Citeseer, under varying $r_{\boldsymbol{X}_{\text{add}}}$ using Zügner & Günnemann (2019)'s base certificate. Our certificate yields significantly larger certified ratios and radii.

Simultaneous attribute and graph perturbations. To evaluate the multi-dimensional version of the certificate, we visualize the certified ratio of randomly smoothed GCN classifiers for different combinations of $r_{\boldsymbol{X}_{\text{del}}}$ and $r_{\boldsymbol{A}_{\text{del}}}$ on Cora-ML. For an additional experiment on simultaneous attribute additions and deletions, see § A. Fig. 3 shows that we achieve high certified ratios even when the attacker is allowed to simultaneously perturb the attributes and the structure. Comparing the contour lines at $50\,\%$ the naïve certiface can only certify much smaller radii, e.g. at most 6 attribute deletions compared to 39 for our approach. The average wall-clock time per certificate is $106.90\,\text{s}$.

Different classifiers. Our method is agnostic towards classifier architectures, as long as they are compatible with the base certificate and their receptive fields can be determined. In Fig. 4 we compare the certified collective robustness of GAT, GCN, APNP and RGCN, using the sparse smoothing certificate on Cora-ML.[5] Better base certificates translate into better collective certificates. While RGCN is supposed to be more robust to adversarial attacks, it has a lower certified ratio than GCN.

Different base certificates. Our method is also agnostic to the base certificate type. We show that it works equally well with base certificates other than randomized smoothing. Specifically, we use the method from Zügner & Günnemann (2019). We certify a GCN model for varying $r_{\boldsymbol{X}_{\text{add}}}$ on Citeseer. Unlike randomized smoothing, this base certificate models local budget constraints. Using the recommended value (default in the reference implementation), we limit the number of attribute additions per node to $\lfloor 0.01D \rfloor = 21$ for both the base and the collective certificate. Fig. 5 shows that the proposed collective certificate is again significantly stronger. The average certified radius $\hat{r}$ increases from 17.12 to 971.36. The average wall-clock time per certificate is $0.39\,\text{s}$.

[5] Our method can certify larger radii, $r \geq 10^3$ (see Fig. 2). Here we show $r \leq 20$ to highlight the difference.

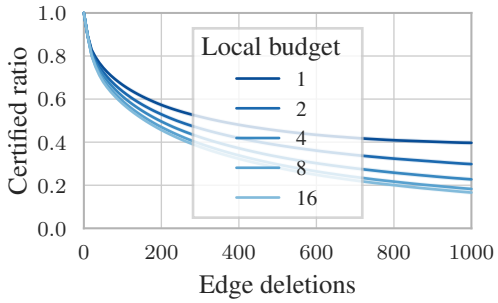


Figure 6: Certified ratios for smoothed GCN on Cora-ML, under varying $r_{\boldsymbol{A}_{\text{del}}}$ and $r_{\boldsymbol{A}_{\text{del,loc}}}$. Stricter local budgets yield larger certified ratios.

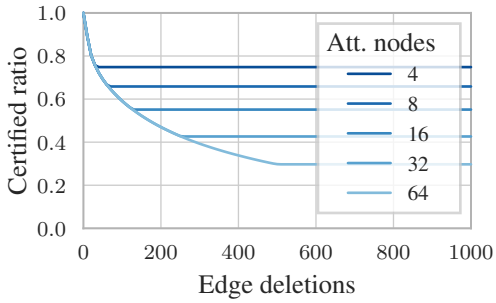


Figure 7: Certified ratios for smoothed GCN on Cora-ML. We vary $r_{\boldsymbol{X}_{\text{del}}}$ and σ. The certified ratios remain constant and non-zero for large $r_{\boldsymbol{X}_{\text{del}}}$.

Local constraints. We evaluate the effect of additional constraints in our threat model. We can enforce local budget constraints and limit the number of attacker-controlled nodes, even if they are not explicitly modeled by the base certificate. In Fig. 6, we use a smoothed GCN on Cora-ML and vary both the global budget for edge deletions, $r_{\boldsymbol{A}_{\text{del}}}$, and the local budgets $\boldsymbol{r}_{\boldsymbol{A}_{\text{del,loc}}}$. Even though the base certificate does not support local budget constraints, reducing the number of admissible deletions per node increases the certified ratio as expected. For example, limiting the adversary to one deletion per node more than doubles the certified ratio at $r_{\boldsymbol{A}_{\text{del}}} = 1000$. In Fig. 7, we fix a relatively large local budget of 16 edge deletions per node (only $\sim 5\%$ of nodes on Cora-ML have a degree > 16) and vary the number of attacker-controlled nodes. We see that for any given number of attacker nodes, there is some point $r_{\boldsymbol{A}_{\text{del}}}$ after which the certified ratio curve becomes constant. This constant value is an an upper limit on the number of classifiers that can be attacked with a given local budget and number of attacker-controlled nodes, and is independent of the global budget.

6 Conclusion

We propose the first collective robustness certificate. Assuming predictions based on a single shared input, we leverage the fact that an adversary must use a single adversarial example to attack all predictions. We focus on Graph Neural Networks, whose locality guarantees that perturbations to the input graph only affect predictions in a close neighborhood. The proposed method combines many weak base certificates into a provably stronger collective certificate. It is agnostic towards network architectures and base certification procedures. We evaluate it on multiple semi-supervised node classification datasets with different classifier architectures and base certificates. Our empirical results show that the proposed collective approach yields much stronger certificates than existing methods, which assume that an adversary can attack predictions independently with different graphs.

References

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, 2018.

Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *Proceedings of Machine Learning and Systems 2020*, pp. 11647–11657, 2020.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Workshop on Artificial Intelligence and Security, AISec*, 2017.

Ping-yeh Chiang, Michael J. Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection by median smoothing. *arXiv preprint arXiv:2007.03730*, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery. Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, pp. 509–516. American Association for Artificial Intelligence, 1998.

Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, and Evangelos E. Papalexakis. All you need is low (rank): Defending against adversarial attacks on graphs. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pp. 169–177, New York, NY, USA, 2020. Association for Computing Machinery.

Boyuan Feng, Yuke Wang, Zheng Wang, and Yufei Ding. Uncertainty-aware attention graph neural network for defending adversarial attacks. *arXiv preprint arXiv:2009.10235*, 2020.

Fuli Feng, Xiangnan He, Jie Tang, and Tat-Seng Chua. Graph adversarial training: Dynamically regularizing based on graph structure. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2019.

Simon Geisler, Daniel Zügner, and Stephan Günnemann. Reliable graph neural networks via robust aggregation. *arXiv preprint arXiv:2010.15651*, 2020.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research, 2017.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Wenjuan Han, Liwen Zhang, Yong Jiang, and Kewei Tu. Adversarial attack and defense of structured prediction models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2327–2338. Association for Computational Linguistics, November 2020.

Han Xu Yao Ma Hao-Chen, Liu Debayan Deb, Hui Liu Ji-Liang Tang Anil, and K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

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

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations (ICLR)*, 2019.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems 32*, pp. 4910–4921. Curran Associates, Inc., 2019.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000. doi: 10.1023/a:1009953814988.

Galileo Mark Namata, Ben London, Lise Getoor, and Bert Huang. Query-driven active surveying for collective classification. In *Workshop on Mining and Learning with Graphs*, 2012.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93, sep 2008. doi: 10.1609/aimag.v29i3.2157.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018.

X Xu, Y Yu, L Song, C Liu, B Kailkhura, C Gunter, and B Li. Edog: Adversarial edge detection for graph neural networks. 3 2020a.

Xiaogang Xu, Hengshuang Zhao, and Jiaya Jia. Dynamic divide-and-conquer adversarial training for robust semantic segmentation. *arXiv preprint arXiv:2003.06555*, 2020b.

Ao Zhang and Jinwen Ma. Defensevgae: Defending against adversarial attacks on graph data via a variational graph autoencoder. *arXiv preprint arXiv:2006.08900*, 2020.

Yingxue Zhang, Sakif Hossain Khan, and Mark Coates. Comparing and detecting adversarial attacks for graph deep learning. In *Representation Learning on Graphs and Manifolds Workshop, International Conference on Learning Representations*, 2019.

Kai Zhou and Yevgeniy Vorobeychik. Robust collective classification against structural attacks. volume 124 of *Proceedings of Machine Learning Research*, pp. 250–259, Virtual, 03–06 Aug 2020.

Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 1399–1407, 2019.

Daniel Zügner and Stephan Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, jul 2019. doi: 10.1145/3292500.3330905.

Daniel Zügner and Stephan Günnemann. Adversarial attacks on graph neural networks via meta learning. In *International Conference on Learning Representations (ICLR)*, 2019.

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.

A ADDITIONAL EXPERIMENTS

Attribute perturbations on additional datasets In addition to citation graphs, we also use graphs constructed from the Reuters-21578 and WebKB corpora to evaluate the proposed certificate. As in the main text, we use randomized smoothing as a base certificate for GCN classifiers and assess the certified ratio for varying global attribute deletion budgets. Fig. 8 shows that the certified ratio increases and that much larger $r_{\boldsymbol{X}_{\text{del}}}$ (up to approximately 10^3) can be certified when using the collective approach. The average certifiable radius for Reuters and WebKB increases from 6.54 and 8.08 to 265.62 and 309.32, respectively. With less than 900 nodes each, both datasets are smaller than our three citation graphs. This leads to even shorter average wall-clock times per certificate: $0.105\,\text{s}$ and $0.116\,\text{s}$.

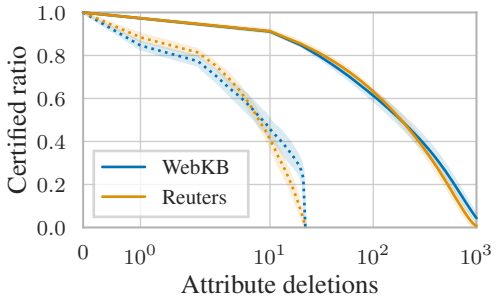
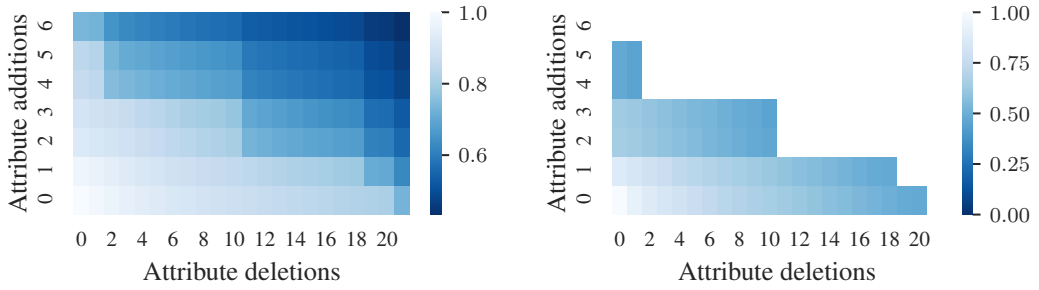


Figure 8: Certified ratios for smoothed GCN on WebKB and Reuters-21578, under varying $r_{\boldsymbol{X}_{\text{del}}}$, comparing the proposed certificate (solid lines) to the naïve certificate (dotted lines).

Simultaneous attribute deletions and additions. In the main experiments section, we applied our collective certificate to simultaneous certification of attribute and adjacency deletions. Here, we assess how it performs for simultaneous deletions and additions of attributes. We again use a randomly smoothed GCN classifiers on Cora-ML, and perform collective certification for different combinations of $r_{\boldsymbol{X}_{\text{add}}}$ and $r_{\boldsymbol{X}_{\text{del}}}$ on Cora-ML. As shown in Fig. 3, the collective certificate is again much stronger than the naïve collective certificate. For example, we obtain certified ratios between 30% and 60% at radii for which the naïve collective certificate cannot certify any robustness at all. The average wall-clock time per certificate is $40.51\,\text{s}$.



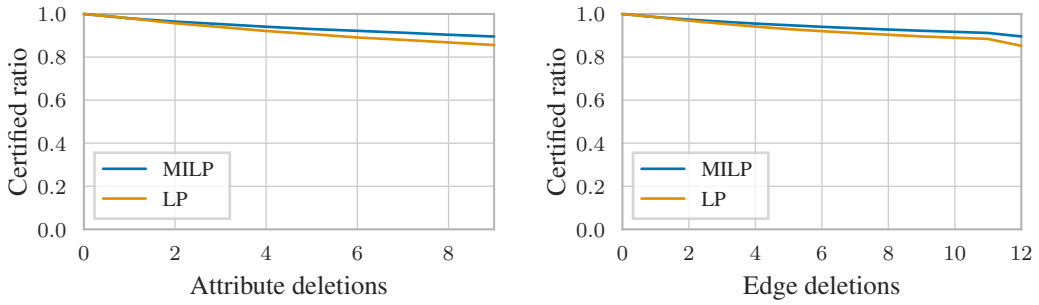
(a) Proposed collective certificate

(b) Naïve collective certificate

Figure 9: Comparison of the proposed collective certificate (Fig. 9a to the naïve collective certificate (Fig. 9b for certification of smoothed GCN on Cora-ML, under varying $r_{\boldsymbol{X}_{\text{add}}}$ and $r_{\boldsymbol{X}_{\text{del}}}$. Our method achieves much larger certified ratios for all combinations of attack radii.

Integrality gap. For all previous experiments, we used the relaxed linear programming version of the certificate to reduce the compute time. To assess the integrality gap (i.e. the difference between the mixed-integer linear programming and the linear programming based certificates), we apply both versions of the certificate to a single smoothed GCN on Cora. We certify both robustness to attribute deletions (Fig. 10a) and edge deletions (Fig. 10b). The wall-clock time per certificate for the MILP increased from $0.24\,\text{s}$ to $64\,\text{h}$ with increasing edge deletion budget ($0.35\,\text{s}$ to $94\,\text{h}$ for attribute deletions). Due to the exploding runtime for the MILP, we cannot compute the integrality gap for radii larger than 8 and 12, respectively.[6] The integrality gap is small (at most 4% for attribute deletions, 4.3% for edge deletions), relative to the certified ratio, and appears to be slightly increasing with increasing global budget.

[6] As discussed in the main paper the relaxed LP is fast and efficient to solve, even for radii larger than 1000.



(a) (b)

Figure 10: Certified ratios for smoothed GCN on Cora, under varying $r_{\boldsymbol{X}_{\text{del}}}$ (Fig. 10a) and $r_{\boldsymbol{A}_{\text{del}}}(Fig.\ 10b)$, using the mixed-integer collective certificate (blue line) and the relaxed linear programming certificate (orange line). The integrality gap is small, relative to the certified ratio.

B Formal definition of threat model parameters and proofs

Here we define formally the set of admissible perturbed graphs $\mathbb{B}_{\mathcal{G}}$ described in § 2. Recall that we have an unperturbed graph $(\boldsymbol{X}, \boldsymbol{A}) \in \mathbb{G}$, global budget parameters $r_{\boldsymbol{X}_{\text{add}}}, r_{\boldsymbol{X}_{\text{del}}}, r_{\boldsymbol{A}_{\text{add}}}, r_{\boldsymbol{A}_{\text{del}}} \in \mathbb{N}_0$, local budget parameters $\boldsymbol{r}_{\boldsymbol{X}_{\text{add,loc}}}, \boldsymbol{r}_{\boldsymbol{X}_{\text{del,loc}}}, \boldsymbol{r}_{\boldsymbol{A}_{\text{add,loc}}}, \boldsymbol{r}_{\boldsymbol{A}_{\text{del,loc}}} \in \mathbb{N}_0^N$ and at most σ adversary-controlled nodes. Let $\gamma : \mathbb{N}_0^4 \mapsto \mathcal{P}(\mathbb{G})$ be a function that maps from global budget parameters to a set of perturbed graphs fulfilling the corresponding global budget constraints:

$$\begin{aligned}(\boldsymbol{X}', \boldsymbol{A}') \in \gamma(b_1, b_2, b_3, b_4) \implies \\ \left|\left\{(n,d) : X_{n,d} = 0 \neq X'_{n,d}\right\}\right| \leq b_1 \wedge \left|\left\{(n,d) : X_{n,d} = 1 \neq X'_{n,d}\right\}\right| \leq b_2 \\ \wedge \left|\left\{(n,m) : A_{n,m} = 0 \neq A'_{n,m}\right\}\right| \leq b_3 \wedge \left|\left\{(n,m) : A_{n,m} = 1 \neq A'_{n,m}\right\}\right| \leq b_4.\end{aligned} \tag{24}$$

$\mathbb{B}_{\mathcal{G}}$ is then defined as follows:

$$\begin{aligned}(\boldsymbol{X}', \boldsymbol{A}') \in \mathbb{B}_{\mathcal{G}} \implies \\ (\boldsymbol{X}', \boldsymbol{A}') \in \gamma(r_{\boldsymbol{X}_{\text{add}}}, r_{\boldsymbol{X}_{\text{del}}}, r_{\boldsymbol{A}_{\text{add}}}, r_{\boldsymbol{A}_{\text{del}}}) \\ \wedge \Big(\forall n \in \{1, \dots, N\} : \\ \left|\left\{d : X_{n,d} = 0 \neq X'_{n,d}\right\}\right| \leq r_{\boldsymbol{X}_{\text{add,loc}\,n}} \wedge \left|\left\{d : X_{n,d} = 1 \neq X'_{n,d}\right\}\right| \leq r_{\boldsymbol{X}_{\text{del,loc}\,n}} \\ \wedge \left|\left\{m : A_{n,m} = 0 \neq X'_{n,m}\right\}\right| \leq r_{\boldsymbol{A}_{\text{add,loc}\,n}} \wedge \left|\left\{m : A_{n,m} = 1 \neq X'_{n,m}\right\}\right| \leq r_{\boldsymbol{A}_{\text{del,loc}\,n}} \Big) \\ \wedge \Big(\exists \mathbb{S} \subseteq \{1, \dots, N\} : |\mathbb{S}| \leq \sigma \wedge \forall d \in \{1, \dots, D\}, n, m \in \{1, \dots, N\} : \\ \left(X'_{n,d} \neq X_{n,d} \implies n \in \mathbb{S}\right) \wedge \left(A'_{n,m} \neq A_{n,m} \implies n \in \mathbb{S} \vee m \in \mathbb{S}\right) \Big)\end{aligned} \tag{25}$$

Next, we show the proof of Lemma 1 delegated from the main paper.

Proof: By the definition of receptive fields (Eq. 2) changes outside the receptive field do not influence the prediction, i.e. $f_n(\boldsymbol{X}', \boldsymbol{A}') = f_n(\boldsymbol{X}'', \boldsymbol{A}'')$. Since $\zeta_n(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}'_{\mathcal{G}}) = 1$ and $(\boldsymbol{X}'', \boldsymbol{A}'') \in \mathbb{B}'_{\mathcal{G}}$, we know that $f_n(\boldsymbol{X}, \boldsymbol{A}) = f_n(\boldsymbol{X}'', \boldsymbol{A}'')$. By transitivity $f_n(\boldsymbol{X}, \boldsymbol{A}) = f_n(\boldsymbol{X}', \boldsymbol{A}')$ $\square$.

C Full collective certificate

Here we discuss how to incorporate local budget constraints and constraints on the number of attacker-controlled nodes into the collective certificate for global budget constraints (see Eq. 15). We also discuss how to adapt it to undirected adjacency matrices (see § C.1) As before, we lower-bound the true objective

$$\min_{(\boldsymbol{X}', \boldsymbol{A}') \in \mathbb{B}_{\mathcal{G}}} \sum_{n \in \mathbb{T}} \mathbf{1}_{f_n(\boldsymbol{X}, \boldsymbol{A}) = f_n(\boldsymbol{X}', \boldsymbol{A}')}. \tag{26}$$

by replacing the indicator functions with base certificates and optimizing over the number of perturbations per node / the perturbed edges. The only difference is that instead of

$$\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b_1, b_2, b_3, b_4\right)\right) \tag{27}$$

we can use

$$\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}_{\mathcal{G}} \cap \gamma\left(b_1, b_2, b_3, b_4\right)\right), \tag{28}$$

where $\mathbb{B}_{\mathcal{G}}$ is the set of admissible perturbed graphs constrained by all parameters of the threat model. This is to account for certificates that feature local budget constraints of their own. In practice, this just means setting the local budget parameters of the base certificate to values larger than or equal to those of the collective threat model. Aside from that, the derivation proceeds as before: Lemma 1 still holds, meaning we only have to consider perturbations within f_n's receptive field when evaluating ζ_n. We can further construct base certificate functions that are monotonically decreasing in the number of these perturbations:

$$\begin{aligned} &\zeta_n'\left(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}_{\mathcal{G}} \cap \gamma\left(b_1, b_2, b_3, b_4\right)\right) = \\ &\max_{\boldsymbol{c} \in \mathbb{N}_0^4} \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}_{\mathcal{G}} \cap \gamma\left(b_1 + c_1, b_2 + c_2, b_3 + c_3, b_4 + c_4\right)\right) \end{aligned} \tag{29}$$

and characterize them by the pareto front of budgets for which they returns 0, i.e. cannot certify robustness.

$$\mathbb{P}^{(n)} = \left\{ \boldsymbol{p} \in \mathbb{P}'^{(n)} | \neg \exists \boldsymbol{p}' \in \mathbb{P}'^{(n)} : \boldsymbol{p}' \neq \boldsymbol{p} \wedge \forall d \in \{1,2,3,4\} : p'_d \leq p_d \right\}. \tag{30}$$

with $\mathbb{P}'^{(n)} = \left\{ p \in \{1, \ldots, N\}^4 | \zeta_n'\left(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}_{\mathcal{G}} \cap \gamma\left(p_1, p_2, p_3, p_4\right)\right) = 0 \right\}$. After encoding $\mathbb{P}^{(n)}$ as a matrix $\boldsymbol{P}^{(n)} \in \mathbb{N}_0^{|\mathbb{P}^{(n)} \times 4}$ we can solve the following optimization problem to obtain a lower bound on Eq. 26:

$$\min_{\left(\boldsymbol{Q}^{(n)}, \boldsymbol{s}^{(n)}\right)_{n=1}^N, \boldsymbol{b}_{\boldsymbol{X}_{\text{add}}}, \boldsymbol{b}_{\boldsymbol{X}_{\text{del}}}, \boldsymbol{B}_{\boldsymbol{A}}, \boldsymbol{t}} |\mathbb{T}| - \sum_{n \in \mathbb{T}} t_n \tag{31}$$

$$\text{s.t.} \qquad ||\boldsymbol{s}^{(n)}||_1 \geq t_n, \qquad Q_{p,d}^{(n)} \geq s_p^{(n)}, \tag{32}$$

$$(\boldsymbol{b}_{\boldsymbol{X}_{\text{add}}})^T \boldsymbol{\psi}^{(n)} \geq Q_{i,1}^{(n)} P_{p,1}^{(n)}, \qquad (\boldsymbol{b}_{\boldsymbol{X}_{\text{del}}})^T \boldsymbol{\psi}^{(n)} \geq Q_{p,2}^{(n)} P_{p,2}^{(n)}, \tag{33}$$

$$\sum_{m,m' \leq N} (1 - A_{m,m'})(\boldsymbol{\Psi}^N \odot B_{\boldsymbol{A}})_{m,m'} \geq Q_{p,3}^{(n)} P_{p,3}^{(n)}, \tag{34}$$

$$\sum_{m,m' \leq N} (1 - A_{m,m'})(\boldsymbol{\Psi}^N \odot B_{\boldsymbol{A}})_{m,m'} \geq Q_{p,4}^{(n)} P_{p,3}^{(n)}, \tag{35}$$

$$||b_{\boldsymbol{X}_{\text{add}}}||_1 \leq r_{\boldsymbol{X}_{\text{add}}}, \qquad ||b_{\boldsymbol{X}_{\text{add}}}||_1 \leq r_{\boldsymbol{X}_{\text{del}}}, \tag{36}$$

$$\sum_{(i,j): A_{i,j}=0} B_{\boldsymbol{A}i,j} \leq r_{\boldsymbol{A}_{\text{add}}}, \qquad \sum_{(i,j): A_{i,j}=1} B_{\boldsymbol{A}i,j} \leq r_{\boldsymbol{A}_{\text{del}}}, \tag{37}$$

$$b_{\boldsymbol{X}_{\text{add}}\, n} \leq a_n r_{\boldsymbol{X}_{\text{add,loc}}\, n}, \qquad b_{\boldsymbol{X}_{\text{del}}\, n} \leq a_n r_{\boldsymbol{X}_{\text{del,loc}}\, n}, \tag{38}$$

$$\sum_{m: A_{n,m}=0} B_{\boldsymbol{A}n,m} + B_{\boldsymbol{A}m,n} \leq r_{\boldsymbol{A}_{\text{add,loc}}\, n}, \qquad \sum_{m: A_{n,m}=1} B_{\boldsymbol{A}n,m} + B_{\boldsymbol{A}m,n} \leq r_{\boldsymbol{A}_{\text{del,loc}}\, n} \tag{39}$$

$$B_{i,j} \leq a_i + a_j \; \forall i, j \in \{1, \ldots, N\}, \tag{40}$$

$$||\boldsymbol{a}||_1 \leq \sigma, \tag{41}$$

$$\boldsymbol{s}^{(n)} \in \{0,1\}^{|\mathbb{P}^{(n)}|}, \qquad \boldsymbol{Q}^{(n)} \in \{0,1\}^{|\mathbb{P}^{(n)}| \times 4}, \qquad \boldsymbol{t} \in \{0,1\}^N, \tag{42}$$

$$\boldsymbol{a} \in \{0,1\}^N, \qquad \boldsymbol{b}_{\boldsymbol{X}_{\text{add}}}, \boldsymbol{b}_{\boldsymbol{X}_{\text{del}}} \in \mathbb{N}_0^N, \qquad \boldsymbol{B}_{\boldsymbol{A}} \in \{0,1\}^{N \times N}, \tag{43}$$

$\forall n \in \{1, \ldots, N\}, p \in \{1, \ldots, |\mathbb{P}^{(n)}|\}, d \in \{1, \ldots, 4\}$.

The constraints from Eq. 32 to Eq. 35 are identical to constraints Eq. 16 to Eq. 19 of our collective certificate for global budget constraints. They simply implement boolean logic to determine whether there is some pareto-optimal $\boldsymbol{p} \in \mathbb{P}(n)$ such that the perturbation in f_n's receptive field matches or exceeds $\boldsymbol{p}$ in all four dimensions. If this is the case, i.e. ζ_n' cannot certify the robustness of f_n, then t_n can be set to 1. Eq. 36 and Eq. 37 enforce the global budget constraints. The difference to global

budget certificate lies in Eq. 38 to Eq. 41. We introduce an additional variable vector $\boldsymbol{a} \in \{0,1\}^N$ that indicates which nodes are attacker controlled. Eq. 38 enforces that the attributes of node n remain unperturbed, unless $a_n = 1$. If $a_n = 1$, the adversary can add or delete at most $r_{\boldsymbol{X}_{\text{del,loc}\,n}}$ or $r_{\boldsymbol{X}_{\text{del,loc}\,n}}$ attribute bits. With edge perturbations, it is sufficient for either incident node to be attacker-controlled. This is expressed via Eq. 40. The number of added or deleted edges incident to node n is constrained via Eq. 39. Finally, Eq. 41 ensures that at most σ nodes are attacker-controlled.

C.1 Undirected adjacency matrix

To adapt our certificate to undirected graphs, we simply change the interpretation of the indicator matrix $\boldsymbol{B}_{\boldsymbol{A}}$. Now, s either $B_{\boldsymbol{A}i,j}$ or $B_{\boldsymbol{A}j,i}$ to 1 perturbs the undirected edge $\{i,j\}$. An edge should not be perturbed twice, which we express through an additional constraint:

$$B_{\boldsymbol{A}i,j} + B_{\boldsymbol{A}j,i} \leq 1 \; \forall i,j \in \{1,\dots,N\}. \tag{44}$$

We further combine Eq. 39, and Eq. 40, which enforced that at least one of the incident nodes of a perturbed edge has to be attacker controlled, which enforced the local budgets for edge perturbations into following constraints:

$$\sum_{m:A_{n,m}=0} B_{\boldsymbol{A}n,m} \leq a_n r_{\boldsymbol{A}_{\text{add,loc}\,n}} \tag{45}$$

$$\sum_{m:A_{n,m}=1} B_{\boldsymbol{A}n,m} \leq a_n r_{\boldsymbol{A}_{\text{del,loc}\,n}} \tag{46}$$

This change does not affect the optimal value of the mixed-integer linear program. But they are more effective than Eq. 39 and Eq. 40 when solving the relaxed linear program and the nodes' local budgets are small relative to their degree.

D Constructing monotonically decreasing base certificates

Our approach requires transforming base certificates ζ_n into base certificates ζ'_n that are monotonically decreasing in their global budget parameters. While deriving our collective certificate, we defined ζ'_n as follows

$$\zeta'_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b_1, b_2, b_3, b_4\right)\right) = \max_{\boldsymbol{c} \in \mathbb{N}_0^4} \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b_1 + c_1, b_2 + c_2, b_3 + c_3, b_4 + c_4\right)\right). \tag{47}$$

This calculation is not necessary, if ζ_n is already monotonically decreasing. In practice, this is often the case (e.g. also the base certificates we have analyzed in our experiments). Base certificates usually correspond to minimization problems constrained by the adversary's global budget. Increasing the global budget means relaxing the constraint set, yielding a smaller optimal value, i.e. the base certificate is monotonically decreasing.

If ζ_n is not monotonically decreasing and we have no additional knowledge to evaluate Eq. 47, we can use the following (potentially sub-optimal) monotonic base certificate

$$\begin{aligned} \zeta''_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b_1, b_2, b_3, b_4\right)\right) = \max_{\boldsymbol{b}' \in \mathbb{N}_0^4} & \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b'_1, b'_2, b'_3, b'_4\right)\right) \\ & \text{s.t. } \forall d \in \{1,2,3,4\} : b_d \leq b'_d \leq r_d, \end{aligned} \tag{48}$$

with parameters $(r_1, \dots, r_4)$. For inputs larger than these r_d, we let ζ' return 0.

Eq. 48 can be solved efficiently as follows: Let $\mathcal{N}' : \mathbb{N}_0^4 \mapsto \mathcal{P}(\mathbb{N}_0^4)$ with

$$\mathcal{N}'(b) = \{b + \boldsymbol{b}' | \boldsymbol{b}' \in 0, 1^4 \wedge ||\boldsymbol{b}'||_1 = 1\}. \tag{49}$$

The function $\mathcal{N}$ maps any input grid coordinate to adjacent coordinates that only differ in one component. For compactness, we allow four-dimensional grids to be indexed by four-element vectors. We can then use Algorithm 1, which exploits dynamic programming, to construct a grid corresponding to ζ''. Algorithm 1 requires $\mathcal{O}(r_1 r_2 r_3 r_4)$ evaluations of ζ.

Algorithm 1: Constructing monotonic base-certificate

Result: 4D grid corresponding to monotonic base certificate ζ' constructed from base certificate ζ, evaluated for all points within budget given by (r_1, r_2, r_3, r_4) .

$\boldsymbol{Z}' \leftarrow \mathbf{0}_{r_1+1, r_2+1, r_3+1, r_4+1}$;
$b_1 \leftarrow r_1$;
while *While* $b_1 \geq 0$ **do**
  $b_2 \leftarrow r_2$;
  while *While* $b_2 \geq 0$ **do**
    $b_3 \leftarrow r_3$;
    while *While* $b_3 \geq 0$ **do**
      $b_4 \leftarrow r_4$;
      while *While* $b_4 \geq 0$ **do**
        $\boldsymbol{b} \leftarrow (b_1, b_2, b_3, b_4)$;
        $\hat{z} \leftarrow \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(\boldsymbol{b}\right)\right)$.;
        $Z'_{\boldsymbol{b}} \leftarrow \max\left(\{Z_{\boldsymbol{b}}\} \cup \{Z'_{\boldsymbol{b}'} | \boldsymbol{b}' \in \mathcal{N}(\boldsymbol{b})\}\right)$;
        $b_4 \leftarrow b_4 - 1$;
      end
      $b_3 \leftarrow b_3 - 1$;
    end
    $b_2 \leftarrow b_2 - 1$;
  end
  $b_1 \leftarrow b_1 - 1$;
end

E Determining the pareto front of base certificates

Our collective certificate is based on characterizing monotonically decreasing base certificates ζ'_n via a set of pareto-optimal points for which ζ'_n cannot certify the robustness of f_n. We can calculate the set of pareto-optimal points that are reachable under global budget constraints paratermized by $(r_1, r_2, r_3, r_4 \in \mathbb{N}_0^4$ as follows:

Let $\mathcal{N}' : \mathbb{N}_0^4 \mapsto \mathcal{P}(\mathbb{N}_0^4)$ with

$$\mathcal{N}'(b) = \{b + \boldsymbol{b}' | \boldsymbol{b}' \in \{0,1\}^4 \wedge ||\boldsymbol{b}'||_1 = 1\}. \tag{50}$$

be a mapping from 4D grid coordinates to adjacent coordinates that differ in only element. Let

$$\mathcal{N}' : \mathbb{N}_0^4 \mapsto \mathcal{P}(\mathbb{N}_0^4) \text{s.t. } \mathcal{N}'(b) = \{b - \boldsymbol{b}' | \boldsymbol{b}' \in \{0,1\}^4 \wedge ||\boldsymbol{b}'||_1 \geq 1\} \tag{51}$$

be a mapping from 4D grid coordinates to adjacent coordinates that are smaller or equal in all elements. We can then use Algorithm 2 to calculate the set of pareto optimal points $\mathbb{P}^{(n)}$ for monotonically decreasing base certificate ζ'_n. In the worst case, Algorithm 2 requires $\mathcal{O}(\hat{r}_{\boldsymbol{X}_{\text{add}}} \hat{r}_{\boldsymbol{X}_{\text{del}}} \hat{r}_{\boldsymbol{A}_{\text{add}}} \hat{r}_{\boldsymbol{A}_{\text{del}}})$ evaluations of ζ', with the $\hat{r}$ corresponding to the largest number of perturbations for which ζ' guarantees the robustness of f_n.

F Tightness for randomized smoothing

In this section we prove that if we use the randomized smoothing based certificate from Bojchevski et al. (2020) as our base certificate, then our collective certificate is tight: If we do not make any further assumptions, outside each smoothed classifier's receptive field and its expected output behavior under the smoothing distribution, we cannot obtain a better collective certificate. We define the base certificate and then provide a constructive proof of the resulting collective certificate's tightness.

F.1 Randomized smoothing for sparse data

Bojchevski et al. (2020) provide a robustness certificate for classification of arbitrary sparse binary data. Applied to node classification, it can be summarized as follows:

Algorithm 2: Determining reachable pareto-optimal points

Result: Set $\mathbb{P}'^{(n)}$ of pareto-optimal points of monotonic base certificate ζ'_n (c.f. Eq. 14) that are reachable within budget given by (r_1, r_2, r_3, r_4) .

$\mathbb{P}'^{(n)} \leftarrow \{\}$;

Function *extend(**b**)*

  $\hat{z}' \leftarrow \zeta'_n(f, \boldsymbol{X}, \boldsymbol{A}, \gamma(\boldsymbol{b}))$.;

  if $\hat{z}' = 0 \wedge \boldsymbol{b} \notin \mathbb{P}'^{(n)} \wedge \forall d \in \{1,2,3,4\} : b_d \leq (r_1+1, r_2+1, r_3+1, r_4+1)_d$ **then**

    for $\boldsymbol{b}' \in \mathcal{N}(\boldsymbol{b})$ **do**

      $\mathbb{P}' \leftarrow \mathbb{P}' \cup \{\boldsymbol{b}\}$;

      extend($\boldsymbol{b}'$) ;

    end

  end

extend((0,0,0,0)) ;

for $\boldsymbol{b} \in \mathbb{P}^{(n)'}$ **do**

  pareto_optimal $\leftarrow$ true;

  for $\boldsymbol{b}' \in \mathcal{N}'(\boldsymbol{b})$ **do**

    if $\boldsymbol{b}' \in \mathbb{P}^{(n)'}$ **then**

      pareto_optimal $\leftarrow$ false

    end

  end

  if pareto_optimal **then**

    $\mathbb{P}^{(n)} \leftarrow \mathbb{P}^{(n)} \cup \{\boldsymbol{b}\}$

  end

end

Assume we are given a probabilistic multi-output classifier $h : \mathbb{G} \mapsto (\Delta_C)^N$, where Δ_C is the C-dimensional probability simplex. Define a smoothed classifier $f : \mathbb{G} \mapsto \{1, \ldots, C\}^N$ with

$$f_n(\boldsymbol{X}, \boldsymbol{A}) = \operatorname{argmax}_{c \in \{1,\ldots,C\}} \mathbb{E}\left[h(\phi_{\text{attr}}(\boldsymbol{X}), \phi_{\text{adj}}(\boldsymbol{A}))\right]_{n,c} \; \forall n \in \{1, \ldots, N\}, \tag{52}$$

where ϕ_{attr} and ϕ_{adj} are two independent randomization schemes that assign probability mass to the set of attribute matrices $\{0,1\}^{N \times D}$ and adjacency matrices $\{0,1\}^{N \times N}$, respectively. The randomization schemes are defined as follows:

$$\Pr\left(\phi_{\text{attr}}(\boldsymbol{X})_{m,d} = 1 - X_{m,d}\right) = \theta_{\boldsymbol{X}_{\text{del}}}^{X_{m,d}} \theta_{\boldsymbol{X}_{\text{add}}}^{(1-X_{m,d})} \; \forall m \in \{1, \ldots, N\}, d \in \{1, \ldots, D\} \tag{53}$$

and

$$\Pr\left(\phi_{\text{adj}}(\boldsymbol{A})_{i,j} = 1 - A_{i,j}\right) = \theta_{\boldsymbol{A}_{\text{del}}}^{A_{i,j}} \theta_{\boldsymbol{A}_{\text{add}}}^{(1-A_{i,j})} \; \forall i, j \in \{1, \ldots, N\}. \tag{54}$$

Each bit's probability of being flipped is dependent on its current value, but independent of the other bits.

An adversarially perturbed graph $(\boldsymbol{X}', \boldsymbol{A}')$ is successful in changing prediction $y_n = f_n(\boldsymbol{X}, \boldsymbol{A})$, if

$$y_n \neq \operatorname{argmax}_{c \in \{1,\ldots,C\}} \mathbb{E}\left[h(\phi_{\text{attr}}(\boldsymbol{X}'), \phi_{\text{adj}}(\boldsymbol{A}'))\right]_{n,c}. \tag{55}$$

Evaluating this inequality is usually not tractable. We can however relax the problem: Let $p_n = \mathbb{E}\left[h(\phi_{\text{attr}}(\boldsymbol{X}), \phi_{\text{adj}}(\boldsymbol{A}))\right]_{n,y_n}$ and let $\mathbb{H}$ be the set of all possible probabilistic single-output classifiers for graphs in $\mathcal{G}$. If

$$\left(\min_{\tilde{h}_n \in \mathbb{H}} \mathbb{E}\left[\tilde{h}_n(\phi_{\text{attr}}(\boldsymbol{X}'), \phi_{\text{adj}}(\boldsymbol{A}'))\right]_{y_n}\right) > 0.5 \tag{56}$$

subject to

$$\mathbb{E}\left[\tilde{h}_n(\phi_{\text{attr}}(\boldsymbol{X}), \phi_{\text{adj}}(\boldsymbol{A}))\right]_{y_n} = p_n, \tag{57}$$

then $f_n(\boldsymbol{X}', \boldsymbol{A}') = f_n(\boldsymbol{X}, \boldsymbol{A})$. It is easy to see why: The unsmoothed h_n is in the set defined by Eq. 57, so the result of the optimization problem is a lower bound on $\mathbb{E}\left[h_n(\phi_{\text{attr}}(\boldsymbol{X}'), \phi_{\text{adj}}(\boldsymbol{A}'))\right]_{f_n}$. If this lower bound is larger than 0.5, then f_n is guaranteed to be the argmax class.

Optimizing over the set of all possible classifiers might appear hard. We can however use the approach from Lee et al. (2019) to find an optimum. Let $\boldsymbol{X}', \boldsymbol{A}'$ be a graph that results from $b_{\boldsymbol{X}_{\text{add}}}$ attribute additions, $b_{\boldsymbol{X}_{\text{del}}}$ attribute deletions, $b_{\boldsymbol{A}_{\text{add}}}$ edge additions, and $b_{\boldsymbol{A}_{\text{del}}}$ edge applied to $(\boldsymbol{X}, \boldsymbol{A})$. We can partition the set of graphs into $(b_{\boldsymbol{X}_{\text{add}}} + b_{\boldsymbol{X}_{\text{del}}} + 1)\,(b_{\boldsymbol{A}_{\text{add}}} + b_{\boldsymbol{A}_{\text{del}}} + 1)$ regions that have a constant likelihood ratio under our smoothing distribution:

$$\left\{\mathbb{J}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}} | q_{\boldsymbol{X}}, q_{\boldsymbol{A}} \in \mathbb{N}_0 \wedge q_{\boldsymbol{X}} \leq \left(b^{(n)}_{\boldsymbol{X}_{\text{add}}} + b^{(n)}_{\boldsymbol{X}_{\text{del}}} + 1\right) \wedge q_{\boldsymbol{A}} \leq \left(b^{(n)}_{\boldsymbol{A}_{\text{add}}} + b^{(n)}_{\boldsymbol{A}_{\text{del}}} + 1\right)\right\} \tag{58}$$

with

$$(\boldsymbol{X}'', \boldsymbol{A}'' \in \mathbb{J}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}}) \implies \left(\frac{\Pr\left(\phi_{\text{attr}}(\boldsymbol{X}) = X'' \wedge \phi_{\text{adj}}(\boldsymbol{A}) = A''\right)}{\Pr\left(\phi_{\text{attr}}(\boldsymbol{X}') = X'' \wedge \phi_{\text{adj}}(\boldsymbol{A}') = A''\right)} = \eta_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}}, \right) \tag{59}$$

where the $\eta_{\cdot,\cdot} \in \mathbb{R}_+$ are constants. The regions have a particular semantic meaning, which will be important for our later proof: Any $(\boldsymbol{X}'', \boldsymbol{A}'') \in \mathbb{J}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}}$ has $q_{\boldsymbol{X}}$ attribute bits and $q_{\boldsymbol{A}}$ adjacency bits that have the same value in $(\boldsymbol{X}, \boldsymbol{A})$, and a different value in $(\boldsymbol{X}', \boldsymbol{A}')$:

$$\begin{aligned}(\boldsymbol{X}'', \boldsymbol{A}'' \in \mathbb{J}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}}) \iff \\ \left|\left\{m, d | X''_{m,d} = X_{m,d} \neq X'_{m,d}\right\}\right| = q_{\boldsymbol{X}} \wedge \left|\left\{i, j | A''_{i,j} = A_{i,j} \neq A'_{i,j}\right\}\right| = q_{\boldsymbol{A}}.\end{aligned} \tag{60}$$

As proven by Lee et al. (2019), we can find an optimal solution to Eq. 56 by optimizing over the set of single-output classifiers that have a constant output in each region of constant likelihood ratio. This can be implemented via the following linear program:

$$\Lambda_n\left(b_{\boldsymbol{X}_{\text{add}}}, b_{\boldsymbol{X}_{\text{del}}}, b_{\boldsymbol{A}_{\text{add}}}, b_{\boldsymbol{A}_{\text{del}}}, p_n, \right) :=$$

$$\min_{\boldsymbol{H}^{(n)}} \sum_{q_{\boldsymbol{X}}=0}^{\left(b_{\boldsymbol{X}_{\text{add}}}+b_{\boldsymbol{X}_{\text{del}}}\right)} \sum_{q_{\boldsymbol{A}}=0}^{\left(b_{\boldsymbol{A}_{\text{add}}}+b_{\boldsymbol{A}_{\text{del}}}\right)} H^{(n)}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}} \Pr((\phi_{\text{attr}}(X'), \phi_{\text{adj}}(A')) \in \mathbb{J}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}}) \tag{61}$$

$$\text{s.t.} \sum_{q_{\boldsymbol{X}}=0}^{\left(b_{\boldsymbol{X}_{\text{add}}}+b_{\boldsymbol{X}_{\text{del}}}\right)} \sum_{q_{\boldsymbol{A}}=0}^{\left(b_{\boldsymbol{A}_{\text{add}}}+b_{\boldsymbol{A}_{\text{del}}}\right)} H^{(n)}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}} \Pr((\phi_{\text{attr}}(X), \phi_{\text{adj}}(A)) \in \mathbb{J}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}}) = p_n, \tag{62}$$

$$\boldsymbol{H}^{(n)} \in [0,1]^{\left(r_{X_{\text{add}}}+r_{X_{\text{del}}}\right)\times\left(r_{A_{\text{add}}}+r_{A_{\text{del}}}\right)}. \tag{63}$$

Any optimal solution $\tilde{\boldsymbol{H}}^{(n)}$ corresponds to a single-output classifier $\tilde{h}_n$ that, for some input graph $(\boldsymbol{X}'', \boldsymbol{A}'')$ simply counts the number of attribute bits $q_{\boldsymbol{X}}$ and adjacency bits $q_{\boldsymbol{A}}$ that have the same value in $(\boldsymbol{X}, \boldsymbol{A})$ and a different value $(\boldsymbol{X}', \boldsymbol{A}')$ and then assigns a probability of $H^{(n)}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}}$ to class f_n and $1 - H^{(n)}_{q_{\boldsymbol{X}},q_{\boldsymbol{A}}}$ to the remaining classes.

The optimal value of $Eq.\ 61$ being larger than 0.5 for a fixed perturbed graph $(\boldsymbol{X}', \boldsymbol{A}')$ only proofs that this particular graph is not a successful attack on f_n. For a robustness certificate, we want to know the result for a worst-case graph. However, the result is only dependent on the number of perturbations $b_{\boldsymbol{X}_{\text{add}}}, b_{\boldsymbol{X}_{\text{del}}}, b_{\boldsymbol{A}_{\text{add}}}$ and $b_{\boldsymbol{A}_{\text{del}}}$, and not on which specific bits are perturbed. Therefore, we can solve the problem for an arbitrary fixed perturbed graph with the given number of perturbations, and obtain a valid robustness certificate.

For use in our collective certificate, we define $\zeta_n(\cdot) \mapsto \{0, 1\}$ with

$$\begin{aligned}&\left[\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(b_{\boldsymbol{X}_{\text{add}}}, b_{\boldsymbol{X}_{\text{del}}}, b_{\boldsymbol{A}_{\text{add}}}, b_{\boldsymbol{A}_{\text{del}}}\right)\right) = 1\right] \\ \iff &\left[\Lambda_n\left(b_{\boldsymbol{X}_{\text{add}}}, b_{\boldsymbol{X}_{\text{del}}}, b_{\boldsymbol{A}_{\text{add}}}, b_{\boldsymbol{A}_{\text{del}}}, p_n\right) > 0.5\right],\end{aligned} \tag{64}$$

where $\Lambda_n\left(b_{\boldsymbol{X}_{\text{add}}}, b_{\boldsymbol{X}_{\text{del}}}, b_{\boldsymbol{A}_{\text{add}}}, b_{\boldsymbol{A}_{\text{del}}}, p_n\right)$ is the optimal solution to Eq. 61. For all remaining possible input sets of graphs $\mathbb{B}'_{\mathcal{G}} \in \mathbb{G} : \mathbb{B}'_{\mathcal{G}} \notin \{\gamma(\boldsymbol{b}) \,|\boldsymbol{b} \in \mathbb{N}_0^4\}$, we set $\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \mathbb{B}'_{\mathcal{G}}\right) = 0$. (This is just for completeness, we never evaluate the base certificate on these sets of graphs.)

F.2 TIGHTNESS PROOF

With the definition of our base certificate in place, we can now formalize and prove that the resulting collective certificate is tight. Recall that randomized smoothing is a black-box method. The classifier that is being smoothed is treated as unknown. A robustness certificate based on randomized smoothing has to account for the worst-case (i.e. least robust under the given threat model)

classifier. Our collective certificate lower-bounds the number of predictions that are guaranteed to be simultaneously robust. We show that with the randomized smoothing base certificate from the previous section, it actually yields the exact number of robust classifiers, assuming the worst-case unsmoothed classifier.

Theorem 1 *Let $(\boldsymbol{X}, \boldsymbol{A})$ be an unperturbed graph. Let $h : \mathbb{G} \mapsto (\Delta_C)^N$ a probabilistic multi-output classifier. Let $f : \mathbb{G} \mapsto \{1, \dots, C\}$ be the corresponding smoothed classifier with*

$$f_n(\boldsymbol{X}'', \boldsymbol{A}'') = \text{argmax}_{c \in \{1,\dots,C\}} \mathbb{E}\left[h(\phi_{\text{attr}}(\boldsymbol{X}''), \phi_{\text{adj}}(\boldsymbol{A}''))\right]_{n,c}, \tag{65}$$

$$y_n = f_n(\boldsymbol{X}, \boldsymbol{A}), \tag{66}$$

$$p_n = \mathbb{E}\left[h(\phi_{\text{attr}}(\boldsymbol{X}), \phi_{\text{adj}}(\boldsymbol{A}))\right]_{n,y_n}, \tag{67}$$

and randomization schemes $\phi_{\text{attr}}(\boldsymbol{X}), \phi_{\text{adj}}(\boldsymbol{A})$ defined as in Eq. 53 and Eq. 54. Let $\boldsymbol{\psi} \in \{0,1\}^N, \boldsymbol{\Psi} \in \{0,1\}^{N \times N}$ be receptive field indicators corresponding to f_n (see Eq. 2). Let $\mathbb{B}_{\mathcal{G}}$ be a set of admissible perturbed graphs, constrained by parameters $r_{\boldsymbol{X}_{\text{add}}}, r_{\boldsymbol{X}_{\text{del}}}, r_{\boldsymbol{A}_{\text{add}}}, r_{\boldsymbol{A}_{\text{del}}}, \boldsymbol{r}_{\boldsymbol{X}_{\text{add,loc}}}, \boldsymbol{r}_{\boldsymbol{X}_{\text{del,loc}}}, \boldsymbol{r}_{\boldsymbol{A}_{\text{add,loc}}}, \boldsymbol{r}_{\boldsymbol{A}_{\text{del,loc}}}, \sigma$, as defined in § 2. Let $\mathbb{T}$ be the indices of nodes targeted by an adversary. Under the given parameters, let $o*$ be the optimal value of the optimization problem defined in § C.

Then there are a perturbed graph $(\boldsymbol{X}', \boldsymbol{A}')$, a probabilistic multi-output classifier $\tilde{h}$ and a corresponding smoothed multi-output classifier $\tilde{f}$ with

$$\tilde{f}_n(\boldsymbol{X}, \boldsymbol{A}) = \text{argmax}_{c \in \{1,\dots,C\}} \mathbb{E}\left[\tilde{h}(\phi_{\text{attr}}(\boldsymbol{X}), \phi_{\text{adj}}(\boldsymbol{A}))\right]_{n,c} \quad \forall n \in \{1, \dots, N\} \tag{68}$$

such that

$$\left|\{n \in \mathbb{T} |\, \tilde{f}_n(\boldsymbol{X}', \boldsymbol{A}') = y_n\right\}| = o*, \tag{69}$$

$$\mathbb{E}\left[\tilde{h}(\phi_{\text{attr}}(\boldsymbol{X}), \phi_{\text{adj}}(\boldsymbol{A}))\right]_{n,y_n} = p_n, \tag{70}$$

and each $\tilde{f}_n$ is only dependent on nodes and edges for which $\boldsymbol{\psi}^{(n)}$ and $\boldsymbol{\Psi}^{(n)}$ have value 1.

Proof: The optimization problem from § C has three parameters $\boldsymbol{b}_{\boldsymbol{X}_{\text{add}}}, \boldsymbol{b}_{\boldsymbol{X}_{\text{del}}}, \boldsymbol{B}_{\boldsymbol{A}}$, which specify the budget allocation of the adversary. Let $\boldsymbol{b}^*_{\boldsymbol{X}_{\text{add}}}, \boldsymbol{b}^*_{\boldsymbol{X}_{\text{del}}}, \boldsymbol{B}^*_{\boldsymbol{A}}$ be their value in the optimum. We can construct a perturbed graph $(\boldsymbol{X}', \boldsymbol{A}')$ from the clean graph $(\boldsymbol{X}, \boldsymbol{A})$ as follows: For every node n, set the first $b^*_{\boldsymbol{X}_{\text{add}}\,n}$ zero-valued bits to one and the first $b_{\boldsymbol{X}_{\text{del}}\,n}{}^*$ non-zero bits to zero. Then, flip any entry (n, m) of $A_{n,m}$ for which $B^*_{\boldsymbol{A}\,n,m} = 1$. The parameters $\boldsymbol{b}^*_{\boldsymbol{X}_{\text{add}}}, \boldsymbol{b}^*_{\boldsymbol{X}_{\text{del}}}, \boldsymbol{B}^*_{\boldsymbol{A}}$ are part of a feasible solution to the optimization problem. In particular, they must fulfill constraints Eq. 36 to Eq. 41, which guarantee that the constructed graph is in $\mathbb{B}_{\mathcal{G}}$.

Given the perturbed graph $(\boldsymbol{X}', \boldsymbol{A}')$, we can calculate the amount of perturbation in the receptive field of f_n:

$$u^{(n)}_{\boldsymbol{X}_{\text{add}}} = (\boldsymbol{b}^*_{\boldsymbol{X}_{\text{add}}})^T \boldsymbol{\psi}^{(n)} \tag{71}$$

$$u^{(n)}_{\boldsymbol{X}_{\text{del}}} = (\boldsymbol{b}^*_{\boldsymbol{X}_{\text{del}}})^T \boldsymbol{\psi}^{(n)} \tag{72}$$

$$u^{(n)}_{\boldsymbol{A}_{\text{add}}} = \sum_{(i,j):A_{i,j}=0} \Psi_{i,j} B^*_{\boldsymbol{A}\,i,j} \tag{73}$$

$$u^{(n)}_{\boldsymbol{A}_{\text{add}}} = \sum_{(i,j):A_{i,j}=1} \Psi_{i,j} B^*_{\boldsymbol{A}\,i,j} \tag{74}$$

We can now specify the unsmoothed multi-output classifier $\tilde{h}$. Recall that in the collective certificate's optimization problem, each f_n is associated with a binary variable t_n. In the optimum, $(t^*_n = 1) \iff \zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(u^{(n)}_{\boldsymbol{X}_{\text{add}}}, u^{(n)}_{\boldsymbol{X}_{\text{del}}}, u^{(n)}_{\boldsymbol{A}_{\text{add}}}, u^{(n)}_{\boldsymbol{A}_{\text{del}}}\right)\right) = 0$ and $o^* = \sum_{n \in \mathbb{T}} |\mathbb{T}| - t^*_n$.

Case 1: $n \notin \mathbb{T}$. Choose $\tilde{h}_n = h_n$. Trivially, constraint Eq. 70 is fulfilled $\tilde{f}_n$ is only dependent on nodes and edges for which $\boldsymbol{\psi}^{(n)}$ and $\boldsymbol{\Psi}^{(n)}$ have value 1. Whether $\tilde{f}$ is adversarially attacked or not does not influence Eq. 69, as $n \notin \mathbb{T}$.

Case 2: $n \in \mathbb{T}$ and $t_n^* = 0$. Choose $\tilde{h}_n = h_n$. Again, constraint Eq. 70 is fulfilled $\tilde{f}_n$ is only dependent on nodes and edges for which $\boldsymbol{\psi}^{(n)}$ and $\boldsymbol{\Psi}^{(n)}$ have value 1. Since $t_n^* = 0$, we know that $\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(u_{\boldsymbol{X}_{\text{add}}}^{(n)}, u_{\boldsymbol{X}_{\text{del}}}^{(n)}, u_{\boldsymbol{A}_{\text{add}}}^{(n)}, u_{\boldsymbol{A}_{\text{del}}}^{(n)}\right)\right) = 1$, i.e. $\tilde{f}_n(\boldsymbol{X}', \boldsymbol{A}') = y_n$.

Case 3: $n \in \mathbb{T}$ and $t_n^* = 1$. Since $t_n^* = 1$, we know that f_n is not certified by the base certificate: $\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(u_{\boldsymbol{X}_{\text{add}}}^{(n)}, u_{\boldsymbol{X}_{\text{del}}}^{(n)}, u_{\boldsymbol{A}_{\text{add}}}^{(n)}, u_{\boldsymbol{A}_{\text{del}}}^{(n)}\right)\right) = 0$. Let $\boldsymbol{H}^{*(n)} \in [0,1]^{\left(u_{\boldsymbol{X}_{\text{add}}}^{(n)} + u_{\boldsymbol{X}_{\text{del}}}^{(n)}\right) \times \left(u_{\boldsymbol{A}_{\text{add}}}^{(n)} + u_{\boldsymbol{A}_{\text{del}}}^{(n)}\right)}$ be the optimum of the linear program underlying the base certificate (see Eq. 61 to Eq. 63). Define $\tilde{h}_n$ as follows:

$$\tilde{h}_n(\boldsymbol{X}'', \boldsymbol{A}'')_{y_n} = H^{*(n)}_{q_{\boldsymbol{X}}^{(n)}(\boldsymbol{X}''), q_{\boldsymbol{A}}^{(n)}(\boldsymbol{A}'')} \tag{75}$$

$$\tilde{h}_n(\boldsymbol{X}'', \boldsymbol{A}'')_{y_n'} = 1 - H^{*(n)}_{q_{\boldsymbol{X}}^{(n)}(\boldsymbol{X}''), q_{\boldsymbol{A}}^{(n)}(\boldsymbol{A}'')} \tag{76}$$

for some $y_n' \neq y_n$ and with

$$q_{\boldsymbol{X}}^{(n)}(\boldsymbol{X}'') = \left|\left\{(n,d) | n \in \mathbb{X}_{\mathbb{B}_{\mathcal{G}}}^{(n)} \wedge X_{n,d}'' = X_{n,d} \neq X_{n,d}'\right\}\right| \tag{77}$$

$$q_{\boldsymbol{A}}^{(n)}(\boldsymbol{A}'') = \left|\left\{(n,m) | (n,m) \in \mathbb{A}_{\mathbb{B}_{\mathcal{G}}}^{(n)} \wedge A_{n,m}'' = A_{n,m} \neq A_{n,m}'\right\}\right|. \tag{78}$$

As previously discussed, this classifier simply counts the number of bits in f_n's receptive field that have the same value in the clean graph $(\boldsymbol{X}, \boldsymbol{A})$ and a different value in the perturbed graph $(\boldsymbol{X}', \boldsymbol{A}'')$. Since $\boldsymbol{H}^{*(n)}$ is a valid solution to the linear program underlying the base certificate, we know that Eq. 70 is fulfilled, as it is equivalent to Eq. 62 from the base certificate. Since $\zeta_n\left(f, \boldsymbol{X}, \boldsymbol{A}, \gamma\left(u_{\boldsymbol{X}_{\text{add}}}^{(n)}, u_{\boldsymbol{X}_{\text{del}}}^{(n)}, u_{\boldsymbol{A}_{\text{add}}}^{(n)}, u_{\boldsymbol{A}_{\text{del}}}^{(n)}\right)\right) = 0$ we know that $\mathbb{E}\left[\tilde{h}_n(\phi_{\text{attr}}(\boldsymbol{X}'), \phi_{\text{adj}}(\boldsymbol{A}'))\right]_{y_n'} \geq 0.5$ (c.f. Eq. 64 , i.e. $\tilde{f}_n$ is successfully attacked, $\tilde{f}_n(\boldsymbol{X}', \boldsymbol{A}') \neq y_n$.

By construction, we have exactly o^* nodes for which $\tilde{f}_n(\boldsymbol{X}', \boldsymbol{A}') = y_n$ and the remaining constraints are fulfilled as well. □

G HYPERPARAMETERS

Training schedule for smoothed classifiers. Training is performed in a semi-supervised fashion with 20 nodes per class as a train set. Another 20 nodes per class serve as a validation set. Models are trained with Adam (learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay = 0.001) for 3000 epochs, using the average cross-entropy loss across all training set nodes, with a batch size of 1. We employ early stopping, if the validation loss does not decrease for 50 epochs. In each epoch, a different graph is sampled from the smoothing distribution. We do not use the KL-divergence based regularization loss proposed for RGCN, as we found it to decrease the certifiable robustness of the model.

Training schedule for non-smoothed GCN. Training is performed in a semi-supervised fashion with 20 nodes per class as a train set. Another 20 nodes per class serve as a validation set. For the first 100 of 1000 epochs, models are trained with Adam (learning rate = 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay = 10^-5), using the average cross-entropy loss across all training set nodes, with a batch size of 8. After 100 episodes, we add the robust loss proposed in (Zügner & Günnemann, 2019) (local budget q = 21, global budget Q = 12, training node classification margin = $\log(90/10)$, unlabeled node classification margin = $\log(60/40)$). The gradient for the robust loss term is accumulated over 5 epochs before each weight update in order to simulate larger batch sizes.

Network parameters. For GCN and GAT, we use two convolution layers with 64 hidden activations and 50% dropout after the hidden layer. For GAT, we set the number of attention heads to 8 for the first layer and 1 for the second layer. RGCN uses independent gaussians as its internal representation (i.e. each feature dimension has a mean and a variance). For RGCN, we use one linear layer, followed by two convolution layers. We set the number of hidden activations to 32 for the means and 32 for the variances. Each hidden layer is followed by 50% dropout on both the means and variances. For APPNP, we use two linear layers with 64 hidden activations and 50% dropout after the hidden

layer, followed by a propagation layer based on approximate pagerank (teleport probablity = 0.15, iterations = 10). To ensure locality, we set all but the top 64 of each row in the approximate pagerank matrix to 0.

Randomized smoothing. Randomized smoothing introduces four additional hyperparameters $\theta_{\boldsymbol{X}_{\text{add}}}, \theta_{\boldsymbol{X}_{\text{add}}}, \theta_{\boldsymbol{A}_{\text{add}}}, \theta_{\boldsymbol{A}_{\text{del}}}$, which control the probability of flipping bits in the attribute and adjacency matrix under the smoothing distribution. If we only certify attribute perturbations, we set $\theta_{\boldsymbol{A}_{\text{add}}} = \theta_{\boldsymbol{A}_{\text{del}}} = 0$, $\theta_{\boldsymbol{X}_{\text{add}}} = 0.002$ and $\theta_{\boldsymbol{X}_{\text{del}}} = 0.6$. If we only certify adjacency perturbations, we set $\theta_{\boldsymbol{X}_{\text{add}}} = \theta_{\boldsymbol{X}_{\text{del}}} = 0$, $\theta_{\boldsymbol{A}_{\text{add}}} = 0$ and $\theta_{\boldsymbol{A}_{\text{del}}} = 0.4$. If we jointly certify attribute and adjacency perturbations, we set $\theta_{\boldsymbol{X}_{\text{add}}} = 0.002, \theta_{\boldsymbol{X}_{\text{del}}} = 0.6, \theta_{\boldsymbol{A}_{\text{add}}} = 0$ and $\theta_{\boldsymbol{A}_{\text{del}}} = 0.4$. Exactly evaluating smoothed classifiers is not possible, they have to be approximated via sampling. We use 1000 samples to determine a classifier's majority class (y_n), followed by 10^6 samples to estimate the probability of the majority class (p_n) via a Clopper-Pearson confidence interval. Applying Bonferri correction, the confidence level for each confidence interval is set to $1 - 0.01/N$, to obtain an overall confidence level of 99% for all certificates.