Randomized Message-Interception Smoothing: Gray-box Certificates for Graph Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Randomized smoothing is one of the most promising frameworks for certifying the adversarial robustness of machine learning models, including Graph Neural Networks (GNNs). Yet, existing randomized smoothing certificates for GNNs are overly pessimistic since they treat the model as a black box, ignoring the underlying architecture. To remedy this, we propose novel gray-box certificates that exploit the message-passing principle of GNNs: We randomly intercept messages and carefully analyze the probability that messages from adversarially controlled nodes reach their target nodes. Compared to existing certificates, we certify robustness to much stronger adversaries that control entire nodes in the graph and can arbitrarily manipulate node features. Our certificates provide stronger guarantees for attacks at larger distances, as messages from farther-away nodes are more likely to get intercepted. We demonstrate the effectiveness of our method on various models and datasets. Since our gray-box certificates consider the underlying graph structure, we can significantly improve certifiable robustness by applying graph sparsification.

1 Introduction

The core principle behind the majority of Graph Neural Networks (GNNs) is message passing – the representation of a node is (recursively) computed based on the representations of its neighbors (Gilmer et al., 2017). This allows for information to propagate across the graph, e.g. in a k-layer GNN the prediction for a node depends on the messages received from its k-hop neighborhood. With such models, if an adversary controls a few nodes in the graph, they can manipulate node features to craft adversarial messages that in turn change the prediction of a target node.

For example in social networks, adversaries may arbitrarily manipulate features of nodes in their control as modelled by recent adversarial attacks (Ma et al., 2020; Zou et al., 2021). Such perturbations may be unnoticeable and detecting them is a difficult unsolved task (Carlini and Wagner, 2017a).

How can we limit the influence of such adversarial attacks? We introduce a simple but powerful idea: *intercept* adversarial messages. Specifically, we propose message-interception smoothing where we randomly delete edges and/or randomly ablate (mask) nodes, and analyze the probability that messages from adversarially controlled nodes reach the target nodes. By transforming any message-passing GNN into a smoothed GNN where the prediction is the majority vote under this randomized message interception we can provide robustness certificates (compare Figure 1).

Experimentally we obtain significantly better robustness guarantees compared to previous (smoothing) certificates for GNNs (compare Section 6). This improvement stems from the fact that our certificates take the underlying architecture of the classifier into account. Unlike previous randomized smoothing certificates which treat the GNN as a black-box, our certificates are *gray-box*. By making the certificate message-passing aware we partially open the black-box and obtain stronger guarantees.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

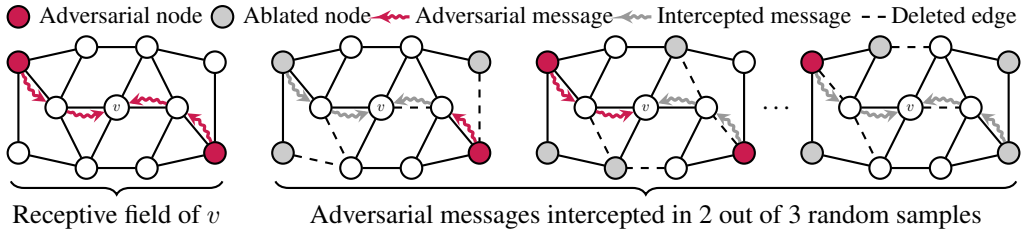

Figure 1: Randomized message-interception smoothing: We model adversaries that can arbitrarily manipulate features of a few nodes in their control (red) to alter the predictions for target nodes v. We randomly delete edges and/or ablate (mask) all features of entire nodes (gray) to intercept messages. We provide certificates based on the majority vote under this randomized message interception.

Our approach is also in contrast to white-box certificates that apply only to very specific models, e.g. Zügner and Günnemann (2019) can only certify the GCN model (Kipf and Welling, 2017). While newly introduced GNNs would require such certificates to be derived from scratch, our approach is model-agnostic and flexible enough to accommodate a large family of message-passing GNNs.

We demonstrate our certificates on node classification datasets and analyze the robustness of existing GNN architectures. By applying simple graph sparsification we further increase the certifiable robustness while retaining high accuracy, as sparsification reduces the number of messages to intercept. In stark contrast to previous probabilistic smoothing-based certificates for GNNs, our certificates require only a few Monte-Carlo samples, are more efficient and practically more relevant: For example, our certificates on Cora-ML saturate in just 17 seconds and assume much stronger adversaries than previous smoothing certificates (Bojchevski et al., 2020) that take up to 25 minutes.

In short, our main contributions are:

- The first gray-box smoothing-based certificates for GNNs that exploit the underlying *message-passing* principle for stronger guarantees.
- Novel randomized smoothing certificates for strong threat models where adversaries can arbitrarily manipulate features of a few nodes in their control.

2 Preliminaries and Background

Threat model. We develop certificates against feature perturbations for *evasion* threat models. Given attributed graphs $G = (\boldsymbol{A}, \boldsymbol{X}) \in \mathbb{G}$ encoded via adjacency matrix $\boldsymbol{A} \in \{0,1\}^{n \times n}$ and feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with n nodes and d features. We formally define the threat model of feature perturbations as a ball $B_\rho(G) \triangleq \{G' : \delta(G, G') \leq \rho\}$ around a given graph $G = (\boldsymbol{A}, \boldsymbol{X})$, where $\delta(G, G') \triangleq \sum_{v=1}^{n} \mathbf{1}_{\boldsymbol{x}_v \neq \boldsymbol{x}'_v}$ denotes the number of nodes whose features differ in at least one dimension when comparing the clean graph G and the perturbed graph $G' = (\boldsymbol{A}', \boldsymbol{X}')$. Intuitively, this means adversaries control up to ρ nodes in the graph and can arbitrarily manipulate node features.

Graph neural networks. We design robustness certificates for message-passing GNNs composed of GNN-layers $\Psi^{(\ell)}$ that aggregate the node representations of previous layers based on the underlying graph structure (Gilmer et al., 2017). Each message-passing layer first gathers messages $\boldsymbol{m}_u^{(\ell-1)}$ from all nodes u in the local neighborhood $\mathcal{N}(v) \triangleq \{u \mid \boldsymbol{A}_{uv} = 1\}$ of node v. The gathered messages are subsequently transformed and aggregated into the hidden representation $\boldsymbol{m}_v^{(\ell)}$ of node v. More formally, a message-passing layer can be described as follows:

$$\boldsymbol{m}_v^{(\ell)} \triangleq \Psi^{(\ell)}_{u \in \mathcal{N}(v) \cup \{v\}} \left(\boldsymbol{m}_v^{(\ell-1)}, \boldsymbol{m}_u^{(\ell-1)} \right)$$

For node classification, message-passing GNNs with k GNN-layers can be described as parametrized functions $f : \mathbb{G} \to \{1, \ldots, C\}^n$ that assign each node v in graph G class $f_v(G)$:

$$f_v(G) \triangleq \arg\max_c \boldsymbol{m}_{v,c}^{(k)}$$

where $\boldsymbol{m}_v^{(0)} \triangleq \boldsymbol{x}_v \in \mathbb{R}^d$ denotes the input and $\boldsymbol{m}_v^{(k)} \in \mathbb{R}^C$ the final representation of node v.

Randomized smoothing. Our robustness certificates for GNNs build upon the randomized smoothing framework (Cohen et al., 2019; Lecuyer et al., 2019): Given any base classifier f, for example a message-passing GNN, we can build a "smoothed" classifier g that classifies randomly perturbed input samples, and then takes the "majority vote" among all predictions. The goal is to construct a smoothed classifier g that behaves similar to f (for example in terms of accuracy), and for which we can prove (probabilistic) robustness certificates.

Randomized ablation (Levine and Feizi, 2020) is a smoothing-based certificate that "ablates" the input: Unlike randomized smoothing where the input in randomly perturbed (e.g. by adding Gaussian noise to images), in randomized ablation the input is randomly masked, for example by replacing parts of the input with a special ablation token that "hides" the original information. If the perturbed input is masked for the majority of predictions, we can issue certificates for the smoothed classifier g.

3 Randomized Message-Interception Smoothing for Graph Neural Networks

The main idea of our gray-box smoothing certificates is to intercept messages from perturbed nodes by (1) deleting edges to disconnect nodes, and/or (2) ablating nodes to mask their features (cf. Figure 1).

To implement this we introduce two independent smoothing distributions $\phi_1(\boldsymbol{A})$ and $\phi_2(\boldsymbol{X})$ that randomly apply those changes to the input graph: The first smoothing distribution $\phi_1(\boldsymbol{A})$ randomly deletes edges in the adjacency matrix ($1 \to 0$) with probability p_d. The second smoothing distribution $\phi_2(\boldsymbol{X})$ randomly ablates all features of nodes with probability p_a by replacing their feature representations with a fixed representation token $\boldsymbol{t} \in \mathbb{R}^d$ for ablated nodes. Introducing two independent smoothing distributions is important since our base classifiers f are GNNs and in practice, GNNs behave differently to structural changes in the graph than to feature ablation of nodes.

We apply this message-interception smoothing distribution $\phi(G) \triangleq (\phi_1(\boldsymbol{A}), \phi_2(\boldsymbol{X}))$ by randomly sampling and then classifying different graphs with a message-passing GNN f. Our smoothed classifier g finally takes the majority vote among the predictions of f for the sampled graphs $\phi(G)$. We formally describe our smoothed g classifier as follows:

$$g_v(G) \triangleq \arg\max_{y \in \mathcal{Y}} p_{v,y}(G) \qquad p_{v,y}(G) \triangleq p(f_v(\phi(G)) = y)$$

where $p_{v,y}(G)$ denotes the probability that the base GNN f classifies node v in graph G as class y under the smoothing distribution $\phi(G) = (\phi_1(\boldsymbol{A}), \phi_2(\boldsymbol{X}))$. Note that the ablation representation $\boldsymbol{t}$ is a trainable parameter of our smoothed classifier and can be optimized during training.

4 Provable Gray-box Robustness Certificates for Graph Neural Networks

We derive provable robustness certificates for the smoothed classifier g by exploiting that perturbed nodes are disconnected or ablated and cannot send messages for the majority of predictions. We therefore develop a condition that guarantees $g_v(G) = g_v(G')$ for any graph $G' \in B_\rho(G)$.

We assume the worst-case, that is adversaries alter the prediction whenever target nodes receive messages from perturbed nodes. Formally, let E denote the event that at least one perturbed message reaches a target node v. Then we can prove that $\Delta \triangleq p(E)$ quantifies how much probability mass of the distribution $p_{v,y}(G)$ over classes y is controlled by the worst-case adversary:

Proposition 1. *Given target node v in graph G, and adversarial budget ρ. Let E denote the event that the prediction $f_v(\phi(G))$ receives at least one messages from perturbed nodes. Then the change in label probability $|p_{v,y}(G) - p_{v,y}(G')|$ is bounded by the probability $\Delta = p(E)$ for all classes $y \in \{1, \dots, C\}$ and graphs G' with $G' \in B_\rho(G)$: $|p_{v,y}(G) - p_{v,y}(G')| \leq \Delta$.*

Proof sketch. Whenever we intercept all adversarial messages, adversaries cannot alter the prediction. Thus $|p_{v,y}(G) - p_{v,y}(G')|$ is bounded by Δ. □

Full proof in Appendix A. Since the probability Δ bounds the worst-case change $|p_{v,y}(G) - p_{v,y}(G')|$ for all classes y, we can utilize Δ to construct robustness certificates: If an adversary cannot shift enough probability mass to change the majority class, our smoothed classifier is robust.

Before we construct our final certificates, we first consider a special case of node ablation smoothing, discussing its relation to randomized ablation for image classifiers (Levine and Feizi, 2020), and then we derive our provably stronger guarantees for the general case of message-interception smoothing.

Proposition 2. *For node feature ablation smoothing only ($p_d = 0$), we have $\Delta = 1 - p_a^\rho$.*

Proof in Appendix B. In this special case of node feature ablation smoothing, our certificates for GNNs are theoretically related to the randomized ablation certificates for image classifiers (Levine and Feizi, 2020). We could apply their certificate to GNNs by randomly ablating features of entire nodes, instead of pixels in an image. However, their approach is specifically designed for image classifiers and comes with serious shortcomings when applied to GNNs. In contrast, our certificates are designed for *graph-structured* data, provably tighter and experimentally stronger. We thoroughly discuss such differences with more technical details in Appendix C. Most importantly, their certificate applied to GNNs ignores the underlying graph structure. Instead, our certificates are message-passing aware and consider the interception of messages via edge deletion as follows:

Message-interception. To model the interception of messages from perturbed nodes $\mathbb{B}$, we take the graph structure $\boldsymbol{A}$ into account: We consider all simple paths $P_{wv}^k = \{(e_1, \dots, e_i) \mid i \leq k\}$ from perturbed nodes $w \in \mathbb{B}$ to target node v of length at most k.[1] Intuitively, if any edge e on path $p \in P_{wv}^k$ is deleted, or the features of w are ablated, messages via path p get intercepted. If all messages from perturbed nodes get intercepted, adversaries cannot alter the prediction (Proof in Appendix A):

Lemma 1. *Given perturbed nodes $\mathbb{B}$. Then $f_v(\phi(G)) = f_v(\phi(G'))$ for any graph $G' \in B_\rho(G)$ if*

$$\forall w \in \mathbb{B} : \big(\forall p \in P_{wv}^k : \exists (i,j) \in p : \phi_1(\boldsymbol{A})_{ij} = 0\big) \vee (\phi_2(\boldsymbol{x}_w) = \boldsymbol{t})$$

Since k-layer message-passing GNNs aggregate information over local neighborhoods, only features of nodes in the *receptive field* affect the prediction of a target node (that is only nodes with a distance of at most k to v). For any perturbed node $w \in \mathbb{B}$ outside of the receptive field we have $P_{wv}^k = \emptyset$ and the message-interception condition of Lemma 1 is always fulfilled.

In practice, however, we do not know the nodes $\mathbb{B}$ controlled by the adversary. To account for this, we assume adversaries control nodes indicated by $\boldsymbol{\rho}_v \in \{0,1\}^n$ that maximize the probability of the event $E(\boldsymbol{\rho}_v)$ that target node v receives perturbed messages:

Theorem 1. *The worst case change in label probability $|p_{v,y}(G) - p_{v,y}(G')|$ is bounded by*

$$\Delta = \max_{||\boldsymbol{\rho}_v||_1 \leq \rho} p\left(E(\boldsymbol{\rho}_v)\right)$$

for all classes $y \in \{1, \dots, C\}$ and any graph $G' \in B_\rho(G)$.

Proof in Appendix A. Note that we show how to compute Δ in practice in the next section (Section 5).

Notably, Δ does not depend on the base GNN f: The probability to receive at least one message from a perturbed node is only characterized by the number of perturbed nodes ρ, and the probabilities p_d for edge deletion and p_a for node feature ablation. Moreover, Δ is monotonously increasing in ρ, since the probability to receive messages from perturb nodes increases the more nodes adversaries control.

Finally, we derive robustness certificates for node classification:

Corollary 1 (Multi-class certificate)**.** *Given Δ as defined in Proposition 1. Then we can certify the robustness $g_v(G) = g_v(G')$ for any graph $G' \in B_\rho(G)$ if*

$$p_{v,y^*}(G) - \Delta > \max_{\tilde{y} \neq y^*} p_{v,\tilde{y}}(G) + \Delta$$

where $y^* \triangleq g_v(G)$ denotes the majority class, and $\tilde{y}$ the follow-up (second best) class.

Proof in Appendix A. We also provide a certificate for binary node classification in Appendix A.

Interestingly, since Δ is monotonously increasing in ρ, there exists a maximally certifiable radius that depends on the graph structure and changes for each target node. In the special case of node feature ablation smoothing, we can directly determine the maximally certifiable radius (Proof in Appendix B):

Proposition 3. *Given fixed $p_a > 0$ and $p_d = 0$, it is impossible to certify a radius ρ if $p_a \leq \sqrt[\rho]{0.5}$.*

[1] We consider simple paths (all nodes appear only once), since we only receive perturbed messages via more complex paths if and only if we receive perturbed messages via the simple part of the complex path.

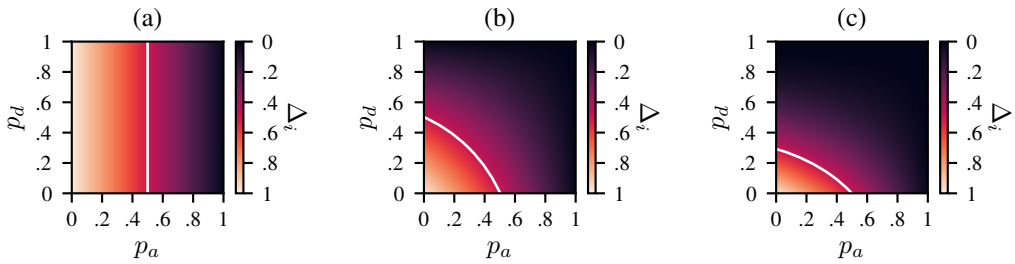

Figure 2: Single source bounding constant Δ_i for different edge deletion probabilities p_d and node feature ablation probabilities p_a. (a) For the target node, p_d does not affect Δ_i. (b) Direct neighbor of target node, single edge. (c) Second-hop neighbor, single path (two edges). White isolines indicate $\Delta_i = 0.5$ and separate the theoretically certifiable region ($\Delta_i < 0.5$) from the uncertifiable region ($\Delta_i \geq 0.5$). (a-c) More distant nodes have smaller theoretically uncertifiable regions.

5 Practical Certificates for Randomized Message-Interception Smoothing

Message-interception certificates constitute two challenges in practice: (1) computing the bounding constant Δ for arbitrary graphs, and (2) computing the label probabilities $p_{v,y^*}(G)$ and $p_{v,\tilde{y}}(G)$. We address the first problem by providing upper bounds on Δ (i.e. lower bounds on the certifiable robustness). For the second problem, we follow existing literature for probabilistic smoothing-based certificates and estimate the smoothed classifier.

Lower bound on certifiable robustness. Computing Δ of Theorem 1 poses a problem: First, finding the worst-case selection of nodes for arbitrary receptive fields involves a challenging optimization problem over graph structure. Second, computing the probability $p(E(\boldsymbol{\rho}_v))$ to receive perturbed messages is challenging even for a fixed selection of nodes $\boldsymbol{\rho}_v$, since in general, it involves evaluating the expensive inclusion-exclusion principle (see Appendix D). We can compute Δ exactly only for special cases such as small or tree-structured receptive fields (Appendix E). For arbitrary graphs, we provide practical upper bounds on Δ by assuming *independence* between paths. Due to Corollary 1, upper bounds on Δ result in lower bounds on the certifiable robustness.

We first derive an upper bound on Δ for a single perturbed node, and then generalize to multiple perturbed nodes. In the following, let E_w denote the event that the target node v receives messages from node w, and $\Delta_w \triangleq p(E_w)$. Then we can derive the following upper bound for single sources:

Theorem 2 (Single Source Multiplicative Bound)**.** *Given target node v and source node w in the receptive field of a k-layer message-passing GNN f with respect to v. Let P^k_{wv} denote all simple paths from w to v of length at most k in graph G. Then $\Delta_w \leq \overline{\Delta}_w$ for upper bound $\overline{\Delta}_w$ defined as:*

$$\overline{\Delta}_w \triangleq \left[1 - \prod_{q \in P^k_{wv}} \left(1 - (1-p_d)^{|q|}\right)\right](1-p_a)$$

where $|q|$ denotes the number of edges on the simple path $q \in P^k_{wv}$ form w to v.

Proof in Appendix F. We visualize Δ_w for different edge deletion and node ablation probabilities in Figure 2. Notably, the upper bound for single sources is tight for one- and two-layer GNNs ($\Delta = \overline{\Delta}_w$), since then all paths from a single source to the target node are independent (Appendix F). The single source multiplicative bound on Δ_w can only be used to certify a radius of $\rho = 1$. In the following, we use Theorem 2 for single sources to bound Δ of Theorem 1 for multiple nodes $\rho > 1$:

Theorem 3 (Generalized multiplicative bound)**.** *Given monotonously decreasing Δ_i for all nodes i in the receptive field so that $\Delta_i \geq \Delta_{i+1}$. Given an adversarial budget of ρ nodes. Then $\Delta \leq \overline{\Delta}_M$ for*

$$\overline{\Delta}_M \triangleq 1 - \prod_{i=1}^{\rho}(1 - \Delta_i)$$

Proof in Appendix F. Notably, the multiplicative bound is tighter than a union bound. We specifically address the approximation error in detail in Appendix G. Moreover, the multiplicative bound is tight in the special case of node feature ablation smoothing (Proposition 4 in Appendix B).

Estimating the smoothed classifier in practice. Computing the probabilities $p_{v,y^*}(G)$ and $p_{v,\tilde{y}}(G)$ exactly is challenging. We thus estimate them similar to previous work by drawing Monte-Carlo samples from ϕ (Cohen et al., 2019; Levine and Feizi, 2020; Bojchevski et al., 2020). We first identify the majority class y^* and follow-up class $\tilde{y}$ using a few samples. We then use more samples to estimate a lower bound $\underline{p_{v,y^*}(G)}$ on $p_{v,y^*}(G)$ and an upper bound $\overline{p_{v,\tilde{y}}(G)}$ on $p_{v,\tilde{y}}(G)$. We use the Clopper-Pearson Bernoulli confidence interval and apply Bonferroni correction to ensure that the bounds hold simultaneously with significance level α. The resulting bounds hold with probability of at least $1-\alpha$. Moreover, our smoothed classifier abstains if $\underline{p_{v,y^*}(G)} \leq \overline{p_{v,\tilde{y}}(G)}$. Finally, our certificates also hold when bounding Δ and the label probabilities (Proof in Appendix H):

Corollary 2. *We guarantee the robustness $g_v(G) = g_v(G')$ with probability of at least $1-\alpha$ for any graph G' with $G' \in B_\rho(G)$ if*

$$\underline{p_{v,y^*}(G)} - \overline{\Delta} > \overline{p_{v,\tilde{y}}(G)} + \overline{\Delta}$$

where y^* denotes the (estimated) majority class, and $\tilde{y}$ the follow-up class.

Limitations. Our certificates require knowledge about the graph structure $\boldsymbol{A}$ and can only account for structural changes with knowledge about the new structure $\boldsymbol{A}'$. We cannot provide certificates without $\boldsymbol{A}'$: While edge deletion potentially increases robustness (due to less messages to intercept), adversaries could arbitrarily increase the number of messages via edge insertion. Other limitations include that the number of simple paths can be huge ($\mathcal{O}(r!)$ for receptive field size r). We argue, however, that (1) graphs are usually sparse, (2) the number of paths can be reduced via sparsification, and (3) we have to compute the paths only once for each graph.

Discussion. Notably, our edge-deletion certificates ($p_d > 0$) also provide guarantees for node insertion and deletion, as disconnected nodes cannot alter the prediction.[2] As discussed above, we can only evaluate such certificates with structural information, that is how inserted/deleted nodes are connected to target nodes: Given clean graphs (as in our evaluation), we know which nodes adversaries *could delete*. Given perturbed graphs, we know which nodes adversaries *could have inserted*.

6 Experimental Evaluation

We evaluate our certificates for different GNN architectures trained on node classification datasets. Our certificates work in standard transductive learning settings used throughout the literature and we report such results in Appendix J. However, combining transductive learning with an evasion threat model comes with serious shortcomings for the evaluation of certificates, since no separate test data is available. For example, we can usually achieve high accuracy by overfitting a Multi-Layer Perceptron (MLP) to labels predicted by GNNs during training. MLPs do not propagate information through the graph at test time and are robust to adversarial messages. Instead, we evaluate our certificates in semi-supervised *inductive* learning settings with hold-out test nodes:

Experimental setup. We randomly draw 20 nodes per class for training and validation, and 10% of nodes for testing. We create separate graphs for training, validation and testing (using the remaining unlabeled nodes for training). We train classifiers on the training graph, assume adversaries control nodes in the graph at test time, and compute certificates for all test nodes. We also randomly delete edges and ablate node features during training (Appendix N). We use $n_0 = 1{,}000$ samples for estimating the majority class, $n_1 = 2{,}000$ samples for certification, and $\alpha = 0.05$. We repeat experiments for five random splits and model initializations, and report averaged results including standard deviation (shaded areas in the plots). We uploaded our implementation as supplemental material.

Datasets and models. We train our models on citation datasets: Cora-ML (Bojchevski and Günnemann, 2018; McCallum et al., 2000) with 2,810 nodes, 7,981 edges and 7 classes; Citeseer (Sen et al., 2008) with 2,110 nodes, 3,668 edges and 6 classes; and PubMed (Namata et al., 2012) with 19,717 nodes, 44,324 edges and 3 classes. We implement smoothed classifiers for four architectures with two message-passing layers: Graph convolutional networks (GCN) (Kipf and Welling, 2017), graph attention networks (GAT and GATv2) (Velickovic et al., 2018; Brody et al., 2021), and soft medoid aggregation networks (SMA) (Geisler et al., 2020). More details in Appendix I. We also compute certificates for the larger graph ogbn-arxiv (Hu et al., 2020) in Appendix K.

[2] We cannot certify node insertion/deletion with feature ablation smoothing, since e.g. new nodes affect the smoothed classifier independent of whether features are ablated or not (unless we delete nodes entirely).

Evaluation metrics. We report the classification accuracy of the smoothed classifier on the test set (*clean accuracy*), and the *certified ratio*, that is the number of test nodes whose predictions are certifiable robust for a given radius. Since all nodes have different receptive fields sizes, we also divide the certifiable radius by the receptive field size. The resulting *normalized* robustness better reflects how much percentage of the "attack surface" (that is the number of nodes the adversary could attack) can be certified. Moreover, we report the area under this (normalized) certified ratio curve (*AUCRC*). For completeness, we also report the *certified accuracy* in Appendix J, that is the number of test nodes that are correctly classified (without abstaining) *and* certifiable robust for a given radius.

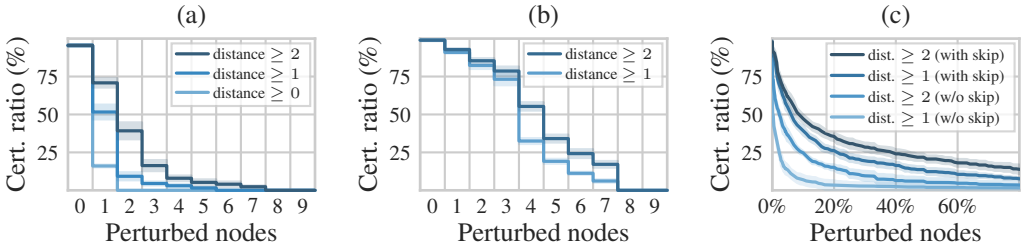

Figure 3: Smoothed GAT on Cora-ML: (a) Robustness at different distances (p_d=0.56, p_a=0.52, ACC=0.8). (b) Skip-connections allow higher ablation probabilities while retaining high accuracy (p_d=0.36, p_a=0.79, ACC=0.8). (c) Robustness normalized by receptive field size ("attack surface").

Message-interception smoothing. In Figure 3 we demonstrate our robustness certificates for specific edge deletion probabilities p_d and node feature ablation probabilities p_a. By making our certificates message-passing aware, we can (1) focus on robustness against feature perturbations of entire nodes, (2) analyze robustness locally in the receptive fields by incorporating the "attack surface", and (3) provide stronger guarantees at larger distances. As we discuss in the following, this yields considerably stronger guarantees compared to previous (smoothing) certificates for GNNs. To the best of our knowledge, we are the first to (1) propose gray-box certificates for GNNs, and (2) certify robustness against such strong adversaries that can arbitrarily manipulate features of a few nodes.

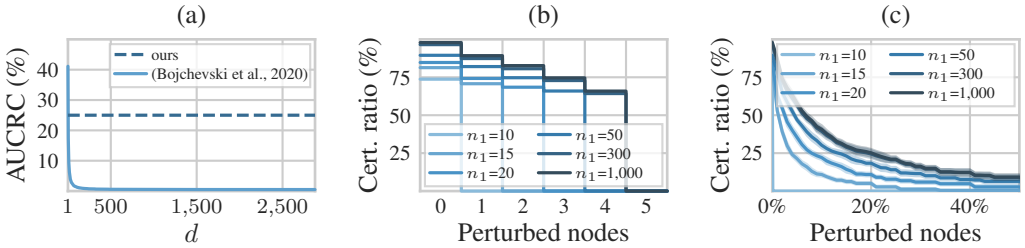

Figure 4: (a) Naïve baseline comparison (base certificate (Bojchevski et al., 2020), 10^5 samples, Cora-ML, $\alpha = 0.05$). (b,c) Our certificate with maximally certifiable radius of 4 with varying samples for certification ($p_d = 0, \alpha = 0.05$). Our certificates saturate faster than existing smoothing certificates.

First certificate for stronger adversaries. Existing certificates for GNNs only certify perturbations to a few attributes $\tilde{\rho}$ in the entire graph. Our certificates are novel as they provide guarantees for stronger adversaries that can arbitrarily manipulate features of a few nodes in the graph. To compare these two approaches, consider a naïve baseline that certifies $\rho = \tilde{\rho}/d$ nodes, where d is the number of attributes per node.[3] If each node in the graph had just a single feature, the number of certifiable nodes ρ is high. As the number of features d per node increases, however, the baseline dramatically deteriorates. In contrast, our certificates are entirely independent of the dimension d and hold regardless of how high-dimensional the underlying node data might be. We show this in Figure 4a for the first smoothing-based certificate for GNNs (Bojchevski et al., 2020), assuming attribute deletions against second-hop nodes ($p_+ = 0, p_- = 0.6$). However, the superiority of our certificate regarding robustness against all features of entire nodes holds for any other GNN certificate proposed so far.

[3]To the best of our knowledge, we are the first to certify such strong adversaries. Thus no baselines exist so far and we compare our method against existing certificates for GNNs using the naïve baseline we propose above.

Efficient message-interception smoothing. We observe that our probabilistic certificates saturate already for small amounts of Monte-Carlo samples from ϕ (see Figure 4 b,c). This is in stark contrast to existing smoothing certificates for GNNs (Bojchevski et al., 2020). We hypothesize that the reasons are twofold: First, the probability Δ to receive messages from perturbed nodes is independent of the base classifier f. Second, the number of nodes is typically lower than the number of attributes in the graph, and we evaluate redundant receptive fields during smoothing as entire subgraphs get disconnected. For a fair comparison, we adopt their transductive setting, significance level $\alpha = 0.01$, and compute certificates for $p_d = 0.3$ and $p_a = 0.85$. Bojchevski et al. (2020) use 10^6 samples for certifying test nodes on Cora-ML, which takes up to 25 minutes. In contrast, our certificates saturate already for 2,000 samples in this setting, which takes only 17 seconds (preprocessing Cora-ML takes 8 additional seconds). Our gray-box certificates are significantly more efficient.

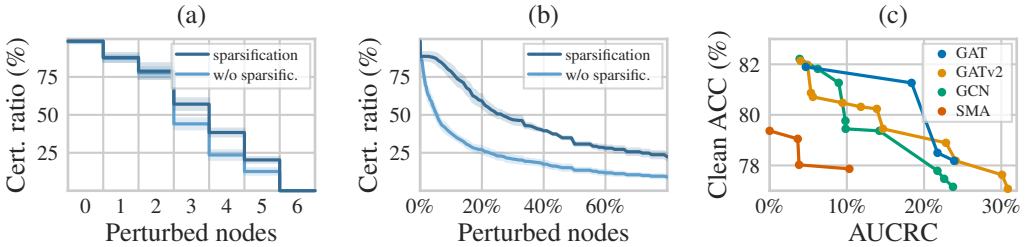

Figure 5: (a,b) Sparsification significantly improves certifiable robustness of our gray-box certificates to second-hop attacks as it reduces (a) messages to intercept, and (b) receptive field sizes and thus the "attack surface". (c) Robustness-accuracy tradeoffs for different GNN architectures on Cora-ML.

Sparsification improves certifiable robustness. Notably, our gray-box certificates incorporate graph structure and become stronger for sparser graphs. This is in contrast to all black-box certificates that ignore the underlying message-passing principles. Specifically, graph sparsification significantly improves robustness while retaining high clean accuracy: First, sparsification reduces the number of paths in the graph and thus reduces the number of messages to intercept. Second, sparsification reduces the number of nodes in the receptive fields and thus the "attack surface", i.e. the number of nodes that send messages. In Figure 5 (a,b) we apply GDC preprocessing (Gasteiger et al., 2019) to the Cora-ML graph at test time. Interestingly, evaluating the model on the sparsified graph yields significantly higher certifiable robustness, although both approaches show high clean accuracy of 79%. GDC preprocessing yields a directed graph and reduces the number of edges in the graph from 15,962 to 14,606 (setting $\epsilon = 0.022$ of GDC). We ignore edge attributes resulting from GDC preprocessing.

Different classifiers. In Figure 5c we compare robustness-accuracy tradeoffs for different GNNs against second-hop attacks. Attention-based message-passing GNNs (Velickovic et al., 2018; Brody et al., 2021) are dominating. We hypothesize that the degree-normalization of GCN (Kipf and Welling, 2017) may be problematic for the performance under randomized edge deletion. Our approach may promote novel message-passing architectures, specifically designed for smoothed classifiers.

Skip-connection. Assuming adversaries do not attack features of target nodes, we can better study robustness in the local neighborhood (Figure 3). We introduce skip-connections from the features of each node to their final representations. We use the transformation weights of the same base classifier, but more complex variations are straightforward. Notably, such skip-connections lead to significantly better robustness-accuracy tradeoffs (Appendix J). Future work could deploy existing data-dependent smoothing methods, which are *orthogonal* to our work, e.g. if nodes represent images, we could additionally deploy Gaussian smoothing (Cohen et al., 2019) for target nodes separately.

Robustness-accuracy trade-offs. We show that edge deletion *and* feature ablation smoothing achieve significantly better robustness-accuracy tradeoffs against attribute attacks to the *second-hop* neighborhood and dominate across all datasets (Figure 6). We compare three settings: (1) edge deletion and feature ablation ($p_d > 0, p_a > 0$), (2) edge deletion only ($p_d > 0, p_a = 0$), and (3) feature ablation only ($p_d = 0, p_a > 0$). We run 1,000 experiments for each setting, drawing random deletion and ablation probabilities from $[0, 1]$ for each experiment (allowing different probabilities for training and inference). We compute dominating points on the Pareto front for each setting. For brevity, we only show points with clean accuracy of at most 5% below of the maximally achieved performance.

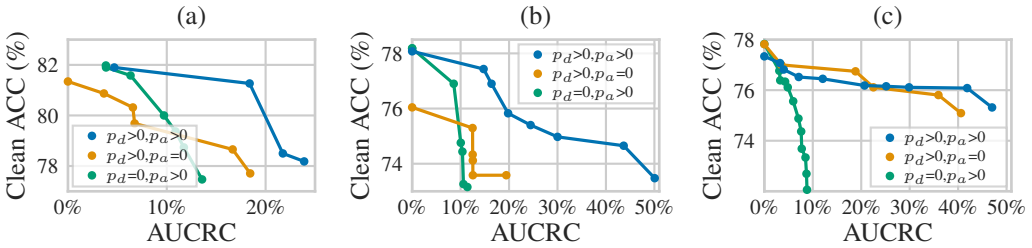

Figure 6: Robustness-accuracy tradeoffs for second-hop attacks against smoothed GAT models trained on (a) Cora-ML, (b) Citeseer, and (c) PubMed. Message-interception smoothing ($p_d > 0, p_a > 0$) dominates across all datasets. Lines connect dominating points on the Pareto front.

7 Related work

The vast majority of work on GNN robustness focuses on heuristic defenses including the detection of adversarial graphs (Zhang and Ma, 2020; Zhang et al., 2019a); architecture modifications (Brody et al., 2021; Zhang et al., 2019b); robust aggregations (Geisler et al., 2020); robust training procedures (Xu et al., 2019; Zügner and Günnemann, 2019), transfer learning (Tang et al., 2020); and graph preprocessing techniques such as edge pruning (Zhang and Zitnik, 2020; Wu et al., 2019), low-rank approximations (Entezari et al., 2020), and graph anomaly detection (Ma et al., 2021).

The effectiveness of such seemingly robust defenses on the adversarial robustness of GNNs can only be assessed against existing adversarial attacks. Heuristic defenses do not guarantee robustness, and may even be broken by stronger attacks later on (Carlini and Wagner, 2017b). Instead, we are interested in robustness certificates that *provably guarantee* the stability of predictions. However, robustness certificates for GNNs are still in their infancy (Günnemann, 2022):

Most certificates for GNNs are designed for specific architectures (Zügner and Günnemann, 2020; Jin et al., 2020; Bojchevski and Günnemann, 2019; Zügner and Günnemann, 2019). Despite providing provable robustness guarantees, their applicability is limited to specific architectures.

Bojchevski et al. (2020) present the first tight and efficient smoothing-based, model-agnostic certificate for graph-structured data. However, their method comes with crucial limitations: First, their method cannot certify robustness against arbitrary feature modifications of entire nodes. Second, their black-box certificate deletes edges but completely ignores the underlying *message-passing* principle. Third, their certificate requires an expensive evaluation of the smoothed classifier, threatening the practicability of their certificate beyond theoretical robustness assessments.

Randomized ablation certificates for image classifiers (Levine and Feizi, 2020) are another approach for discrete data. Such certificates have already been applied to point cloud classifiers (Liu et al., 2021) and even for individual attribute perturbations in GNNs (Bojchevski et al., 2020). However, Bojchevski et al. (2020) show that their method outperforms such ablation certificates for individual attributes. In contrast, we propose to certify entire nodes, instead of only a few of their attributes. As already discussed, applying their ablation certificates for image classifiers directly to GNNs comes with serious shortcomings that we overcome (Section 4 and details in Appendix C).

Recently proposed collective certificates (Schuchardt et al., 2021) are *orthogonal* to ours. Our certificates could lead to significant improvements in such collective settings, as adversaries cannot attack first-hop neighbors of all nodes simultaneously.

8 Conclusion

We propose novel gray-box, message-passing aware robustness certificates for GNNs against strong threat models where adversaries can arbitrarily manipulate all features of a few nodes. Our smoothing-based certificates randomly delete edges and/or mask features of entire nodes to intercept adversarial messages. Our certificates are significantly stronger than existing methods and could lead to novel graph preprocessing techniques that further prevent adversarial examples.

References

Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data Min. Knowl. Discov.*, 29(3):626–688, 2015.

Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *ICLR (Poster)*. OpenReview.net, 2018.

Aleksandar Bojchevski and Stephan Günnemann. Certifiable robustness to graph perturbations. In *NeurIPS*, pages 8317–8328, 2019.

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

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *CoRR*, abs/2105.14491, 2021.

Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *CoRR*, abs/1705.07263, 2017a.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57. IEEE Computer Society, 2017b.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 2019.

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

James D Esary, Frank Proschan, and David W Walkup. Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38(5):1466–1474, 1967.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. In *NeurIPS*, pages 13333–13345, 2019.

Simon Geisler, Daniel Zügner, and Stephan Günnemann. Reliable graph neural networks via robust aggregation. In *NeurIPS*, 2020.

Simon Geisler, Tobias Schmidt, Hakan Sirin, Daniel Zügner, Aleksandar Bojchevski, and Stephan Günnemann. Robustness of graph neural networks at scale. *CoRR*, abs/2110.14038, 2021.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 2017.

Larry Goldstein and Nathakhun Wiroonsri. Stein's method for positively associated random variables with applications to the ising and voter models, bond percolation, and contact process. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 54, pages 385–421. Institut Henri Poincaré, 2018.

Stephan Günnemann. Graph neural networks: Adversarial robustness. In Lingfei Wu, Peng Cui, Jian Pei, and Liang Zhao, editors, *Graph Neural Networks: Foundations, Frontiers, and Applications*, chapter 8, pages 149–176. Springer, 2022.

Stefan Heindorf, Martin Potthast, Benno Stein, and Gregor Engels. Vandalism detection in wikidata. In *CIKM*, pages 327–336. ACM, 2016.

Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. BIRDNEST: bayesian inference for ratings-fraud detection. In *SDM*, pages 495–503. SIAM, 2016.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, 2020.

Hongwei Jin, Zhan Shi, Venkata Jaya Shankar Ashish Peruri, and Xinhua Zhang. Certified robustness of graph convolution networks for graph classification under topological attacks. In *NeurIPS*, 2020.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*. OpenReview.net, 2017.

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

Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4585–4593, 2020.

Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Pointguard: Provably robust 3d point cloud classification. In *CVPR*, pages 6186–6195. Computer Vision Foundation / IEEE, 2021.

Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. Towards more practical adversarial attacks on graph neural networks. In *NeurIPS*, 2020.

Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Quan Z. Sheng, and Hui Xiong. A comprehensive survey on graph anomaly detection with deep learning. *CoRR*, abs/2106.07178, 2021.

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.

Galileo Namata, Ben London, Lise Getoor, Bert Huang, and UMD EDU. Query-driven active surveying for collective classification. In *10th International Workshop on Mining and Learning with Graphs*, volume 8, page 1, 2012.

Jan Schuchardt, Aleksandar Bojchevski, Johannes Klicpera, and Stephan Günnemann. Collective robustness certificates: Exploiting interdependence in graph neural networks. In *ICLR*. OpenReview.net, 2021.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

Xianfeng Tang, Yandong Li, Yiwei Sun, Huaxiu Yao, Prasenjit Mitra, and Suhang Wang. Transferring robustness for graph neural network against poisoning attacks. In *WSDM*, pages 600–608. ACM, 2020.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR (Poster)*. OpenReview.net, 2018.

Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. Adversarial examples for graph data: Deep insights into attack and defense. In *IJCAI*, pages 4816–4823. International Joint Conferences on Artificial Intelligence Organization, 2019.

Kaidi Xu, Hongge Chen, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Mingyi Hong, and Xue Lin. Topology attack and defense for graph neural networks: An optimization perspective. In *IJCAI*, pages 3961–3967. ijcai.org, 2019.

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

Xiang Zhang and Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. In *NeurIPS*, 2020.

Yingxue Zhang, S Khan, and Mark Coates. Comparing and detecting adversarial attacks for graph deep learning. In *Proc. Representation Learning on Graphs and Manifolds Workshop, Int. Conf. Learning Representations, New Orleans, LA, USA*, 2019a.

Yingxue Zhang, Soumyasundar Pal, Mark Coates, and Deniz Üstebay. Bayesian graph convolutional neural networks for semi-supervised classification. In *AAAI*, pages 5829–5836. AAAI Press, 2019b.

Xu Zou, Qinkai Zheng, Yuxiao Dong, Xinyu Guan, Evgeny Kharlamov, Jialiang Lu, and Jie Tang. TDGIA: effective injection attacks on graph neural networks. In *KDD*, pages 2461–2471. ACM, 2021.

Daniel Zügner and Stephan Günnemann. Certifiable robustness and robust training for graph convolutional networks. In *KDD*, pages 246–256. ACM, 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*, pages 1656–1665, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] To the best of our knowledge, all claims in abstract and introduction reflect the contributions and scope of our paper. We also provide a list of our core contributions directly in our introduction.
   (b) Did you describe the limitations of your work? [Yes] We discuss the limitations of our approach in a highlighted paragraph in section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Without doubt, adversarial attacks can have negative impacts on the society. Particularly alarming are recent attacks against GNNs for more realistic threat models (Ma et al., 2020) and attacks that scale to large graphs (Geisler et al., 2021). Robustness certificates are tools to assess robustness and help to (1) better understand robustness, (2) build more robust classifiers, and (3) eventually prevent adversarial attacks including their negative consequences. Our certificates represent a contribution towards diminishing and preventing potential negative impacts of adversarial attacks on the society.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We carefully reviewed the ethics guidelines (`https://neurips.cc/public/EthicsGuidelines`). To the best of our knowledge, our paper conforms to all ethic guidelines. Our certificates represent a contribution to prevent negative social impacts of adversarial attacks, as discussed above.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] To the best of our knowledge, we state all assumptions of our theoretical results.
   (b) Did you include complete proofs of all theoretical results? [Yes] We show all statements, including additional elaborations, in the Appendix. We always link theoretical results in the paper to the corresponding proofs in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We uploaded the code required to reproduce our main results as supplemental material. All required datasets are publicly available, and can be loaded for example with PyTorch Geometric (Fey and Lenssen, 2019).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We describe important training details directly at the beginning of our experiment section (section 6). We further thoroughly list all training details in Appendix I.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We repeat each experiment for five random splits and model initializations, only report averaged results, and our plots explicitly show

the standard deviation over the five experiments (shaded areas in plots). In separate experiments, we tested our method for another set of five randomly drawn seeds, but we did not observe significant differences in the results. We also uploaded all seeds as supplemental material to ensure reproducibility.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We discuss the runtime for each experiment in our experiment section (section 6). We conduct all experiments in an internal cluster with the following GPU type: NVIDIA A100-PCIE-40GB.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] Yes, to the best of our knowledge we cite all authors of the assets we use.

   (b) Did you mention the license of the assets? [N/A] To the best of our knowledge, our datasets are well-established research datasets with MIT License or public domain.

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We uploaded our new assets as supplemental material, specifically the code to reproduce our main results.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] To the best of our knowledge, our assets do not require consent.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] To the best of our knowledge, our assets do not contain any personal, protected or offensive data.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We do not conduct research with human subjects.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We do not conduct research with human subjects.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We do not conduct research with human subjects.