On provable privacy vulnerabilities of graph representations

Ruofan Wu [*1] **Guanhua Fang** [*2] **Qiying Pan** [3] **Mingyang Zhang** [1] **Tengfei Liu** [1] **Weiqiang Wang** [1] **Wenbiao Zhao** [1]

Abstract

Graph representation learning (GRL) is critical for extracting insights from complex network structures, but it also raises security concerns due to potential privacy vulnerabilities in these representations. This paper investigates the structural vulnerabilities in graph neural models where sensitive topological information can be inferred through edge reconstruction attacks. Our research primarily addresses the theoretical underpinnings of cosine-similarity-based edge reconstruction attacks (COSERA), providing theoretical and empirical evidence that such attacks can perfectly reconstruct sparse Erdős–Rényi graphs with independent random features as graph size increases. Conversely, we establish that sparsity is a critical factor for COSERA's effectiveness, as demonstrated through analysis and experiments on stochastic block models. Finally, we explore the resilience of (provably) private graph representations produced via noisy aggregation (NAG) mechanism against COSERA. We empirically delineate instances wherein COSERA demonstrates both efficacy and deficiency in its capacity to function as an instrument for elucidating the trade-off between privacy and utility. [1]

1. Introduction

With the surging developments of graph representation learning (GRL) (Hamilton et al., 2017b), there has been growing apprehensions concerning the security challenges associated with the deployment of graph neural models in real-world scenarios (Dai et al., 2022). GRL models harness the topological information of the underlying graph for producing high-quality predictions or graph representations. Meanwhile, these models bear the risk of inadvertently divulging the same topological information through the graph representations they produce. Such kind of security risks have been empirically validated through the examination of the attacking performance of edge reconstruction algorithms (Duddu et al., 2020; He et al., 2021; Wu et al., 2022a; Zhou et al., 2023), among which the simplest form of attack based solely on the representation similarity of node pairs is shown to achieve strikingly strong performance, without the requirement of additional knowledge like encoder architecture or auxiliary datasets (He et al., 2021).

Despite the empirical evidence of topological vulnerabilities of graph representations, theoretical explanations delineating the effectiveness of such attacks remain largely unexplored: As demonstrated in previous studies (Duddu et al., 2020; He et al., 2021), similarity-based attacks are remarkably effective against *sparse* graphs that exhibit a generalized homophily pattern, i.e., there exists a significant correlation between the similarity of node features and edge adjacency information. This phenomenon posits that *feature similarity* may serve as a confounding factor, potentially impacting the efficacy of similarity-based attacks. It is therefore valuable to understand the influence of graph properties, such as feature similarity and sparsity, on the edge reconstruction process of the attacking procedures.

Beyond their capability in characterizing the vulnerabilities of representations, attacking algorithms may also function as empirical attestations of privacy-preserving inference protocols that fulfill formal privacy guarantees such as differential privacy (Cummings et al., 2023, Section 4). As an illustrative case, membership inference attacks can be employed for auditing differential privacy (Tramer et al., 2022). Since edge reconstruction is equivalent to edge membership inference on graphs (Zhang et al., 2023), it is thus pertinent to explore the performance of similarity-based attacks when confronted with privacy-preserving graph representations (Sajadmanesh et al., 2023; Wu et al., 2023).

In this paper, we take initial steps toward a principled understanding of structural vulnerabilities of graph representations under the **cos**ine-similarity-based **e**dge **r**econstruction **a**ttack (hereafter abbreviated as COSERA) which is arguably the most operable attack in many practical scenarios. In particular, we establish the following theoretical as well as empirical findings:

[*]Equal contribution [1]Ant Group [2]Fudan University [3]Shanghai Jiaotong University. Correspondence to: Ruofan Wu <ruofan.wrf@antgroup.com>.

[1]Code available at https://github.com/Rorschach1989/cosera

(i) **Success modes of COSERA** Through applying COSERA to sparse Erdős–Rényi graphs equipped with independent random node features, we show that COSERA is able to achieve perfect reconstruction as the graph size goes to infinity. The result indicates that feature similarity is not necessary for COSERA to succeed. Synthetic experiments are conducted to empirically validate our theory.

(ii) **Failure modes of COSERA** We show, through theoretical analysis and corroborative synthetic experiments, performance lower bounds when applying COSERA to stochastic block models (SBM) with independent random node features: When the underlying SBM has $\Theta(1)$ intra-group connection probability, edge recovery through graph representations becomes provably hard.

(iii) **COSERA as a privacy auditing tool** We evaluate COSERA against graph representation generated via noisy aggregation (NAG). We show theoretical protection guarantees of NAG and empirically identify instances that manifest the competencies and limitations of COSERA as a mechanism for eliciting the privacy-utility trade-off. Notably, we find cases when NAG offers substantial protection against COSERA even when its privacy guarantee is vacuous.

2. Related works

In this section, we present a concise survey of related literature. An augmented version with more discussions is postponed to appendix A.

2.1. Edge reconstruction attacks on graphs

Recent literature instantiates a variety of edge reconstruction attacks with differing adversary models in terms of capabilities and knowledge. Initially proposed by (Duddu et al., 2020), COSERA demonstrated high success with limited information access. Later studies introduced more potent adversaries with additional knowledge, such as node features and structural data(Zhang et al., 2021; He et al., 2021; Zhou et al., 2023), or even the ability to maliciously manipulate the graph structure (Wu et al., 2022a; Meng et al., 2023). On the theory side, Chanpuriya et al. (2021) proposed an algorithm that provably recovers graph structure based on representations generated via DeepWalk. Zhang et al. (2023) showed that when block structure exists in the underlying graph, the performance of COSERA is uneven across nodes in different blocks. Zhou et al. (2023) use information-theoretic arguments to construct more powerful attacks than COSERA. Nevertheless, the aforementioned studies did not provide a theoretical rationale for the practical vulnerabilities manifested as a result of the COSERA.

2.2. Protection against edge reconstruction attacks

Edge differential privacy (EDP)(Nissim et al., 2007) is a prominent privacy model safeguarding against edge reconstruction attacks. Algorithms like DP-SGD (Abadi et al., 2016) can provide private GNN models that protect individual training sample membership. However, these methods don't ensure privacy during inference(Chien et al., 2023). Current inference-time privacy solutions include edge-wise randomized response (Wu et al., 2022a) and noisy aggregation (NAG) mechanisms (Sajadmanesh et al., 2023; Wu et al., 2023; Chien et al., 2023). Approaches based on the information bottleneck principle, such as regularization or optimization techniques (Wang et al., 2021; Zhou et al., 2023), have been proposed to limit privacy leakage, but they typically rely on crude approximations of mutual information and lack a solid theoretical foundation.

3. Preliminaries

Setup and notations Consider an undirected graph $G = (V, E)$ with $n = |V|$ nodes associated with node features $X \in \mathbb{R}^{n\times d}$. Denote A as the corresponding adjacency matrix and D as the diagonal matrix with the v-th diagonal entry being the degree of node v. In this paper, we will study *victim models* taking a simple form of graph neural encoder, the *linear graph neural network* (Wu et al., 2019) which has been widely adopted in previous theoretical studies on graph neural networks (Awasthi et al., 2021; Xu et al., 2021; Wu et al., 2022b). Specifically, the node representation matrix of an L-layer linear GNN is computed as:

$$H^{(L)} = \left((D+I)^{-1}(A+I)\right)^{L} XW, \tag{1}$$

where the identity matrix is added for ensuring self-loops, and $W \in \mathbb{R}^{d\times d}$ is a learnable weight matrix. Throughout this paper, we will assume the node feature dimension and the hidden dimension to be equal to d and refer to this as the feature dimension, as otherwise we may add an extra input projection step to fulfill this requisite. We further denote $\|W\|_{\text{op}}$ and $\kappa(W)$ as the operator norm (i.e., largest singular value) and condition number (i.e., the ratio of largest and smallest singular value) of matrix W.

Threat model We assume the adversary knows the node set V and is able to inquire node representations of an arbitrary node subset $V_{\text{victim}} \subset V$. Hereafter we will refer to the subgraph induced via V_{victim} as the *victim subgraph* $G_{\text{victim}} = (V_{\text{victim}}, E_{\text{victim}})$. The goal of the adversary is to recover an arbitrary fraction of E_{victim} based on the acquired node representations $H^{(L)}_{\text{victim}} = \{h_v^{(L)}, v \in V_{\text{victim}}\}$. The adversary manifests practically in scenarios such as the deployment of graph neural models (Wu et al., 2022a) and vertical federated learning (Zhou et al., 2020), wherein clients reciprocally exchange node embeddings to facili-

tate collaborative modeling. In our analysis, it is usually convenient to let the victim node set be the entire node set V, and we occasionally drop the subscription without misunderstandings.

Remark 3.1. The proposed adversary is weak in the sense that only black-box access to model outputs is available without any other prior information such as model architecture and auxiliary datasets, which is conventionally deemed indispensable in attack paradigms against privacy (Nasr et al., 2021). Furthermore, the adversary's objective is notably ambitious, as it encompasses the potential targeting of the complete set of edges within the victim subgraph. This stands in contrast to stronger adversary models, such as the one employed in membership inference attacks (Nissim et al., 2007), where the adversary possesses the capacity to compromise all aspects except for ascertaining the presence of a particular victim edge.

The COSERA is conducted in an embarrassingly simple manner, with the adjacence relation between node u and node v inferred as:

$$\widehat{A}_{uv}^{\mathsf{COSERA}}(\tau) = \begin{cases} 1 & \text{if } \frac{\langle h_u^{(L)}, h_v^{(L)} \rangle}{\left\| h_u^{(L)} \right\|_2 \left\| h_v^{(L)} \right\|_2} \geq \tau, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where the cutoff threshold τ is allowed to depend on the embedding set H_{victim} but is uniform across all edge decisions. Hereafter without misunderstandings, we will drop the superscript in (2) and denote $\widehat{A}(\tau)$ as the reconstructed adjacency matrix under threshold τ. To measure the performance of the attack, we use false positive rate (FPR) and false negative rate (FNR) defined as

$$\begin{aligned} \mathsf{FPR}\left(\widehat{A}, A; \tau\right) &= \frac{\sum_{u,v} \mathbf{1}\left(\widehat{A}_{uv}(\tau) = 1, A_{uv} = 0\right)}{\sum_{u,v} \mathbf{1}\left(A_{uv} = 0\right)}, \\ \mathsf{FNR}\left(\widehat{A}, A; \tau\right) &= \frac{\sum_{u,v} \mathbf{1}\left(\widehat{A}_{uv}(\tau) = 0, A_{uv} = 1\right)}{\sum_{u,v} \mathbf{1}\left(A_{uv} = 1\right)}. \end{aligned} \tag{3}$$

We further define the error rate ERR as the summation of FPR and FNR. Employing these metrics facilitates a more nuanced characterization of attack performance, particularly when the underlying graph is sparse.

Intuitively, the success of COSERA is determined by the correlation between node representation similarity and edge presence. Previous empirical observations demonstrate the effectiveness of COSERA against graphs that exhibit strong correlations between node feature similarity and edge presence (He et al., 2021). We will refer to such kinds of graphs as being homophilous in a generalized sense (Jin et al., 2022; Luan et al., 2023). Due to the message-passing nature of GNN encoders, it is intuitively reasonable that recursive aggregation of node representations strengthens the correlation and results in successful edge reconstructions. However, it is non-trivial whether COSERA mechanism may succeed in the absence of the aforementioned generalized homophily pattern, which motivates our first analysis.

4. COSERA against sparse Erdős–Rényi graphs

In this section, we study the behavior of COSERA with the underlying (victim) graph generated following the Erdős–Rényi structure $G \sim \mathcal{G}_{\text{er}}(n, p)$. Here, the adjacency matrix is generated such that each entry is independently distributed (up to symmetric constraints $A_{uv} = A_{vu}$) following a Bernoulli distribution $\text{Ber}(p)$. We focus on the sparse regime when $p = O\left(\frac{\log n}{n}\right)$. We further assume that the node features X_v's are generated i.i.d. according to an isotropic Gaussian distribution $X_v \sim N(0, I_d)$. It follows that the correlation of node feature similarity and edge presence is zero. The following theorem characterizes the effectiveness of COSERA under the Erdős–Rényi setup.

Theorem 4.1. *Let C_1 be a universal constant and $G \sim \mathcal{G}_{er}(n, p)$. Assume the following:*

(i) The graph generation mechanism satisfies $p \leq C_1 \frac{\log n}{n}$.

(ii) The depth of GNN encoder L and the feature dimension d satisfies $d \gg (C_2 \log n)^{6L+2} \log n$ with $C_2 = 1.5C_1$.

(iii) The condition number of the GNN encoder weight satisfies

$$(\kappa(W))^2 \leq \sqrt{\frac{d}{\log n}} / 8(C_2 \log n)^{3L}. \tag{4}$$

Then there exists a threshold $\tau = \Theta\left(\frac{1}{(C_2 \log n)^{2L}}\right)$ such that with probability at least $1 - \frac{2}{n^2}$, the following holds:

$$\mathit{FNR}\left(\widehat{A}, A; \tau\right) = 0, \ \mathit{FPR}\left(\widehat{A}, A; \tau\right) \leq \frac{(C_2 \log n)^{2L}}{n}. \tag{5}$$

Theorem 4.1 implies that, even when COSERA can not borrow strength from the homophily nature of the underlying graph, it is able to produce accurate reconstructions when the graph size is sufficiently large. An additional intriguing implication from theorem 4.1 pertains to the dependence of reconstruction performance on the GNN encoder depth L: Provided that the node feature dimension is sufficiently large, the reconstruction performance degrades when the depth of the encoder increases, which is related to the renowned phenomenon of oversmoothing in GNN literature (Wu et al., 2022b). Intuitively, as the depth of GNN encoders increases, the resulting node representations tend to converge (Oono & Suzuki, 2019), becoming less distinct from one another. This convergence diminishes

the discriminative capacity of similarity metrics, thereby impairing their sensitivity.

Remark 4.2 (Practicality). Theorem 4.1 requires the node feature dimension d to grow in a polylog(n) rate, a condition which may not consistently align with practical scenarios. At present, this requirement is a byproduct of our proof strategy. In section 7.1 we will further examine the implications of feature dimensionality. The existence of a threshold that theorem 4.1 manifests might not guide the choice of threshold in practice. Instead, we may rely on heuristics or side-information (He et al., 2021) to determine the threshold.

Remark 4.3 (Extensions). The consequence of theorem 4.1 extends to setups with looser generative requirements under minimal modifications to the proof. In particular, the isotropic Gaussian distribution assumption of node features can be relaxed to other distribution families like sub-Gaussian type distributions with a weak dependence among distinct coordinates. Moreover, our result also holds even when the edge probability p between nodes u and v depends on node features (i.e., $p = p_{uv}$ is a function of X_u and X_v), as long as $p = O(\frac{\log n}{n})$.

5. COSERA against dense SBMs

In this section, we reveal the limitation of COSERA by constructing a reconstruction problem that is provably hard. We consider the following stochastic block model (Abbe, 2018), where each node is assigned a community membership from one of K groups $k(v) \in [K]$. The (u, v)-th entry of the adjacency matrix is generated as

$$A_{uv} \sim \begin{cases} \text{Ber}(p), & \text{if } k(u) = k(v) \\ \text{Ber}(q), & \text{otherwise} \end{cases} . \tag{6}$$

For ease of presentation, we further assume that the groups share the same size, i.e., n is a multiple of K. Denote the generation mechanism as $G \sim \mathcal{G}_{\text{sbm}}(n, K, p, q)$. We have the following result:

Theorem 5.1. *Let $G \sim \mathcal{G}_{sbm}(n, K, p, q)$ and $p = \Theta(1)$. Assume the GNN encoder to be of depth L and feature dimension $d \gg \max\{\log n/p^2, K^2 \log^3 n\}$ with the weight matrix being the identity matrix. Then with probability at least $1 - 1/n^2$, for any fixed $\tau \in [0, 1]$, one of the following three statements must hold:*

(i) FPR $\left(\widehat{A}, A; \tau\right) \geq \frac{1-p}{2K}$ and FNR $\left(\widehat{A}, A; \tau\right) \geq \frac{q}{2}$.

(ii) FPR $\left(\widehat{A}, A; \tau\right) \geq \frac{1-p}{2K} + \frac{1-q}{2}$.

(iii) FNR $\left(\widehat{A}, A; \tau\right) \geq \frac{p}{2K} + \frac{q}{2}$.

According to theorem 5.1, given any cutoff threshold if the within-group connection probability is of the order $\Theta(1)$ and the number of groups K does not diverge (Otherwise, we will return to the sparse regime in section 4) , the performance of COSERA measured by error rate ERR is lower bounded by non-vanishing constants when the feature dimension is sufficiently large. The theorem characterizes the inherent limitations of COSERA when the underlying graph is dense. As K gets large, the lower bound of false positive/negative rate decreases. It indicates that COSERA is more successful when the graph is less connected.

Remark 5.2. Alternatively, we may interpret theorem 5.1 as manifesting instances where COSERA is constrained to discerning only population-level relational information—such as the affiliation of two nodes to a common group—rather than identifying the existence of specific edges when the underlying graph is dense and admits certain group-wise structures.

6. Provable defense by noisy aggregation

Having demonstrated the susceptibility of GNN representations to COSERA, it becomes an intriguing research question to examine the behavior of COSERA within the context of privacy-preserving GRL: In this paper, we explore the defensive efficacy of noisy aggregation (NAG), which has been proposed recently as a provably privacy-preserving algorithm (Sajadmanesh et al., 2023; Wu et al., 2023) under the edge differential privacy model (Nissim et al., 2007). Concretely, we study an L-layer noisy GNN with the l-th layer computed recursively as:

$$h_v^{(l)} = T\left(\text{AGG}\left(\frac{W_l h_u^{(l-1)}}{\left\|h_u^{(l-1)}\right\|_2}, u \in \overline{N(v)}\right) + \epsilon\right), \tag{7}$$

where $\overline{N(v)} := N(v) \cup \{v\}$ denotes node v's extended neighborhood and $h_v^{(l)}$ denotes the representation of node v at the l-th layer. The aggregation mechanism AGG is a permutation invariant function that defines the message-passing process and T is some (possibly) non-linear transform. The NAG framework is interpreted as a noisy aggregation of l_2-normalized node representations, with the additive perturbation generated from a zero-mean isotropic Gaussian distribution with scale σ, i.e., $\epsilon \sim N(0, \sigma^2 I_d)$. Let $H^{(l)}$ denote the node representation matrix corresponding to the output the l-th GNN layer, and let $\mathbf{H} = \{H^{(l)}\}_{0 \leq l \leq L}$ denote all the intermediate representations produced by the underlying GNN with weights $\mathbf{W} = \{W_l\}_{l \in [L]}$. The following theorem characterizes the defensive capability of NAG under several standard choices of AGG:

Theorem 6.1. *For any adversary $\mathcal{A}$ that has access to the output $\mathbf{H}$ of an L-layer GNN under NAG with weights $\mathbf{W}$ and produces an estimate of the adjacency matrix of the underlying graph $\widehat{A} = \mathcal{A}(\mathbf{H}, \mathbf{W})$, we have the following*

lower bound:

$$\inf_{\widehat{A}} \min_{u\in V, v\in V} \left[\mathbb{P}\left(\widehat{A}_{uv}=1|A_{uv}=0\right)+\mathbb{P}\left(\widehat{A}_{uv}=0|A_{uv}=1\right)\right] \geq 1-\sqrt{1-\exp\left(-C\frac{\sum_{l\in[L]}\|W_l\|_{op}^2}{\sigma^2}\right)}. \tag{8}$$

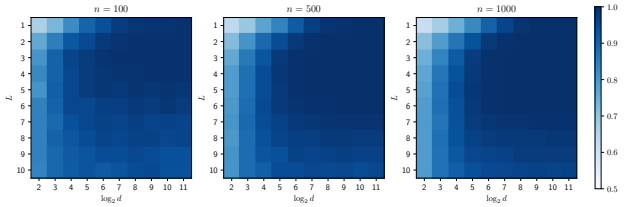
Here the constant C depends on the AGG mechanism of the GNN. In particular, if AGG is summation pooling (Xu et al., 2018), mean pooling (Hamilton et al., 2017a) or GCN pooling (Kipf & Welling, 2016) then $C = 1$; If AGG is max pooling (Hamilton et al., 2017a) or attentive pooling (Veličković et al., 2018) then $C = 4$.

The adversary in theorem 6.1 is much stronger than the COSERA adversary which only has access to $H^{(L)}$. In fact, the adversary in theorem 6.1 might be further strengthened to having the prior knowledge of the precise (probabilistic) generative mechanism of the target graph and the result still holds. Theorem 6.1 indicates that for *any* node pairs, the summation of type-I error and type-II error (in the language of binary hypothesis testing (Lehmann et al., 1986)) incurred by *any* such adversary is lower bounded by a constant, which will be significantly above zero when the noise scale is of the same order to the operator norms of the weight matrices of the GNN encoder. However, implementing NAG with a large noise scale essentially destroys model utility. Moreover, according to recent empirical observations (Carlini et al., 2019), in practice even when the formal privacy guarantee is vacuous, i.e., the lower bound is close to zero, we may still get decent protection against practical adversaries with limited knowledge. It is therefore of interest to examine how NAG protects the edge privacy against COSERA in practice. This investigation could provide additional empirical evidence regarding the potential of COSERA as an instrument for auditing private GRL algorithms such as NAG. Specifically, we consider two noisy training schemes for obtaining the model weights $\mathbf{W}$:

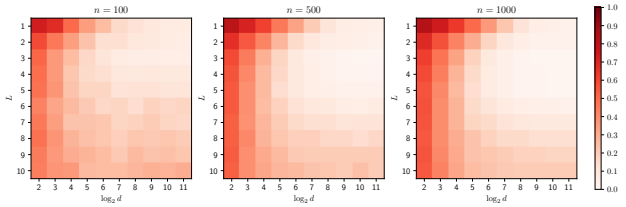
Unconstrained scheme We choose a fixed noise scale σ during both training and inference without explicit control over the weights $\mathbf{W}$. The resulting model might not produce meaningful privacy guarantees in the sense of theorem 6.1 as the operator norms of weights $\mathbf{W}$ are determined by the training dynamics.

Constrained scheme We choose a fixed noise scale σ during both training and inference and use normalization techniques (Miyato et al., 2018) to provide a priori control of model weights $\mathbf{W}$, thereby providing tighter control of formal privacy level according to theorem 6.1.

We will empirically inspect the protection of NAG representations trained via both unconstrained and constrained schemes against COSERA in section 7.2.



(a) Measured in AUROC metric, darker color implies higher attacking performance



(b) Measured in ERR metric, lighter color implies higher attacking performance

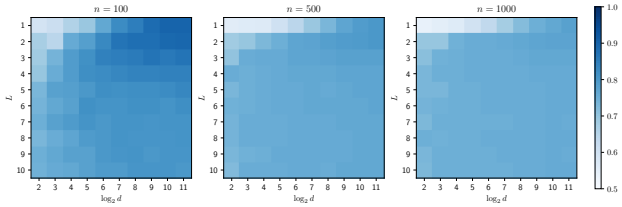
Figure 1: Attacking efficacy of COSERA over sparse Erdős–Rényi graphs, with each grid's value indicating COSERA's performance measured in either AUROC (first row) or ERR (second row) metric.

7. Experiments

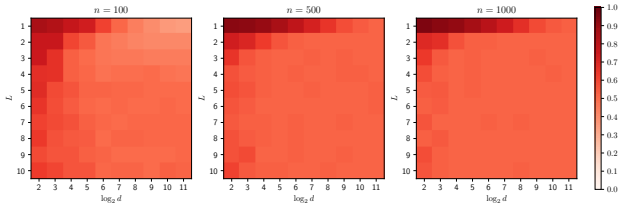
7.1. Synthetic experiments

In this section, we conduct experiments using synthetic datasets to empirically verify the theoretical developments in section 4 and section 5.

Erdős–Rényi experiments In our first experiment, we test COSERA on graph representations produced by (1) over Erdős–Rényi graphs with edge probability $p = \frac{\log n}{n}$ with graph size $n \in \{100, 500, 1000\}$. We set the weight to be the identity matrix and further present results under random weights in appendix C.1. We vary the feature dimension $d \in \{2^j, 2 \leq j \leq 11\}$ and network depth $1 \leq L \leq 10$ in order to obtain a fine-grained assessment of COSERA. We measure the performance of the attack under two metrics: AUROC and the ERR as defined in section 3, minimized over the choice of threshold τ. We present the evaluations in figure 1. The results corroborate with our theoretical developments: We demonstrate that COSERA is able to achieve near-perfect reconstruction of all edges *only* in the "large d, small L" regime. Notably, we find COSERA to be less successful under relatively deep network architectures (i.e., $L \geq 5$) when the feature dimension is sufficiently large. Yet the behaviors in small d regimes appear to be less predictable, a phenomenon we hypothesize may be attributable to an inadequate concentration of inner products in instances where the feature dimension is relatively small. Furthermore, the influence of the feature dimension appears to be more pronounced than that of the network depth. This



(a) Measured in AUROC metric, darker color implies higher attacking performance



(b) Measured in ERR metric, lighter color implies higher attacking performance

Figure 2: Attacking efficacy of COSERA over dense SBM graphs, with each grid's value indicating COSERA's performance measured in either AUROC (first row) or ERR (second row) metric.

suggests that a greater number of features, despite their independence from graph topology, lead to potentially more privacy risks as transmitted through GNN representations. Conversely, augmenting the network depth does not necessarily correlate with an elevation in the success rate of COSERA.

SBM experiments In our second experiment, we test COSERA graph representations over SBM graphs with $K = 3, p = 0.3, q = 0.05$, with the rest of the experimental setups analogous to that in the Erdős–Rényi experiments. The evaluations are presented in figure 2. The results reveal the presence of a pronounced barrier that hinders the success of the attack across a wide range of configurations corresponding to different network depths and feature dimensions. Furthermore, we observe that the results tend to stabilize as the size of the graph increases.

To investigate the impact of SBM structure on the performance of COSERA, we fix the GNN architecture at $L = 1$ and evaluate on a graph with 100 nodes and node feature dimension $d = 2048$. Note that we choose a relatively large node feature dimension to ensure that the assumption listed in theorem 5.1 is approximately met. We vary the SBM within-group probability according to $p \in \{0.1, 0.3, 0.5, 0.7\}$ and the number of groups according to $1 \leq K \leq 20$. The results, plotted in figure 3, suggest that in general, the attacking performance is positively correlated with the number of groups K since more groups yield stronger sparsity according to the SBM generation law. This phenomenon is also in accordance with theorem 5.1.

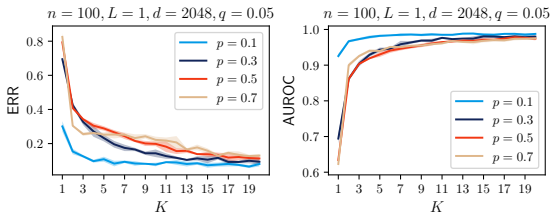


Figure 3: Performance of COSERA on SBM with varying K and p. All plots are based on 5 independent trials with shades indicating one standard deviation.

7.2. Empirical protection of NAG against COSERA

In this section, we empirically study the defensive performance of noisy aggregation (7) against COSERA under the 5 analyzed AGG mechanisms listed in theorem 6.1. We will use the Planetoid datasets (Yang et al., 2016) for evaluation. Due to space limits, we report results on the Cora and Citeseer datasets in the main text and postpone the complete report in appendix C.2.

Experimental setup We consider a transductive node classification setting and use the standard train-test splits. The GNN models are trained using the training labels and evaluated on the test nodes. The performances of COSERA are evaluated on the subgraphs induced by the test nodes. We report the configuration of GNN encoding, as well as the attacking pipeline and training hyperparameters in appendix C.2.1. We use the following two types of training configurations as proposed in section 6:

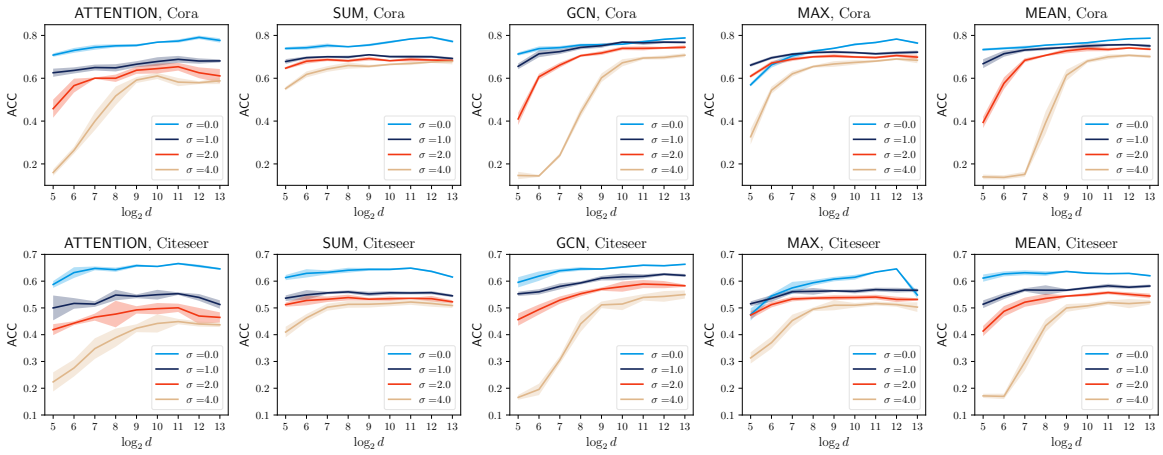
Unconstrained scheme Under the unconstrained scheme, we use aggressive perturbation plans by applying noise with scale range $\sigma \in \{0, 1, 2, 4\}$, with $\sigma = 0$ indicating no protection, and $d \in \{2^i, 5 \leq i \leq 13\}$.

Constrained scheme Under the constrained scheme, we adopt the spectral normalization technique (Miyato et al., 2018) to control the spectral norm of each layer at approximately 1 (with relative error $< 10\%$). We use conservative perturbation plans by applying noise with scale range $\sigma \in \{0, 0.01, 0.05, 0.1, 0.5, 1\}$, and $d \in \{2^i, 5 \leq i \leq 13\}$. Note that with $\sigma = 1$, we obtain a non-vacuous lower bound according to (8).

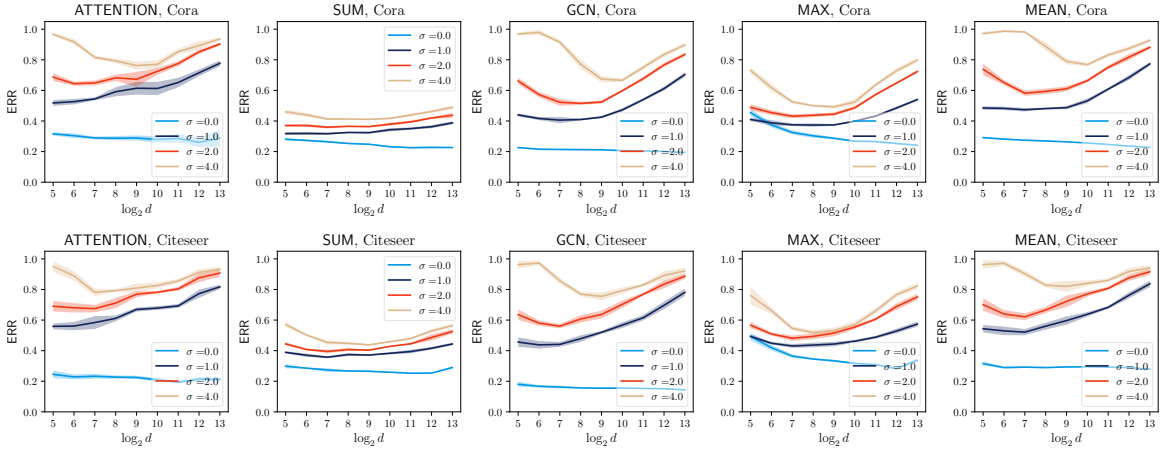
The model utilities are measured using classification accuracy and attack performances are measured by ERR. We will additionally investigate the attack performance under the AUROC metric in appendix C.2.

Results and observations The experimental evaluations are presented in figure 4 for the unconstrained scheme and in figure 5 for the constrained scheme. We summarize our observations and findings as follows:

Without protections, COSERA is more effective for



(a) GNN model performance over Cora and Citeseer datasets under 5 different aggregation types.



(b) Attacking performance of COSERA over Cora and Citeseer datasets under 5 different aggregation types.

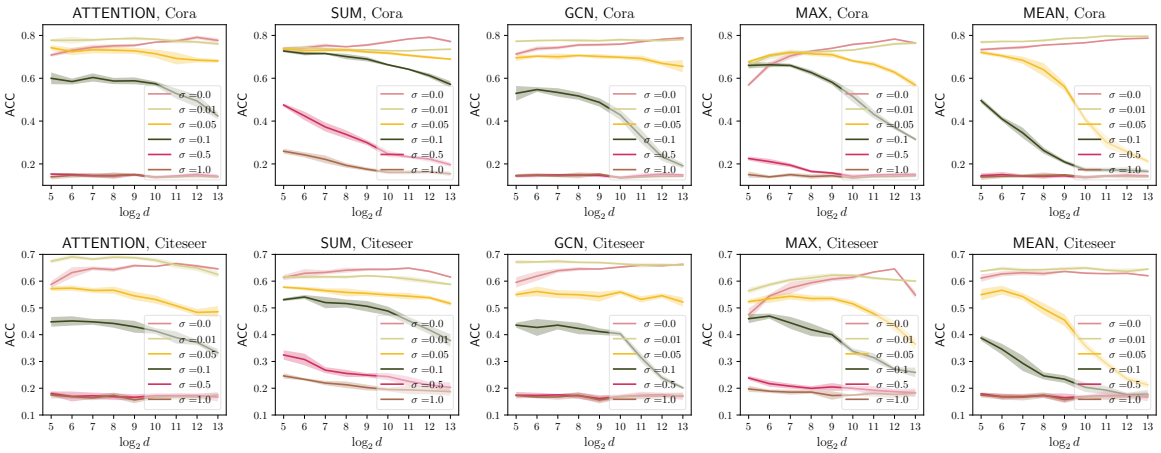
Figure 4: Privacy-utility trade-off on Cora and Citeseer datasets using the unconstrained training scheme. The horizontal axes measure feature dimension d in $\log_2$ scale and the vertical axes stand for performance measures All plots are based on 5 independent trials with shades indicating one standard deviation.

larger d While previous works (Duddu et al., 2020; He et al., 2021) has elucidated the vulnerability of graph representations in the Planetoid datasets, our results further augment previous research by demonstrating a monotonic relationship between COSERA efficacy and d.

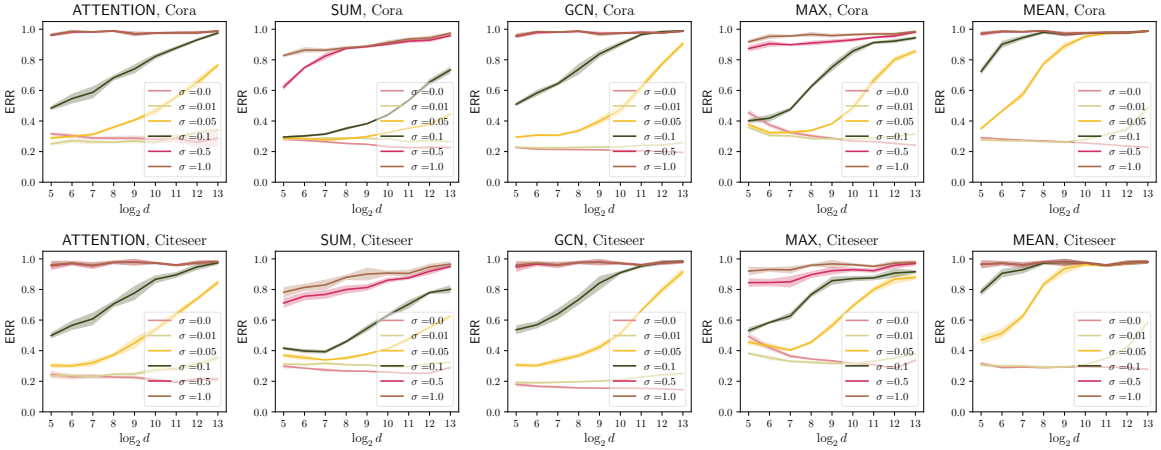
COSERA empirically elicits privacy-utility trade-off under the constrained scheme When the noise level is moderate, i.e., $\sigma \in \{0.01, 0.05\}$. The result demonstrates that privacy and utility are, at least to some extent, at odds: Under lower noise level, COSERA is able to achieve non-trivial success especially when d is small. Furthermore, raising the feature dimension d results in both a decrease in utility as well as an increase in privacy. This is actually predictable: Since we explicitly control the operator norm to be around 1, a larger d implies a smaller "signal-to-noise ratio" with the signal being (loosely) defined as the magnitude of the aggregated node representations.

COSERA losses power against NAG using larger ds in the unconstrained scheme A surprising evidence according to figure 4 and 5 is that when the feature dimension d is sufficiently raised, i.e., $d > 1024$, the attacking performances exhibit U-shaped ERR curves. Consequently, we are able to achieve decent protection against COSERA (AUROC < 0.6 or ERR > 0.8) while at the same time incurring slight degradation in model utility (> 0.7 Accuracy in Cora and Pubmed) Moreover, the phenomenon is more evident for higher noise levels. The outcome seems favorable insofar as we have identified GNN solutions that manifest both high performance and a degree of privacy since the training procedure is ostensibly unrelated to the attacking mechanism. But does that sound too good to be true?

A closer look at GNN solutions obtained via NAG in the



(a) GNN model performance over Cora and Citeseer datasets under 5 different aggregation types.



(b) Attacking performance of COSERA over Cora and Citeseer datasets under 5 different aggregation types.

Figure 5: Privacy-utility trade-off on Cora and Citeseer datasets using the constrained training scheme. The semantic interpretation of the axes is analogous to that depicted in figure 4

unconstrained scheme As COSERA is just one form of attack mechanism under a weak adversary, protecting against COSERA does not necessarily imply strict notions of privacy. Motivated by theorem 6.1, we conduct a spectrum study regarding the GNN solutions obtained via NAG in the unconstrained scheme. Specifically, we plot the operator norm of the weight matrices corresponding to the GNN layers across all scenarios and report them in the last column in figure 13 in appendix C.3. The results exhibit a rapidly growing trend of weights' operator norms regarding the increase of both feature dimension d and noise level σ. For GNN models trained using noisy aggregation under large ds, the corresponding bounds according to (7) become vacuous, i.e., practically zero. Additionally, these solutions may exhibit diminished robustness, as the corresponding Lipschitz constants are likely to be inadequately regulated (Yang et al., 2020). To conclude, we have found successful empirical defenses against COSERA without satisfying strict notions of privacy, suggesting that **COSERA has limitations as a tool for auditing private GRL training procedures**.

Impact of different AGG mechanisms According to figure 4 and 5, the previously discovered phenomenons are present for all the 5 aggregation types. Nevertheless, the degree to which these phenomena exhibit varies with the specific type of aggregation employed. Notably, the behaviors of ATTENTION, MEAN, and GCN pooling display similarities attributable to their shared mechanism in (weighted) average aggregation. Conversely, the efficacy of the COSERA against Noisy Aggregation (NAG) when SUM and MAX pooling are utilized appears less susceptible to changes in d.

8. Discussion and conclusion

In this paper, we have studied the behavior of the COSERA adversary by characterizing its performance against different

kinds of underlying graph structures as well as encoding mechanisms, both in theory and through extensive empirical evaluations. Notwithstanding, several research problems warrant further study, which we discuss in appendix D.

9. Impact statements

The pervasive integration of graph representation learning (GRL) into various sectors, from social networks to bioinformatics, underscores the necessity of addressing the security and privacy risks inherent in these technologies. This paper contributes to the understanding of such risks by dissecting the structural vulnerabilities of graph representations under cosine-similarity-based edge reconstruction attacks (COSERA). Our work has significant ethical implications and societal consequences, as we aim to balance the need for advanced data analytics with the imperative of safeguarding individual and community privacy.

Theoretically articulating the success and failure modes of COSERA, our research offers a framework for evaluating GRL models against potential privacy breaches. The insights gained can guide the development of more secure algorithms that resist inadvertent information disclosure. By highlighting the efficacy of COSERA in various settings, this paper also underscores the potential for such attacks to serve as auditing tools for privacy-preserving mechanisms, thereby fostering the creation of more trustworthy GRL systems.

As GRL technologies continue to evolve, our work calls attention to the importance of proactive privacy research in the field. It encourages the industry to adopt privacy-by-design principles and serves as a reminder to policymakers to consider the implications of GRL in legislation around data protection. Future societal consequences hinge on our ability to reconcile the benefits of GRL with the privacy rights of individuals, necessitating ongoing research, transparent practices, and informed governance to navigate this complex landscape.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Abbe, E. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Awasthi, P., Das, A., and Gollapudi, S. A convergence analysis of gradient descent on graph neural networks. *Advances in Neural Information Processing Systems*, 34: 20385–20397, 2021.

Canonne, C. L. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Chanpuriya, S., Musco, C., Sotiropoulos, K., and Tsourakakis, C. Deepwalking backwards: from embeddings back to graphs. In *International Conference on Machine Learning*, pp. 1473–1483. PMLR, 2021.

Chien, E., Chen, W.-N., Pan, C., Li, P., Ozgur, A., and Milenkovic, O. Differentially private decoupled graph convolutions for multigranular topology protection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Jagielski, M., Huang, Y., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., et al. Challenges towards the next frontier in privacy. *arXiv preprint arXiv:2304.06929*, 2023.

Dai, E., Zhao, T., Zhu, H., Xu, J., Guo, Z., Liu, H., Tang, J., and Wang, S. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *arXiv preprint arXiv:2204.08570*, 2022.

Duddu, V., Boutet, A., and Shejwalkar, V. Quantifying privacy leakage in graph embedding. In *MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 76–85, 2020.

Hamilton, W., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017a.

Hamilton, W. L., Ying, R., and Leskovec, J. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017b.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

He, X., Jia, J., Backes, M., Gong, N. Z., and Zhang, Y. Stealing links from graph neural networks. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2669–2686, 2021.

Jin, D., Wang, R., Ge, M., He, D., Li, X., Lin, W., and Zhang, W. Raw-gnn: Random walk aggregation based graph neural network. *arXiv preprint arXiv:2206.13953*, 2022.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Lehmann, E. L., Romano, J. P., and Casella, G. *Testing statistical hypotheses*, volume 3. Springer, 1986.

Luan, S., Hua, C., Xu, M., Lu, Q., Zhu, J., Chang, X.-W., Fu, J., Leskovec, J., and Precup, D. When do graph neural networks help with node classification? investigating the homophily principle on node distinguishability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=kJmYu3Ti2z.

Meng, L., Bai, Y., Chen, Y., Hu, Y., Xu, W., and Weng, H. Devil in disguise: Breaching graph neural networks privacy through infiltration. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1153–1167, 2023.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Nasr, M., Songi, S., Thakurta, A., Papernot, N., and Carlin, N. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pp. 866–882. IEEE, 2021.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

Sajadmanesh, S., Shamsabadi, A. S., Bellet, A., and Gatica-Perez, D. Gap: Differentially private graph neural networks with aggregation perturbation. In *USENIX Security 2023-32nd USENIX Security Symposium*, 2023.

Tramer, F., Terzis, A., Steinke, T., Song, S., Jagielski, M., and Carlini, N. Debugging differential privacy: A case study for privacy auditing, 2022.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Wang, B., Guo, J., Li, A., Chen, Y., and Li, H. Privacy-preserving representation learning on graphs: A mutual information perspective. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1667–1676, 2021.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. Simplifying graph convolutional networks. In *International conference on machine learning*, pp. 6861–6871. PMLR, 2019.

Wu, F., Long, Y., Zhang, C., and Li, B. Linkteller: Recovering private edges from graph neural networks via influence analysis. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2005–2024. IEEE, 2022a.

Wu, R., Zhang, M., Lyu, L., Xu, X., Hao, X., Fu, X., Liu, T., Zhang, T., and Wang, W. Privacy-preserving design of graph neural networks with applications to vertical federated learning. *arXiv preprint arXiv:2310.20552*, 2023.

Wu, X., Chen, Z., Wang, W., and Jadbabaie, A. A non-asymptotic analysis of oversmoothing in graph neural networks. *arXiv preprint arXiv:2212.10701*, 2022b.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Xu, K., Zhang, M., Jegelka, S., and Kawaguchi, K. Optimization of graph neural networks: Implicit acceleration by skip connections and more depth. In *International Conference on Machine Learning*, pp. 11592–11602. PMLR, 2021.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Yang, Z., Cohen, W., and Salakhudinov, R. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

Zhang, H., Wu, B., Wang, S., Yang, X., Xue, M., Pan, S., and Yuan, X. Demystifying uneven vulnerability of link stealing attacks against graph neural networks. In *International Conference on Machine Learning*, pp. 41737–41752. PMLR, 2023.

Zhang, Z., Liu, Q., Huang, Z., Wang, H., Lu, C., Liu, C., and Chen, E. Graphmi: Extracting private graph data from graph neural networks. *arXiv preprint arXiv:2106.02820*, 2021.

Zhang, Z., Liu, Q., Huang, Z., Wang, H., Lee, C.-K., and Chen, E. Model inversion attacks against graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Zhou, J., Chen, C., Zheng, L., Wu, H., Wu, J., Zheng, X., Wu, B., Liu, Z., and Wang, L. Vertically federated graph neural network for privacy-preserving node classification. *arXiv preprint arXiv:2005.11903*, 2020.

Zhou, Z., Zhou, C., Li, X., Yao, J., Yao, Q., and Han, B. On strengthening and defending graph reconstruction attack with markov chain approximation. *arXiv preprint arXiv:2306.09104*, 2023.

A. Related works

Typically, there exist two categories of private information that may potentially be compromised during the training or deployment phases of graph neural network models: The (sensitive) node attributes and the adjacency relation between nodes. In this paper we focus on the later category since edge adjacency relations are less informative, i.e., for each pair of nodes, the existence of an edge constitutes only a single bit of information.

A.1. Edge reconstruction attacks on graph-structured data

Contemporary developments on edge reconstruction attacks differ significantly in their conceptualization of adversaries, particularly in terms of their capabilities (Zhang et al., 2021; 2022) and the extent of prior knowledge they possess about the GRL model and the underlying graph dataset (He et al., 2021). The mechanism of COSERA was first proposed in (Duddu et al., 2020) and later studies in (He et al., 2021). Empirical evidences suggest that with only black-box access to node representations, the COSERA mechanism obtains a high success rate (AUC > 0.9 for the Citeseer dataset). Subsequent developments have explored stronger attacks under more powrful adversaries. In (He et al., 2021) the authors investigated the impact of an adversary's prior knowledge, including the possession of node features, partial graph structure, and access to a shadow dataset, on the success rate of corresponding attack strategies. Inspire by information bottleneck,(Zhou et al., 2023) improves COSERA via carefully exploiting intermediate representations produced by GNNs. Notably, despite the adversaries in (He et al., 2021; Zhou et al., 2023) being equipped with substantially more information compared to COSERA, the resulting enhancement in attack performance exhibited by these adversaries demonstrates only marginal improvements relative to COSERA. The GraphMI attack (Zhang et al., 2021) disables the adversary from being able to acquire node representations but instead requires access to node features and labels, as well as white-box access to the GNN model. Recent works explored influence-based attacking schemes, wherein the adversary is allowed to alter the graph information: The LinkTeller attack (Wu et al., 2022a) manipulates node features while (Meng et al., 2023) infiltrates the underlying graph with malicious nodes.

A.2. Theoretical explorations in graph recovery from neural representations

In (Chanpuriya et al., 2021), the authors proposed an algorithm that provably recovers graph structure based on representations generated via DeepWalk, which is a factorizaton-based procedure and different from GNN-produced representations. In (Zhang et al., 2023) the authors showed that when block structure exists in the underlying graph, the performance of COSERA is uneven across node in different blocks. In (Zhou et al., 2023), the authors use information-theoretic arguments to construct more powerful attacks than COSERA. Nevertheless, the aforementioned studies did not provide a theoretical rationale for the practical vulnerabilities manifested as a result of the COSERA.

A.3. Privacy protection against edge reconstruction attacks

Edge differential privacy (EDP) (Nissim et al., 2007) is the most popular privacy notion that offers a formal protection against edge reconstruction attacks. Standard private training algorithms like DPSGD (Abadi et al., 2016) may produce GNN models that is provably private in the sense that membership information of any individual training sample is limitly disclosed. [2] However, such approaches do not provide privacy during *inference* time (Chien et al., 2023). Protection mechanisms against inference-time adversaries are mostly based on noisy version of GNN encoding such as edge-wise randomized response (Wu et al., 2022a) that provides very strong privacy protection yet being overly destructive to model utility. Noisy aggregation (NAG) mechansims (Sajadmanesh et al., 2023; Wu et al., 2023; Chien et al., 2023) are recently proposed that empirically achieves better privacy-utility trade-offs. Inspired by the information bottleneck principle, (Wang et al., 2021; Zhou et al., 2023) proposed to use regularization or saddle-point optimization techniques to control privacy leakage. Yet these proposals are not principled in theory.

[2] Note that this require a careful sensitivity analysis with respect to the correct privacy model like EDP.

B. Proofs of Main Theorems

B.1. Proof of theorem 4.1

In the proof, for notational simplicity, we abuse notation by treating $A = A + I$ and $D = D + I$ (i.e., self-edge is included in the edge graph). We then define $A^{(L)} := A \cdot \underbrace{...}_{L \text{ times}} \cdot A$ and $p_{ij}^{(L)} := ((D^{-1}A)^L)_{ij}$.

To prove the result, we require the following lemmas.

Lemma B.1. *Let $B(n, p)$ denote the binomial distribution with probability p and size n.*

1. *Suppose X dominates $B(n, p)$. For any $a > 0$, we have*

$$\mathbb{P}(X < np - a) \leq \exp\{-a^2/2np\}. \tag{9}$$

2. *Suppose X is dominated by $B(n, p)$. For any $a > 0$, we have*

$$\mathbb{P}(X > np + a) \leq \min\{\exp\{-a^2/2np + a^3/(np)^3\}, \exp\{-\frac{a^2}{2np + 2a/3}\}\}. \tag{10}$$

Proof of Lemma B.1 is standard and we omit it here.

Lemma B.2. *Given a graph with edge probability p ($p \leq C_1 \frac{\log n}{n}$), then*

$$\mathbb{P}(\sum_j A_{i_1 j} A_{i_2 j} \geq 1) \leq C_2 \frac{(\log n)^2}{n}, \tag{11}$$

where i_1, i_2 are two nodes uniformly randomly sampled from the graph ($C_2 = 1.5C_1$).

Proof of Lemma B.2. By the monotonicity, we only need to consider the case when $p \equiv C_1 \frac{\log n}{n}$. By Lemma B.1, we know that $\frac{C_1}{2} \log n \leq \sum_j A_{ij} \leq \frac{3C_1}{2} \log n$ with high probability at least $1 - 1/n^2$ all $i \in [n]$ when C_1 is a sufficiently large constant.

According to the independence between $A_{i_1 j}$ and $A_{i_2 j}$ ($i_1 \neq i_2$), we could easily compute that

$$\begin{aligned} \mathbb{E}[\sum_j A_{i_1 j} A_{i_2 j}] &= \mathbb{E}[\mathbb{E}[\sum_j A_{i_1 j} A_{i_2 j} | \sum_j A_{i_1 j}]] \\ &\leq \frac{3C_1}{2} \log n \cdot \frac{C_1 \log n}{n} + \frac{1}{n^2} \cdot n \cdot \frac{C_1 \log n}{n} \\ &\leq 3C_1^2 \frac{\log^2 n}{n}. \end{aligned} \tag{12}$$

We then have $\mathbb{P}(\sum_j A_{i_1 j} A_{i_2 j} \geq 1) \leq 3C_1^2 \frac{\log^2 n}{n}$ by Markov inequality. Note that $\sum_j A_{i_1 j} A_{i_2 j}$ only takes integer value. In other words, $\mathbb{P}(\sum_j A_{i_1 j} A_{i_2 j} = 0) \geq 1 - 3C_1^2 \frac{\log^2 n}{n}$. □

Lemma B.3. *Given a graph with edge probability p ($p \leq C_1 \frac{\log n}{n}$), then*

$$\mathbb{P}(\sum_j A_{i_1 j}^{(L)} A_{i_2 j}^{(L)} \geq 1) \leq \frac{(C_2 \log n)^{2L}}{n}, \tag{13}$$

where i_1, i_2 are two nodes uniformly randomly sampled from the graph.

Proof of Lemma B.3. By recalling the definition of $A_{ij}^{(L)}$ that $A_{ij}^{(L)}$ equals one only when node i and node j can be connected within a path of length L. Therefore, with probability at least $1 - 1/n^2$, it holds $|\mathcal{N}_j^{(L)}| \leq (\frac{3C_1 \log n}{2})^L$, where $\mathcal{N}_j^{(L)} := \{i : A_{ij}^{(L)} = 1\}$

Note that, given fixed j, $A_{i_1 j}A_{i_2 j}$ is greater than 0 only if $i_1, i_2 \in \mathcal{N}_j^{(L)}$. By the symmetry, we know that this happens with probability at most $\frac{|\mathcal{N}_j^{(L)}|(|\mathcal{N}_j^{(L)}|-1)}{n(n-1)}$ when i_1, i_2 are uniformly randomly sampled. Therefore, by union bound, we have

$$\begin{aligned}\mathbb{P}(\sum_j A_{i_1 j}^{(L)} A_{i_2 j}^{(L)} \geq 1) &\leq \sum_j \frac{|\mathcal{N}_j^{(L)}|(|\mathcal{N}_j^{(L)}|-1)}{n(n-1)} \\ &\leq \frac{(1.5C_1 \log n)^{2L}}{n-1}, \end{aligned} \tag{14}$$

which concludes the proof. □

Proof of the theorem. For any pair of two nodes i and j, we next recall the formula of cosine similarity, $\cos\theta(H_i^{(L)}, H_j^{(L)})$,

$$\cos\theta(H_i^{(L)}, H_j^{(L)}) := \frac{\langle H_i^{(L)}, H_j^{(L)}\rangle}{\sqrt{\langle H_i^{(L)}, H_i^{(L)}\rangle \cdot \langle H_j^{(L)}, H_j^{(L)}\rangle}}, \tag{15}$$

which will be used recurrently in the following main proof.

According to the generation mechanism of node features (i.e., isotropic Gaussian assumption), we have that $|\frac{1}{d}\|X_j\|^2 - 1| \leq 3\sqrt{\frac{\log n}{d}}$ and $|\frac{1}{d}\langle X_j, X_{j'}\rangle| \leq 3\sqrt{\frac{\log n}{d}}$ for all j, j' with probability at least $1 - 1/n^2$.

Case 1: without considering the learnable weight matrix W. For the numerator in $\cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)})$, when i_1 and i_2 are truly connected, we have

$$\begin{aligned}\langle H_{i_1}^{(L)}, H_{i_2}^{(L)}\rangle &= \sum_{j=1}^n p_{i_1 j}^{(L)} p_{i_2 j}^{(L)} \|X_j\|^2 + \sum_{j\neq j'} p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} \langle X_j, X_{j'}\rangle \\ &\geq p_{i_1 i_1}^{(L)} p_{i_2 i_1}^{(L)} \|X_{i_1}\|^2 + \sum_{j\neq j'} p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} \langle X_j, X_{j'}\rangle \\ &\geq \frac{1}{|\mathcal{N}_{i_1}^{(L)}||\mathcal{N}_{i_2}^{(L)}|}\|X_{i_1}\|^2 + \sum_{j\neq j'} p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} \langle X_j, X_{j'}\rangle \\ &\geq \frac{1}{(C_2 \log n)^{2L}} - 3\sqrt{\frac{\log n}{d}} \ \text{(use the fact that } \sum_{j\neq j'} p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} \leq 1) \\ &\geq \frac{2}{3}\cdot\frac{1}{(C_2 \log n)^{2L}} \end{aligned} \tag{16}$$

when $d > 9(C_2 \log n)^{4L+2} \cdot \log n$. On the other hand, when i_1 and i_2 are not connected, by Lemma B.3, we know there are at most $\frac{(C_2 \log n)^{2L}}{n} \cdot n(n-1)/2$ pairs of i_1, i_2 such that $\sum_j p_{i_1 j}^{(L)} p_{i_2 j}^{(L)} > 0$. For the rest of pairs, we have

$$\begin{aligned}\langle H_{i_1}^{(L)}, H_{i_2}^{(L)}\rangle &= \sum_{j\neq j'} p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} \langle X_j, X_{j'}\rangle \\ &\leq 3\sqrt{\frac{\log n}{d}} \\ &< \frac{1}{3}\cdot\frac{1}{(C_2 \log n)^{3L+1}}, \end{aligned} \tag{17}$$

when $d > 9(C_2 \log n)^{6L+2} \cdot \log n$.

For the denominator $(\|H_{i_1}^{(L)}\| \cdot \|H_{i_2}^{(L)}\|)^{1/2}$, we give the upper and lower bounds of $\|H_i^{(L)}\|$. We can compute

$$\begin{aligned}\langle H_i^{(L)}, H_i^{(L)} \rangle &= \sum_{j=1}^{n} p_{ij}^{(L)} p_{ij}^{(L)} \|X_j\|^2 + \sum_{j \neq j'} p_{ij}^{(L)} p_{ij'}^{(L)} \langle X_j, X_{j'} \rangle \\ &\leq 1 + 3\sqrt{\frac{\log n}{d}} \\ &< 1 + \frac{1}{3} \cdot \frac{1}{(C_2 \log n)^{2L+1}}, \end{aligned} \tag{18}$$

where we use the fact that $\sum_j p_{ij}^{(L)} p_{ij}^{(L)} \leq 1$. Conversely, we have

$$\begin{aligned}\langle H_i^{(L)}, H_i^{(L)} \rangle &= \sum_{j=1}^{n} p_{ij}^{(L)} p_{ij}^{(L)} \|X_j\|^2 + \sum_{j \neq j'} p_{ij}^{(L)} p_{ij'}^{(L)} \langle X_j, X_{j'} \rangle \\ &\geq \frac{1}{(C_2 \log n)^L} - 3\sqrt{\frac{\log n}{d}} \\ &\geq \frac{1}{(C_2 \log n)^L} - \frac{1}{3} \cdot \frac{1}{(C_2 \log n)^{2L+1}}, \end{aligned} \tag{19}$$

where we use the fact that $\sum_j p_{ij}^{(L)} p_{ij}^{(L)} \geq 1/(C_2 \log n)^L$ when $|N_i^{(L)}| \leq (C_2 \log n)^L$.

To sum up, $\cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)})$ is at least

$$\frac{2}{3} \cdot \frac{1}{(C_2 \log n)^{2L}} / (1 + \frac{1}{3(C_2 \log n)^{2L+1}}) \geq \frac{1}{2} \cdot \frac{1}{(C_2 \log n)^{2L}} \tag{20}$$

when node i_1 and i_2 are connected. On the other hand, $\cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)})$ is at most

$$\frac{1}{3} \cdot \frac{1}{(C_2 \log n)^{3L+1}} / (\frac{1}{(C_2 \log n)^L} - \frac{1}{3} \cdot \frac{1}{(C_2 \log n)^{2L+1}}) < \frac{1}{2} \cdot \frac{1}{(C_2 \log n)^{2L}} \tag{21}$$

for all pairs (except at most $\frac{(C_2 \log n)^{2L}}{n} \cdot n(n-1)/2$ pairs) of disconnected nodes i_1 and i_2.

By choosing the cutoff $\tau = \frac{1}{2} \cdot \frac{1}{(C_2 \log n)^{2L}}$, with probability at least $1 - 2/n^2$, we have the false negative is zero and the false positive is $\frac{(C_2 \log n)^{2L}}{n}$.

Case 2: with considering the learnable weight matrix W. Additionally, if the learnable weight W is taken into account, we can derive the following results. We define κ_1 and κ_2 to be the largest and smallest positive constants such that

$$\kappa_1 \langle X, X' \rangle \leq \langle WX, WX' \rangle \leq \kappa_2 \langle X, X' \rangle$$

holds. It is easy to see that $\kappa_2/\kappa_1 = (\kappa(W))^2$. Then the parallel version of (16) becomes

$$\langle H_{i_1}^{(L)}, H_{i_2}^{(L)} \rangle \geq \kappa_1 \frac{2}{3} \frac{1}{(C_2 \log n)^{2L}}. \tag{22}$$

The parallel version of (17) becomes

$$\langle H_{i_1}^{(L)}, H_{i_2}^{(L)} \rangle \leq 3\kappa_2 \sqrt{\frac{\log n}{d}}. \tag{23}$$

The parallel version of (18) becomes

$$\langle H_i^{(L)}, H_i^{(L)} \rangle \leq \kappa_2 (1 + 3\sqrt{\frac{\log n}{d}}). \tag{24}$$

The parallel version of (19) becomes

$$\langle H_{i_1}^{(L)}, H_{i_2}^{(L)} \rangle \geq \kappa_1 (\frac{1}{(C_2 \log n)^L} - 3\sqrt{\frac{\log n}{d}}). \tag{25}$$

Combining above results, we have that

$$\begin{aligned} \cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)}) &\geq \frac{\kappa_1 \frac{2}{3} \frac{1}{(C_2 \log n)^{2L}}}{\kappa_2 (1 + 3\sqrt{\frac{\log n}{d}})} \\ &\geq \frac{\kappa_1}{2\kappa_2} \frac{1}{(C_2 \log n)^{2L}} =: \text{cut}_1(L) \end{aligned} \tag{26}$$

when i_1, i_2 are connected and $d \gg \log^2 n$ and

$$\begin{aligned} \cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)}) &\leq \frac{3\kappa_2 \sqrt{\frac{\log n}{d}}}{\kappa_1 (\frac{1}{(C_2 \log n)^L} - 3\sqrt{\frac{\log n}{d}})} \\ &\leq \frac{4\kappa_2}{\kappa_1} \frac{\sqrt{\frac{\log n}{d}}}{\frac{1}{(C_2 \log n)^L}} =: \text{cut}_2(L) \end{aligned} \tag{27}$$

when i_1, i_2 are not connected and $d \gg (C_2 \log n)^{6L+2} \log n$.

Therefore as long as $d \gg (C_2 \log n)^{6L+2} \log n$ and

$$(\frac{\kappa_1}{\kappa_2})^2 \geq 8(C_2 \log n)^{3L} \cdot \sqrt{\frac{\log n}{d}} \tag{28}$$

holds, we can choose any cutoff τ between $\text{cut}_1(L)$ and $\text{cut}_2(L)$ so that false negative rate is zero and false positive rate is no larger than $(C_2 \log n)^{2L}/n$. This completes the proof.

B.2. Proof of theorem 5.1

To prove the desired result, we first need the following lemmas. In the rest of proof, we abuse the notation by treating p as p_0 and q as q_0.

By applying the Hoeffding's inequality, we can obtain the following two lemmas.

Lemma B.4. *It holds* $|\sum_{j:i,j \text{ in the same group}} A_{ij} - \frac{n}{K} \cdot p_0| \leq 3\log n =: \epsilon_1$ *for all* i *with probability at least* $1 - 1/n^2$.

Lemma B.5. *Suppose* i *is in group* k*, it holds* $|\sum_{j:i,j \text{ in the group } k'(\neq k)} A_{ij} - \frac{n}{K} \cdot q_0| \leq \min\{\frac{1}{2}\frac{n}{K} \cdot q_0, 3\log n\} =: \epsilon_2$ *for all* i *with probability at least* $1 - 1/n^2$.

Combining Lemma B.4 and Lemma B.5, we have the following lemma.

Lemma B.6. *It holds* $|\sum_j A_{ij} - (\frac{n}{K} \cdot p_0 + (n - \frac{n}{K}) \cdot q_0)| \leq \epsilon_1 + (K-1)\epsilon_2$ *with probability at least* $1 - 2/n^2$.

In summary, with high probability confidence, Lemma B.6 gives the characterization of degree (i.e. number of neighbours) of every node i.

We then make a step forward and characterize the normalized degree $p_{ij}^{(L)}$ for $L \geq 2$ in the following lemmas.

Lemma B.7. *With probability at least* $1 - 1/n^2$*, it holds that* $|A_{ij}^{(2)} - (\frac{n}{K}p_0^2 + (n - n/K)q_0 p_0)| \leq 6\log n + \frac{1}{2}\frac{n}{K} \cdot q_0$ *for* i, j *from the same group and* $|A_{ij}^{(2)} - (\frac{n}{K}p_0 q_0 + (n - n/K)q_0^2)| \leq \min\{\frac{2}{3}(\frac{n}{K}p_0 q_0 + (n - n/K)q_0^2), 3(K-1)\log n\}$ *for* i, j *from different groups.*

Lemma B.8. *For* $L \geq 2$*, suppose there exist constants* $a_1^{(L)}$ *and* $a_2^{(L)}$ *such that* $|A_{ij}^{(L)} - a_1^{(L)}| \leq \epsilon_1^{(L)}$ *when* i, j *are in the same group and* $|A_{ij}^{(L)} - a_2^{(L)}| \leq \epsilon_2^{(L)}$ *when* i, j *are not in the same group. It holds that*

$$\begin{aligned} &|A_{ij}^{(L+1)} - a_1^{(L+1)}| \leq \epsilon_1^{(L)} \quad i, j \text{ in the same group} \\ &|A_{ij}^{(L+1)} - a_2^{(L+1)}| \leq \epsilon_2^{(L)} \quad i, j \text{ not in the same group}, \end{aligned} \tag{29}$$

with

$$
\begin{aligned}
a_1^{(L+1)} &:= (a_1^{(L)}\frac{n}{K}p_0 + a_2^{(L)}(n-n/K)q_0),\\
a_2^{(L+1)} &:= a_1^{(L)}\frac{n}{K}q_0 + a_2^{(L)}\frac{n}{K}p_0 + a_2^{(L)}(n-2n/K)q_0\\
\epsilon_1^{(L+1)} &:= \epsilon_1^{(L)}\frac{n}{K}p_0 + \epsilon_1 a_1^{(L)} + \epsilon_1\epsilon_1^{(L)} + \epsilon_2^{(L)}(n-n/K)q_0 + (K-1)\epsilon_2 a_2^{(L)} + (K-1)\epsilon_2\epsilon_2^{(L)},\\
\epsilon_2^{(L+1)} &:= \epsilon_1^{(L)}\frac{n}{K}q_0 + \epsilon_2 a_1^{(L)} + \epsilon_2\epsilon_1^{(L)} + \epsilon_2^{(L)}\frac{n}{K}p_0 + \epsilon_1 a_2^{(L)} + \epsilon_1\epsilon_2^{(L)} + \epsilon_2^{(L)}(n-2n/K)q_0 + (K-2)\epsilon_2 a_2^{(L)} + (K-2)\epsilon_2\epsilon_2^{(L)}.
\end{aligned}
$$

Proof of Lemma B.7 is a special case of that of Lemma B.8. In the following, we prove Lemma B.8.

Proof of Lemma B.8. By the definition, we know $A_{ij}^{(L+1)} = \sum_{j'} A_{ij'}^{(L)} A_{j'j}$.

When i, j are from the same class (w.l.o.g, we denote it as class 1), then it holds

$$
\begin{aligned}
& |A_{ij}^{(L+1)} - (a_1^{(L)}\frac{n}{K}p_0 + a_2^{(L)}(n-n/K)q_0)|\\
= \;& |\sum_{j'} A_{ij'}^{(L)} A_{j'j} - (a_1^{(L)}\frac{n}{K}p_0 + a_2^{(L)}(n-n/K)q_0)|\\
\le \;& |\sum_{j' \text{ in class 1}} A_{ij'}^{(L)} A_{j'j} - a_1^{(L)}\frac{n}{K}p_0| + |\sum_{j' \text{ not in class 1}} A_{ij'}^{(L)} A_{j'j} - a_2^{(L)}(n-n/K)q_0|\\
= \;& \epsilon_1^{(L)}\frac{n}{K}p_0 + \epsilon_1 a_1^{(L)} + \epsilon_1\epsilon_1^{(L)} + \epsilon_2^{(L)}(n-n/K)q_0 + (K-1)\epsilon_2 a_2^{(L)} + (K-1)\epsilon_2\epsilon_2^{(L)}. \qquad (30)
\end{aligned}
$$

Therefore, we can let $a_1^{(L+1)} := (a_1^{(L)}\frac{n}{k}p_0 + a_2^{(L)}(n-n/K)q_0)$ and $\epsilon_1^{(L+1)} := \epsilon_1^{(L)}\frac{n}{K}p_0 + \epsilon_1 a_1^{(L)} + \epsilon_1\epsilon_1^{(L)} + \epsilon_2^{(L)}(n-n/K)q_0 + (K-1)\epsilon_2 a_2^{(L)} + (K-1)\epsilon_2\epsilon_2^{(L)}$.

When i, j are not from the same class (w.l.o.g. we assume i is from class 1 and j is from class 2), then it holds

$$
\begin{aligned}
& |A_{ij}^{(L+1)} - (a_1^{(L)}\frac{n}{K}q_0 + a_2^{(L)}\frac{n}{K}p_0 + a_2^{(L)}(n-2n/K)q_0)|\\
= \;& |\sum_{j'} A_{ij'}^{(L)} A_{j'j} - (a_1^{(L)}\frac{n}{K}q_0 + a_2^{(L)}\frac{n}{K}p_0 + a_2^{(L)}(n-2n/K)q_0)|\\
\le \;& |\sum_{j' \text{ in class 1}} A_{ij'}^{(L)} A_{j'j} - a_1^{(L)}\frac{n}{k}q_0| + |\sum_{j' \text{ in class 2}} A_{ij'}^{(L)} A_{j'j} - a_2^{(L)}\frac{n}{K}p_0|\\
& + |\sum_{j' \text{ not in class 1 \& 2}} A_{ij'}^{(L)} A_{j'j} - a_2^{(L)}(n-2n/K)q_0|\\
\le \;& \epsilon_1^{(L)}\frac{n}{K}q_0 + \epsilon_2 a_1^{(L)} + \epsilon_2\epsilon_1^{(L)} + \epsilon_2^{(L)}\frac{n}{K}p_0 + \epsilon_1 a_2^{(L)} + \epsilon_1\epsilon_2^{(L)}\\
& + \epsilon_2^{(L)}(n-2n/K)q_0 + (K-2)\epsilon_2 a_2^{(L)} + (K-2)\epsilon_2\epsilon_2^{(L)}. \qquad (31)
\end{aligned}
$$

Therefore, we can let $a_2^{(L+1)} := a_1^{(L)}\frac{n}{K}q_0 + a_2^{(L)}\frac{n}{K}p_0 + a_2^{(L)}(n-2n/K)q_0$ and $\epsilon_2^{(L+1)} := \epsilon_1^{(L)}\frac{n}{K}q_0 + \epsilon_2 a_1^{(L)} + \epsilon_2\epsilon_1^{(L)} + \epsilon_2^{(L)}\frac{n}{K}p_0 + \epsilon_1 a_2^{(L)} + \epsilon_1\epsilon_2^{(L)} + \epsilon_2^{(L)}(n-2n/K)q_0 + (K-2)\epsilon_2 a_2^{(L)} + (K-2)\epsilon_2\epsilon_2^{(L)}$. □

By above induction, it can be seen that, for any fixed L, $\epsilon_1^{(L)}/a_1^{(L)} = O_p(\frac{\log n}{n})$, $\epsilon_2^{(L)}/a_1^{(L)} = O_p(\frac{\log n}{n})$. It also holds $a_2^{(L)}/a_1^{(L)} = O_p(\frac{\log n}{n})$ when true edge probability satisfies $q_0 = O_p(\frac{\log n}{n})$, and $\epsilon_2^{(L)}/a_2^{(L)} = O_p(\frac{\log n}{n})$ when $q_0 \gg \frac{\log n}{n}$.

Recall the definition that $p_{ij}^{(L)} = ((D^{-1}A)^L)_{ij}$, therefore $p_{ij}^{(L)} \propto A_{ij}^{(L)}$ for any fixed i. In other words, for fixed $L \ge 2$, we

have

$$p_{ij}^{(L)} := \bar{p}_{ij}^{(L)} + O_p(\frac{k\log n}{n^2}) = \underbrace{\frac{a_1^{(L)}}{\frac{n}{K}\cdot a_1^{(L)} + (n-\frac{n}{K})\cdot a_2^{(L)}}}_{=:p_1^{(L)}} + O_p(\frac{k\log n}{n^2}), \quad i,j \text{ in the same group},$$

$$p_{ij}^{(L)} = \bar{p}_{ij}^{(L)} + O_p(\frac{k\log n}{n^2}) = \underbrace{\frac{a_2^{(L)}}{\frac{n}{K}\cdot a_1^{(L)} + (n-\frac{n}{K})\cdot a_2^{(L)}}}_{=:p_2^{(L)}} + O_p(\frac{k\log n}{n^2}), \quad i,j \text{ not in the same group}. \tag{32}$$

Here, on a very high level, we can treat $\bar{p}_{ij}^{(L)}$ as the population version of $((D^{-1}A)^L)_{ij}$. When i,j in the same group, then $\bar{p}_{ij}^{(L)} \equiv p_1^{(L)}$. Otherwise, $\bar{p}_{ij}^{(L)} \equiv p_2^{(L)}$. With above preparations, we are ready to prove the theorem as follows.

Proof of the theorem. We need to consider the case $L \geq 2$ and $L = 1$ separately.

Case 1: $L \geq 2$. We define $\bar{X}_k := \sum_{i\in \text{ group k}} X_i$ and $r^{(L)} := a_2^{(L)}/a_1^{(L)}$. For the numerator in $\cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)})$, when i_1 and i_2 are in the same group (w.l.o.g, suppose it is group 1), we have

$$\begin{aligned}
\langle H_{i_1}^{(L)}, H_{i_2}^{(L)}\rangle &= \sum_{j=1}^n p_{i_1 j}^{(L)} p_{i_2 j}^{(L)} \|X_j\|^2 + \sum_{j\neq j'} p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} \langle X_j, X_{j'}\rangle \\
&= \sum_{j=1}^n \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j}^{(L)} \|X_j\|^2 + \sum_{j\neq j'} \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j'}^{(L)} \langle X_j, X_{j'}\rangle + \text{error} \qquad (33)\\
&= p_1^{(L)} p_1^{(L)} \langle \bar{X}_1, \bar{X}_1\rangle + 2\sum_{k\neq 1} p_1^{(L)} p_2^{(L)} \langle \bar{X}_1, \bar{X}_k\rangle + \sum_{k\neq 1} p_2^{(L)} p_2^{(L)} \langle \bar{X}_k, \bar{X}_k\rangle \\
&\quad + \sum_{k\neq k'\neq 1} p_2^{(L)} p_2^{(L)} \langle \bar{X}_k, \bar{X}_{k'}\rangle + \text{error} \\
&= p_1^{(L)} p_1^{(L)} \frac{n}{K} + (K-1) p_2^{(L)} p_2^{(L)} \frac{n}{K} + O_p((K-1) p_1^{(L)} p_2^{(L)} \frac{n}{K}\frac{1}{\sqrt{d}} \qquad (34)\\
&\quad + (K-1)(K-2) p_2^{(L)} p_2^{(L)} \frac{n}{K}\frac{1}{\sqrt{d}}) + \text{error},
\end{aligned}$$

where (34) uses the property of node feature generation mechanism that $\langle \bar{X}_k, \bar{X}_k\rangle = \frac{n}{K}(1+\sqrt{1/d})$ for any k and $\langle \bar{X}_k, \bar{X}_{k'}\rangle = O_p(\frac{n}{K\sqrt{d}})$ for $k \neq k'$. Here the error term in (34) is error $:= \sum_{j=1}^n (p_{i_1 j}^{(L)} p_{i_2 j}^{(L)} - \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j}^{(L)}) \|X_j\|^2 + \sum_{j\neq j'} (p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} - \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j'}^{(L)}) \langle X_j, X_{j'}\rangle$, which can be controlled as follows.

$$\begin{aligned}
|\text{error}| &= |\sum_{j=1}^n (p_{i_1 j}^{(L)} p_{i_2 j}^{(L)} - \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j}^{(L)}) \|X_j\|^2 + \sum_{j\neq j'} (p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} - \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j'}^{(L)}) \langle X_j, X_{j'}\rangle| \\
&\leq |\sum_{j=1}^n (p_{i_1 j}^{(L)} p_{i_2 j}^{(L)} - \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j}^{(L)}) \|X_j\|^2| + |\sum_{j\neq j'} (p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} - \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j'}^{(L)}) \langle X_j, X_{j'}\rangle| \\
&\leq C(\frac{k\log n}{n^2} + \sum_j \frac{k\log n}{n^2}\sqrt{\frac{\log n}{d}}) \\
&= O_p(\frac{k\log n}{n^2} + \frac{k\log n}{n}\sqrt{\frac{\log n}{d}}), \qquad (35)
\end{aligned}$$

where (35) utilizes the fact that $\sum_j \bar{p}_{ij} \equiv 1$ for any i and (32) by adjusting the constant.

When i_1, i_2 are not in the same group (w.l.o.g, suppose i_1 in group 1 and i_2 in group 2), we have

$$\begin{aligned}
\langle H_{i_1}^{(L)}, H_{i_2}^{(L)} \rangle &= \sum_{j=1}^{n} p_{i_1 j}^{(L)} p_{i_2 j}^{(L)} \|X_j\|^2 + \sum_{j \neq j'} p_{i_1 j}^{(L)} p_{i_2 j'}^{(L)} \langle X_j, X_{j'} \rangle \\
&= \sum_{j=1}^{n} \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j}^{(L)} \|X_j\|^2 + \sum_{j \neq j'} \bar{p}_{i_1 j}^{(L)} \bar{p}_{i_2 j'}^{(L)} \langle X_j, X_{j'} \rangle + \text{error} \\
&= p_1^{(L)} p_2^{(L)} (\langle \bar{X}_1, \bar{X}_1 \rangle + \langle \bar{X}_2, \bar{X}_2 \rangle) + (p_1^{(L)} p_1^{(L)} + p_2^{(L)} p_2^{(L)}) \langle \bar{X}_1, \bar{X}_2 \rangle \\
&\quad + (p_1^{(L)} p_2^{(L)} + p_2^{(L)} p_2^{(L)}) \sum_{k \neq 1,2} \langle \bar{X}_1 + \bar{X}_2, \bar{X}_k \rangle + \sum_{k \neq 1,2} p_2^{(L)} p_2^{(L)} \langle \bar{X}_k, \bar{X}_k \rangle \\
&\quad + \sum_{k \neq k' \neq 1,2} p_2^{(L)} p_2^{(L)} \langle \bar{X}_k, \bar{X}_{k'} \rangle + \text{error} \\
&= 2 p_1^{(L)} p_2^{(L)} \frac{n}{K} + (K-2) p_2^{(L)} p_2^{(L)} \frac{n}{K} \\
&\quad + O_p((p_1^{(L)} p_1^{(L)} + K p_1^{(L)} p_2^{(L)} + K^2 p_2^{(L)} p_2^{(L)}) \frac{n}{K} \sqrt{\frac{1}{d}}) + \text{error}.
\end{aligned} \tag{36}$$

To sum up, if i_1, i_2 are in the same group, $\cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)})$ satisfies

$$\begin{aligned}
&\cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)}) \\
=&\frac{\langle H_{i_1}^{(L)}, H_{i_2}^{(L)} \rangle}{\sqrt{\langle H_{i_1}^{(L)}, H_{i_1}^{(L)} \rangle \cdot \langle H_{i_2}^{(L)}, H_{i_2}^{(L)} \rangle}} \\
=&\frac{p_1^{(L)} p_1^{(L)} \frac{n}{k} + (K-1) p_2^{(L)} p_2^{(L)} \frac{n}{K}}{p_1^{(L)} p_1^{(L)} \frac{n}{K} + (K-1) p_2^{(L)} p_2^{(L)} \frac{n}{K} + C\big((K p_1^{(L)} p_2^{(L)} + K^2 p_2^{(L)2}) \frac{n}{K\sqrt{d}} + O_p(\frac{K \log n}{n}(\frac{1}{n} + \sqrt{\frac{\log n}{d}}))\big)} \\
=&\underbrace{1}_{\text{cut}_1(L)} - o_p(1)
\end{aligned} \tag{37}$$

as long as $d \gg K^2 \log^3 n / b^2$.

If i_1, i_2 are not in the same group, $\cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)})$ satisfies

$$\begin{aligned}
&\cos\theta(H_{i_1}^{(L)}, H_{i_2}^{(L)}) \\
=&\frac{\langle H_{i_1}^{(L)}, H_{i_2}^{(L)} \rangle}{\sqrt{\langle H_{i_1}^{(L)}, H_{i_1}^{(L)} \rangle \cdot \langle H_{i_2}^{(L)}, H_{i_2}^{(L)} \rangle}} \\
=&\frac{2 p_1^{(L)} p_2^{(L)} \frac{n}{K} + (K-2) p_2^{(L)} p_2^{(L)} \frac{n}{K} + C\big((K p_1^{(L)} p_2^{(L)} + K^2 p_2^{(L)2}) \frac{n}{K\sqrt{d}} + O_p(\frac{K \log n}{n}(\frac{1}{n} + \sqrt{\frac{\log n}{d}}))\big)}{p_1^{(L)} p_1^{(L)} \frac{n}{K} + (K-1) p_2^{(L)} p_2^{(L)} \frac{n}{K}} \\
=&\underbrace{\frac{2r^{(L)} + (K-2) r^{(L)2}}{1 + (K-1) r^{(L)2}}}_{\text{cut}_2(L)} + o_p(1).
\end{aligned} \tag{38}$$

Remark. As $L \to \infty$, $r^{(L)}$ will converge to 1. Therefore, $\text{cut}_2(L)$ will eventually equal $1 \equiv \text{cut}_1(L)$.

Case 2: $L = 1$. For notational convenience, we define $\tilde{X}_{k,1}^{(i)} := \sum_{i \in \text{ group } k} b_{i,1}^{(i)} X_i$ where $b_{i,1}^{(i)}$'s are i.i.d. Bernoulli random variables with success probability p_0 and $\tilde{X}_{k,2}^{(i)} := \sum_{i \in \text{ group } k} b_{i,2}^{(i)} X_i$ where $b_{i,2}^{(i)}$'s are i.i.d. Bernoulli random variables with success probability q_0.

Then it is straightforward to calculate that, if i_1, i_2 are in the same group 1, it holds

$$\begin{aligned}\langle H_{i_1}^{(1)}, H_{i_2}^{(1)}\rangle \quad =_d \quad & \frac{1}{D_{i_1}D_{i_2}}\big(\langle \tilde{X}_{1,1}^{(i_1)}, \tilde{X}_{1,1}^{(i_2)}\rangle + \sum_{k\neq 1}\langle \tilde{X}_{1,1}^{(i_1)}, \tilde{X}_{k,2}^{(i_2)}\rangle + \sum_{k\neq 1}\langle \tilde{X}_{k,2}^{(i_1)}, \tilde{X}_{1,1}^{(i_2)}\rangle \\ & + \sum_{k\neq 1}\langle \tilde{X}_{k,2}^{(i_1)}, \tilde{X}_{k,2}^{(i_2)}\rangle + \sum_{k,k'\neq 1}\langle \tilde{X}_{k,2}^{(i_1)}, \tilde{X}_{k',2}^{(i_2)}\rangle\big) \\ = \quad & \frac{1}{D_{i_1}D_{i_2}}\big(\frac{n}{k}p_0^2 + (K-1)\frac{n}{K}q_0^2 + O_p(p_0\frac{n\sqrt{\log n}}{K\sqrt{d}} + \sqrt{p_0q_0}\frac{n\sqrt{\log n}}{\sqrt{d}} + Kq_0\frac{n\sqrt{\log n}}{\sqrt{d}})\big). \quad (39)\end{aligned}$$

Similarly, if i_1, i_2 are not in the same group (w.l.o.g, they are in group 1 and 2 respectively), it holds

$$\begin{aligned}\langle H_{i_1}^{(1)}, H_{i_2}^{(1)}\rangle \quad =_d \quad & \frac{1}{D_{i_1}D_{i_2}}\big(\langle \tilde{X}_{1,1}^{(i_1)}, \tilde{X}_{1,2}^{(i_2)}\rangle + \langle \tilde{X}_{2,2}^{(i_1)}, \tilde{X}_{1,1}^{(i_2)}\rangle + \langle \tilde{X}_{1,1}^{(i_1)}, \tilde{X}_{2,1}^{(i_2)}\rangle + \langle \tilde{X}_{2,2}^{(i_1)}, \tilde{X}_{1,2}^{(i_2)}\rangle \\ & + \sum_{k\neq 1,2}\langle \tilde{X}_{1,1}^{(i_1)} + \tilde{X}_{2,2}^{(i_1)}, \tilde{X}_{k,2}^{(i_2)}\rangle + \sum_{k\neq 1,2}\langle \tilde{X}_{k,2}^{(i_1)}, \tilde{X}_{1,2}^{(i_2)} + \tilde{X}_{2,1}^{(i_2)}\rangle \\ & + \sum_{k\neq 1,2}\langle \tilde{X}_{k,2}^{(i_1)}, \tilde{X}_{k,2}^{(i_2)}\rangle + \sum_{k,k'\neq 1,2}\langle \tilde{X}_{k,2}^{(i_1)}, \tilde{X}_{k',2}^{(i_2)}\rangle\big) \\ = \quad & \frac{1}{D_{i_1}D_{i_2}}\big(2\frac{n}{k}p_0q_0 + (K-2)\frac{n}{K}q_0^2 + O_p(p_0\frac{n\sqrt{\log n}}{K\sqrt{d}} + \sqrt{p_0q_0}\frac{n\sqrt{\log n}}{\sqrt{d}} + Kq_0\frac{n\sqrt{\log n}}{\sqrt{d}})\big). \quad (40)\end{aligned}$$

Moreover, if $i_1 = i_2$, it holds

$$\begin{aligned}\langle H_{i_1}^{(1)}, H_{i_1}^{(1)}\rangle \quad =_d \quad & \frac{1}{D_{i_1}D_{i_1}}\big(\langle \tilde{X}_{1,1}^{(i_1)}, \tilde{X}_{1,1}^{(i_1)}\rangle + 2\sum_{k\neq 1}\langle \tilde{X}_{1,1}^{(i_1)}, \tilde{X}_{k,2}^{(i_1)}\rangle\rangle \\ & + \sum_{k\neq 1}\langle \tilde{X}_{k,2}^{(i_1)}, \tilde{X}_{k,2}^{(i_1)}\rangle + \sum_{k,k'\neq 1}\langle \tilde{X}_{k,2}^{(i_1)}, \tilde{X}_{k',2}^{(i_1)}\rangle\big) \\ = \quad & \frac{1}{D_{i_1}D_{i_2}}\big(\frac{n}{k}p_0 + (K-1)\frac{n}{K}q_0 + O_p(p_0\frac{n\sqrt{\log n}}{K\sqrt{d}} + \sqrt{p_0q_0}\frac{n\sqrt{\log n}}{\sqrt{d}} + Kq_0\frac{n\sqrt{\log n}}{\sqrt{d}})\big). \quad (41)\end{aligned}$$

To sum up, we arrive at

$$\begin{aligned}& \cos\theta(H_{i_1}^{(1)}, H_{i_2}^{(1)}) \\ = \quad & \frac{\langle H_{i_1}^{(1)}, H_{i_2}^{(1)}\rangle}{\sqrt{\langle H_{i_1}^{(1)}, H_{i_1}^{(1)}\rangle \cdot \langle H_{i_2}^{(1)}, H_{i_2}^{(1)}\rangle}} \\ = \quad & \frac{\frac{n}{k}p_0^2 + (K-1)\frac{n}{K}q_0^2}{\frac{n}{k}p_0 + (K-1)\frac{n}{K}q_0 + O_p(p_0\frac{n\sqrt{\log n}}{K\sqrt{d}} + \sqrt{p_0q_0}\frac{n\sqrt{\log n}}{\sqrt{d}} + Kq_0\frac{n\sqrt{\log n}}{\sqrt{d}})} \\ = \quad & \underbrace{\frac{\frac{n}{k}p_0^2 + (K-1)\frac{n}{K}q_0^2}{\frac{n}{k}p_0 + (K-1)\frac{n}{K}q_0}}_{\mathrm{cut}_1(1)} + o_p(1) \qquad (42)\end{aligned}$$

for i_1, i_2 from the same group, when $d \gg \log n$. Similarly, we have

$$\begin{aligned}& \cos\theta(H_{i_1}^{(1)}, H_{i_2}^{(2)}) \\ = \quad & \frac{\langle H_{i_1}^{(1)}, H_{i_2}^{(1)}\rangle}{\sqrt{\langle H_{i_1}^{(1)}, H_{i_1}^{(1)}\rangle \cdot \langle H_{i_2}^{(1)}, H_{i_2}^{(1)}\rangle}} \\ = \quad & \frac{\frac{n}{k}p_0^2 + (K-1)\frac{n}{K}q_0^2 + O_p(p_0\frac{n\sqrt{\log n}}{K\sqrt{d}} + \sqrt{p_0q_0}\frac{n\sqrt{\log n}}{\sqrt{d}} + Kq_0\frac{n\sqrt{\log n}}{\sqrt{d}})}{\frac{n}{k}p_0 + (K-1)\frac{n}{K}q_0} \\ = \quad & \underbrace{\frac{2\frac{n}{k}p_0q_0 + (K-2)\frac{n}{K}q_0^2}{\frac{n}{k}p_0 + (K-1)\frac{n}{K}q_0}}_{\mathrm{cut}_2(1)} + o_p(1) \qquad (43)\end{aligned}$$

for i_1, i_2 from different groups, when $d \gg \log n/p_0^2$.

Therefore, for any fixed $L \geq$ and any fixed cutoff $\tau \geq \mathrm{cut}_1(L)$, then COSERA will predict at least $pK\frac{n}{K} \cdot (\frac{n}{K}-1)/2 + qK(K-1)/2 + \frac{n}{K} \cdot \frac{n}{K}$ truly connected pairs as dis-connected. In other words, we have the false negative rate is at least $p/(2k) + q/2$. If the cutoff τ is between $\mathrm{cut}_2(L)$ and $\mathrm{cut}_1(L)$, then COSERA will predict at least $(1-p)K\frac{n}{K} \cdot (\frac{n}{K}-1)/2$ truly dis-connected pairs as connected and predict at least $qK(K-1)/2 + \frac{n}{K} \cdot \frac{n}{K}$ truly connected pairs as dis-connected. That is, false positive rate is at least $(1-p)/(2k)$ and false negative rate is at least $(1-q)/2$. If the cutoff τ is less than $\mathrm{cut}_2(L)$, then COSERA will predict at least $(1-p)K\frac{n}{K} \cdot (\frac{n}{K}-1)/2 + (1-q)K(K-1)/2 + \frac{n}{K} \cdot \frac{n}{K}$ truly connected pairs as dis-connected. That is, false positive rate is at least $(1-p)/(2k) + (1-q)/2$. This completes the proof.

B.3. Proof of theorem 6.1

Before starting the proof, we restate the concrete definitions of the (NAG enabled)-GNNs involved in the theorem in the message-passing form as in (7):

Mean pooling (Hamilton et al., 2017a) This is the most standard form of message passing GNN. With the un-normalized and un-perturbed version analyzed in section 4 and 5:

$$h_v^{(l)} = \mathsf{ReLU}\left(\frac{1}{d_v+1}\sum_{u\in N(v)\cup\{v\}} \frac{W_l h_u^{(l-1)}}{\left\|h_u^{(l-1)}\right\|_2} + \epsilon\right) \tag{SAGE-meanpool}$$

Summation pooling (Xu et al., 2018) This is a simplified version of the GIN model which is also analyzed in (Wu et al., 2023):

$$h_v^{(l)} = \mathsf{ReLU}\left(\sum_{u\in N(v)\cup\{v\}} \frac{W_l h_u^{(l-1)}}{\left\|h_u^{(l-1)}\right\|_2} + \epsilon\right) \tag{GIN}$$

Max pooling (Hamilton et al., 2017a) In its un-normalized and un-perturbed version, this corresponds to the mostly used SAGE model:

$$h_v^{(l)} = \mathsf{ReLU}\left(\max_{u\in N(v)\cup\{v\}} \frac{W_l h_u^{(l-1)}}{\left\|h_u^{(l-1)}\right\|_2} + \epsilon\right) \tag{SAGE-maxpool}$$

GCN pooling (Kipf & Welling, 2016) The GCN pooling takes the form

$$h_v^{(l)} = \mathsf{ReLU}\left(\frac{1}{\sqrt{d_v+1}}\sum_{u\in N(v)\cup\{v\}} \frac{W_l h_u^{(l-1)}}{\sqrt{d_u+1}\left\|h_u^{(l-1)}\right\|_2} + \epsilon\right) \tag{GCN}$$

Attentive pooling (Veličković et al., 2018) This is also know as the GAT model. To simplify notations, let $\tilde{h}_v^{(l)} = h_v^{(l)}/\left\|h_v^{(l)}\right\|_2$, then the GAT model is recursively defined as

$$\begin{aligned} h_v^{(l)} &= \mathsf{ReLU}\left(\sum_{u\in N(v)\cup\{v\}} \alpha_{uv} W_l \tilde{h}_u^{(l-1)} + \epsilon\right) \\ \alpha_{uv} &= \frac{\exp\left(\mathsf{LeakyReLU}\left(\langle \beta_{\text{src}}, W_l\tilde{h}_u^{(l-1)}\rangle + \langle \beta_{\text{dst}}, W_l\tilde{h}_v^{(l-1)}\rangle\right)\right)}{\sum_{u\in N(v)\cup\{v\}}\exp\left(\mathsf{LeakyReLU}\left(\langle \beta_{\text{src}}, W_l\tilde{h}_v^{(l-1)}\rangle + \langle \beta_{\text{dst}}, W_l\tilde{h}_v^{(l-1)}\rangle\right)\right)} \end{aligned} \tag{GAT}$$

where $\beta_{\text{src}}, \beta_{\text{dst}} \in \mathbb{R}^d$ are learnable vector parameters.

The theorem is a consequence of the following lemma:

Lemma B.9. *Fix an arbitrary node pair (u, v). Let $\mathbf{H}_1$ and $\mathbf{H}_0$ be the collection of node representations generated under $A_{uv} = 1$ and $A_{uv} = 0$, respectively. It follows that the Kullback-Leibler divergence between $\mathbf{H}_1$ and $\mathbf{H}_0$ is bounded:*

$$D_{KL}\left(\mathbf{H}_1 \parallel \mathbf{H}_0\right) \leq C\frac{\sum_{l\in[L]}\|W_l\|_{op}^2}{\sigma^2}. \tag{44}$$

Here the constant $C = 1$ for (SAGE-meanpool)*,* (GIN) *and* (GCN)*; and $C = 4$ for* (SAGE-maxpool) *and* (GAT)*.*

Proof of lemma B.9. The proof is essentially a proof of Rényi differential privacy similar to that in (Wu et al., 2023). First we fix a single l-th layer of GNN defined in (7). We rewrite (7) as:

$$h_v^{(l)} = T\left(\mathsf{AGG}\left(\frac{W_l h_u^{(l-1)}}{\left\|h_u^{(l-1)}\right\|_2}, u \in N(v)\cup\{v\}\right) + \epsilon\right) := T\left(\tilde{h}_v^{(l-1)} + \epsilon\right) \tag{45}$$

Let the corresponding representation matrix be $H_1^{(l)}$ for $A_{uv} = 1$ and $H_0^{(l)}$ for $A_{uv} = 0$ for any $l \in [L]$. Further denote $\widetilde{H}_a^l = \{\tilde{h}_{v,a}^{(l)}\}_{v\in V}$ as the intermediate representation defined as in (45) with $A_{uv} = a, a \in \{0, 1\}$. Then by standard results on Rényi divergence (Mironov, 2017), we have

$$\mathrm{D}_{\mathrm{KL}}\left(H_1^l \parallel H_0^l\right) = \frac{\left\|\widetilde{H}_1^{(l)} - \widetilde{H}_0^{(l)}\right\|_2^2}{2\sigma^2} \tag{46}$$

For some input H^{l-1}. It follows that given all the other edges, the only terms that contributes to $\left\|\widetilde{H}_1^{(l)} - \widetilde{H}_0^{(l)}\right\|_2^2$ are $\left\|\tilde{h}_{v,1}^{(l)} - \tilde{h}_{v,0}^{(l)}\right\|_2^2$ and $\left\|\tilde{h}_{u,1}^{(l)} - \tilde{h}_{u,0}^{(l)}\right\|_2^2$. Next we give the derivation of various GNN architectures:

The case of (SAGE-meanpool) We let d_v to be the degree of v assuming $A_{uv} = 1$. Further let $g_v^{(l)} = \frac{W_l h_u^{(l-1)}}{\left\|h_u^{(l-1)}\right\|_2}$ We have:

$$\left\|\tilde{h}_{u,1}^{(l)} - \tilde{h}_{u,0}^{(l)}\right\|_2 = \left\|\frac{1}{d_v+1}\left(g_v^{(l-1)} - \frac{1}{d_v}\sum_{u\in\overline{N}(v)\setminus\{v\}} g_u^{(l-1)}\right)\right\|_2 \tag{47}$$

$$\leq \frac{1}{2}\left(\left\|g_v^{(l-1)}\right\|_2 + \frac{1}{d_v}\sum_{u\in\overline{N}(v)\setminus\{v\}}\left\|g_u^{(l-1)}\right\|_2\right) \tag{48}$$

$$\leq \frac{1}{2}\left(\|W_l\|_{\mathrm{op}} + \frac{1}{d_v}\sum_{u\in\overline{N}(v)\setminus\{v\}}\|W_l\|_{\mathrm{op}}\right) \tag{49}$$

$$= \|W_l\|_{\mathrm{op}} \tag{50}$$

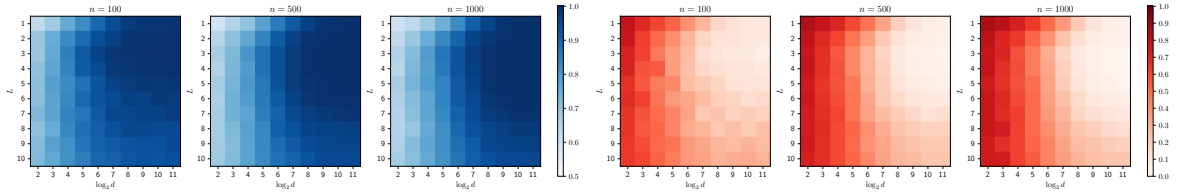
Analogously we have $\left\|\tilde{h}_{u,1}^{(l)} - \tilde{h}_{u,0}^{(l)}\right\|_2^2 \leq \|W_l\|_{\mathrm{op}}^2$ and thus $\mathrm{D}_{\mathrm{KL}}\left(H_1^{(l)} \parallel H_0^{(l)}\right) \leq \frac{\|W_l\|_{\mathrm{op}}^2}{\sigma^2}$. The result follows from adaptive composition as in (Mironov, 2017, Proposition 1).

The case of (GIN) This follows by combining the preceding argument with (Wu et al., 2023, Proposition 1).

The case of (GCN) This follows by combining the preceding argument with (Wu et al., 2023, Proposition 2).

The case of (SAGE-maxpool) The result follows from the following fact that $\left\|\max_{u\in\overline{N}(v)} g_u - \max_{u\in\overline{N}\setminus\{v\}} g_u\right\|_2$ attains its maximum when $g_v = -g_u, \forall u \in \overline{N}\setminus\{v\}$ since all the g_us are unit vectors.

□



(a) Measured in AUROC metric, darker color implies higher attacking performance

(b) Measured in ERR metric, lighter color implies higher attacking performance

Figure 6: Attacking efficacy of COSERA over sparse Erdős–Rényi graphs, with each grid's value indicating COSERA's performance measured in either AUROC (first row) or ERR (second row) metric.

Proof of theorem 6.1. We view the reconstruction problem regarding A_{uv} as a binary hypothesis testing problem

$$H_0 : A_{uv} = 0 \qquad \text{v.s.} \qquad H_1 : A_{uv} = 1. \tag{51}$$

Then according to hypothesis testing theory (Lehmann et al., 1986), we have

$$\inf_{\widehat{A}} \left[\mathbb{P}\left(\widehat{A}_{uv} = 1 | A_{uv} = 0 \right) + \mathbb{P}\left(\widehat{A}_{uv} = 0 | A_{uv} = 1 \right) \right] \geq 1 - d_{\mathrm{TV}}\left(\mathbf{H}_1, \mathbf{H}_0 \right), \tag{52}$$

where we use $d_{\mathrm{TV}}\left(\mathbf{H}_1, \mathbf{H}_0 \right)$ to denote the total variation distance of distributions induced by $\mathbf{H}_1$ and $\mathbf{H}_0$ respectively. By the Bretagnolle–Huber bound Canonne (2022, Theorem 1), we have

$$d_{\mathrm{TV}}\left(\mathbf{H}_1, \mathbf{H}_0 \right) \leq \sqrt{1 - \exp\left(-\mathrm{D}_{\mathrm{KL}}\left(\mathbf{H}_1 \parallel \mathbf{H}_0 \right) \right)} \tag{53}$$

The result then follows by combining (52), (53) and lemma B.9. □

C. Further experiments

C.1. Synthetic dataset with random GNN weights

The experimental setup in this section is basically the same as that in section 7.1, except that the model weights are generated by the following process: For an L-layer Linear GNN, we generate the weight matrix as:

$$W = W_1 \times \cdots \times W_L. \tag{54}$$

Here each $W_l, 1 \leq l \leq 10$ is a random matrix generated using the initialization method proposed in (He et al., 2015). The evaluations are shown in figure 6. The results exhibit a similar pattern to figure 1 where the weight matrix is set to identity. However, the attacking performance differs between the two scenarios: When the matrix W is poorly conditioned (a consequence of the construction (54)), the attacking performance degrades especially when the feature dimension d is not sufficiently large.

C.2. A complete report of privacy-utility assessments on Planetoid datasets

C.2.1. Training configurations and attacking pipeline

Network design For node v with label y_v, the prediction is defined as

$$\hat{y}_v = \arg\max_{c \in [C]} \mathsf{dec}\left(\mathsf{enc}\left(G, \mathbf{W} \right)[v] \right)[c], \tag{55}$$

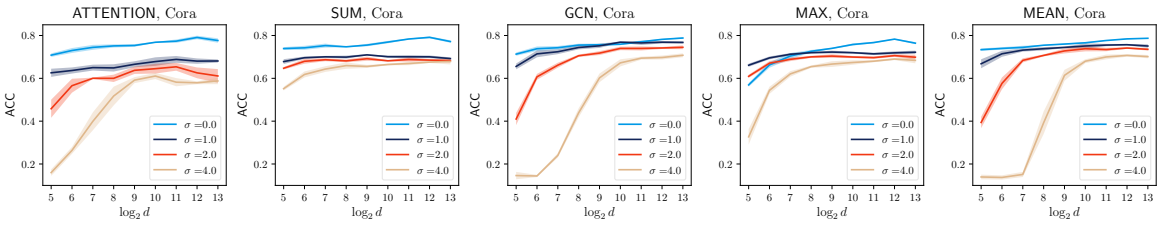
where we use $[\cdot]$ to denote the operation of vector index. Here the encoder enc is designed via stacking L noisy GNN layers (in the sense of NAG) with aggregation mechanism AGG $\in \{\mathsf{MEAN}, \mathsf{SUM}, \mathsf{GCN}, \mathsf{ATTENTION}, \mathsf{MAX}\}$ as defined above. Note that the encoder maps input node features into node representations of dimension d, which might be larger than the number of classes C. The decoder dec is a linear map that maps node representations to predictions.

Attacking paradigm The attacking procedure of COSERA will be based on the node representations produced by the GNN encoder enc under a dimension of d. The attack is conducted over the node representations corresponding to the test subset, i.e., the victim subgraph is the subgraph induced by the test nodes.

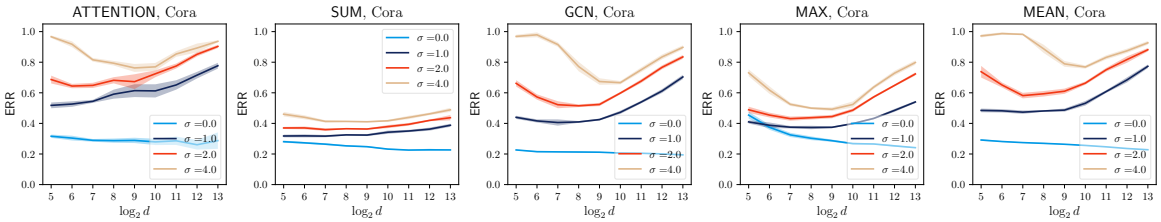
Training configurations Across all the experiments, we fix the GNN model to be of depth 2 and use full-batch training for 1000 steps(epochs) using the Adam optimizer with a learning rate of 0.001.

C.2.2. UNCONSTRAINED SCHEME

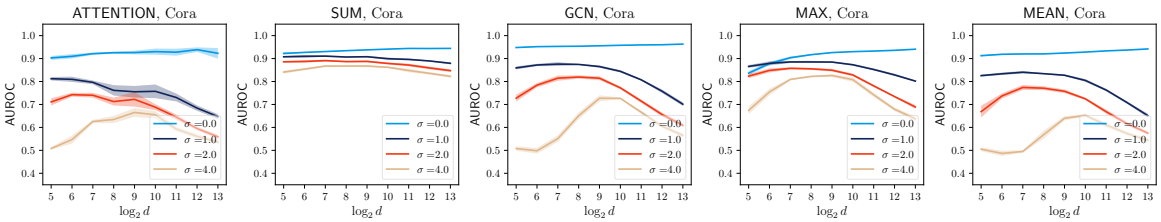
We plot the full experimental results under the unconstrained scheme for the Cora, Citeseer and Pubmed datasets in figure 7, figure 8 and figure 9, respectively, where we evaluate the performance of COSERA under both ERR and AUROC metrics. The result is consistent with those findings listed in section 7.2.



(a) GNN model performance over Cora dataset under 5 different aggregation types.



(b) Attacking performance of COSERA over Cora dataset (measured by ERR) under 5 different aggregation types.



(c) Attacking performance of COSERA over Cora dataset (measured by AUROC) under 5 different aggregation types.

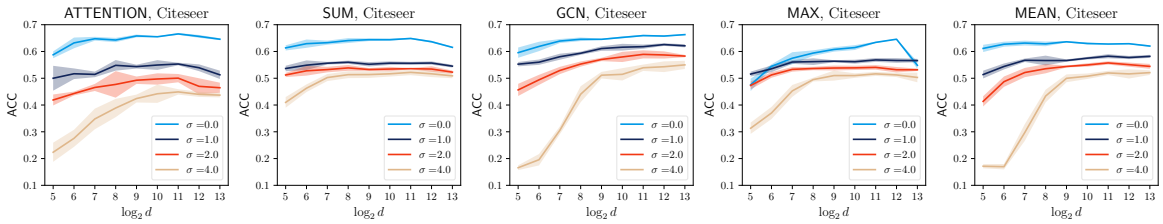
Figure 7: Privacy-utility trade-off on Cora dataset using the unconstrained training scheme. The horizontal axes measure feature dimension d in $\log_2$ scale and the vertical axes stands for performance measures All plots are based on 5 independent trials with shades indicating one standard deviation.

C.2.3. CONSTRAINED SCHEME

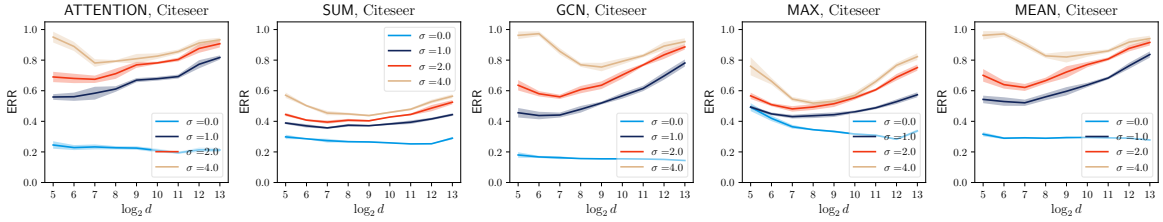
We plot the full experimental results under the constrained scheme for the Cora, Citeseer and Pubmed datasets in figure 10, figure 11 and figure 12, respectively, where we evaluate the performance of COSERA under both ERR and AUROC metrics. The result is consistent with those findings listed in section 7.2.

C.3. Spectrum study of GNN solutions obtained under the unconstrained scheme

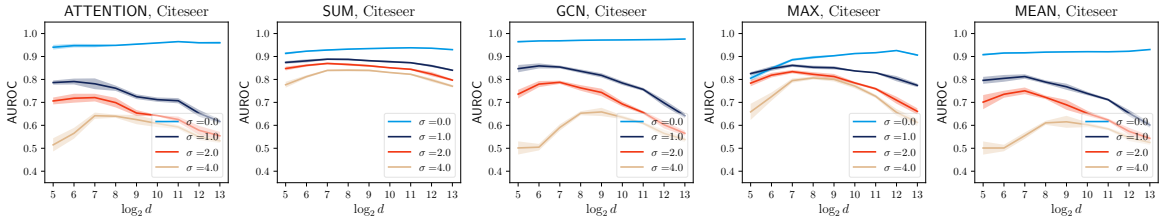
We plot the operator norms of the GNN layers in figure 13. The results demonstrate that the operator norms grow at a rapid rate with the increase of the feature dimension d, rendering strict privacy guarantee vacuous.



(a) GNN model performance over Citeseer dataset under 5 different aggregation types.



(b) Attacking performance of COSERA over Citeseer dataset (measured by ERR) under 5 different aggregation types.



(c) Attacking performance of COSERA over Citeseer dataset (measured by AUROC) under 5 different aggregation types.

Figure 8: Privacy-utility trade-off on Citeseer dataset using the unconstrained training scheme. The horizontal axes measure feature dimension d in $\log_2$ scale and the vertical axes stands for performance measures All plots are based on 5 independent trials with shades indicating one standard deviation.

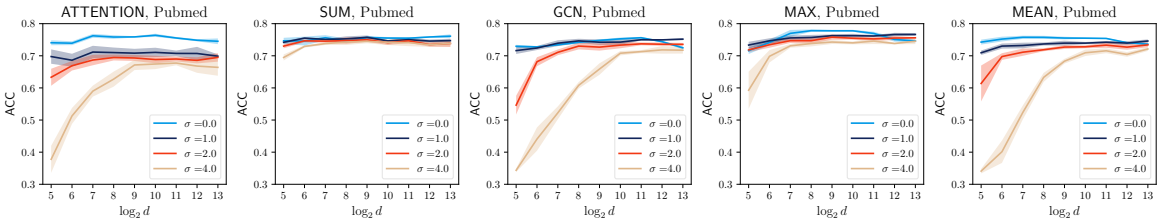
D. Discussions

D.1. Stronger adversary for dense graphs or deep encoders

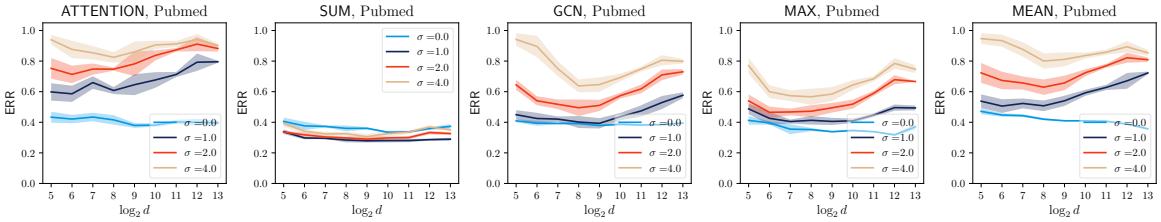
We have shown the limitations of COSERA over dense SBM graphs as well as deep GNN encoders. As our analysis applies to the specific COSERA adversary, it is thus of interest to ask whether there exists stronger attacking paradigms that is provably effective against dense graphs or deep GNN encoders. On the flipside, it is also valuable to understand whether the phenomenon of oversmoothing may fundamentally affect the performance of *any* black-box adversary.

D.2. Quantifying the advantage of adversaries with more knowledge

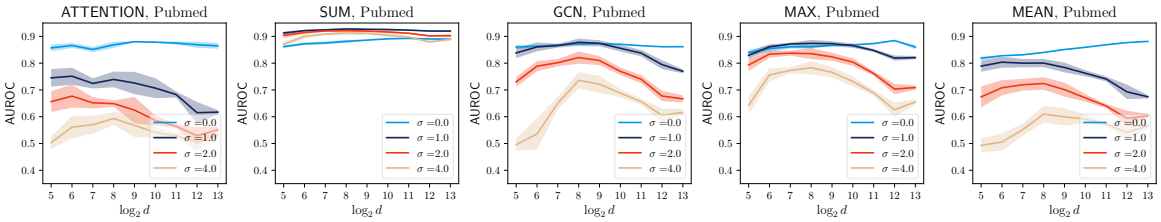
Despite its effectiveness, the knowledge available to COSERA is rather limited. Although previous study (He et al., 2021) has shown empirical evidences that equipping the adversary with more capability may results in stronger attacking algorithms, theoretical explication of these enhancements has yet to be articulated. In particular, it is of interest to quantify the amplification of adversarial capacity afforded by scenarios in which the adversary is granted white-box access to the model weights or node features.



(a) GNN model performance over Pubmed dataset under 5 different aggregation types.

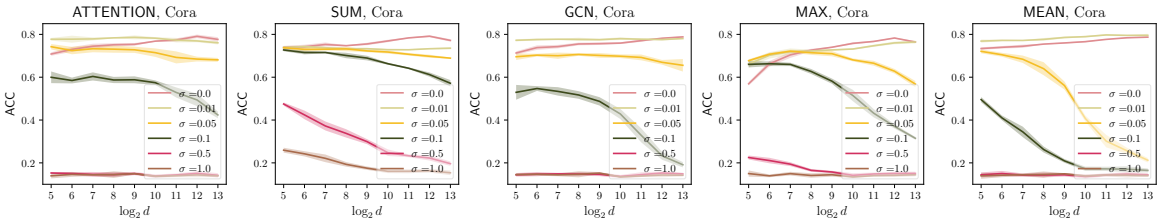


(b) Attacking performance of COSERA over Pubmed dataset (measured by ERR) under 5 different aggregation types.

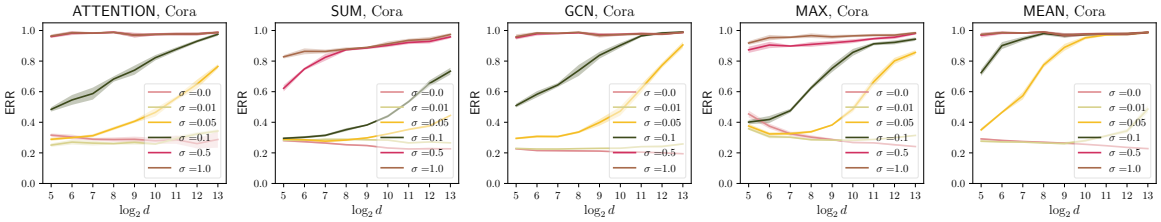


(c) Attacking performance of COSERA over Pubmed dataset (measured by AUROC) under 5 different aggregation types.

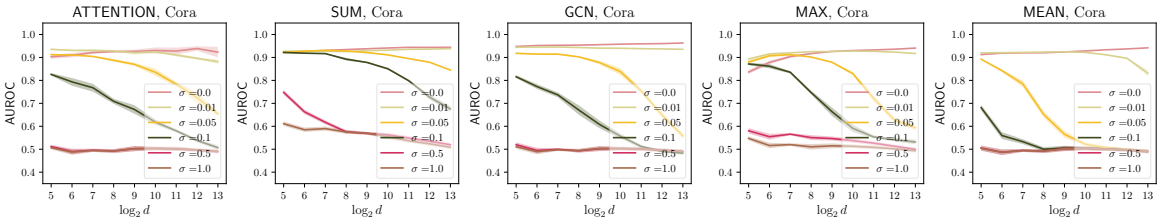
Figure 9: Privacy-utility trade-off on Pubmed dataset using the unconstrained training scheme. The horizontal axes measure feature dimension d in $\log_2$ scale and the vertical axes stands for performance measures All plots are based on 5 independent trials with shades indicating one standard deviation.



(a) GNN model performance over Cora dataset under 5 different aggregation types.

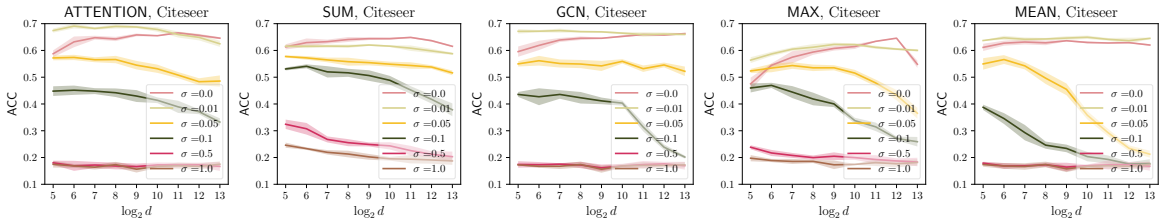


(b) Attacking performance of COSERA over Cora dataset (measured by ERR) under 5 different aggregation types.

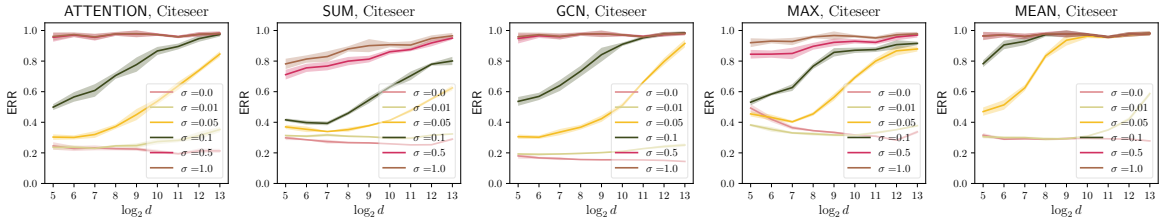


(c) Attacking performance of COSERA over Cora dataset (measured by AUROC) under 5 different aggregation types.

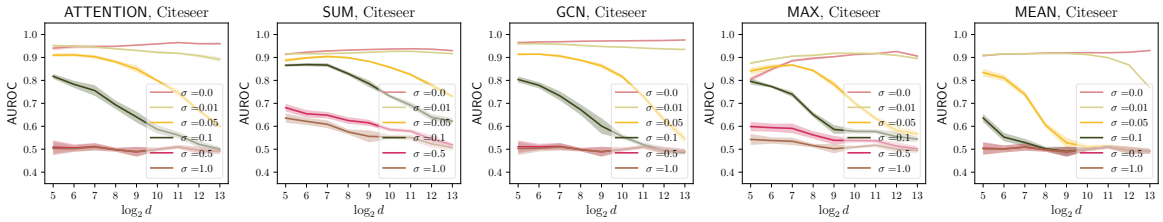
Figure 10: Privacy-utility trade-off on Cora dataset using the constrained training scheme. The horizontal axes measure feature dimension d in $\log_2$ scale and the vertical axes stands for performance measures All plots are based on 5 independent trials with shades indicating one standard deviation.



(a) GNN model performance over Citeseer dataset under 5 different aggregation types.

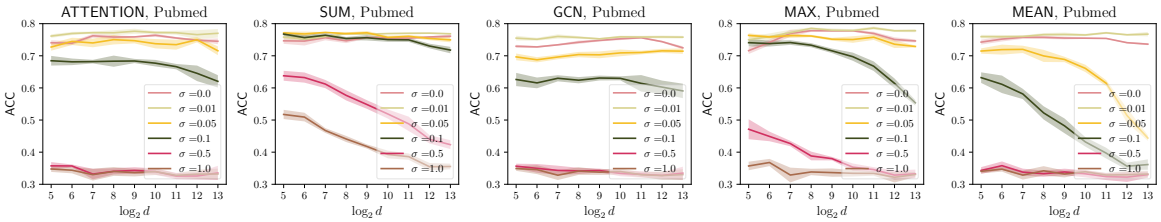


(b) Attacking performance of COSERA over Citeseer dataset (measured by ERR) under 5 different aggregation types.

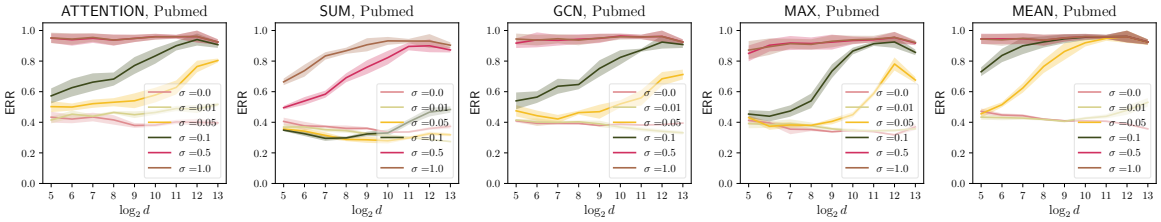


(c) Attacking performance of COSERA over Citeseer dataset (measured by AUROC) under 5 different aggregation types.

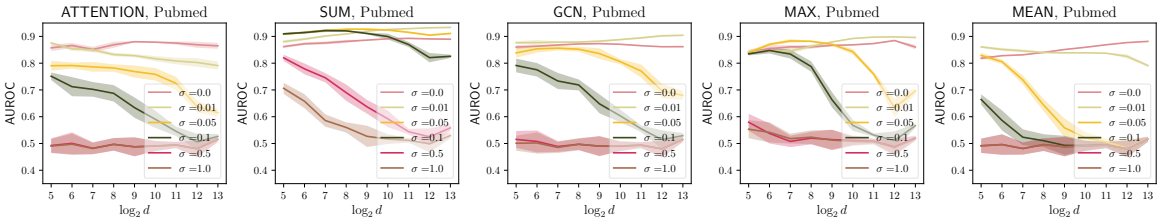
Figure 11: Privacy-utility trade-off on Citeseer dataset using the constrained training scheme. The horizontal axes measure feature dimension d in $\log_2$ scale and the vertical axes stands for performance measures All plots are based on 5 independent trials with shades indicating one standard deviation.



(a) GNN model performance over Pubmed dataset under 5 different aggregation types.

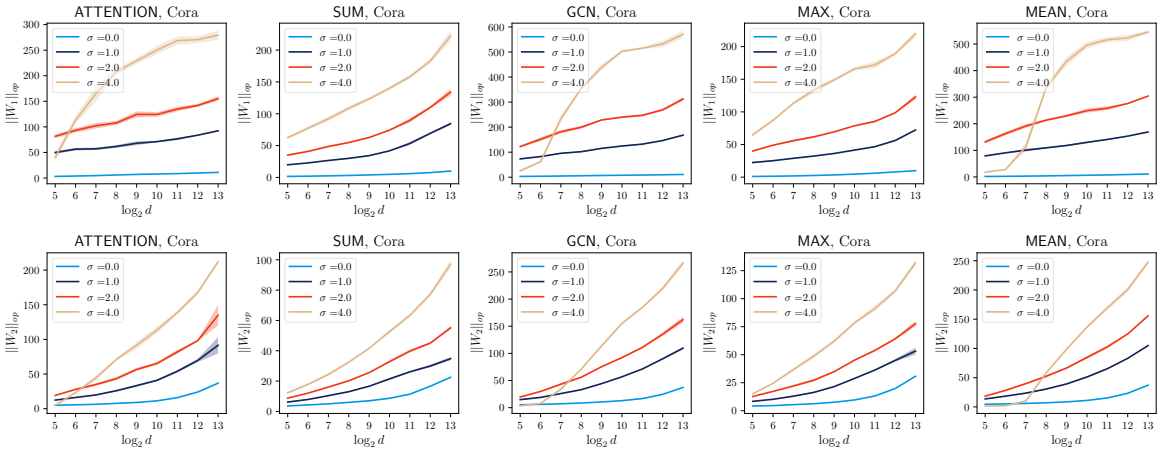


(b) Attacking performance of COSERA over Pubmed dataset (measured by ERR) under 5 different aggregation types.

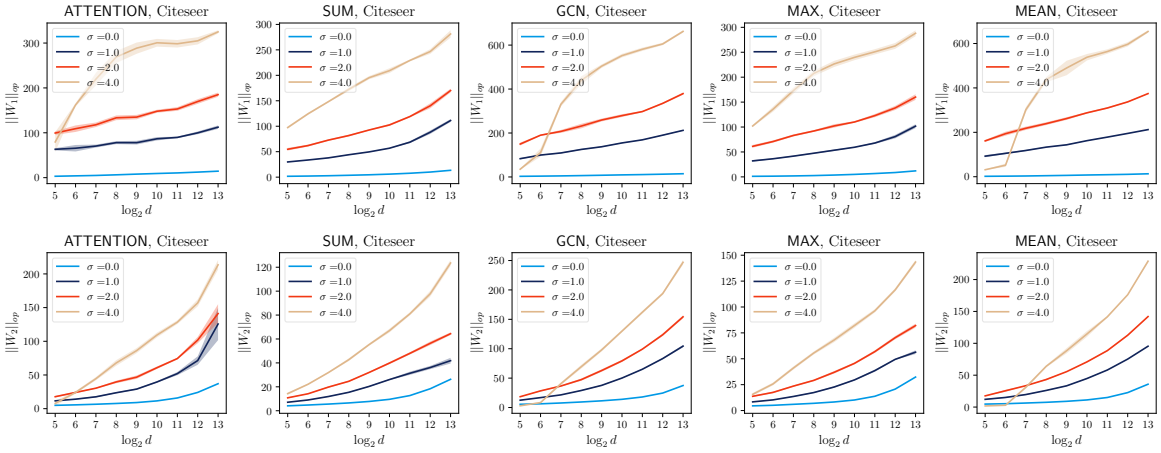


(c) Attacking performance of COSERA over Pubmed dataset (measured by AUROC) under 5 different aggregation types.

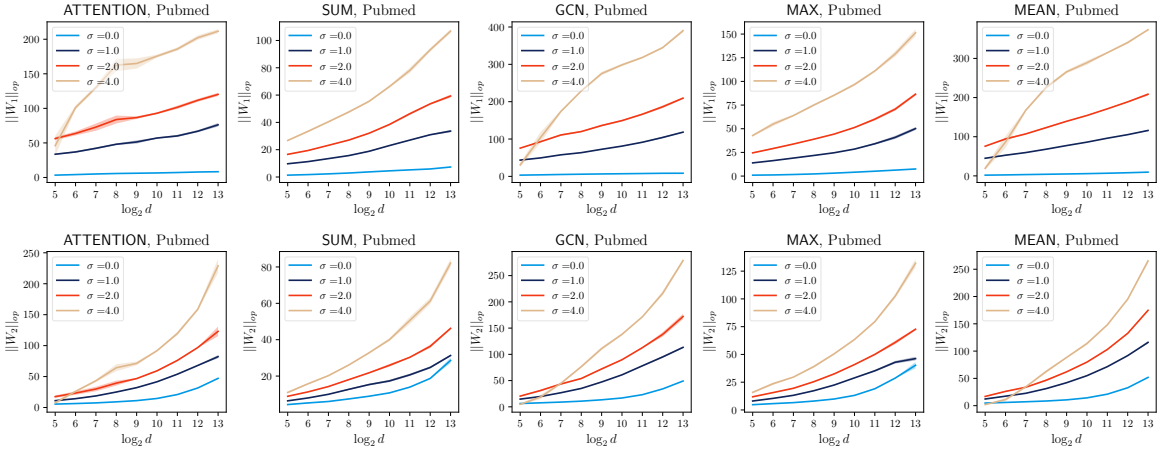
Figure 12: Privacy-utility trade-off on Pubmed dataset using the constrained training scheme. The horizontal axes measure feature dimension d in $\log_2$ scale and the vertical axes stands for performance measures All plots are based on 5 independent trials with shades indicating one standard deviation.



(a) Spectrum study on the Cora dataset under 5 different aggregation types.



(b) Spectrum study on the Citeseer dataset under 5 different aggregation types.



(c) Spectrum study on the Pubmed dataset under 5 different aggregation types.

Figure 13: Spectrum study on the Planetoid datasets under the unconstrained training scheme. The horizontal axes measure feature dimension d in $\log_2$ scale and the vertical axes measures the operator norm of the projection weights of the GNN. All plots are based on 5 independent trials with shades indicating one standard deviation.