Private and Expressive Graph Representations

Patrick Indri, Tamara Drucks, and Thomas Gärtner

Machine Learning Research Unit, TU Wien, Vienna, Austria,
{firstname.lastname}@tuwien.ac.at

Abstract. We propose using homomorphism density vectors to obtain graph embeddings that are both private and expressive. Homomorphism densities are provably highly discriminative and offer a powerful tool for distinguishing non-isomorphic graphs. By adding noise calibrated to each density's sensitivity, we ensure that the resulting embeddings satisfy formal differential privacy guarantees. Our construction preserves expressivity in expectation, as each private embedding remains unbiased with respect to the true homomorphism densities. We study the trade-off between privacy, utility, and expressivity, both theoretically and empirically, and show that our private embeddings match the accuracy of their non-private counterparts with increased resilience to privacy attacks.

Keywords: Graph Machine Learning · Differential Privacy · Expressivity.

1 Introduction

We investigate the tradeoff between expressivity and privacy in the context of graph representation learning. In graph learning, expressivity analysis studies the ability to distinguish pairs of non-isomorphic graphs, while private algorithms ensure that similar graphs yield similar outputs. Therefore, requiring algorithms to be both expressive and private can be challenging, as there may be a tension between these desiderata. So far, there has been little investigation towards a better theoretical understanding of the trade-offs which exist between privacy, expressivity, and utility, i.e., predictive performance. We fill this research gap and propose to guarantee expressivity *in expectation* while using *differential privacy* (DP) to provide privacy guarantees. We focus on *graph-level* learning tasks while providing *edge-level* privacy guarantees. We build upon existing work that relies on homomorphism counts, either as standalone graph representations or to increase the expressive power of GNNs [23,34,12]. Specifically, we use homomorphism densities, the normalized homomorphism counts, as graph embeddings. We obtain homomorphism density vectors that are DP and expressive in expectation. Empirically, our private embeddings achieve performance comparable to that of their non-private counterpart, while being more resilient to privacy attacks. An advantage of our approach is that the private homomorphism embedding we obtain can be then used for any downstream task without incurring further privacy cost, thanks to the post-processing property of DP [7].

To summarize, we address the following research questions: Can we construct graph embeddings based on noisy homomorphism density vectors that satisfy formal DP guarantees while remaining, in expectation, as expressive as their non-private counterparts? How does the choice of privacy budget quantitatively impact the utility of these embeddings on downstream tasks? What is the trade-off between the privacy guarantees and the expressivity of these embeddings?

2 Related work

Recent literature in graph representation learning has analyzed the *expressive power* of learning algorithms, i.e., their ability to learn different representations for non-isomorphic graphs. One line of work analyzes the expressive power of GNNs through the lens of k-Weisfeiler-Leman (k-WL) tests, a hierarchy of increasingly expressive color refinement algorithms [37,19]. Different approaches [23,34] rely on graph representations built using homomorphism counts [1,17] to obtain arbitrarily expressive representations, at least in expectation [23,34]. While some recent work has investigated the interactions between expressivity and robustness in GNNs [3,14], there is a lack of research on the relationship between privacy and expressivity, from both a theoretical and an empirical standpoint. The structural information in graphs is, in fact, often of sensitive nature. Privacy attacks can target the edges [27] or the nodes [13] of a graph, which should therefore be protected [20,15,39]. Graph reconstruction attacks can effectively recover the sensitive information from trained models [40,36] and a number of DP graph learning approaches have therefore been proposed [29,31,30,32,26]. More specifically, recent work has considered the problem of private subgraph counting, with a focus on triangle counting [4,11,22]. Nevertheless, as recently highlighted in Sajadmanesh et al. [32], there is a lack of research efforts that investigate the expressive power of DP graph learning algorithms.

3 Preliminaries

In this section, we introduce the relevant preliminaries on graph homomorphisms, expressivity, and differential privacy, with more details in Appendix A.

3.1 Graph theory

Let $G = (V, E) \in \mathcal{G}$ be a graph where $\mathcal{G}$ is the set of graphs with bounded number of nodes. G has node set $V(G)$ with $|V(G)| = n$ and edge set $E(G)$ with cardinality $e(G) = |E(G)|$. For two sets $S, T \subseteq V(G)$, let $e_G(S, T)$ denote the number of edges with one endpoint in S and one endpoint in T. For a graph G with n nodes and adjacency matrix A_G, let $\|A_G\|_1 = \frac{1}{n^2}\sum_{i,j}^{n} |A_{ij}|$ be the ℓ_1 norm of A_G. We refer to a graph $F \in \mathcal{F} \subseteq \mathcal{G}$ as a *pattern* when we compute the homomorphisms from F to some graph G. Given two graphs F, G, a *homomorphism* from F to G is an adjacency-preserving map $\psi : V(F) \to V(G)$. We call ψ an *isomorphism* in case it is bijective. For two graphs $G, G' \in \mathcal{G}$, let $G \simeq G'$ denote that the two graphs are *isomorphic*.

Definition 1 (Homomorphism density). *Let* $\hom(F, G)$ *denote the number of homomorphisms from* F *to* G*. Then, we define the homomorphism density as*

$$t(F, G) = \frac{\hom(F, G)}{|V(G)|^{|V(F)|}}.$$

For a given vector of patterns $\boldsymbol{F} = (F_1, \ldots, F_d)$ we can consider the homomorphism density vector $\boldsymbol{t}(\boldsymbol{F}, G) := (t(F_1, G), \ldots, t(F_d, G))$.

Definition 2 (Edge edit distance [17,9]). *For two graphs* G, G' *with the same number of nodes, the edge distance* d_{edge} *is defined as*

$$d_{edge}(G, G') = \frac{1}{2} \|A_G - A_{G'}\|_1 . \tag{1}$$

Definition 3 (Cut distance [17]). *For two graphs* G, G' *with the same number of nodes, the cut distance* $d_\square$ *is defined as*

$$d_\square(G, G') = \max_{S,T \subseteq V(G)} \frac{|e_G(S,T) - e_{G'}(S,T)|}{n^2}. \tag{2}$$

It holds that $d_\square(G, G') \leq 2d_{\text{edge}}(G, G')$ [17]. The *counting lemma* provides an upper bound on the absolute difference in the homomorphism densities of two graphs with respect to the same pattern.

Lemma 1 (Counting Lemma [17]). *For any three simple graphs* F*,* G*, and* G'*, with* G, G' *having the same number of nodes,*

$$|t(F, G) - t(F, G')| \leq e(F) d_\square(G, G'). \tag{3}$$

As presented by Lovász [17, Lemma 10.22], the counting lemma relies on a slightly different notion of cut distance which allows to consider graphs with node sets of different cardinalities. We provide further details in Appendix A.

3.2 Completeness and expressivity in graph learning

The expressive power of graph learning algorithms is commonly measured as their ability to distinguish between pairs of non-isomorphic graphs. Let $\varphi : \mathcal{G} \to \mathbb{R}^d$ be a *graph embedding*. We assume φ to be permutation invariant, i.e., that for all $G, G' \in \mathcal{G}$, $G \simeq G'$ implies $\varphi(G) = \varphi(G')$. This is trivially true for homomorphism counts and homomorphism densities. A seminal result by Lovász asserts that homomorphism counts enjoy strong distinguishing properties.

Theorem 1 (Lovász, [16]). *Two graphs* G, G' *are isomorphic if and only if* $\hom(F, G) = \hom(F, G')$ *for all simple graphs* F*.*

We are interested in the ability of an embedding to distinguish non-isomorphic graphs and thus introduce the concept of *completeness* as follows.

Definition 4 (Completeness). *An embedding $\varphi : \mathcal{G} \to \mathbb{R}^d$ is complete if for all $G, G' \in \mathcal{G}$, $G \simeq G'$ if and only if $\varphi(G) = \varphi(G')$.*

We next introduce an *expressivity* notion where we restrict our patterns to some specific graph class $\mathcal{F} \subseteq \mathcal{G}$.

Definition 5 ($\mathcal{F}$-expressivity). *An embedding $\varphi : \mathcal{G} \to \mathbb{R}^d$ is $\mathcal{F}$-expressive if, for all $G, G' \in \mathcal{G}$ and for all $F \in \mathcal{F}$, $\hom(F, G) = \hom(F, G')$ if and only if $\varphi(G) = \varphi(G')$.*

Consider now an embedding parametrized by a random variable $X \sim \mathcal{D}$ for some distribution $\mathcal{D}$ and denote it by $\varphi_X : \mathcal{G} \to \mathbb{R}^d$. We introduce notions of completeness and expressivity *in expectation* as follows.

Definition 6 (Expectation-completeness). *An embedding $\varphi_X : \mathcal{G} \to \mathbb{R}^d$ is expectation-complete if the embedding $\mathbb{E}_X[\varphi_X]$ is complete.*

Definition 7 ($\mathcal{F}$-expectation-expressivity). *An embedding $\varphi_X : \mathcal{G} \to \mathbb{R}^d$ is $\mathcal{F}$-expectation-expressive if the embedding $\mathbb{E}_X[\varphi_X]$ is $\mathcal{F}$-expressive.*

An expectation-complete embedding is also $\mathcal{F}$-expectation-expressive for any graph class $\mathcal{F}$.

3.3 Differential privacy

Differential privacy (DP) is a formal notion of privacy that protects individual training points. DP is defined in terms of *neighboring databases*. A *database* is a collection of *points*, where a point in a database may be, e.g., a row in a table or an edge in a graph. Two databases x, x' are *neighboring* if they differ in a single point, that is, if one single point is present in one database but not in the other. We denote this as $x \sim x'$. DP guarantees that an attacker cannot confidently determine from which of two neighboring databases the output of a DP *mechanism* has been obtained from. We introduce two notions of DP and briefly describe how to achieve DP according to these notions.

Definition 8 ((ϵ, δ)-DP, [6]). *Let $\epsilon > 0$ and $\delta \in [0, 1)$. A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies δ-approximate ϵ-indistinguishability differential privacy, denoted as (ϵ, δ)-DP, if, for all neighboring $x, x' \in \mathcal{X}$,*

$$\Pr[\mathcal{M}(x) \in \mathcal{Y}] \leq e^{\epsilon} \Pr[\mathcal{M}(x') \in \mathcal{Y}] + \delta, \tag{4}$$

where probabilities are taken over the randomness of $\mathcal{M}$.

In DP, we refer to ϵ as the *privacy budget* of a mechanism, with larger values of ϵ providing less privacy, and a value of $\epsilon = 0$ providing perfect privacy. To make a given function f private, one can add noise proportional to its *global sensitivity* $GS_f = \max_{x \sim x'} \|f(x) - f(x')\|$. Appendix A.2 provides a quick overview of how to use noise proportional to GS_f to achieve DP for f. The standard DP mechanism described in Appendix A.2 that relies on noise proportional to the global sensitivity can however result in very poor performance. GS_f considers in fact the worst case behavior of f around an arbitrary point x, even though the sensitivity of f around most of the points of interest may be smaller.

A distributional flavor of DP can be formalized in terms of the divergence of a randomized mechanism when applied to two neighboring databases.

Definition 9 ((ρ, ω)-tCDP, [2]). *Let $\rho > 0$ and $\omega > 1$. Let $D_\alpha(\cdot \,\|\, \cdot)$ denote the Rényi divergence of order α [28,33]. A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies ω-truncated ρ-concentrated differential privacy, denoted as (ρ, ω)-tCDP, if, for all neighboring $x, x' \in \mathcal{X}$, for all $\alpha \in (1, \omega)$ it holds that*

$$D_\alpha(\mathcal{M}(x) \,\|\, \mathcal{M}(x')) \leq \rho\alpha. \tag{5}$$

Definition 8 and Definition 9 can be formally related as tCDP implies (ϵ, δ)-DP (see Lemma 3 in Appendix A.2). It is convenient to consider tCDP as, in contrast to the standard mechanisms described in Appendix A.2, it allows to achieve DP while considering a *local* notion of sensitivity for a function f at a point x.

Proposition 1. *(tCDP with Gaussian noise, [2]) Let $f, g : \mathcal{X} \to \mathbb{R}$ satisfy, for every pair of neighboring databases $x, x' \in \mathcal{X}$ and for $\varDelta_f, \varDelta_g \geq 0$,*

$$|f(x) - f(x')| \leq \varDelta_f \cdot e^{g(x)/2}, \quad |g(x) - g(x')| \leq \varDelta_g. \tag{6}$$

Let $\mathcal{M}(x) = f(x) + \mathcal{N}(0, e^{g(x)})$. The randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathbb{R}$ satisfies $(\varDelta_f^2 + \varDelta_g^2, \frac{1}{2\varDelta_g})$-tCDP.

In Proposition 1, $\varDelta_f \cdot e^{g(x)/2}$ is a *smooth* upper bound on the *local sensitivity* of f in x. This is consistent with the smooth sensitivity framework introduced by Nissim et al. [24], which we describe in Appendix A.2.

4 Private and expressive homomorphism densities

In this section, we propose a general method to obtain private and, in expectation, expressive graph representations. We present an informal version of our main result as follows. We defer all missing proofs to Appendix B.

Theorem 2 (Informal). *Let $\mathcal{D}$ be a distribution on $\mathcal{F} \subseteq \mathcal{G}$ with full support. Let $G \in \mathcal{G}$ be a graph and $\boldsymbol{F} = (F_1, \ldots, F_d) \sim \mathcal{D}^d$ be a vector of patterns. Let*

$\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ be the noisy homomorphism density vector obtained by adding Gaussian noise to $\boldsymbol{t}(\boldsymbol{F}, G)$. Then, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is differentially private and, for large enough d, $\mathcal{F}$-expressive with high probability. If we have $\mathcal{F}^d = \mathcal{G}^d$, then $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is also complete with high probability.

Our graph embedding can be used for any downstream graph learning task without incurring further privacy cost, thanks to the post-processing property of DP [7]. Moreover, we can control the expressive power of our embedding by considering different graph classes $\mathcal{F}$ and precisely characterize the trade-off between privacy and expressivity. In the next sections, we discuss how we obtain our privacy and expressivity guarantees.

4.1 Privacy guarantees

In this section, we provide DP guarantees for the homomorphism density embeddings. First, we discuss how to bound the sensitivity of $\boldsymbol{t}(\boldsymbol{F}, G)$, which is essential to determine how much noise must be added to the vector for provable privacy. Then, we show how to obtain a tCDP homomorphism density embedding.

We are interested in obtaining a DP embedding of a graph G by means of its homomorphism density $t(F, G)$ for a given pattern F. In most of the following discussion, we may consider F to be fixed, but we remark that to achieve completeness in expectation (see Section 4.2) the patterns are sampled from a distribution as $F \sim \mathcal{D}$. We focus on *edge* privacy, and strive to protect the presence/absence of individual edges in a graph. We then interpret neighboring graphs, according to the following definition, as two neighboring databases.

Definition 10 (Neighboring graphs). *Two graphs G, G' with the same number of nodes are* neighboring *graphs, written $G \sim G'$, if $d_{edge}(G, G') = 1$.*

A first bound to the *global* sensitivity of the homomorphism densities can be directly obtained from the counting lemma.

Corollary 1. *For any two neighboring graphs $G \sim G'$ with n nodes and for any pattern F it holds that*

$$|t(F, G) - t(F, G')| \leq e(F) d_{\square}(G, G') = \frac{2e(F)}{n^2}. \tag{7}$$

Proof. The proof follows from Lemma 1 and Definition 3 by direct computation, with the reminder that $e_G(S, S) = 2e(S)$ for any $S \subseteq V(G)$.

In most practical settings the amount of noise one needs to add to obtain DP guarantees using the bound in Corollary 1 is too large. In many cases, domain knowledge allows to assume that the degree of the graphs is bounded and thus to obtain a smaller bound on the sensitivity of the homomorphism densities.

Theorem 3 (Sensitivity of homomorphism density for bounded degree graphs). *Let $G \sim G'$ be two graphs with n nodes and maximum degree $\Delta_{\max}$. For any pattern F with $m > 1$ nodes, it holds that*

$$|t(F,G) - t(F,G')| \leq \frac{2e(F)}{n^2} \left(\frac{\Delta_{\max}}{n}\right)^{m-2}. \tag{8}$$

For large graphs and large patterns $\left(\frac{\Delta_{\max}}{n}\right)^{m-2} \ll 1$. Therefore, the bound provided by Equation (8) is often tighter in practice than the one we could obtain from the counting lemma as in Corollary 1. We recover the upper bound provided by the counting lemma by setting $\Delta_{\max} = n$. As a corollary, we can obtain an upper bound for the difference of the homomorphism density vector of two graphs G, G' with a given distance $d_{\text{edge}}(G, G') = k \geq 1$.

Corollary 2. *Let $G \sim G'$ be two graphs with n nodes, $d_{edge}(G, G') = k$, and maximum degree $\Delta_{\max}$. For any pattern F with $m > 1$ nodes, it holds that*

$$|t(F,G) - t(F,G')| \leq \frac{2ke(F)}{n^2} \left(\frac{\Delta_{\max}}{n}\right)^{m-2}. \tag{9}$$

For a given vector of patterns $\boldsymbol{F}$, we can use Theorem 3 to compute an upper bound on the global sensitivity of $\boldsymbol{t}(\boldsymbol{F}, G)$ by taking, e.g., the maximum over the patterns in $\boldsymbol{F}$. With this upper bound on the global sensitivity, we obtain DP guarantees with the standard Gaussian mechanism as in Appendix A.2. In practice, however, the global sensitivity computed in this way is much too large to obtain useful embeddings as it often has a magnitude comparable to that of the homomorphism densities themselves. To provide practically useful bounds, we set to obtain bounds that rely on a *local* notion of sensitivity.

We present the notions of local and smooth sensitivities (see also Appendix A) for the homomorphism densities for a given graph G. For each pattern F, the local sensitivity of the homomorphism density $t(F, G)$ for k edge edits can be written as $LS^{(k)}_{t,F}(G) = \max_{G' \in \mathcal{G}: d_{\text{edge}}(G,G') \leq k} |t(F,G) - t(F,G')|$. Let $\beta > 0$. The *β-smooth* sensitivity of $t(F, G)$ in G can then be written as

$$S_{t,F} = \max_{k \geq 0} \left(e^{-\beta k} \cdot LS^{(k)}_{t,F}(G)\right). \tag{10}$$

For $G \in \mathcal{G}$, it holds that $LS^{(1)}_{t,F}(G) \leq GS_{t,F} = \max_{G \sim G'} \|t(F,G) - t(F,G')\|$. We thus expect a method that relies on smooth sensitivities to provide better utility compared to one that relies on global sensitivities. The pattern-wise smooth sensitivity can be used to upper-bound the smooth sensitivity of $\boldsymbol{t}(\boldsymbol{F}, G)$.

Proposition 2. *Let $S^*_t(G) = \|S_{t,F_1}, \ldots, S_{t,F_d}\|_2$ and $\beta > 0$. Let*

$$S_t(G) = \max_{k \geq 0} \left(e^{-\beta k} \max_{G' \in \mathcal{G}: d_{edge}(G,G') \leq k} \|\boldsymbol{t}(\boldsymbol{F}, G) - \boldsymbol{t}(\boldsymbol{F}, G')\|_2\right) \tag{11}$$

be the β-smooth sensitivity of $\boldsymbol{t}(\boldsymbol{F}, G)$ at G. Then, it holds that $S^*_t(G) \geq S_t(G)$.

For each given pattern F we can use Theorem 3 and Corollary 2 to upper bound the smooth sensitivity of $t(F, G)$, and use Proposition 1 to obtain a private homomorphism density value. As we are interested in a private version of $\boldsymbol{t}(\boldsymbol{F}, G) \in \mathbb{R}^d$, we need to derive a d-dimensional version of Proposition 1 to take advantage of the bound in Proposition 2 for the entire density vector.

Theorem 4 (tCDP with Gaussian noise in $\mathbb{R}^d$). *Let $f : \mathcal{X} \to \mathbb{R}^d$ and $g : \mathcal{X} \to \mathbb{R}$ satisfy, for every pair of neighboring databases x, x' and for $\Delta_f, \Delta_g \geq 0$,*

$$\|f(x) - f(x')\|_2 \leq \Delta_f e^{g(x)/2}, \quad |g(x) - g(x')| \leq \Delta_g. \tag{12}$$

Let $\mathcal{M}(x) = f(x) + \mathcal{N}\left(0, e^{\,g(x)}\, I_d\right)$. The randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathbb{R}^d$ satisfies $\left(\Delta_f^2 + d \cdot \Delta_g^2, \frac{1}{2\Delta_g}\right)$-tCDP.

With Theorem 4, we obtain a private mechanism scaled to the smooth sensitivity of the homomorphism density of each graph using Gaussian noise.

Theorem 5. *Let $\mathbf{t}(\boldsymbol{F}, G)$ be the homomorphism density vector for graph G and pattern set $\boldsymbol{F}$ with $|\boldsymbol{F}| = d$, $\rho' > 0$, and $S_t^*(G)$ be a β-smooth upper bound to the local sensitivity as per Proposition 2. Then, the mechanism*

$$\tilde{\boldsymbol{t}}(\boldsymbol{F}, G) = \boldsymbol{t}(\boldsymbol{F}, G) + \mathcal{N}\left(\mathbf{0}, \frac{[S_t^*(G)]^2}{2\rho'} I_d\right). \tag{13}$$

is $\left(2\rho' + d \cdot 4\beta^2, \frac{1}{4\beta}\right)$-tCDP for neighboring graphs as per Definition 10.

As discussed, the smooth sensitivities are by definition upper-bounded by the global sensitivity. Therefore, we expect the amount of noise added by the procedure in Theorem 4 to be significantly smaller than the one we would need to apply for the standard Gaussian mechanism described in Appendix A.2, leading to a better privacy-utility trade-off.

4.2 Expressivity

In this section, we show that the private homomorphism densities we obtain with Theorem 5 are, in expectation, complete and therefore expressive. Furthermore, we discuss how to achieve a determined level of expressivity and characterize the trade-off between privacy and expressivity.

As a first observation, recall that the notion of completeness introduced in Definition 4 assumes permutation invariance. It is easy to see that $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is *not* permutation invariant, a necessary consequence of the fact that DP requires a *randomized* mechanism. This observation, however, does not affect the possibility to obtain expressive or even complete graph embeddings *in expectation*.[1]

[1] For these results, we need to further premise the fact that homomorphism densities, in contrast to homomorphism counts, do not distinguish G and a *blowup* of G. This subtlety stems from our notational conventions and can be easily addressed while maintaining DP. We address this technicality in Appendix B.

Theorem 6. *For any $G \in \mathcal{G}$, $\tilde{t}(F, G)$ is $\mathcal{F}$-expectation-expressive for $F \sim \mathcal{D}$ if $\mathcal{D}$ has full support on $\mathcal{F} \subseteq \mathcal{G}$. If $\mathcal{F} = \mathcal{G}$, then $\tilde{t}(F, G)$ is expectation-complete.*

In our setting, we sample a vector of patterns and thus show the following.

Theorem 7. *Let $\mathcal{D}$ be a distribution on $\mathcal{F} \subseteq \mathcal{G}$ with full support. Let $G \in \mathcal{G}$, $\boldsymbol{F} \sim \mathcal{D}^d$, and $\theta \in [0, 1]$. For large enough d, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is $\mathcal{F}$-expressive with probability at least $1 - \theta$. If $\mathcal{F}^d = \mathcal{G}^d$, then, for large enough d, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is complete with probability at least $1 - \theta$.*

Theorem 6 and Theorem 7 demonstrate that, despite the noise required for DP, our homomorphism density embeddings retain full discriminative power in expectation and, with enough patterns, with high probability.

We can characterize the expressive power of our embedding more precisely by recalling that expectation-completeness implies $\mathcal{F}$-expectation-expressivity for all $\mathcal{F}$ (see Section 3.2). For a specific graph class $\mathcal{F}$, the expectation-complete embedding $t(F, G)$ therefore also distinguishes, in expectation, exactly those graphs which are distinguished by $F \in \mathcal{F}$. We thus propose to sample from a graph class that precisely determines a certain level of expressivity in expectation [23]. For instance, it is well known that 1-WL [37,19] is equivalent to representing graphs via their homomorphism count vectors restricted to *trees* as patterns. In other words, two graphs have the same color multiset [37] if and only if they have the same homomorphism counts for all trees. This equivalence can be generalized for many popular GNN architectures by determining their *homomorphism-distinguishing closed* graph class [38,21]. For instance, the k-WL hierarchy corresponds to the homomorphism-distinguishing closed graph classes of treewidth k [19,18].

Table 1. Common GNNs with their homomomorphism-distinguishing closed graph classes $\mathcal{F}$ and their maximum edge counts for a given number of nodes. For more details, see Zhang et al. [40] and Appendix B.3.

GNN	Graph class $\mathcal{F}$	$\max_{F \in \mathcal{F}, m=\lvert V(F)\rvert} e(F)$
MPNNs (1-WL)	Trees	$m - 1$
r-ℓMPNNs (r-ℓWL) [25]	Fan-cactus graphs	$2m - 3$
Subgraph k-GNNs	$\{F : \exists U \subset V(F)$ s.t. $\lvert U\rvert \leq k$ and $F \setminus U$ is a forest$\}$	$m(k + 1) - 1 - \frac{k^2+3k}{2}$
k-FGNNs (k-WL)	$\{F : \mathrm{tw}(F) \leq k\}$	$km - \frac{1}{2}k(k + 1)$

In Table 1, we provide the graph class $\mathcal{F}$ and its maximum number of edges for some well-known GNN architectures that have expressive power precisely characterized by $\mathcal{F}$, i.e., that can distinguish all non-isomorphic graphs in $\mathcal{F}$. We can therefore determine the $\mathcal{F}$-expectation-expressivity of our embedding and illustrate the trade-off between expressivity and the noise required for a given privacy guarantee as follows.

Proposition 3. *Fix a tCDP privacy parameter $\rho' > 0$ and a graph G with n nodes. Let $\mathcal{F}$ be a class of patterns. The Gaussian noise necessary to obtain the tCDP guarantee in Theorem 4 has variance $\sigma^2 = \mathcal{O}\left((\max_{F\in\mathcal{F}} e(F))^2/n^4\right)$.*

Proof. From Theorem 3, the local sensitivity of each pattern is $\mathcal{O}(e(F)/n^2)$. The vector-wise smooth sensitivity in Proposition 2 is, in turn, not smaller than the largest local sensitivity and is therefore $S_t^*(G) = \mathcal{O}\left(\max_{F\in\mathcal{F}} e(F)/n^2\right)$. For a fixed ρ', the variance of the noise in Theorem 5 is $\sigma^2 = \mathcal{O}\left((\max_{F\in\mathcal{F}} e(F))^2/n^4\right)$.

With reference to Table 1, more expressive GNN architectures often have greater bounds on $e(F)$ for $F \in \mathcal{F}$. From Proposition 3 we can therefore conclude that with patterns sampled from more expressive graph classes, more noise is required to achieve a given privacy guarantee. Thus, we have identified an explicit trade-off between privacy and expressivity.

5 Experiments

We evaluate the private homomorphism density vectors obtained with our approach by using them to perform a graph classification task on three commonly used `OGBG` benchmark datasets: `MOLHIV`, `MOLBACE`, and `MOLTOX21` [10]. Our goal is to make the trade-off between the desiderata of privacy, utility, and expressivity explicit, as well as to show that our private homomorphism densities can provide a good balance between these desiderata. As we focus on how the graph structure can be privately leveraged to have expressive representations, our experiments rely on the graph structure *only* as encoded by the homomorphism density vectors and do not consider any node or edge features. For our results, we experiment with values $\rho' \in [10^{-8}, 1]$ and pick $\beta = \rho'/5$. We upper bound smooth sensitivities by evaluating Equation (10) up to $k = 6$. For visualization purposes, we convert our tCDP guarantees into (ϵ, δ)-DP guarantees using Lemma 3; (ϵ, δ)-DP guarantees are easier to interpret as privacy budgets roughly in the range $\epsilon \in (0, 10]$ are generally understood to provide meaningful privacy protection in graph machine learning [35,30]. We use $\delta = 10^{-6}$ for all our guarantees, and take $\Delta_{\max} = 6$. For each experiment we sample a pattern vector $\boldsymbol{F}$ of $d = 50$ patterns, with the sampling strategy described in Welke et al. [34] and sample patterns with treewidth of 1. For each experimental setting, we sample patterns 3 times and perform 3 runs for each sample, for a total of 9 runs. We use nearest neighbor classifiers trained on the private homomorphism densities to predict the class of unseen graphs. We consider the 500, 10, and 100 nearest neighbors for `MOLHIV`, `MOLBACE`, and `MOLTOX21` respectively. We compare our results with classifiers trained on the noise-free, non-private homomorphism densities. We evaluate the performance of our classifiers reporting the classification AUC for different privacy budgets. In a binary classification setting, we use can formalize the trade-off between privacy and the classification AUC as follows.

Proposition 4. *In a binary classification setting with separable classes, the AUC curve follows the error function* erf *for embeddings perturbed with additive Gaussian noise.*

As our private mechanism relies on Gaussian noise, Proposition 4 applies. In a practical setting, even though we may not have perfectly separated classes we still expect the AUC to roughly follow the erf function.

Privacy attacks. To empirically test our privacy guarantees, we consider the following attack scenario. We assume a strong attacker that has access to the vector of patterns $\boldsymbol{F}$ and to the original set of graphs $\{G_1, \ldots, G_N\}$. For each $G_i \in \{G_1, \ldots, G_N\}$, the attacker can compute the true homomorphism density vector $\boldsymbol{t}(\boldsymbol{F}, G_i)$. The attacker has access to the private homomorphism densities and its goal is to recover an unknown graph G from the private $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ by matching it with one of the computed $\boldsymbol{t}(\boldsymbol{F}, G_i)$. Concretely, we train a nearest neighbor classifier on the (noise-free) homomorphism densities and use this classifier to perform the attack. We compute the Top-1 attack accuracy by recording whether the nearest neighbor of $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is the true graph's density $\boldsymbol{t}(\boldsymbol{F}, G)$, which allows the attacker to identify G. We compute the Top-10 attack accuracy by recording wether the true graph appears in the 10 nearest neighbors. Note that this experiment provides an empirical lower bound to the attacker's abilities, but the possibility of a stronger attacker is not excluded.

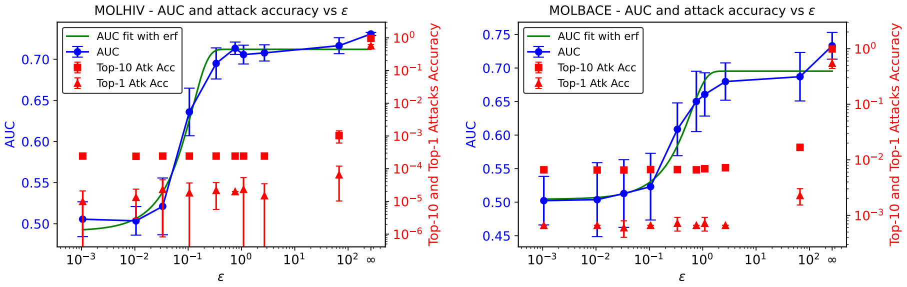

Fig. 1. Classification AUC and attack accuracy against privacy budget ϵ. We report average results with error bars of 2 standard deviations across 9 runs.

Our experiments, which we display in Figure 1, show that our approach successfully obtains a private embedding which retains discrimination abilities that are comparable to that of a non-private embedding ($\epsilon = \infty$). In the high privacy regime, with $\epsilon < 1$, we obtain an AUC comparable to the one obtained using the non-private homomorphism densities. At the same time, the attacker performance drastically decreases and the Top-1 attack accuracy is consistently below 10^{-4} for `MOLHIV` and 10^{-3} for `MOLBACE` for $\epsilon < 1$, while being close to

1 for $\epsilon = \infty$. Moreover, the classification AUC closely follows the error function, empirically confirming the formal connection between privacy and AUC discussed in Proposition 4. Additional results for MOLTOX21 (Appendix C) invoke similar considerations. This result is of great practical utility, as it allows to predictably determine the maximum privacy budget for a given desired AUC, and vice-versa the predicted AUC for a given privacy budget. We remark that our private embeddings can be used with any machine learning algorithm, and are not specifically tailored for the nearest neighbor classifier we used.

Privacy-expressivity trade-off. We perform a small set of experiments to empirically assess to which degree the result in Proposition 3 practically affects the performance of a classifier. With a setup as above, we use a nearest neighbor classifier on MOLHIV, sampling patterns with a maximum treewidth of $\{1, 2, 3\}$. Note that our sampling strategy does not guarantee that the sampled graphs match the maximum treewidth [34]. We compare the performance of the classifiers for a fixed $\rho' = 0.01$.

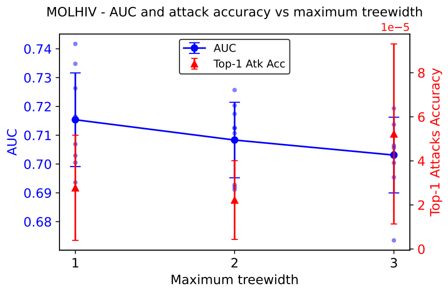

Fig. 2. Classification AUC and attack accuracy against maximum treewidth. We report average results with error bar of 2 standard deviations across 9 runs.

The results in Figure 2 show a slight downward trend of the AUC as the maximum treewidth of the patterns increases. Moreover, we observe how, despite the same formal guarantee, the accuracy of the privacy attack slightly increases for larger maximum treewidth. This result suggests that using more expressive patterns may lead to worse utility for a fixed privacy guarantee, suggesting that there is indeed a practical trade-off between expressivity and privacy.

6 Conclusion

We propose a method to obtain private and expressive graph embeddings using noisy homomorphism density vectors. By calibrating noise to the sensitivity of each density, our approach provides formal DP guarantees while preserving expressivity in expectation. Experimental results show that these private embeddings retain high classification performance, suggesting that privacy and expressivity can be effectively balanced in graph representation learning.

A key limitation of our approach is that it inherits the strengths and weaknesses of homomorphism densities themselves—if these embeddings are not well-suited for a task, the private embeddings might also perform poorly. A promising direction for future work is to refine the noise calibration by more precisely analyzing the sensitivity of specific graph classes, and to privately encode node features to further improve the privacy–utility trade-off.

References

1. Böker, J.: Graph similarity and homomorphism densities. In: 48th International Colloquium on Automata, Languages, and Programming (ICALP 2021). pp. 32–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik (2021)
2. Bun, M., Dwork, C., Rothblum, G.N., Steinke, T.: Composable and versatile privacy via truncated cdp. In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. pp. 74–86 (2018)
3. Campi, F., Gosch, L., Wollschläger, T., Scholten, Y., Günnemann, S.: Expressivity of graph neural networks through the lens of adversarial robustness. In: The Second Workshop on New Frontiers in Adversarial Machine Learning
4. Ding, X., Zhang, X., Bao, Z., Jin, H.: Privacy-preserving triangle counting in large graphs. In: Proceedings of the 27th ACM international conference on information and knowledge management. pp. 1283–1292 (2018)
5. Dwork, C.: Differential privacy. In: International colloquium on automata, languages, and programming. pp. 1–12. Springer (2006)
6. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: Privacy via distributed noise generation. In: Advances in cryptology-EUROCRYPT 2006: 24th annual international conference on the theory and applications of cryptographic techniques, st. Petersburg, Russia, May 28-June 1, 2006. proceedings 25. pp. 486–503. Springer (2006)
7. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3–4), 211–407 (2014)
8. Gil, M., Alajaji, F., Linder, T.: Rényi divergence measures for commonly used univariate continuous distributions. Information Sciences **249**, 124–131 (2013)
9. Grohe, M.: word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data. In: proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems. pp. 1–16 (2020)
10. Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., Leskovec, J.: Open graph benchmark: Datasets for machine learning on graphs. Advances in neural information processing systems **33**, 22118–22133 (2020)
11. Imola, J., Murakami, T., Chaudhuri, K.: {Communication-Efficient} triangle counting under local differential privacy. In: 31st USENIX security symposium (USENIX Security 22). pp. 537–554 (2022)
12. Jin, E., Bronstein, M., Ceylan, İ.İ., Lanzinger, M.: Homomorphism counts for graph neural networks: all about that basis. In: Proceedings of the 41st International Conference on Machine Learning. pp. 22075–22098 (2024)
13. Kasiviswanathan, S.P., Nissim, K., Raskhodnikova, S., Smith, A.: Analyzing graphs with node differential privacy. In: Theory of Cryptography: 10th Theory of Cryptography Conference, TCC 2013, Tokyo, Japan, March 3-6, 2013. Proceedings. pp. 457–476. Springer (2013)
14. Kummer, L., Gansterer, W.N., Kriege, N.M.: On the relationship between robustness and expressivity of graph neural networks (2025), https://arxiv.org/abs/2504.13786
15. Li, Y., Purcell, M., Rakotoarivelo, T., Smith, D., Ranbaduge, T., Ng, K.S.: Private graph data release: A survey. ACM Computing Surveys **55**(11), 1–39 (2023)
16. Lovász, L.: Operations with structures. Acta Mathematica Hungarica **18**(3-4), 321–328 (1967)
17. Lovász, L.: Large networks and graph limits, vol. 60. American Mathematical Soc. (2012)

18. Morris, C., Lipman, Y., Maron, H., Rieck, B., Kriege, N.M., Grohe, M., Fey, M., Borgwardt, K.: Weisfeiler and leman go machine learning: The story so far. Journal of Machine Learning Research **24**(333), 1–59 (2023)
19. Morris, C., Ritzert, M., Fey, M., Hamilton, W.L., Lenssen, J.E., Rattan, G., Grohe, M.: Weisfeiler and leman go neural: Higher-order graph neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 4602–4609 (2019)
20. Mueller, T.T., Usynin, D., Paetzold, J.C., Rueckert, D., Kaissis, G.: Sok: Differential privacy on graph-structured data. arXiv preprint arXiv:2203.09205 (2022)
21. Neuen, D.: Homomorphism-distinguishing closedness for graphs of bounded treewidth. arXiv preprint arXiv:2304.07011 (2023)
22. Nguyen, D., Halappanavar, M., Srinivasan, V., Vullikanti, A.: Faster approximate subgraph counts with privacy. Advances in Neural Information Processing Systems **36**, 70402–70432 (2023)
23. Nguyen, H., Maehara, T.: Graph homomorphism convolution. In: International Conference on Machine Learning. pp. 7306–7316. PMLR (2020)
24. Nissim, K., Raskhodnikova, S., 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)
25. Paolino, R., Maskey, S., Welke, P., Kutyniok, G.: Weisfeiler and Leman go loopy: A new hierarchy for graph representational learning. Advances in Neural Information Processing Systems **37**, 120780–120831 (2024)
26. Pei, X., Deng, X., Tian, S., Liu, J., Xue, K.: Privacy-enhanced graph neural network for decentralized local graphs. IEEE Transactions on Information Forensics and Security **19**, 1614–1629 (2024). https://doi.org/10.1109/TIFS.2023.3329971
27. Raskhodnikova, S., Smith, A.: Differentially private analysis of graphs. Encyclopedia of Algorithms (2016)
28. Rényi, A.: On measures of entropy and information. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics. vol. 4, pp. 547–562. University of California Press (1961)
29. Sajadmanesh, S., Gatica-Perez, D.: Locally private graph neural networks. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. p. 2130–2145. CCS '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3460120.3484565, https://doi.org/10.1145/3460120.3484565
30. Sajadmanesh, S., Gatica-Perez, D.: Locally private graph neural networks. In: Proceedings of the 2021 ACM SIGSAC conference on computer and communications security. pp. 2130–2145 (2021)
31. Sajadmanesh, S., Gatica-Perez, D.: Progap: Progressive graph neural networks with differential privacy guarantees. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining. pp. 596–605 (2024)
32. Sajadmanesh, S., Shamsabadi, A.S., Bellet, A., Gatica-Perez, D.: GAP: Differentially private graph neural networks with aggregation perturbation. In: 32nd USENIX Security Symposium (USENIX Security 23). pp. 3223–3240. USENIX Association, Anaheim, CA (Aug 2023), https://www.usenix.org/conference/usenixsecurity23/presentation/sajadmanesh
33. Van Erven, T., Harremos, P.: Rényi divergence and Kullback-Leibler divergence. IEEE Transactions on Information Theory **60**(7), 3797–3820 (2014)

34. Welke, P., Thiessen, M., Jogl, F., Gärtner, T.: Expectation-complete graph representations with homomorphisms. In: International Conference on Machine Learning. pp. 36910–36925. PMLR (2023)
35. Wu, F., Long, Y., Zhang, C., 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 (2022)
36. Wu, R., Fang, G., Zhang, M., Pan, Q., Liu, T., Wang, W.: On provable privacy vulnerabilities of graph representations. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 90891–90933. Curran Associates, Inc. (2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/a526cc8f6ffb74bedb6ff313e3fdb450-Paper-Conference.pdf
37. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)
38. Zhang, B., Gai, J., Du, Y., Ye, Q., He, D., Wang, L.: Beyond Weisfeiler-Lehman: A quantitative framework for GNN expressiveness. In: The Twelfth International Conference on Learning Representations (2024)
39. Zhang, Y., Zhao, Y., Li, Z., Cheng, X., Wang, Y., Kotevska, O., Yu, P.S., Derr, T.: A survey on privacy in graph neural networks: Attacks, preservation, and applications. IEEE Transactions on Knowledge and Data Engineering (2024)
40. Zhang, Z., Chen, M., Backes, M., Shen, Y., Zhang, Y.: Inference attacks against graph neural networks. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 4543–4560 (2022)

A Additional Preliminaries

In this section we provide additional details on the preliminaries.

A.1 Cut Norm

In our preliminaries we have implicitly assumed that G and G' are defined on the same node set, i.e., the nodes of G and G' have some fixed labeling $\in [n]$ which minimizes the cut distance. If, instead, two graphs G and G' have the same cardinality n but on different node sets, their distance is defined as

$$\hat{\delta}_\square(G,G') = \min_{\hat{G},\hat{G}'} d_\square(\hat{G},\hat{G}'), \tag{14}$$

with $\hat{G}$ and $\hat{G}'$ ranging over all possible labelings of G and G' by $1,\dots,n$.

For two graphs G and G' on different node sets, we define the cut distance using *fractional overlays*. A fractional overlay of two graphs G of order n and G' of order n' is a nonnegative $n \times n'$ matrix $X = [X_{iu}]_{n\times n'}$ such that $\sum_{u=1}^{n'} = \frac{1}{n}$ and $\sum_{i=1}^{n} = \frac{1}{n'}$. If $n = n'$, let $\sigma : V(G) \to V(G')$ be a bijection. Then, $X_{iu} = \frac{1}{n}\mathbb{1}(\sigma(i) = u)$ is a fractional overlay. For a fixed fractional overlay X, we define the labeled cut distance as

$$d_\square(G,G',X) = \max_{Q,R\subseteq V(G)\times V(G')} \Big| \sum_{\substack{iu\in Q\\ jv\in R}} X_{iu}X_{jv}\big(\mathbb{1}(ij\in E(G)) - \mathbb{1}(uv\in E(G'))\big)\Big|.$$

The cut distance between G and G' is defined over all possible overlays $\mathcal{X}(G,G')$:

$$\delta_\square(G,G') = \min_{X\in\mathcal{X}(G,G')} d_\square(G,G',X). \tag{15}$$

Note that, in general, for two graphs with the same cardinality $\delta_\square$ may not coincide with $\hat{\delta}_\square$ and it holds that $\delta_\square(G,G') \le \hat{\delta}_\square(G,G')$ [17]. We can now re-state the counting lemma with more precise notation.

Lemma 2 (Counting Lemma [17, Lemma 10.22]). *For any three simple graphs F, G, and G', it holds that:*

$$|t(F,G) - t(F,G')| \le e(F)\delta_\square(G,G'). \tag{16}$$

As in our setting we consider pairs of graphs G, G' with the same number of nodes which share the same node set, we have that $d_\square(G,G') = \delta_\square(G,G')$ and we thus do not need to consider the cut distance defined over fractional overlays.

A.2 Differential Privacy

We provide here additional preliminaries on DP, with a focus on how to achieve DP with additive noise scaled to the *global* sensitivity of a function.

Definition 11 (ϵ-DP, [5]). *Let $\epsilon > 0$. A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies ϵ-indistinguishability differential privacy, denoted as ϵ-DP, if, for all neighboring $x, x' \in \mathcal{X}$,*

$$\Pr[\mathcal{M}(x) \in \mathcal{Y}] \leq e^{\epsilon} \Pr[\mathcal{M}(x) \in \mathcal{Y}], \tag{17}$$

where probabilities are taken over the randomness of $\mathcal{M}$.

Definition 12 ((ϵ, δ)-DP, [6]). *Let $\epsilon > 0$ and $\delta \in [0, 1)$. A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies δ-approximate ϵ-indistinguishability differential privacy, denoted as (ϵ, δ)-DP, if, for all neighboring $x, x' \in \mathcal{X}$,*

$$\Pr[\mathcal{M}(x) \in \mathcal{Y}] \leq e^{\epsilon} \Pr[\mathcal{M}(x) \in \mathcal{Y}] + \delta, \tag{18}$$

where probabilities are taken over the randomness of $\mathcal{M}$.

In the literature, ϵ-DP is also referred to as *pure* DP while (ϵ, δ)-DP is also referred to as *approximate* DP. Given a deterministic function f, one can build a private mechanism from f by means of additive noise calibrated to its global sensitivity $GS_{f,p} = \max_{x \sim x'} \|f(x) - f(x')\|_p$, where $\|\cdot\|_p$ is a ℓ_p-norm. When p is omitted, we consider ℓ_2 norms.

Theorem 8 (Laplace mechanism for pure DP, [5,7]). *Let $f : \mathcal{X} \to \mathbb{R}$ have ℓ_1 sensitivity GS_{f,ℓ_1}. The randomized mechanism $\mathcal{M}(x) = f(x) + Lap\left(\frac{GS_{f,\ell_1}}{\epsilon}\right)$ satisfies ϵ-DP, where $Lap(b)$ denotes Laplacian noise with mean 0 and scale b.*

Theorem 9 (Gaussian mechanism for approximate DP, [6,5]). *Let $f : \mathcal{X} \to \mathbb{R}$ have ℓ_2 sensitivity GS_{f,ℓ_2}. The randomized mechanism $\mathcal{M}(x) = f(x) + \mathcal{N}(0, \sigma^2)$ satisfies (ϵ, δ)-DP for $\sigma \geq \frac{GS_{f,\ell_2}\sqrt{2\ln(1.25/\delta)}}{\epsilon}$.*

Lemma 3 (tCDP implies (ϵ, δ)-DP, [2]). *Suppose mechanism $\mathcal{M}$ satisfies (ρ, ω)-tCDP with a Rényi divergence of order α. Then, for all $\delta \in [0, 1)$, $1 < \alpha \leq \omega$, $\mathcal{M}$ satisfies (ϵ, δ)-DP with*

$$\epsilon = \begin{cases} \rho + 2\sqrt{\rho \ln(1/\delta)} & \textit{if} \quad \ln(1/\delta) \geq (\omega - 1)^2 \rho \\ \rho\omega + ln(1/\delta)/(\omega - 1) & \textit{if} \quad \ln(1/\delta) \geq (\omega - 1)^2 \rho. \end{cases}$$

Definition 13 (Smooth Sensitivity, [24]). *For a function $f : \mathcal{X} \to \mathbb{R}$, let $d(x, x')$ measure the distance between x and x', where $d(x, x') = 1$ indicates that $x \sim x'$. We define the local sensitivity of f in x as:*

$$LS_f^{(k)}(x) = \max_{x' \in \mathcal{X} : d(x, x') \leq k} |f(x) - f(x')| \,. \tag{19}$$

The β-smooth *sensitivity of* f *in* x *is then defined as*

$$S_f = \max_{k \geq 0} \left(e^{-\beta k} \cdot LS_f^{(k)}(x) \right) . \tag{20}$$

It is immediate to see that for all $x \in \mathcal{X}$, it holds that $LS_f^{(1)}(x) \leq GS_f$. Therefore, we expect a method that relies on smooth sensitivities to provide better utility, compared to one that relies on global sensitivities.

B Missing Proofs

Theorem 2 (Formal). *Let* $\mathcal{D}$ *be a distribution on* $\mathcal{F} \subseteq \mathcal{G}$ *with full support. Let* $G \in \mathcal{G}$ *be a graph and* $\boldsymbol{F} = (F_1, \ldots, F_d) \sim \mathcal{D}^d$ *be a vector of patterns. Then, the graph representation* $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G) = \boldsymbol{t}(\boldsymbol{F}, G) + \mathcal{N}\left(\boldsymbol{0}, \frac{[S_t^*(G)]^2}{2\rho'} I_d\right)$ *is* $\mathcal{F}$*-expectation-expressive and* $(2\rho' + d \cdot 4\beta, \frac{1}{4\beta})$*-tCDP, where* $\rho' > 0$ *and* $S_t^*(G)$ *is a* β*-smooth upper-bound on the local sensitivity of* $\boldsymbol{t}(\boldsymbol{F}, G)$*.*

If $\mathcal{F}^d = \mathcal{G}^d$*, then* $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ *is also expectation-complete.*

Proof. From Theorem 5, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is $(2\rho' + d \cdot 4\beta, \frac{1}{4\beta})$-tCDP. From Theorem 6, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is $\mathcal{F}$-expectation-expressive and expectation-complete if $\mathcal{F}^d = \mathcal{G}^d$. □

B.1 Privacy

Theorem 3 (Sensitivity of homomorphism density for bounded degree graphs). *Let* $G \sim G'$ *be two graphs with* n *nodes and maximum degree* $\Delta_{\max}$*. For any pattern* F *with* $m > 1$ *nodes, it holds that*

$$|t(F, G) - t(F, G')| \leq \frac{2e(F)}{n^2} \left(\frac{\Delta_{\max}}{n} \right)^{m-2} . \tag{8}$$

Proof. Without loss of generality, let $\{u, v\} \in E(G)$ and $\{u, v\} \notin E(G')$. We can explicitly compute an upper bound on $|t(F, G) - t(F, G')|$ by counting how many homomorphisms involve $\{u, v\}$. Note that we do not need to consider homomorphisms that *do not* involve $\{u, v\}$ as their count is equal for both G and G'. First, we can pick any edge of F and map it onto $\{u, v\}$. For this first step, we have a total of $2e(F)$ choices, as we take into account either order of the endpoints of each edge of F. We now map the remaining $m - 2$ nodes of F. A third node of F can now be mapped in a total of at most $\Delta_{\max}$ ways, as at most $\Delta_{\max}$ nodes are adjacent to either u or v. We can proceed similarly with the remaining nodes. After the first two nodes of F have been mapped, there are then a total of $(\Delta_{\max})^{m-2}$ ways to map the remaining $m - 2$ nodes of F. In total, there are therefore at most $2e(F)(\Delta_{\max})^{m-2}$ counts which differ for G and G'. Taking the normalization into account, we get $|t(F, G) - t(F, G')| \leq \frac{2e(F)(\Delta_{\max})^{m-2}}{n^m} = \frac{2e(F)}{n^2} \left(\frac{\Delta_{\max}}{n} \right)^{m-2}$. □

Corollary 2. *Let $G \sim G'$ be two graphs with n nodes, $d_{edge}(G, G') = k$, and maximum degree $\Delta_{\max}$. For any pattern F with $m > 1$ nodes, it holds that*

$$|t(F, G) - t(F, G')| \leq \frac{2ke(F)}{n^2} \left(\frac{\Delta_{\max}}{n} \right)^{m-2} . \tag{9}$$

Proof. The statement follows by applying, iteratively, the triangle inequality as

$$|t(F, G) - t(F, G')| \leq \underbrace{\left| t(F, G) - t(F, G^{(1)}) \right| + \cdots + \left| t(F, G^{(k-1)}) - t(F, G') \right|}_{k \text{ entries}} . \tag{21}$$

The superscript is a notation to index the graphs, and where for each of the entries in the sum $\left| t(F, G^{(i)}) - t(F, G^{(i+1)}) \right|$, $G^{(i)} \sim G^{(i+1)}$ and therefore there are k contributions as per Theorem 3. □

Proposition 2. *Let $S_t^*(G) = \|S_{t,F_1}, \ldots, S_{t,F_d}\|_2$ and $\beta > 0$. Let*

$$S_t(G) = \max_{k \geq 0} \left(e^{-\beta k} \max_{G' \in \mathcal{G} : d_{edge}(G,G') \leq k} \|\boldsymbol{t}(\boldsymbol{F}, G) - \boldsymbol{t}(\boldsymbol{F}, G')\|_2 \right) \tag{11}$$

be the β-smooth sensitivity of $\boldsymbol{t}(\boldsymbol{F}, G)$ at G. Then, it holds that $S_t^*(G) \geq S_t(G)$.

Proof. Fix a k. For each pattern F_i and for each $G' : d_{\text{edge}}(G, G') \leq k$ it holds that $|t(F_i, G) - t(F_i, G')| \leq e^{\beta k} S_{t,F_i}$. Then,

$$\|\boldsymbol{t}(\boldsymbol{F}, G) - \boldsymbol{t}(\boldsymbol{F}, G)\|_2^2 = \sum_{i=1}^{d} |t(F_i, G) - t(F_i, G')|^2 \tag{22}$$

$$\leq e^{2\beta k} \sum_{i=1}^{d} (S_{t,F_i})^2 = e^{2\beta k} (S_t^*(G))^2 \tag{23}$$

It then follows that $S_t(G) \leq \max_{k \geq 0} \left(e^{-\beta k} \sqrt{e^{2\beta k} (S_t^*(G))^2} \right) = S_t^*(G)$. □

Lemma 4 (Adapted from Gil et al. [8], Table 2). *Consider two multivariate Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, where $\boldsymbol{\Sigma}_0 = \sigma^2 I_d$ and $\boldsymbol{\Sigma}_1 = e^s \sigma^2 I_d$. Then,*

$$D_\alpha(\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \parallel \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) \tag{24}$$

$$= \frac{\alpha \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2^2}{2\left[\alpha e^s + (1 - \alpha)\right] \sigma^2} - \frac{d}{2(\alpha - let1)} \left[\alpha s - \ln(\alpha e^s + 1 - \alpha)\right] \tag{25}$$

Proof. Let, for shortness, $(\boldsymbol{\Sigma}_\alpha)^* = \alpha \boldsymbol{\Sigma}_1 + (1 - \alpha) \boldsymbol{\Sigma}_2$. From [2], it holds that

$$D_\alpha(\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \parallel \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)) \tag{26}$$

$$= \underbrace{\frac{\alpha}{2} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\intercal \left[(\boldsymbol{\Sigma}_\alpha)^*\right]^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)}_{(\star)} - \underbrace{\frac{1}{2(\alpha - 1)} \ln \frac{\det(\boldsymbol{\Sigma}_\alpha)^*}{(\det \boldsymbol{\Sigma}_0)^{1-\alpha} (\det \boldsymbol{\Sigma}_1)^\alpha}}_{(\star\star)} . \tag{27}$$

Note that $(\boldsymbol{\Sigma}_\alpha)^* = [\alpha e^s + (1-\alpha)]\, \sigma^2 I_d$, and therefore

$$(\star) = \frac{\alpha \left\| \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right\|_2^2}{2\left[\alpha e^s + (1-\alpha)\right] \sigma^2} \quad \text{and} \tag{28}$$

$$(\star\star) = -\frac{1}{2(\alpha-1)} \ln \frac{\left[\alpha e^s + (1-\alpha)\right]^d \sigma^{2d}}{(\alpha^{2d})^{1-\alpha} e^{sd\alpha} (\alpha^{2d})^{\alpha}} = -\frac{1}{2(\alpha-1)} \ln \frac{\left[\alpha e^s + (1-\alpha)\right]^d}{e^{sd\alpha}} \tag{29}$$

$$= \frac{d}{2(\alpha-1)} \left[\alpha s - \ln(\alpha e^s + 1 - \alpha)\right], \tag{30}$$

which concludes the derivation. □

Theorem 4 (tCDP with Gaussian noise in $\mathbb{R}^d$). *Let $f : \mathcal{X} \to \mathbb{R}^d$ and $g : \mathcal{X} \to \mathbb{R}$ satisfy, for every pair of neighboring databases x, x' and for $\Delta_f, \Delta_g \geq 0$,*

$$\left\| f(x) - f(x') \right\|_2 \leq \Delta_f e^{g(x)/2}, \quad |g(x) - g(x')| \leq \Delta_g. \tag{12}$$

Let $\mathcal{M}(x) = f(x) + \mathcal{N}\left(0, e^{\,g(x)}\, I_d\right)$. The randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathbb{R}^d$ satisfies $\left(\Delta_f^2 + d \cdot \Delta_g^2, \frac{1}{2\Delta_g}\right)$-tCDP.

Proof. We bound the Rényi divergence of two neighboring databases following Lemma 4, under the conditions in Theorem 4. Similarly to Bun et al. [2], we consider $\alpha, s, \gamma \in \mathbb{R}$ with $\alpha(e^s - 1) + 1 \geq \gamma = 1/2$. Note first that $s = g(x') - g(x)$, as $\boldsymbol{\Sigma}_1 = e^{g(x')} I_d = e^{g(x') - g(x)} e^{g(x)} I_d = e^s \boldsymbol{\Sigma}_0$. Due to the Δ_g-lipschitzness of g, $s > -\Delta_g$. We can ensure $\alpha(e^s - 1) + 1 \geq \gamma = 1/2$ by noting that $e^s - 1 \geq e^{-\Delta_g} - 1 \geq -\Delta_g$, and we can therefore set $\alpha \leq 1/2\Delta_g$ to get $\alpha(e^s - 1) + 1 \geq 1 - \alpha\Delta_g \geq 1/2$.

The first term in Equation (25) is bounded as

$$\frac{\alpha \left\| \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right\|_2^2}{2\left[\alpha e^s + (1-\alpha)\right] \sigma^2} \leq \frac{\alpha \left\| \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \right\|_2^2}{2\gamma\sigma^2} \leq \alpha \Delta_f^2. \tag{31}$$

The second term in Equation (25) can be bounded via a Taylor expansion of the function $h(s) = \ln\left[\alpha e^s + (1-\alpha)\right]$. First, compute

$$h(0) = 0, \quad h'(s) = \frac{\alpha e^s}{\alpha e^s + (1-\alpha)}, \quad h'(0) = \alpha, \quad h''(s) = \frac{\alpha(1-\alpha) e^s}{\left[\alpha e^s + (1-\alpha)\right]^2}. \tag{32}$$

As in Bun et al. [2], for $\alpha > 1$ and $\alpha(e^s - 1) + 1 \geq \gamma$ it holds that $0 \leq h''(s) \leq \frac{\alpha(\alpha-1)}{\gamma^2}$. Considering a Taylor expansion in $s = 0$, $h(s) = \alpha s + \frac{1}{2} h''(\zeta) s^2$ for some $\zeta \in [0, s]$, and so

$$\alpha s - h(s) = -\frac{1}{2} h''(\zeta) s^2 \leq \frac{\alpha(\alpha-1) s^2}{2\gamma^2}. \tag{33}$$

Thus, for $\gamma = 1/2$ the second term in Equation (25) reduces to

$$\frac{d}{2(d-1)} \left[\alpha s - h(s)\right] \leq \frac{\alpha d s^2}{4\gamma^2} \leq \alpha d \Delta_g^2. \tag{34}$$

Equation (31) and Equation (34) together complete the proof. □

Theorem 5. *Let $\boldsymbol{t}(\boldsymbol{F}, G)$ be the homomorphism density vector for graph G and pattern set $\boldsymbol{F}$ with $|\boldsymbol{F}| = d$, $\rho' > 0$, and $S_t^*(G)$ be a β-smooth upper bound to the local sensitivity as per Proposition 2. Then, the mechanism*

$$\tilde{\boldsymbol{t}}(\boldsymbol{F}, G) = \boldsymbol{t}(\boldsymbol{F}, G) + \mathcal{N}\left(\boldsymbol{0}, \frac{[S_t^*(G)]^2}{2\rho'} I_d\right). \tag{13}$$

is $\left(2\rho' + d \cdot 4\beta^2, \frac{1}{4\beta}\right)$-tCDP for neighboring graphs as per Definition 10.

Proof. Following the notation in Theorem 4, let $e^{g(G)} = \frac{[S_t^*(G)]^2}{2\rho'}$ and thus $g(G) = \ln\left(\frac{[S_t^*(G)]^2}{2\rho'}\right) = 2\ln(S_t^*(G)) - \ln(2\rho')$. Therefore, for two adjacent graphs $G \sim G'$, $\Delta_g = |g(G) - g(G')| = 2\left|S_t^*(G) - S_t^*(G')\right| \leq 2\beta$ as S_t^* is β-smooth. Setting $\|\boldsymbol{t}(\boldsymbol{F}, G) - \boldsymbol{t}(\boldsymbol{F}, G')\|_2 \leq S_t^*(G) = \Delta_f e^{g(G)/2} = \Delta_f \left(\frac{[S_t^*(G)]^2}{2\rho'}\right)^{1/2} = \Delta_f \frac{S_t^*(G)}{\sqrt{2\rho'}}$, it follows that $\Delta_f = \sqrt{2\rho'}$.

From Theorem 4, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is thus $\left(2\rho' + d \cdot 4\beta^2, \frac{1}{4\beta}\right)$-tCDP. □

B.2 Expressivity

Remark 1 (On graph blowups). For the following proofs, it is necessary to address the fact that two graphs G,G', where G' is a *blowup* of G, have the same homomorphism density for any pattern F [17, Theorem 5.32]. A p-blowup of G can be obtained by replacing each node of G by $p \geq 1$ twin copies [17]. Therefore, homomorphism densities cannot be used to distinguish all non-isomoprhic graphs. We can address this in two ways. We can either rely on homomorphism *counts*, which do not present the same problem and can be used to obtain a complete embedding [16,34]. As our DP statements consider pairs of graphs with the same number of nodes, this only requires to rescale the definitions of sensitivity and leads to equivalent statements about the privacy of the embeddings. This does not affect the utility of our embeddings which are, simply, rescaled. Alternatively, we can append the node count $|V(G)|$ to the homomorphism density embedding of G to distinguish it from all its blowups. This operation is trivially DP with respect to the neighboring graph notion in Definition 10 and costs no further privacy budget. As we rely on the counting lemma to derive our sensitivity bounds, we choose to present our results in terms of homomorphism densities. Therefore, we will assume that, if necessary, the node count is appended to the embedding so that the following statements hold. We stress that this is simply a choice of presentation, as all our privacy and expressivity statements could be easily rephrased in terms of homomorphism counts.

Theorem 6. *For any $G \in \mathcal{G}$, $\tilde{t}(F, G)$ is $\mathcal{F}$-expectation-expressive for $F \sim \mathcal{D}$ if $\mathcal{D}$ has full support on $\mathcal{F} \subseteq \mathcal{G}$. If $\mathcal{F} = \mathcal{G}$, then $\tilde{t}(F, G)$ is expectation-complete.*

Proof. Consider

$$\tau = \mathbb{E}_F[t(F, G)] = \sum_{F' \in \mathcal{F}} \Pr_{\mathcal{D}}(F = F')t(F', G)e_{F'}, \tag{35}$$

where $e_{F'} \in \mathbb{R}^{|\mathcal{F}|}$ is a standard basis unit vector of $\mathbb{R}^{|\mathcal{F}|}$. We can write $\tilde{t}(F, G) = t(F, G) + Y$ where $Y \sim \mathcal{N}(\mu_Y = 0, \sigma^2)$ for some variance σ^2. Note that Y and F are independent random variables. It then holds that

$$\begin{aligned} \mathbb{E}[\tilde{t}(F, G)] &= \mathbb{E}[t(F, G) + Ye_F] = \mathbb{E}_F[t(F, G)] + \mathbb{E}_Y[Ye_F] && (36)\\ &= \mathbb{E}_F[t(F, G)] + \mathbb{E}_Y[Y]\,\mathbb{E}_Y[e_F] = \mathbb{E}_F[t(F, G)] + \mu_Y\,\mathbb{E}_Y[e_F] && (37)\\ &= \mathbb{E}_F[t(F, G)] = \tau. && (38) \end{aligned}$$

It remains to show that τ is $\mathcal{F}$-expressive. Let G, G' be two graphs for which there exists $F' \in \mathcal{F}$ such that $\hom(F', G) \neq \hom(F', G')$, and let τ, τ' be the corresponding vector representations. If $|V(G)| \neq |V(G')|$ and G' is a blowup of G or vice-versa, simply append the node counts to τ, τ' to get $(\tau, |V(G)|) \neq (\tau', |V(G')|)$. If $|V(G)| = |V(G')|$, then $\hom(F', G) \neq \hom(F', G')$ implies that $t(F', G) \neq t(F', G')$. As $\mathcal{D}$ has full support on $\mathcal{F}$, then $\Pr(F = F') > 0$ and therefore $\Pr(F = F')t(F', G) \neq \Pr(F = F')t(F', G')$, which implies $\tau \neq \tau'$. This shows that τ is $\mathcal{F}$-expressive. If $\mathcal{F} = \mathcal{G}$, then $\tau \neq \tau'$ for any two $G \not\simeq G'$, with analogous argument. Therefore, τ is in this case complete. □

Theorem 7. *Let $\mathcal{D}$ be a distribution on $\mathcal{F} \subseteq \mathcal{G}$ with full support. Let $G \in \mathcal{G}$, $\boldsymbol{F} \sim \mathcal{D}^d$, and $\theta \in [0, 1]$. For large enough d, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is $\mathcal{F}$-expressive with probability at least $1 - \theta$. If $\mathcal{F}^d = \mathcal{G}^d$, then, for large enough d, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is complete with probability at least $1 - \theta$.*

Proof. Let G, G' be any two graphs for which there exists $F' \in \mathcal{F}$ such that $\hom(F', G) \neq \hom(F', G')$. First, we consider the noise-free homomorphism density vectors and want to show that

$$\boldsymbol{t}(\boldsymbol{F}, G) = (t(F_1, G), \ldots, t(F_d, G)) \neq (t(F_1, G'), \ldots, t(F_d, G')) = \boldsymbol{t}(\boldsymbol{F}, G') \tag{39}$$

with probability at least $1 - \theta$, where $F_1, \ldots, F_d \sim \mathcal{D}$ iid. To show this, we adapt the proof of Lemma 3 by Welke et al. [34]. Since $t(F, G)$ is $\mathcal{F}$-expressive for $F \sim \mathcal{D}$, then $\mathbb{E}_F[t(F, G)] \neq \mathbb{E}_F[t(F, G')]$. In particular, there exists a set $\mathfrak{F}_{G,G'}$ of outcomes of F with $\Pr(F \in \mathfrak{F}_{G,G'}) = p > 0$ such that for all $F^* \in \mathfrak{F}$ it holds that $t(F^*, G) \neq t(F^*, G')$. We want that $\Pr[\exists\, i \in \{1, \ldots, d\} : F_i \in \mathfrak{F}_{G,G'}] \geq 1 - \theta$, and thus it must hold that $1 - (1 - p)^d \geq 1 - \theta$. Solving for d, we obtain that if $d \geq \lceil \frac{\ln(1/\theta)}{\ln\left(\frac{1}{1-p}\right)} \rceil$, then $\boldsymbol{t}(\boldsymbol{F}, G)$ is $\mathcal{F}$-expressive with probability at least $1 - \theta$.

Considering now $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$, note that if $t(F^*, G) \neq t(F^*, G')$, then, for any variance σ^2, it also holds that $\tilde{t}(F^*, G) = t(F^*, G) + \mathcal{N}(0, \sigma^2) \neq t(F^*, G') + \mathcal{N}(0, \sigma^2) =$

$\tilde{t}(F^*, G')$ with probability 1. That is, the patterns for which the noise-free homomorphism densities will distinguish G and G', also work with additive noise. Therefore, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is $\mathcal{F}$-expressive with probability at least $1 - \theta$.

If $\mathcal{F} = \mathcal{G}$, then $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G) \neq \tilde{\boldsymbol{t}}(\boldsymbol{F}, G')$ any two $G \not\simeq G'$, with analogous argument. Therefore, $\tilde{\boldsymbol{t}}(\boldsymbol{F}, G)$ is in this case complete with probability at least $1 - \theta$. □

B.3 Homomorphism-distinguishing closed graph classes

In Table 1, we report homomorphism-distinguishing closed graph classes for known GNN architectures [38]. For r-ℓMPNNs, we upper bound the number of edges by the maximum number of edges in outerplanar graphs since fan-cactus graphs are outerplanar [25]. For k-FGNNs, we can upper bound the number of edges for graphs of bounded treewidth k by considering the number of edges in a k-tree, as formalized in the following proposition.

Proposition 5. *Let $\mathcal{F} = \{F : \operatorname{tw}(F) \leq k\}$. Then, any $F \in \mathcal{F}$ with $|V(F)| = m$ has at most $km - \frac{1}{2}k(k+1)$ edges.*

Proof. A k-tree is a maximal graph of treewidth k and can be constructed by expanding a $(k+1)$-clique with new nodes such that each new node is connected to exactly k existing nodes. The initial $(k+1)$-clique has $\frac{1}{2}k(k+1)$ edges. We add $m - (k+1)$ new nodes, where each new node is connected to exactly k existing nodes, thus introducing $k(m - (k+1))$ new edges. Thus, any $F \in \mathcal{F}$ has at most $km - \frac{1}{2}k(k+1)$ edges. □

Remark 2. Maximal outerplanar graphs are 2-trees. Indeed, if we set $k = 2$, we recover our upper bound on the number of edges for outerplanar graphs.

Proposition 6. *Let $\mathcal{F} = \{F : \exists U \subset V(F)$ such that $|U| \leq k$ and $F \setminus U$ is a forest$\}$. Then, any $F \in \mathcal{F}$ with $|V(F)| = m$ has at most $m(k+1) - 1 - \frac{1}{2}(k^2 + 3k)$ edges.*

Proof. $F \setminus U$ is a forest and has thus at most $m - k - 1$ edges. Let $F[U]$ denote the subgraph induced by vertex set U. $F[U]$ has at most $\frac{1}{2}(k(k-1))$ edges. Every node in $F[U]$ is connected to at most every node in $F \setminus U$. Thus, any $F \in \mathcal{F}$ has at most $m - k - 1 + \frac{1}{2}(k(k-1)) + k(m-k) = m(k+1) - 1 - \frac{1}{2}(3k + k^2)$ many edges. □

B.4 Experiments

Proposition 4. *In a binary classification setting with separable classes, the AUC curve follows the error function* erf *for embeddings perturbed with additive Gaussian noise.*

Proof. In a binary classification setting, let C_0 and C_1 be the two classes with means μ_0 and μ_1. Assume a one-dimensional setting and that the classes are separated by $\mu_1 - \mu_0 = ♤ > 0$. If the points in each class are perturbed by additive noise $\mathcal{N}(0, \sigma^2)$, the distance B between points from the two classes is $B \sim \mathcal{N}(♤, 2\sigma^2)$. With these assumptions, the AUC is the probability that points are not misranked and thus $\text{AUC} = \Pr[B > 0] = \Pr[\mathcal{N}(♤, 2\sigma^2) > 0] = \Phi(\frac{♤}{\sigma\sqrt{2}}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{♤}{2\sigma}\right)\right]$, where Φ is the Gaussian cumulative density function. □

C Additional Results

In this section, we report additional results from the experimental evaluation.

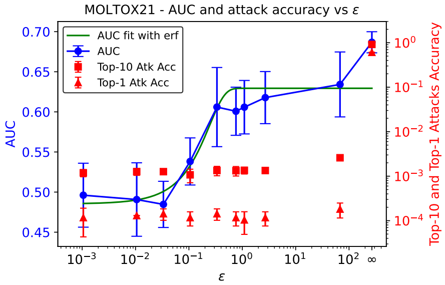

Fig. 3. Classification AUC and attack accuracy against privacy budget ϵ. We report average results with error bar of 2 standard deviations across 9 runs. For this experiment, we consider only the first task for the `MOLTOX21` benchmark.