Privacy Auditing of Multi-Domain Graph Pre-Trained Model Under Membership Inference Attacks

Jiayi Luo[1], Qingyun Sun[1], Yuecen Wei[2], Haonan Yuan[1], Xingcheng Fu[3], Jianxin Li[1]*

[1]SKLCCSE, School of Computer Science and Engineering, Beihang University, Beijing, China
[2]School of Software, Beihang University, China
[3]Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, China
{luojy, sunqy, yuanhn, lijx}@buaa.edu.cn, weiyc@buaa.edu.cn, fuxc@gxnu.edu.cn

Abstract

Multi-domain graph pre-training has emerged as a pivotal technique in developing graph foundation models. While it greatly improves the generalization of graph neural networks, its privacy risks under membership inference attacks (MIAs), which aim to identify whether a specific instance was used in training (member), remain largely unexplored. However, effectively conducting MIAs against multi-domain graph pre-trained models is a significant challenge due to: *(i) Enhanced Generalization Capability*: Multi-domain pre-training reduces the overfitting characteristics commonly exploited by MIAs. *(ii) Unrepresentative Shadow Datasets*: Diverse training graphs hinder the obtaining of reliable shadow graphs. *(iii) Weakened Membership Signals*: Embedding-based outputs offer less informative cues than logits for MIAs. To tackle these challenges, we propose **MGP-MIA**, a novel framework for **M**embership **I**nference **A**ttacks against **M**ulti-domain **G**raph **P**re-trained models. Specifically, we first propose a membership signal amplification mechanism that amplifies the overfitting characteristics of target models via machine unlearning. We then design an incremental shadow model construction mechanism that builds a reliable shadow model with limited shadow graphs via incremental learning. Finally, we introduce a similarity-based inference mechanism that identifies members based on their similarity to positive and negative samples. Extensive experiments demonstrate the effectiveness of our proposed MGP-MIA and reveal the privacy risks of multi-domain graph pre-training.

1 Introduction

Multi-domain graph pre-training (Zhao et al. 2024; Yu et al. 2024, 2025; Wang et al. 2025c; Yuan et al. 2025) has emerged as a critical technique for developing graph foundation models (Zhao et al. 2025; Shi et al. 2024; Mao et al. 2024; Liu et al. 2025; Wang et al. 2025d). By pre-training Graph Neural Networks (GNNs) across graphs from diverse domains using self-supervised learning paradigms such as link prediction (Zhang and Chen 2018) and contrastive learning (You et al. 2020), multi-domain graph pre-training enables the pre-trained GNNs to capture transferable structural and semantic patterns, thereby improving the generalization across a wide range of downstream tasks.

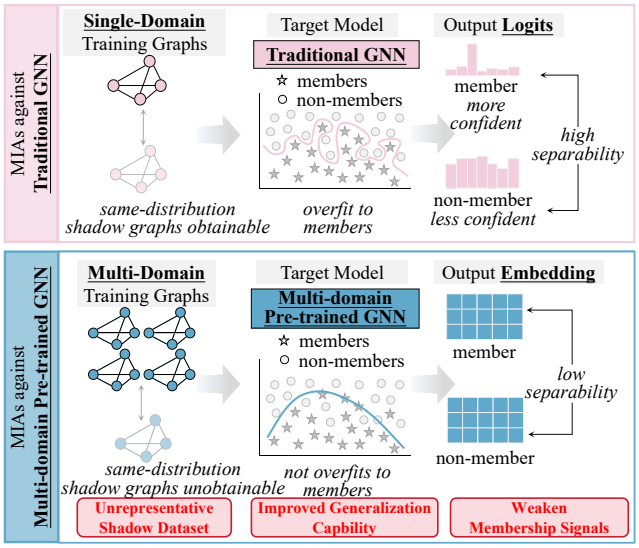

Figure 1: A comparison between graph MIAs against Traditional GNNs and Multi-domain Pre-trained GNNs.

Although multi-domain graph pre-training significantly improves graph learning performance (Zhao et al. 2024; Yu et al. 2024, 2025; Wang et al. 2025c; Yuan et al. 2025), its privacy vulnerabilities under Membership Inference Attacks (MIAs) remain largely unexplored. MIAs are privacy attacks that aim to determine whether a specific data sample was included in a model's training set (Olatunji, Nejdl, and Khosla 2021; He et al. 2021; Wei et al. 2024, 2025; Wang et al. 2025b; Song et al. 2025; Luo et al. 2025). When developers publish pre-trained models on open-source platforms to promote reuse and enable downstream users to build their own graph foundation models, they may inadvertently expose these models to privacy risks. Adversaries can exploit the released models to infer the presence of sensitive records in the training data, leading to the disclosure of user-related information.

Typical MIAs involve two stages: the attacker first trains shadow models on shadow data from the target distribution to mimic the target model, then queries the shadow models with member and non-member samples in the shadow

*Corresponding Author

data to collect prediction logits, from which it learns a rule to infer whether a target sample was in training, based on the fact that models often give higher confidence to members than non-members. However, as shown in Figure 1, performing effective MIAs against multi-domain graph pre-trained models remains significantly challenging due to three key factors: **(i) Improved Generalization Capability:** Multi-domain graph pre-training can capture transferable knowledge that improves generalization, thereby mitigating the overfitting signals commonly exploited by MIAs. **(ii) Unrepresentative Shadow Datasets:** Multi-domain pre-training scenarios involve diverse training graphs from multiple domains, making it difficult to obtain domain-aligned shadow graphs and thus hindering the construction of a reliable and effective shadow model for inference attack. **(iii) Weakened Membership Signals:** Since pre-trained encoders output embeddings rather than logits, the resulting representations carry weaker overfitting signals, making it harder to distinguish members from non-members.

The aforementioned challenges limit the effectiveness of existing graph MIAs, as further discussed in Section 4. To address these limitations, we propose **MGP-MIA** with code is available at https://github.com/RingBDStack/MGP-MIA, a novel framework for **M**embership **I**nference **A**ttacks targeting **M**ulti-domain **G**raph **P**re-trained models. Specifically, *to amplify membership signals in the target model and counteract the generalization effect introduced by multi-domain pre-training*, we propose a membership signal amplification mechanism that increases the overfitting degree of the target model via machine unlearning, motivated by the observation that machine unlearning can release parameter capacity and induce stronger memorization on the remaining data. *To construct a reliable shadow model without access to domain-aligned shadow data*, we design an incremental shadow model construction mechanism. Instead of training a shadow model from scratch, this mechanism constructs the shadow model by fine-tuning the target model using parameter regularization in an incremental learning manner, allowing it to better approximate the membership inference characteristics of the target model only with a limited shadow graph. *To extract membership signals from output embeddings in the absence of explicit cues like logits*, we introduce a similarity-based inference mechanism. This mechanism constructs attack features by measuring the similarity between a target node and its positive and negative samples, and infers membership status based on these similarity patterns. This design is motivated by the fact that self-supervised pre-training paradigms generally encourage embeddings to move closer to positive samples and farther from negative ones. We conduct extensive experiments targeting the representative multi-domain graph pre-training methods and demonstrate the superior performance of MGP-MIA.

In summary, our main contributions are as follows:

- We audit the privacy leakage of multi-domain graph pre-trained models under Membership Inference Attacks (MIAs) and propose a novel approach named MGP-MIA, for performing such attacks. To the best of our knowledge, this is the first work to investigate this problem.
- We propose a membership signal amplification mechanism to counteract the generalization effects of multi-domain pre-training. We design an incremental shadow model construction mechanism to build reliable shadow models without domain-aligned shadow data. We introduce a similarity-based inference mechanism to extract membership signals from output embeddings.
- Extensive experiments against multi-domain graph pre-training models demonstrate the effectiveness of MGP-MIA and reveal their privacy vulnerabilities to MIAs.

2 Related Work

Multi-domain Graph Pre-training. Multi-domain graph pre-training serves as the basis for graph foundation models (Shi et al. 2024; Mao et al. 2024; Liu et al. 2025; Wang et al. 2025d) by applying self-supervised learning to graphs from diverse domains (Zhao et al. 2024; Yu et al. 2024, 2025; Yuan et al. 2025; Wang et al. 2025c), facilitating the extraction of transferable knowledge. Existing approaches adopt either graph contrastive learning (You et al. 2020) or link prediction (Zhang and Chen 2018) objectives. In contrastive learning, GCOPE (Zhao et al. 2024) adds virtual nodes to link domains and optimizes them dynamically during training, while SAMGPT (Yu et al. 2025) enhances structural consistency using structure tokens to unify message aggregation. MDGFM (Wang et al. 2025c) adaptively balances features and topology, refining graphs to reduce noise and align structures. In link prediction, MDGPT (Yu et al. 2024) uses domain tokens to align semantic features, and BRIDGE (Yuan et al. 2025) introduces a domain aligner to extract shared representations and suppress noise.

Graph Membership Inference Attack. Membership inference attacks (MIAs) aim to determine whether a specific data instance was used during the training of a target model (Hu et al. 2022). He et al. (2021) and Olatunji, Nejdl, and Khosla (2021) first extend MIAs to GNNs by incorporating structural information into both target and shadow models. ProIA (Wei et al. 2025) enhanced inference performance by introducing prompt-based techniques to augment the attack model's background knowledge. Meanwhile, GCL-LEAK (Wang and Wang 2024) is the first graph MIA against contrastive learning, but its focus on the federated setting with access to the training process makes it inapplicable to our scenario. Subsequent work Zhang et al. (2022) and Wu et al. (2021) targeted whole-graph membership, and Wang and Wang (2022) conducted a systematic study on inferring the membership of particular groups of nodes and links. To operate under more constrained adversarial settings, Conti et al. (2022) and Dai and Lu (2025) propose label-only MIAs.

3 Problem Formulation

The goal of MIAs is to infer whether a given instance was part of the training set $\mathcal{D}_{\text{Train}}$ of a target model $\mathcal{F}_{\text{Target}}$. Typically, the adversary uses a shadow dataset $\mathcal{D}_{\text{Shadow}}$ from the same distribution as $\mathcal{D}_{\text{Train}}$ to train a shadow model $\mathcal{F}_{\text{Shadow}}$ that mimics the target. Based on the outputs of $\mathcal{F}_{\text{Shadow}}$ on

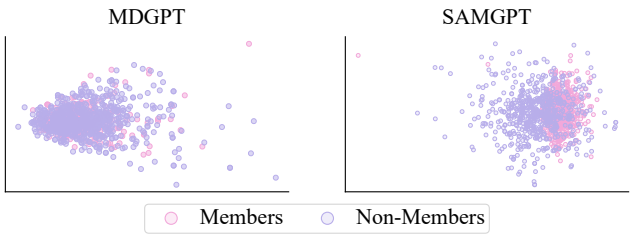

Figure 2: Separability analysis of output node embeddings.

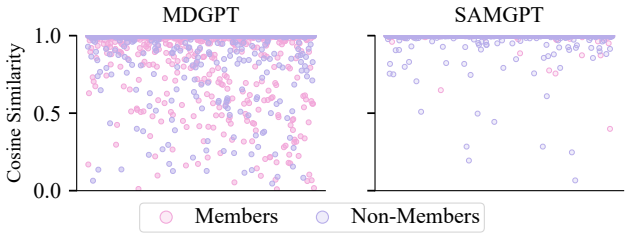

Figure 3: Robustness analysis of output node embeddings.

$\mathcal{D}_{\text{Train}}$ and $\mathcal{D}_{\text{Test}}$, the adversary constructs an attack model $\mathcal{F}_{\text{Attack}}$ to infer whether a given instance is a member.

Attacker's Goal. We focus on the node-level membership inference attacks, following prior work (Hu et al. 2022; Olatunji, Nejdl, and Khosla 2021; Wei et al. 2025). Given a target GNN $\mathcal{F}_{\text{Train}}$ pre-trained across graphs $\{\mathcal{G}_i\}_{i=1}^{m}$ from multiple domains, the goal is to determine whether a specific node v was included in the training set. Formally, the attacker aims to learn an attack model $\mathcal{F}_{\text{Attack}}$ that maximizes the expected accuracy of membership prediction as follows:

$$\max_{\mathcal{F}_{\text{Attack}}} \mathbb{E}_v \left[\mathbb{I} \left(\mathcal{F}_{\text{Attack}}(v, \mathcal{F}_{\text{Target}}) = y_v \right) \right], \tag{1}$$

where $y_v \in \{0, 1\}$ indicates whether node v belongs to the training set of $\mathcal{F}_{\text{Target}}$ (1 for member, 0 for non-member), and $\mathbb{I}(\cdot)$ denotes the indicator function.

Attacker's Knowledge. We assume a white box adversary with full access to the target model $\mathcal{F}_{\text{Target}}$, including its architecture, parameters, and training algorithm. However, the attacker does not have access to the original training process or the complete multi-domain training dataset. Instead, the attacker is assumed to have a shadow graph from one domain that shares a similar distribution with the target node v under inference. *This setting reflects practical scenarios where costly pre-trained models are publicly released to support downstream GFM development (Liu et al. 2025), potentially exposing privacy vulnerabilities.*

4 Pre-attack Analysis

In this section, we analyze *why existing graph MIAs are not applicable to multi-domain graph pre-trained models*. Existing methods can be categorized into two types: confidence-based and stability-based. Confidence-based attacks exploit confidence scores in the prediction logits, while Stability-based attacks assess the stability of predicted labels. Both rely on the observation that member instances typically yield more confident (for confidence-based) or more stable (for stability-based) predictions than non-members. However, the outputs of multi-domain graph pre-trained models are node embeddings. To assess whether these embeddings exhibit separability similar to prediction logits or stability similar to output labels for members, we conduct toy experiments on the Cora (Yang, Cohen, and Salakhudinov 2016) dataset using MDGPT (Yu et al. 2024) (link prediction) and SAMGPT (Yu et al. 2025) (contrastive learning) as representative victims. We summarize two key observations below (More experimental details are provided in Appendix C):

★*Takeaway 1: Embeddings are weakly separable between members and non-members.* To assess whether the embeddings exhibit separability between members and non-members similar to the logits, we visualize the embeddings using PCA (Abdi and Williams 2010) in Figure 2. As shown, the embeddings themselves are not clearly separable.

★*Takeaway 2: Embeddings are not more stable for members.* We perturb the graph edges and visualize the similarity between original and perturbed embeddings. As shown in Figure 3, member embeddings are not more stable under perturbation, particularly in link-prediction-based methods.

In summary, both the weak separability and the lack of enhanced stability for member nodes in the output embeddings clearly highlight the inherent limitations of directly applying existing graph membership inference attacks against multi-domain graph pre-trained models.

5 MGP-MIA

In this section, we present **MGP-MIA**, a novel framework for **M**embership **I**nference **A**ttacks against **M**ulti-domain **G**raph **P**re-trained models. As illustrated in Figure 4, our proposed MGP-MIA consists of three key components. First, the membership signal amplification mechanism enhances membership signals by mitigating the generalization effect of multi-domain training through machine unlearning. Second, the incremental shadow model construction mechanism employs incremental learning to build a shadow model that replicates the overfitting behavior of the target model only with the limited shadow dataset. Finally, the similarity-based inference mechanism constructs attack features by measuring the similarity between a target sample and its positive and negative samples, which are then used to train an attack model for membership inference.

5.1 Membership Signal Amplification Mechanism

Multi-domain pre-training trains GNNs on multiple graphs to learn transferable knowledge that generalizes across diverse downstream tasks and domains. However, the resulting improved generalization reduces the model's overfitting behavior, which in turn weakens the critical membership signals and makes effective attacks much more difficult.

To address this challenge, we propose a membership signal amplification mechanism that enhances membership signals through machine unlearning. Machine unlearning is a privacy-preserving strategy that removes the influence of specific data instances from a trained model, making it be-

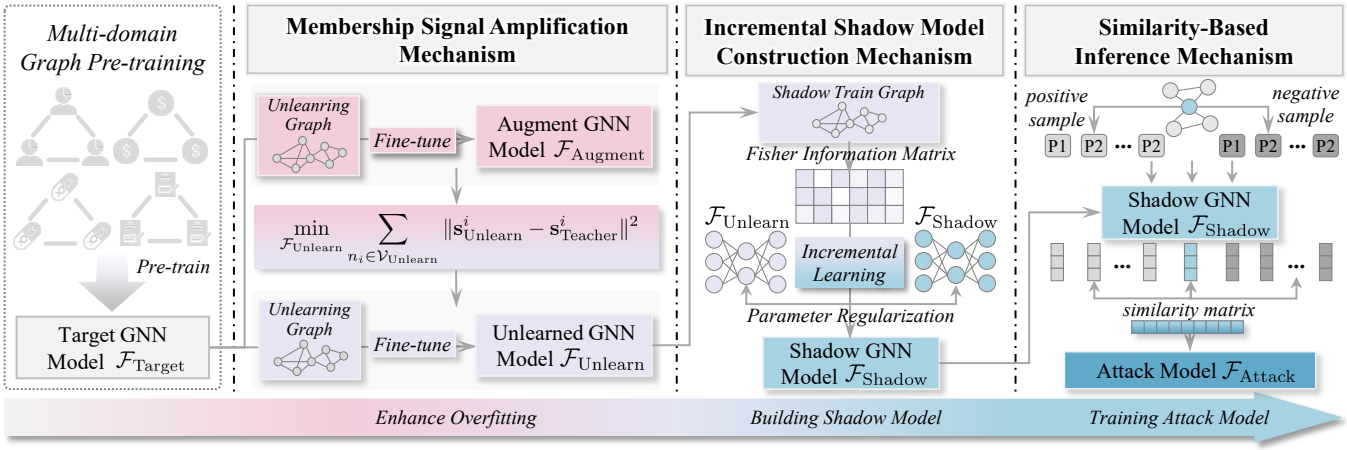

Figure 4: Overview of MGP-MIA. The membership signal amplification mechanism first leverages machine unlearning to enhance overfitting. The shadow model is then built from the unlearned model using incremental learning. Finally, the attack features are derived from similarities between each sample and its positive and negative counterparts to train the attack model.

have as if those instances were never used during training (Chen et al. 2022). However, prior work has shown that inexact machine unlearning can lead to stronger overfitting on the remaining samples as it erases specific learned information and reallocates model capacity (Hayes et al. 2025). Motivated by this phenomenon, we repurpose this privacy protection strategy as a tool to help us enhance the effectiveness of membership inference privacy attacks. While vanilla machine unlearning methods that do not require access to the remaining data typically perform unlearning through gradient ascent (Jang et al. 2022; Rashid et al. 2025). Directly applying this approach ignores the diverse characteristics of different nodes and can severely degrade model utility. Inspired by recent advances in unlearning for large language models (Wang et al. 2025a), we propose a selective and controllable unlearning strategy that employs an argument model to identify disproportionately memorized nodes and guide the model to suppress overfitted components while maintaining the utility of the unlearned model.

Specifically, we randomly extract a subgraph $\mathcal{G}_{\text{Unlearn}}$ from the given shadow graph $\mathcal{G}_{\text{Shadow}}$ as the target for inexact machine unlearning. The target model $\mathcal{F}_{\text{Target}}$ is fine-tuned on $\mathcal{G}_{\text{Unlearn}}$ for a few epochs to produce the augment model $\mathcal{F}_{\text{Augment}}$. We then compare the output embeddings of $\mathcal{F}_{\text{Target}}$ and $\mathcal{F}_{\text{Augment}}$ to evaluate how the similarity between each node and its positive and negative neighbors changes under the two models. Specifically, for a given node v_i, we define the similarity score vector as follows:

$$\mathbf{s}^i = \begin{bmatrix} \text{sim}(\mathbf{h}_i, \mathbf{h}_{i_1^+}), \ldots, \text{sim}(\mathbf{h}_i, \mathbf{h}_{i_P^+}), \\ \text{sim}(\mathbf{h}_i, \mathbf{h}_{i_1^-}), \ldots, \text{sim}(\mathbf{h}_i, \mathbf{h}_{i_N^-}) \end{bmatrix}, \tag{2}$$

where $\mathbf{h}_i$ is the out embedding of node i produced by model, $\{i_p^+\}_{p=1}^P$ are its positive samples, $\{i_n^-\}_{n=1}^N$ are its negative samples, and $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity function. Let $\mathbf{s}^i_{\text{Target}}$ and $\mathbf{s}^i_{\text{Augment}}$ denote the similarity between the output embeddings of node i and its positive and negative samples under the target and augment models, respectively. The difference $(\mathbf{s}^i_{\text{Target}} - \mathbf{s}^i_{\text{Augment}})$ captures the memorization sensitivity of node i, indicating the extent to which its associated knowledge should be unlearned. This difference is used to compute the teacher similarity score $\mathbf{s}^i_{\text{Teacher}}$, which serves as the guiding signal for the unlearning process:

$$\mathbf{s}^i_{\text{Teacher}} = \mathbf{s}^i_{\text{Target}} - \lambda \cdot (\mathbf{s}^i_{\text{Target}} - \mathbf{s}^i_{\text{Augment}}), \tag{3}$$

where λ is the hyperparameter controlling the strength of machine unlearning. Finally, we fine-tune the target model $\mathcal{F}_{\text{target}}$ to minimize the deviation between the student similarity score $\mathbf{s}^i_{\text{Unlearn}}$ and the teacher score $\mathbf{s}^i_{\text{Teacher}}$, resulting in the unlearned model $\mathcal{F}_{\text{Unlearn}}$:

$$\min_{\mathcal{F}_{\text{Unlearn}}} \sum_{n_i \in \mathcal{V}_{\text{Unlearn}}} \|\mathbf{s}^i_{\text{Unlearn}} - \mathbf{s}^i_{\text{Teacher}}\|^2, \tag{4}$$

where $\mathcal{V}_{\text{unlearn}}$ is the node set of the unlearning graph $\mathcal{G}_{\text{Unlearn}}$. The unlearned model $\mathcal{F}_{\text{Unlearn}}$ releases the model capacity previously occupied, allowing the remaining data to be memorized more strongly and thereby enhancing overfitting.

5.2 Incremental Shadow Model Construction Mechanism

The shadow model is a surrogate trained by an attacker to mimic the target model's behavior using data from a distribution similar to the target model's training set. As the attacker controls its training process, each instance can be labeled as a member or non-member. By recording the shadow model's outputs with these labels, the attacker builds an attack dataset capturing behavioral differences between seen and unseen data. This dataset is then used to train an attack model to infer the real membership of the target model.

However, in multi-domain graph pre-training, it is impractical to obtain a shadow dataset covering all training domains. A realistic assumption is that the attacker only has access to a shadow graph $\mathcal{G}_{\text{Shadow}}$ from the same domain as the target node v under inference. To construct a reliable shadow model that mimics the target model's membership-related behavior using limited shadow data, we propose an

incremental shadow model construction mechanism. Specifically, we randomly split the shadow graph into a training graph $\mathcal{G}^{\text{Train}}_{\text{Shadow}}$ and a test graph $\mathcal{G}^{\text{Test}}_{\text{Shadow}}$. We use the shadow dataset as a proxy to estimate the Fisher Information Matrix $\mathbf{I}_{\text{Unlearn}}$ (Jastrzebski et al. 2021), which quantifies the importance of each parameter in $\mathcal{F}_{\text{Unlearn}}$ to the target domain:

$$\mathbf{I}_{\text{Unlearn}}(\theta) = \mathbb{E}_{v \sim \mathcal{G}^{\text{Train}}_{\text{Shadow}}} \left[\frac{\partial^2 \mathcal{L}_{\text{task}}(\mathcal{F}_{\text{Unlearn}}; v)}{\partial \theta^2} | \theta \right], \quad (5)$$

where $\theta \in \mathbf{\Theta}_{\text{Unlearn}}$ denotes the parameter of the unlearned model $\mathcal{F}_{\text{Unlearn}}$, and $\mathcal{L}_{\text{task}}$ is the loss function associated with the pre-training task. We then fine-tune the unlearned model $\mathcal{F}_{\text{Unlearn}}$ on the shadow training graph to obtain the shadow model with the following objective:

$$\min_{\mathbf{\Theta}_{\text{Shadow}}} \sum_{v \in \mathcal{G}^{\text{Train}}_{\text{Shadow}}} \mathcal{L}_{\text{task}}(\mathcal{F}_{\text{Shadow}}; v) + \alpha \sum_i \mathbf{I}^{(i)}_{\text{Unlearn}} \left(\mathbf{\Theta}^{(i)}_{\text{Shadow}} - \mathbf{\Theta}^{(i)}_{\text{Unlearn}} \right)^2, \quad (6)$$

where $\mathcal{L}_{\text{task}}$ is the pre-training loss, $\mathbf{I}^{(i)}_{\text{Unlearn}}$ denotes the i-th element of the Fisher Information Matrix $\mathbf{I}_{\text{Unlearn}}$, and α controls the regularization strength.

5.3 Similarity-Based Inference Mechanism

As discussed in Section 4, the output embeddings from multi-domain graph pre-trained models themselves lack enough membership signals for MIAs. This is because the embeddings are primarily correlated with the intrinsic features of the node and its local neighborhood, rather than with whether the node was included in the model's training data.

To extract membership signals from output embeddings, we draw on the principles of self-supervised pre-training, which guide models to encode semantic relationships by pulling positive samples closer and pushing negative ones apart. Following this intuition, we construct attacker features by measuring the embedding similarity between the target node and its associated positive and negative samples. Specifically, for each target node v, we randomly select m positive samples $\{v_i^+\}_{i=1}^m$ and m negative samples $\{v_i^-\}_{i=1}^m$. In contrastive learning, each v_i^+ is an augmented view of v, while v_i^- is a randomly sampled node unrelated to v. In link prediction, v_i^+ shares a ground-truth edge with v, and v_i^- does not. These similarity scores form the feature vector $\mathbf{s}_v$ constructed as Eq. (2) to serve as the attack feature. To construct the attack dataset $\mathcal{D}_{\text{Attack}}$, we generate similarity-based feature vectors $\mathbf{s}_v$ for each node v with shadow model $\mathcal{F}_{\text{Shadow}}$ in the shadow training graph $\mathcal{G}^{\text{Train}}_{\text{Shadow}}$ and shadow testing graphs $\mathcal{G}^{\text{Test}}_{\text{Shadow}}$, respectively. Each node v is associated with a membership label $y_v \in \{0, 1\}$, where $y_v = 1$ if $v \in \mathcal{G}^{\text{Train}}_{\text{Shadow}}$ (i.e., a member), and $y_v = 0$ if $v \in \mathcal{G}^{\text{Test}}_{\text{Shadow}}$ (i.e., a non-member). The attack dataset is thus defined as:

$$\mathcal{D}_{\text{Attack}} = \left\{ (\mathbf{s}_v, y_v) \middle| v \in \mathcal{G}^{\text{Train}}_{\text{Shadow}} \cup \mathcal{G}^{\text{Test}}_{\text{Shadow}} \right\}. \quad (7)$$

Finally, we adopt a two-layer MLP as the attack model $\mathcal{F}_{\text{Attack}}$, trained on $\mathcal{D}_{\text{Attack}}$ using the cross-entropy loss. Once trained, the model can be used during inference to predict the membership status of a target node in the real target model $\mathcal{F}_{\text{target}}$ by its similarity-based attack feature vector. The overall procedure and time complexity of our proposed MGP-MIA are summarized in Appendix B.

6 Experiments

6.1 Experimental Settings

Datasets. To evaluate the membership inference performance of MGP-MIA, we conduct experiments on five widely used benchmark datasets Zhao et al. (2024). Cora, CiteSeer, and PubMed (Yang, Cohen, and Salakhudinov 2016) are citation networks where nodes represent publications and edges denote citations. Computers and Photos (Shchur et al. 2018) are Amazon co-purchase graphs, where nodes are products and edges indicate frequent co-purchases. Each dataset is randomly split into two equal halves: one for the target graph $\mathcal{G}^{\text{Train}}_{\text{Target}}$ and $\mathcal{G}^{\text{Test}}_{\text{Target}}$, respectively. The target model is trained on $\mathcal{G}^{\text{Train}}_{\text{Target}}$, and membership inference treats nodes in $\mathcal{G}^{\text{Train}}_{\text{Target}}$ as members and nodes in $\mathcal{G}^{\text{Test}}_{\text{Target}}$ as non-members.

Victims. We evaluate the effectiveness of MGP-MIA against four multi-domain graph pre-trained models, categorized by their self-supervised learning objectives. (1) Contrastive learning: GCOPE (Zhao et al. 2024) introduces virtual nodes to connect different domains and dynamically optimizes them during training; SAMGPT (Yu et al. 2025) enhances structural consistency across domains via structure tokens that unify message aggregation. (2) Link prediction: MDGPT (Yu et al. 2024) uses domain tokens to align semantic features across domains, while BRIDGE (Yuan et al. 2025) employs a domain aligner to capture shared patterns and suppress domain-specific noise.

Baselines. We compare our proposed MGP-MIA with several related baseline methods. Since no existing approach specifically targets multi-domain graph pre-trained models, we adapt suitable variants of relevant baseline methods. Implementation details are provided in Appendix D. The evaluated baselines include: (1) Embed-MIA (Duddu, Boutet, and Shejwalkar 2020), which trains an attack model directly on the output embeddings of the nodes in the shadow graph to distinguish members from non-members. (2) Grad-MIA (Nasr, Shokri, and Houmansadr 2019), which leverages the loss gradients with respect to the input node to train an attack model, based on the intuition that member samples produce smaller and more distinctive gradients. (3) NLO-MIA (Conti et al. 2022), which is a label-only attack that trains an attack model to infer membership based on embedding robustness under structural perturbations. (4) GLO-MIA (Dai and Lu 2025): which is a black-box attack that perturbs the input graph and infers membership by adopting a threshold over the output similarity. (5) GE-MIA (Duddu, Boutet, and Shejwalkar 2020), which assumes access to a few real member and non-member samples, and infers membership based on their distances to embedding cluster centers. (6) GPIA (Wang and Wang 2022), which fine-tunes the model on individual samples to capture parameter changes for training an attack model, assuming access to a few real member and non-member samples.

Dataset		Cora		CiteSeer		PubMed		Photo		Computers	
Victims	Method	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
MDGPT	Embed-MIA	68.89±1.83	60.31±4.27	66.53±1.26	53.25±2.90	60.60±0.80	61.93±0.68	60.81±2.52	64.99±1.16	61.54±0.51	64.70±0.73
	Grad-MIA	51.51±1.41	22.03±7.08	50.76±0.49	14.29±1.84	49.21±2.79	35.29±4.61	50.74±5.16	16.65±6.53	55.15±4.12	41.17±7.13
	NLO-MIA	59.39±0.69	50.04±1.76	60.75±0.74	53.42±2.11	54.31±0.52	54.51±0.83	55.46±2.76	54.70±3.49	60.85±3.15	61.55±1.99
	GLO-MIA	50.00±0.00	66.67±0.00	50.00±0.00	66.67±0.00	50.00±0.00	66.67±0.00	50.00±0.00	66.67±0.00	50.00±0.00	66.67±0.00
	GE-MIA	60.79±1.63	67.63±1.97	54.90±1.60	61.06±1.99	51.69±0.83	55.04±3.70	53.34±2.44	58.83±7.07	53.16±1.58	59.99±3.41
	GPIA	72.20±16.41	76.41±15.07	68.58±17.65	48.45±38.29	65.75±4.34	62.19±15.12	61.95±16.17	73.13±8.84	68.35±4.30	65.84±15.79
	MGP-MIA	**81.79**±0.94	**83.99**±0.87	**77.36**±1.31	**80.06**±2.08	**74.77**±0.47	**77.09**±1.06	**74.05**±0.33	**77.23**±0.55	**80.66**±1.43	**82.05**±1.33
BRIDGE	Embedding	66.62±1.02	58.93±1.74	65.61±1.50	52.31±2.95	55.46±0.48	58.65±1.07	59.43±1.15	64.47±1.01	58.94±1.59	64.20±1.17
	Gradient	49.91±1.75	32.35±2.33	50.00±2.30	40.33±11.27	51.45±3.02	51.39±3.68	49.06±5.35	49.60±3.89	45.44±3.32	47.14±3.13
	NLO-MIA	60.97±1.02	53.83±1.53	61.48±1.63	53.53±2.28	52.17±0.56	52.13±1.05	53.38±4.49	55.48±4.72	54.54±3.83	56.39±3.23
	GLO-MIA	50.92±1.28	66.25±0.93	50.52±0.72	66.06±0.94	49.98±0.07	66.62±0.12	50.02±0.04	65.47±2.67	48.16±4.01	66.63±0.08
	GE-MIA	55.75±3.43	59.34±4.37	52.86±2.36	52.77±8.76	50.66±0.30	52.13±2.74	52.63±1.22	62.41±1.84	53.20±1.85	57.58±7.01
	GPIA	66.76±11.87	66.51±13.16	62.77±15.71	46.22±42.40	59.34±5.59	55.36±10.67	53.28±5.64	47.08±30.14	54.38±4.59	38.15±26.29
	MGP-MIA	**81.20**±1.10	**79.97**±1.26	**79.57**±1.25	**80.94**±1.46	**74.93**±0.69	**79.05**±0.28	**70.36**±1.74	**73.06**±2.75	**73.39**±0.43	**76.13**±1.05

Table 1: Membership inference attack performance against **link-prediction-based** multi-domain graph pre-trained models. Best results are in **bold**, and runner-ups are underlined.

Dataset		Cora		CiteSeer		PubMed		Photo		Computers	
Victims	Method	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
GCOPE	Embed-MIA	60.00±22.36	73.33±14.91	50.00±0.00	66.67±0.00	70.32±27.10	76.73±22.27	60.00±22.36	20.00±44.72	70.00±27.39	40.00±54.77
	Grad-MIA	48.42±4.94	50.71±7.92	51.02±3.30	54.16±0.47	57.81±3.86	57.11±6.99	52.47±7.20	55.51±7.53	57.97±4.67	60.53±2.22
	NLO-MIA	54.58±2.80	53.93±3.64	53.97±2.09	53.10±1.54	51.94±0.44	48.64±1.83	55.21±4.53	57.11±4.39	55.22±3.28	57.76±6.51
	GLO-MIA	41.89±7.82	33.12±27.55	44.98±0.69	60.81±1.17	48.64±2.34	65.42±2.16	45.46±6.10	62.30±6.06	49.87±0.16	53.21±29.75
	GE-MIA	51.19±1.36	55.32±9.01	50.63±0.55	38.08±16.77	50.85±0.44	32.04±6.76	50.96±0.94	35.96±27.88	50.71±0.30	16.62±8.18
	GPIA	80.00±27.39	60.00±54.77	70.00±27.39	66.67±40.82	60.00±41.83	40.00±54.77	80.00±27.39	60.00±54.77	**90.00**±22.36	**93.33**±14.91
	MGP-MIA	**87.21**±1.04	**88.13**±0.83	**85.86**±0.75	**87.44**±0.60	**80.20**±1.77	**83.37**±1.17	**83.19**±1.05	**85.04**±0.69	84.80±1.43	86.35±0.84
SAMGPT	Embedding	54.39±9.81	14.10±31.53	51.18±10.02	43.49±33.11	48.61±5.00	16.54±29.76	50.17±1.43	40.22±36.62	49.80±1.14	42.90±32.76
	Gradient	61.82±2.33	60.71±2.50	52.19±2.19	47.71±3.34	50.03±1.22	51.76±2.20	48.22±1.82	52.86±7.42	54.70±3.70	61.21±4.21
	NLO-MIA	52.87±1.97	52.66±3.14	49.84±2.58	49.63±4.49	49.51±0.18	51.07±2.62	49.11±4.85	49.68±4.85	50.53±2.80	50.67±0.77
	GLO-MIA	61.02±0.61	63.52±3.11	55.16±5.31	47.16±30.81	53.71±2.09	59.74±11.10	51.18±1.93	63.45±7.48	50.98±1.23	35.32±30.08
	GE-MIA	73.32±3.88	74.99±4.52	73.97±4.44	75.67±5.64	55.21±0.35	51.03±6.16	50.61±0.66	44.41±11.64	55.77±0.68	55.34±3.39
	GPIA	58.55±2.76	59.11±7.01	55.31±2.30	57.25±5.80	54.59±1.23	61.83±3.24	84.54±4.89	83.94±3.92	73.33±16.28	77.20±11.04
	MGP-MIA	**99.91**±0.20	**99.88**±0.27	**98.83**±1.17	**98.86**±1.14	**91.30**±8.61	**92.37**±7.20	**98.11**±3.22	**98.12**±2.93	**91.72**±17.29	**93.92**±12.38

Table 2: Membership inference attack performance against **contrastive-learning-based** multi-domain graph pre-trained models. Best results are in **bold**, and runner-ups are underlined.

Metrics We adopt Accuracy (ACC) and F1-score (F1) as the evaluation metrics to assess the effectiveness of MIAs. Accuracy measures the overall proportion of correctly classified member and non-member samples, while F1-score provides a balanced evaluation of precision and recall.

Implements. All the victims are pre-trained across five graph datasets. The attacker can only access a shadow graph with a similar distribution to the target node under inference and does not have access to graphs from all domains. For the attacker model, both MGP-MIA and all baselines use a two-layer MLP with a latent dimension of 256. Baseline hyperparameters follow the settings recommended in their original papers and are further fine-tuned for fairness. Full hyperparameter configurations and implementation details are provided in Appendix D. All experiments are conducted on a single NVIDIA V100 GPU and repeated 5 times.

6.2 Attacking Link-Prediction-Based Multi-Domain Graph Pre-trained Models

We evaluate the performance of MGP-MIA against two link-prediction–based multi-domain graph pre-trained models: MDGPT and BRIDGE. Membership inference results in terms of accuracy and F1 score are reported in Table 1.

Analysis. Table 1 shows that our proposed MGP-MIA consistently achieves the best performance across all datasets and both link prediction-based multi-domain graph pretraining models, MDGPT and BRIDGE. Compared to the strongest baseline (GPIA), MGP-MIA improves accuracy and F1 by 9.6% and 7.5% respectively on Cora with MDGPT, even though GPIA benefits from access to real member and non-member samples. Similar improvements are observed on other datasets, with accuracy and F1 gains reaching up to 10.8% and 16.3%. These results highlight the

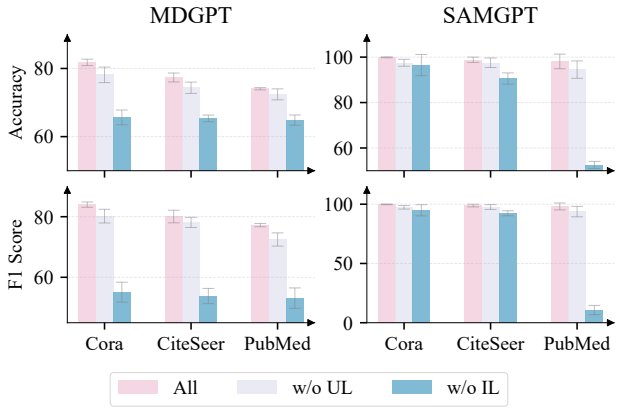

Figure 5: The ablation study of MGP-MIA.

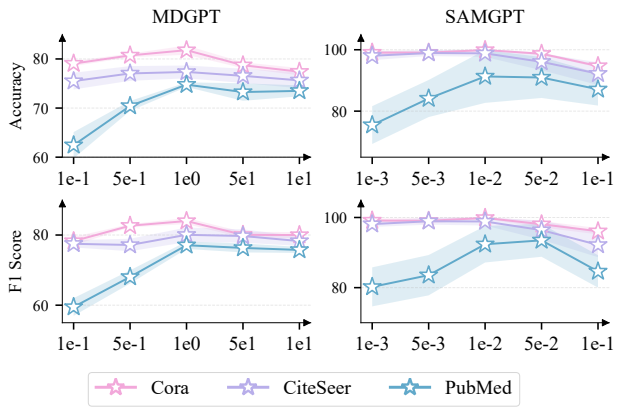

Figure 6: The study of hyperparameter α

effectiveness of MGP-MIA even using only output embeddings, without requiring privileged supervision.

6.3 Attacking Contrastive-Learning-Based Multi-Domain Graph Pretrained Models

We evaluate the performance of our proposed MGP-MIA against the contrastive-learning-based multi-domain graph pre-trained models, specifically GCOPE and SAMGPT. Membership inference results in terms of accuracy and F1 score are reported in Table 2.

Analysis. As shown in Table 2, MGP-MIA consistently outperforms all baselines across most datasets for both GCOPE and SAMGPT. On GCOPE, it achieves the highest accuracy and F1 in all cases except the Computers dataset, where GPIA slightly leads. However, GPIA assumes access to some real member and non-member nodes, while MGP-MIA requires no such privileged information. On SAMGPT, MGP-MIA achieves the best performance across all datasets. For example, it improves over the closest baseline by 26.6% in accuracy and 25.0% in F1 on Cora, and by over 20% in both metrics on Computers. These results highlight the effectiveness and generality of MGP-MIA in attacking contrastive-learning-based pre-trained models under more practical and constrained settings.

6.4 Ablation Study

To evaluate the contribution of each component in our proposed MGP-MIA, we construct two ablated variants:

- **MGP-MIA (w/o UL):** This variant removes the Membership Signal Amplification mechanism described in Section 5.1, which is designed to enhance membership signals. Instead, it fine-tunes the target model directly to obtain the shadow model, omitting the unlearning step.
- **MGP-MIA (w/o IL):** This variant removes the Incremental Shadow Model Construction mechanism, which aims to replicate the overfitting behavior of the target model using limited shadow data. Instead, it trains the shadow model from scratch on the shadow dataset.

We select MDGPT, which is based on the link prediction objective, and SAMGPT, which is based on contrastive learning, as the target victims. Experiments are conducted on the Cora, CiteSeer, and PubMed datasets. The results of the ablation study are shown in Figure 5. As illustrated, both components contribute to the overall performance of MGP-MIA. The incremental learning module enables the shadow model to more closely replicate the overfitting behavior of the target model, providing substantial intrinsic gains. The unlearning component further improves performance by amplifying membership signals, offering additional benefits.

6.5 Hyperparameter Study

In this section, we analyze the sensitivity of MGP-MIA to the hyperparameter α, which controls the regularization strength during the incremental learning stage by constraining important parameters identified by the Fisher Information Matrix in Eq. (5). Using the same victim models and dataset settings as the ablation analysis in Section 6.4, the results are presented in Figure 6. We observe that MGP-MIA is relatively robust to changes in α, maintaining stable performance across a wide range of values. Moreover, the optimal setting remains nearly the same across different datasets.

7 Conclusion

This paper investigates the privacy risks of multi-domain graph pre-trained models via Membership Inference Attacks (MIAs). We show that existing graph-based MIA methods are largely ineffective in this setting, as output embeddings contain limited overfitting signals. To address this, we propose MGP-MIA, a novel MIA framework tailored for multi-domain graph pre-training. MGP-MIA amplifies membership signals through unlearning to counteract the generalization from multi-domain training, adopts an incremental shadow model construction strategy to replicate the target model's overfitting behavior with limited shadow data, and infers membership by measuring similarity to positive and negative samples. Extensive experiments on representative models demonstrate the effectiveness of MGP-MIA and reveal the privacy risks of multi-domain graph pre-training.

Acknowledgements

The corresponding author is Jianxin Li. This work was supported by the National Natural Science Foundation of China under Grants No. 62225202 and No. 62302023, and by the Fundamental Research Funds for the Central Universities. We express our sincere gratitude to all reviewers for their valuable efforts and contributions.

References

Abdi, H.; and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4): 433–459.

Chen, M.; Zhang, Z.; Wang, T.; Backes, M.; Humbert, M.; and Zhang, Y. 2022. Graph unlearning. In *SIGSAC*, 499–513.

Conti, M.; Li, J.; Picek, S.; and Xu, J. 2022. Label-only membership inference attack against node-level graph neural networks. In *AISec*, 1–12.

Dai, J.; and Lu, Y. 2025. Graph-Level Label-Only Membership Inference Attack Against Graph Neural Networks. *Applied Sciences*, 15(9): 5086.

Duddu, V.; Boutet, A.; and Shejwalkar, V. 2020. Quantifying privacy leakage in graph embedding. In *MobiQuitous*, 76–85.

Hayes, J.; Shumailov, I.; Triantafillou, E.; Khalifa, A.; and Papernot, N. 2025. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *SaTML*, 497–519. IEEE.

He, X.; Wen, R.; Wu, Y.; Backes, M.; Shen, Y.; and Zhang, Y. 2021. Node-level membership inference attacks against graph neural networks. *arXiv*.

Hu, H.; Salcic, Z.; Sun, L.; Dobbie, G.; Yu, P. S.; and Zhang, X. 2022. Membership inference attacks on machine learning: A survey. *CSUR*, 54(11s): 1–37.

Jang, J.; Yoon, D.; Yang, S.; Cha, S.; Lee, M.; Logeswaran, L.; and Seo, M. 2022. Knowledge unlearning for mitigating privacy risks in language models. *arXiv*.

Jastrzebski, S.; Arpit, D.; Astrand, O.; Kerg, G. B.; Wang, H.; Xiong, C.; Socher, R.; Cho, K.; and Geras, K. J. 2021. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *ICML*, 4772–4784. PMLR.

Liu, J.; Yang, C.; Lu, Z.; Chen, J.; Li, Y.; Zhang, M.; Bai, T.; Fang, Y.; Sun, L.; Yu, P. S.; et al. 2025. Graph foundation models: Concepts, opportunities and challenges. *TPAMI*.

Luo, J.; Sun, Q.; Yuan, H.; Fu, X.; and Li, J. 2025. Robust Graph Learning Against Adversarial Evasion Attacks via Prior-Free Diffusion-Based Structure Purification. In *Proceedings of the ACM on Web Conference 2025*, 2098–2110.

Mao, H.; Chen, Z.; Tang, W.; Zhao, J.; Ma, Y.; Zhao, T.; Shah, N.; Galkin, M.; and Tang, J. 2024. Graph foundation models. *CoRR*.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *SP*, 739–753. IEEE.

Olatunji, I. E.; Nejdl, W.; and Khosla, M. 2021. Membership inference attack on graph neural networks. In *TPS-ISA*, 11–20. IEEE.

Rashid, M. R. U.; Liu, J.; Koike-Akino, T.; Wang, Y.; and Mehnaz, S. 2025. Forget to flourish: Leveraging machine-unlearning on pretrained language models for privacy leakage. In *AAAI*, volume 39, 20139–20147.

Shchur, O.; Mumme, M.; Bojchevski, A.; and Günnemann, S. 2018. Pitfalls of graph neural network evaluation. *arXiv*.

Shi, C.; Chen, J.; Liu, J.; and Yang, C. 2024. Graph foundation model. *Frontiers of Computer Science*, 18(6): 186355.

Song, Y.; Wei, Y.; Lu, Y.; Sun, Q.; Shao, M.; Wang, L.-e.; Hu, C.; Li, X.; and Fu, X. 2025. Mitigating message imbalance in fraud detection with dual-view graph representation learning. *arXiv preprint arXiv:2507.06469*.

Wang, B.; Zi, Y.; Sun, Y.; Zhao, Y.; and Qin, B. 2025a. Balancing Forget Quality and Model Utility: A Reverse KL-Divergence Knowledge Distillation Approach for Better Unlearning in LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1306–1321.

Wang, J.; Yang, L.; Wei, Y.; Si, J.; Guo, C.; Sun, Q.; Li, X.; and Fu, X. 2025b. An Out-Of-Distribution Membership Inference Attack Approach for Cross-Domain Graph Attacks. *arXiv preprint arXiv:2505.20074*.

Wang, S.; Wang, B.; Shen, Z.; Deng, B.; and Kang, Z. 2025c. Multi-domain graph foundation models: Robust knowledge transfer via topology alignment. *ICML*.

Wang, X.; and Wang, W. H. 2022. Group property inference attacks against graph neural networks. In *SIGSAC*, 2871–2884.

Wang, X.; and Wang, W. H. 2024. GCL-Leak: Link Membership Inference Attacks against Graph Contrastive Learning. *Proceedings on Privacy Enhancing Technologies*.

Wang, Z.; Liu, Z.; Ma, T.; Li, J.; Zhang, Z.; Fu, X.; Li, Y.; Yuan, Z.; Song, W.; Ma, Y.; et al. 2025d. Graph Foundation Models: A Comprehensive Survey. *arXiv*.

Wei, Y.; Fu, X.; Liu, L.; Sun, Q.; Peng, H.; and Hu, C. 2025. Prompt-based Unifying Inference Attack on Graph Neural Networks. In *AAAI*, volume 39, 12836–12844.

Wei, Y.; Yuan, H.; Fu, X.; Sun, Q.; Peng, H.; Li, X.; and Hu, C. 2024. Poincaré differential privacy for hierarchy-aware graph embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 9160–9168.

Wu, B.; Yang, X.; Pan, S.; and Yuan, X. 2021. Adapting membership inference attacks to GNN for graph classification: Approaches and implications. In *ICDM*, 1421–1426. IEEE.

Yang, Z.; Cohen, W.; and Salakhudinov, R. 2016. Revisiting semi-supervised learning with graph embeddings. In *ICML*, 40–48. PMLR.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *NeurIPS*, 33: 5812–5823.

Yu, X.; Gong, Z.; Zhou, C.; Fang, Y.; and Zhang, H. 2025. Samgpt: Text-free graph foundation model for multi-domain pre-training and cross-domain adaptation. In *WWW*, 1142–1153.

Yu, X.; Zhou, C.; Fang, Y.; and Zhang, X. 2024. Text-free multi-domain graph pre-training: Toward graph foundation models. *arXiv*.

Yuan, H.; Sun, Q.; Shi, J.; Fu, X.; Hooi, B.; Li, J.; and Yu, P. S. 2025. How Much Can Transfer? BRIDGE: Bounded Multi-Domain Graph Foundation Model with Generalization Guarantees. In *ICML*.

Zhang, M.; and Chen, Y. 2018. Link prediction based on graph neural networks. *NeurIPS*, 31.

Zhang, Z.; Chen, M.; Backes, M.; Shen, Y.; and Zhang, Y. 2022. Inference attacks against graph neural networks. In *USENIX Security*, 4543–4560.

Zhao, H.; Chen, A.; Sun, X.; Cheng, H.; and Li, J. 2024. All in one and one for all: A simple yet effective method towards cross-domain graph pretraining. In *SIGKDD*, 4443–4454.

Zhao, Z.; Su, Y.; Li, Y.; Zou, Y.; Li, R.; and Zhang, R. 2025. A Survey on Self-Supervised Graph Foundation Models: Knowledge-Based Perspective. *TKDE*.