An Out-Of-Distribution Membership Inference Attack Approach for Cross-Domain Graph Attacks

Jinyan Wang[1,2] , **Liu Yang**[1,2], **Yuecen Wei**[3,4*], **Jiaxuan Si**[1,2], **Chenhao Guo**[1,2], **Qingyun Sun**[4], **Xianxian Li**[1,2], **Xingcheng Fu**[1,2*]
[1]Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, China
[2]Guangxi Key Lab of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin, China
[3]School of Software, Beihang University, Beijing, China
[4]SKLCCSE, School of Computer Science and Engineering, Beihang University, China
{wangjy612, ylzyg, sijiaxuan03, guochenhao03, fuxc, lixx}@gxnu.edu.cn,
{weiyc, sunqy}@buaa.edu.cn

Abstract

Graph Neural Network-based methods face privacy leakage risks due to the introduction of topological structures about the targets, which allows attackers to bypass the target's prior knowledge of the sensitive attributes and realize membership inference attacks (MIA) by observing and analyzing the topology distribution. As privacy concerns grow, the assumption of MIA, which presumes that attackers can obtain an auxiliary dataset with the same distribution, is increasingly deviating from reality. In this paper, we categorize the distribution diversity issue in real-world MIA scenarios as an Out-Of-Distribution (OOD) problem, and propose a novel **G**raph **OOD** **M**embership **I**nference **A**ttack (GOOD-MIA) to achieve cross-domain graph attacks. Specifically, we construct shadow subgraphs with distributions from different domains to model the diversity of real-world data. We then explore the stable node representations that remain unchanged under external influences and consider eliminating redundant information from confounding environments and extracting task-relevant key information to more clearly distinguish between the characteristics of training data and unseen data. This OOD-based design makes cross-domain graph attacks possible. Finally, we perform risk extrapolation to optimize the attack's domain adaptability during attack inference to generalize the attack to other domains. Experimental results demonstrate that GOOD-MIA achieves superior attack performance in datasets designed for multiple domains.

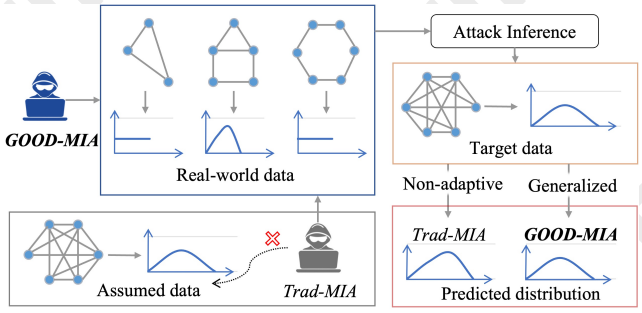

Figure 1: Traditional MIA (Trad-MIA) vs GOOD-MIA.

1 Introduction

Graph Neural Networks (GNNs) [Wu *et al.*, 2019; Fu *et al.*, 2023] have been widely applied in various practical and potentially high-risk scenarios, such as social networks [Sharma *et al.*, 2024], bioinformatics networks [Zhang *et al.*, 2021], and medical diagnosis [Boll *et al.*, 2024]. Existing researches [Wu *et al.*, 2020b; Veličković *et al.*, 2017] leverage the ability of GNNs to capture structural information and node features to address diverse downstream tasks [Tu *et al.*, 2021]. However, the in-depth mining of data and the powerful representation capabilities of the model also raise serious privacy concerns [Zhang *et al.*, 2025; Zhang *et al.*, 2024a].

With the growing concern for personal privacy security [Zhang *et al.*, 2024b; Wei *et al.*, 2024; Li *et al.*, 2025], graph-structured data has been proven to be highly susceptible to Membership Inference Attacks (MIA) due to its rich associative semantics. In MIA, attackers attempt to infer whether a specific node belongs to the training set of the target GNN model, which can lead to severe privacy leakage, especially when the model is trained on sensitive domain datasets. For example, when medical diagnosis data is modeled as a graph for training, MIA allows attackers to obtain an individual's health information without having specific details. However, the success of traditional MIA [Wei

*Co-corresponding Authors.

et al., 2025] is usually based on the assumption that attackers can access shadow datasets with the same distribution as the target. In reality, as shown in Fig. 1, it is often difficult to obtain similarly distributed data, and there can be biases in data distribution across different environments [Liu *et al.*, 2021]. These issues lead to diminished attack efficacy due to significant gaps in training data distribution. Therefore, cross-domain graph attacks are more practical as they better reflect scenarios where data access is restricted and the data comes from diverse distributions across different domains.

Out-of-distribution (OOD) [Liu *et al.*, 2021] methods show excellent capabilities in domain adaptation tasks. Existing approaches [Arjovsky *et al.*, 2019; Krueger *et al.*, 2021] make the model's learned representations consistent across different data distributions through invariant learning, thus providing robustness and generalization to unseen distributions. However, graph-structured data may not exhibit high connectivity but a scale-free power-law distribution [Wu *et al.*, 2020a]. It indicates that the structural differences among nodes and edges in different domains may be significant, requiring the shadow model to account for graph data with substantial distributional differences while mimicking the target model's decisions. Therefore, arbitrarily changing the training domain or roughly incorporating graph OOD [Liu *et al.*, 2023; Wu *et al.*, 2022] methods may lead to a shadow model that fails to adapt to multiple domains, resulting in posterior distribution shifts and affecting the effectiveness of the attack.

Therefore, to explore the privacy risks faced by models when dealing with data from different distributions, an intuitive idea is to reveal the privacy vulnerabilities of GNNs in OOD inference scenarios by studying the cross-domain graph attack. Overall, our research faces the following two challenges: Based on the MIA assumptions, existing methods overly rely on specific data features and lack adaptability to the topology of graph OOD data. This results in an inability to distinguish the intrinsic characteristics of training data when faced with confounded distributions, leading to suboptimal attack performance. Therefore, **the key issue lies in simultaneously capturing the common representations of graph features and graph structures across domains, and extending the attack.** Due to the distribution discrepancies between the shadow and target datasets, existing attack models tend to overfit specific features of the shadow dataset, failing to adequately learn the attributes and structural information directly relevant to downstream tasks. **This necessitates joint invariant learning to reinforce the acquisition of critical topological information in graph structures.**

To address the above problem, we propose a novel out-of-distribution MIA approach for cross-domain graph attacks, named GOOD-MIA. Specifically, to acquire knowledge from multiple domains during the training, we generate multiple graphs in different environments using an augmentation method [Fu *et al.*, 2024]. Then, we extract invariant features of the data in multi-domain generalization training to depict the distribution of the training data. Moreover, in the model's inference, we constrain sufficient and critical beneficial information for downstream classification tasks and further reinforce invariant representations to maintain the mimicry of the target model's behavior. For the attack model, we encourage the equalization of training risks to minimize the likelihood of risk changes when the distribution shifts. Finally, extensive experiments validate the effectiveness of cross-domain attacks. Our contributions are summarized as follows:

- To the best of our knowledge, this is the first work to conduct cross-domain attacks against GNNs. It breaks the conventional settings of MIA and reveals the privacy leakage risks of graph models on unknown distributions.
- We propose a novel **G**raph **O**ut-**O**f-**D**istribution **M**embership **I**nference **A**ttack (GOOD-MIA) to achieve cross-domain graph attacks. By capturing invariant representations of cross-domain graph data and constraining the training direction of the model, we mitigate distribution shifts during the risk extrapolation process.
- Comprehensive experiments on multiple real-world datasets demonstrate that GOOD-MIA leads in cross-domain graph attack performance.

2 Related Work

2.1 Membership Inference Attacks on GNN

MIAs aim to infer whether or not a data sample was used to train a target model [Shokri *et al.*, 2017]. Later works [Hayes *et al.*, 2017; Song and Shmatikov, 2019; He *et al.*, 2020] further investigate the feasibility of MIAs in other types of models, such as image generative and segmentation models. [Olatunji *et al.*, 2021] first migrated membership inference attack to the graph data, using a shadow training technique. The proposed scheme is based on node-level tasks performed on graph data. [He *et al.*, 2021] proposed a scheme for the membership inference attack using the 0-hop subgraph and the 2-hop subgraph, which combined the membership inference attack with the structure of the graph. However, the adversary requires all needs a shadow dataset that is identically distributed with the target dataset as the auxiliary dataset in the above membership inference attacks. In practical application scenarios, due to the diversity of data, it is almost impossible to obtain identically distributed data as the auxiliary dataset. Therefore, the effectiveness of the aforementioned MIAs is likely overestimated [Hintersdorf *et al.*, 2021].

2.2 Out-Of-Distribution Generalization on Graphs

Previous work [Beery *et al.*, 2018; Recht *et al.*, 2019] has shown that the performance of neural networks is sensitive to distribution changes and exhibits unsatisfactory performance in new environments. It is difficult to solve this generalization problem because the observations in the training data cannot cover all real-world environments. Several kinds of strategies can be applied to tackle OOD generalization [Liu *et al.*, 2021]. Causal inference aims to learn causal representations in causal graphs. By capturing causal representations, the model can obtain potential direct associations, which can help resist distribution changes caused by interventions. Invariant learning methods represented by invariant risk minimization (IRM) [Arjovsky *et al.*, 2019], which are proposed based on causal inference, have extended generalization models to more practical environments. Although the above methods have improved the generalization ability of existing ML

models, it is difficult to identify invariance due to the complex topological structure of the graph. GIL [Li *et al.*, 2022] captures the invariant relationships within the graph structure and the environment by jointly optimizing three modules. EERM [Wu *et al.*, 2022] trains multiple context generators to explore invariance within environments. FLOOD [Liu *et al.*, 2023] combines invariant representation learning and contrastive learning to train a more flexible framework to deal with different environments. However, the features captured by the existing invariant learning are relatively generic. Although they can render the model more generalizable, these features may not be beneficial for downstream tasks when dealing with data from other distributions.

3 Preliminaries

In this section, we introduce the notation used in this paper, as well as some related methods.

3.1 Problem Statement

The goal of an adversary is to determine whether a given node is used to train a target GNN model or not. Formally, let $\mathcal{G} = (\mathcal{V}, A, \mathbf{X})$ represents the graph dataset, $\mathcal{V}$ is the node set with size $n = |\mathcal{V}|$, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the edge set. We denote the adjacency matrix of $\mathcal{G}$ as $A \in \{0,1\}^{n \times n}$, where $A_{ij} = 1$ if node v_i connects to node v_j, otherwise $A_{ij} = 0$, and $\mathcal{N}(v)$ is the neighbor set of node v. $\mathbf{X} \in \mathbb{R}^{n \times f}$ is the matrix of node attributes where each row vector $\mathbf{X}_v \in \mathbb{R}^f$ is the corresponding attributes of node v. Given a target node v in the target dataset $\mathcal{G}_t$, a target GNN model $\mathcal{M}_t$, and the adversary's background knowledge $\mathcal{K}$. Membership inference attack $\mathcal{A}$ is defined as:

$$\mathcal{A} : v, \mathcal{M}_t, \mathcal{K} \mapsto \{ \text{member, non-membe} \}. \tag{1}$$

3.2 Invariant Learning

Invariant learning is used to capture invariant relationships among different distributions. Therefore, when conducting cross-domain attacks, we can utilize invariant learning to capture invariant representations between different datasets.

Empirical Risk Minimization (ERM) [Vapnik, 1991] solution is found by minimizing the global risk, expressed as the expected loss over the observational distribution, but it does not generalize well to other domains in the testing[Liu *et al.*, 2023], ERM as:

$$\mathcal{R}_{\mathrm{ERM}}\left(f_w\right) = \mathbb{E}_{P_{\mathrm{obs}}(x,y,e)}\left[\ell\left(f_w(x), y\right)\right], \tag{2}$$

where $\mathbb{E}_{P_{\mathrm{obs}}(x,y,e)}$ indicates that the expectation is taken with respect to the observed data distribution $P_{\mathrm{obs}}(x, y, e)$, ℓ is the loss function and ω is the parameter.

Invariant Risk Minimization (IRM) [Arjovsky *et al.*, 2019] includes a regularization objective that enables the classifier $f(\cdot)$ to achieve optimality across all environments, following:

$$\mathcal{R}_{\mathrm{IRM}}\left(f_w\right) = \sum_{e \in \mathcal{E}^{\mathrm{obs}}} \mathcal{R}_e\left(f_w\right) + \beta \left\|\nabla_\omega \mathcal{R}_e\left(f_\omega\right)\right\|_2^2, \tag{3}$$

where $\mathcal{E}^{\mathrm{obs}} is$ is the observed environment, β is penalty weight, $\mathcal{R}_e$ is short for $\mathcal{R}_{\mathrm{ERM}}$ in environment e, and $\|\cdot\|_2^2$ is the square of the L_2 norm.

Risk Extrapolation (REx) [Krueger *et al.*, 2021] is a form of robust optimization over a perturbation set of extrapolated domains. That means that reducing differences in risk across training domains can reduce a model's sensitivity to distribution shifts. The REx is defined as:

$$\mathcal{R}_{\mathrm{REx}}\left(f_\omega\right) = \max_{\substack{\sum_e \lambda_e = 1 \\ \lambda_e \geq \lambda_{\min}}} \sum_{e \in \mathcal{E}^{\mathrm{obs}}} \lambda_e \mathcal{R}_e\left(f_\omega\right), \tag{4}$$

where λ is the extrapolated weight.

3.3 Graph Information Bottleneck

The information bottleneck (IB) [Fu *et al.*, 2025a; Fu *et al.*, 2025b] principle uses mutual information $I(X;Y)$ as a cost function and regularization. The GIB is defined as follows:

$$\mathrm{GIB}_\xi(\mathcal{G}, Y; Z) \triangleq [-I(Y;Z) + \xi I(\mathcal{G};Z)], \tag{5}$$

where ξ is the balancing coefficient.

4 Graph Out-Of-Distribution Membership Inference Attack

In this section, we first define cross-domain MI attacks against GNNs. Then, we discuss the threat model and present the attack methodology.

4.1 Threat Model

In this paper, we study MIAs under the black-box settings, which means the adversary can't access the target model's parameters but can only observe the input and output of the target model, We then analyze the adversary's background knowledge $\mathcal{K}$ along two dimensions, i.e., shadow dataset, shadow model.

Our setting. As mentioned above, it is very difficult to obtain datasets with the same distribution in real life, so our setting is different from previous work [He *et al.*, 2021; Olatunji *et al.*, 2021]. We assume that the attacker uses a dataset with a different distribution from the target dataset for auxiliary training. Using the shadow dataset, the attacker needs to train a shadow model that can learn invariant features and structures. However, the generalized representations and structures learned from data in different domains may not fully mimic the target model. Therefore, based on the above discussion, the purpose of training the shadow model is not only to mimic the behavior of the target model but also to summarize the membership status of data points in the training set of the ML model.

4.2 Attack Methodology

According to the traditional procedure of membership inference attacks on previous ML models and GNN models [Olatunji *et al.*, 2021; Shokri *et al.*, 2017], our GOOD-MIA also models the attack model as a binary classification task where the goal is to determine if a given node $v \in V_t$. We illustrate the pipeline of GOOD-MIA in Fig. 2, which consists of two modules: (1) **Shadow model training**: we adopt IRM and GIB to train the GNN model for invariant representation learning and privacy-sensitive learning. The training environments are constructed by data augmentation on graphs. (2) **Attack model training**: we adopt REx to train the binary classification model for OOD generalization.

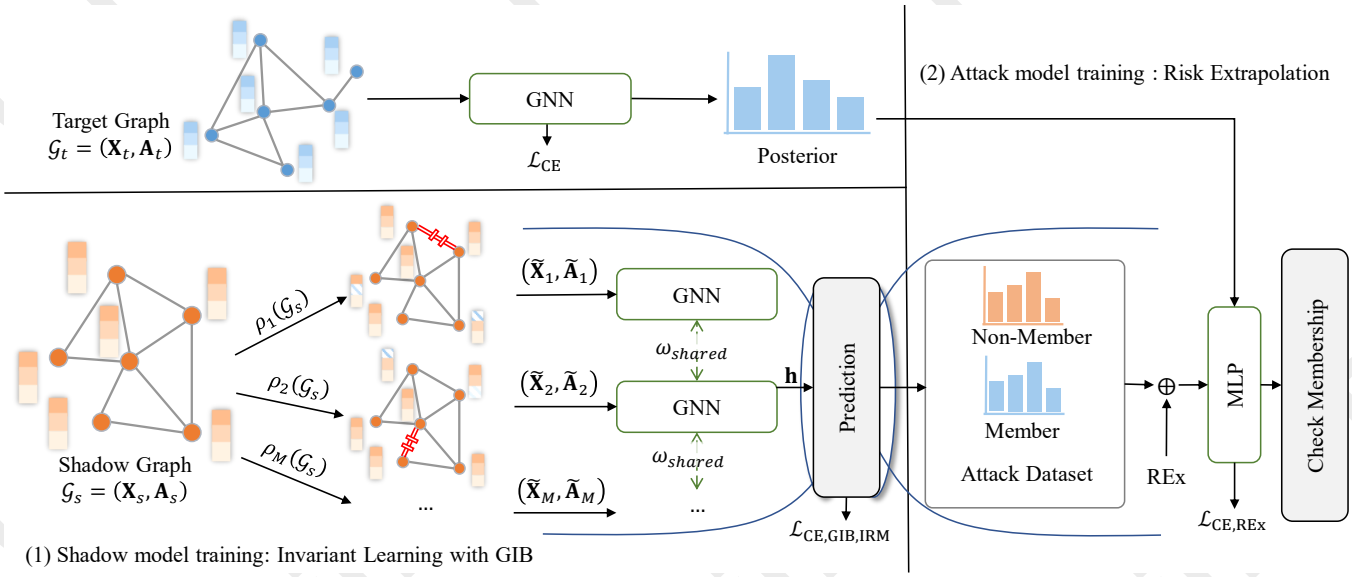

Figure 2: Framework of GOOD-MIA. (1) The input graph is augmented to construct M training environments. Then, we employ IRM and GIB to extract cross-domain graph features and structural that can facilitate cross-domain attacks. Next, (2) the output posteriors of the shadow model are used to construct the attack training set, and REx is employed to enable the attack model to conduct cross-domain attacks.

Shadow Model Training

To endow the shadow model with generalization ability, we need a shadow model that can capture the invariant properties of graphs, such as features and structures, enabling the posterior distribution learned by the model to be as close as possible to the output posterior of the target model. For a shadow dataset $\mathcal{G}_{\text{S}}$, the adversary first constructs multiple training environments from the original shadow graph $\mathcal{G}_{\text{S}}$. Then, each augmented graph $\mathcal{G}_{\text{S}}$ is divided into two disjoint subgraph, including $\mathcal{G}_{\text{S}}^{\text{Train}}$ and $\mathcal{G}_{\text{S}}^{\text{Test}}$. We perform two typical graph augmentations, namely node feature masking and DropEdge [You *et al.*, 2020]:

$$\rho_e(\mathbf{X}, \mathbf{A}) = \left(\widetilde{\mathbf{X}}_e, \widetilde{\mathbf{A}}_e\right), \quad e = 1, \ldots, M, \tag{6}$$

where $\widetilde{\mathbf{X}}_e$ represents the features after data augmentation, $\widetilde{\mathbf{A}}_e$ is the adjacency matrix after data augmentation, and e represents different environments.

Next, we train a GNN encoder $f_\omega(\cdot)$ to extract invariant features and to capture information-sensitive representations from the graphs in different training environments. $f_\omega : (\text{X}, \text{A}) \to \mathbb{R}^d$ is a L-layer graph neural networks and outputs d-dimension representation for each node. In layer $l (l = 1, ..., L)$, the representation for node i under environment e is defined by:

$$\begin{aligned} \mathbf{z}_{e,i}^{(l)} &= \text{AGG}\left(\mathbf{h}_{e,i}^{(l-1)}, \left\{\mathbf{h}_{e,j}^{(l-1)} \mid j \in \mathcal{N}_e(i)\right\}\right), \\ \mathbf{h}_{e,i}^{(l+1)} &= \text{UPDATE}\left(z_i^{(l)}, \arg\min_{\mathbf{h}_{e,i}} \mathcal{L}_{\text{GIB}}\right), \end{aligned} \tag{7}$$

where $\mathcal{N}_e(i)$ indicates the neighbor set of node i decided by $\widetilde{\text{A}}_e$, and $\mathbf{h}_{e,i}^{(0)} = \widetilde{\text{X}}_e$.

Finally, a softmax layer is applied to the node representations in the last layer for the final prediction of the node classes. The GNN parameterized by ω is trained by minimizing the cross-entropy loss defined by:

$$\mathcal{R}_e(\omega) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \text{Y}_{ij} \log\left[\sigma\left(f_\omega\left(\widetilde{\text{X}}_e, \widetilde{\text{A}}_e\right)\right)\right]_{ij}, \tag{8}$$

where σ is the activation function.

To learn a better invariant representation, we use IRM to capture the invariant representations X_c in the graph structure during across different training domains, when $\forall e_1 \neq e_2$, $P_{e_1}(Y \mid X_c) = P_{e_2}(Y \mid X_c)$, IRM constructs a linear combination with penalty weights as:

$$\mathcal{R}_{\text{IRM}}(f_\omega) = \sum_{e=1}^{M} \mathcal{R}_e(f_\omega) + \beta_1 \left\|\nabla_\omega \mathcal{R}_e(f_\omega)\right\|_2^2, \tag{9}$$

where $\beta_1 \in [0, +\infty)$ controls the balance between reducing average risk and penalty weights of risks.

Overall, in the shadow model training phase, both invariant learning and information bottleneck are jointly optimized under the overall loss as:

$$\min_{\omega} \mathcal{L}_{\text{train}} = \alpha\text{GIB} + (1-\alpha)\mathcal{R}_{\text{IRM}}, \tag{10}$$

where $\alpha \in [0, 1)$ is the weight factor used to balance the constant risk and the graph information bottleneck.

Attack Model Training

The attack model is a binary machine learning classifier and its input is derived from a node's posteriors provided by a GNN. The MLP parameterized by (w, b) is trained by minimizing the cross-entropy loss defined as.

$$\mathcal{R}_e(w, b) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} \mathbf{y}_{ij} \log(p_{ij}). \tag{11}$$

To enable the attack model to obtain a good generalization ability of the attack, we adopt the REx principle. The goal of using REx is to reduce the differences in risks across different domains, thereby enhancing the model's robustness against distribution shifts. It encourages the equality of training risks and when a distribution shift occurs at test time, the risks are more likely to change less. Minimax-REx builds a linear affine combination of training risks, as represented by:

$$\begin{aligned}\mathcal{R}_{\mathrm{MM-REx}}(\psi, b) \doteq & \max_{\substack{\sum_e \lambda_e = 1 \\ \lambda_e \geq \lambda_{\min}}} \sum_{e=1}^{M} \lambda_e \mathcal{R}_e(\psi, b) \\ & = (1 - M\lambda_{\min}) \max_e \mathcal{R}_e(\psi, b) \\ & + \lambda_{\min} \sum_{e=1}^{M} \mathcal{R}_e(\psi, b),\end{aligned} \tag{12}$$

where M is the number of environments, which consists of a posteriori of the enhanced graph.

Practically, as there is a maximum in Eq. (12), it is hard and unstable to optimize $\mathcal{R}_{\mathrm{MM-REx}}$. To tackle the problem, we replace Eq. (12) with the variance of risks as:

$$\begin{aligned}\mathcal{R}_{\mathrm{V-REx}}(\psi, b) \doteq & \beta_2 \operatorname{Var}\left(\{\mathcal{R}_1(\psi, b), \ldots, \mathcal{R}_M(\psi, b)\}\right) \\ & + \sum_{e=1}^{M} \mathcal{R}_e(\psi, b),\end{aligned} \tag{13}$$

where $\beta_2 \in [0, +\infty)$ controls the balance between reducing average risk and enforcing equality of risks.

4.3 Overall Algorithm and Complexity Analysis

The overall training algorithm is summarized in Algorithm 1. Given a shadow graph $\mathcal{G}_{\mathrm{S}}$, we first construct M training environments by data augmentation (Line 1). With the GNN initialized parameters (Line 2), we get the node representation under each environment (Line 5). After that, the model is trained until convergence by minimizing Eq.(8) (Line 7). In the attack model training phase, We use the output posterior of the shadow model to construct the attack dataset and regard the posterior outputs of different augmented graphs as different environments in the attack dataset. Finally, the attack model is updated by minimizing Eq. (13) (Line 11).

Consider a graph with N nodes and E edges, the average degree is $\bar{d}$, an MLP with N samples, D input dimensions, and H dimensions of the hidden layer. GNN with L layers computes embeddings in time $O\left(NL\bar{d}^2\right)$ and the time complexity of the MLP is $O\left(NDH\right)$. GOOD-MIA does M encoder computations per update step (M for training environment) plus a prediction step. The overall time complexity has a linear relationship with the previous work.

5 Experiments

In this section, we investigate the effectiveness of the proposed attack model in the face of three cross-domain settings with practical significance, aiming to address the following research questions.

Algorithm 1 GOOD-MIA: Out-of-Distribution Membership Inference Attack Approach for Cross-domain Graphs Attack

Input: target graph $\mathcal{G}_{\text{target}}$, shadow graph $\mathcal{G}_{\text{shadow}}$, Number of training environments M
Parameter: ω, ψ
Output: Membership prediction

1: Construct M shadow training environments by Eq. (6);
2: Initialization parameters ω, ψ;
3: **while** Shadow model training epoch $< N^{S}_{epoch}$ **do**
4: **for** $e = 1, \ldots, M$ **do**
5: Get the representation $\boldsymbol{h}$ for nodes of $\mathcal{G}_e$ w.r.t Eq. (7);
6: **end for**
7: Train ω by minimizing Eq. (10);
8: **end while**
9: **return** Posterior probability for shadow model;
10: **while** Attack model training epoch $< N^{Attack}_{epoch}$ **do**
11: Train ω by minimizing Eq. (13);
12: **end while**

Datasets	#Nodes	#Edges	#Classes	#Features
Cora	2708	5429	7	1433
Citeseer	3327	4732	6	3703
Pubmed	19717	44338	3	500
Twitch	1912 - 9498	31299 - 153138	2	2545
FB-100	769 - 41536	16656 - 1590655	2	8319

Table 1: Statistics for experimental datasets

5.1 Experimental Settings

Datasets. We adopt five node property prediction datasets of different sizes and properties, including Cora, Citeseer, Pubmed, Twitch and Facebook-100. For Cora, Citeseer and Pubmed [Sen *et al.*, 2008], we keep the original node labels and synthetically create spurious node features to introduce distribution shifts between different domain data. The Twitch and Facebook-100 [Rozemberczki and Sarkar, 2021; Traud *et al.*, 2012] represent canonical real-world social networks. For Twitch, we consider subgraph-level data splits: nodes in subgraph DE are used as target model datasets, while nodes in ENGB, ES, FR, PTBR, RU and TW are used as shadow model datasets to set cross-domain attacks in different domain environments. For Facebook-100, we use John Hopkins, Amherst and Cornell5 as the target datasets, Penn and Reed as the shadow datasets. We summarize the dataset information in Tab. 1.

Baselines. We established three backbone models (GCN, GAT, SGC) to test the expressiveness of GOOD-MIA in different scenarios. We compare the proposed GOOD-MIA with the state-of-the-art attack methods TSTS [Olatunji *et al.*, 2021] on GNNs. The baselines in our experiment have the same setting as our method, that is, the shadow dataset has a different distribution from the target dataset.

Evaluation Metrics. We use Accuracy, AUC score, and Recall [He *et al.*, 2021; Olatunji *et al.*, 2021; Salem *et al.*, 2018] to evaluate the performance of the attack model.

	Model	Cora			CiteSeer			Pubmed		
		ACC	AUC	Recall	ACC	AUC	Recall	ACC	AUC	Recall
TSTS	GCN	67.31	67.34	67.18	79.34	79.39	78.35	50.57	50.58	50.60
	GAT	65.81	65.33	65.83	77.49	77.57	77.96	50.18	50.19	50.18
	SGC	66.36	66.28	66.33	80.21	81.52	80.09	51.19	51.22	51.19
GOOD-MIA	GCN	**74.15**	**73.99**	**73.02**	**84.10**	**84.04**	84.26	53.79	53.81	53.79
	GAT	71.23	71.14	70.99	81.81	82.48	81.87	**54.47**	**54.39**	**54.48**
	SGC	71.07	71.80	71.40	82.72	82.83	**84.67**	53.94	53.92	53.95
Ablation	GOOD-MIA\IRM	62.87	62.96	62.87	62.33	62.34	62.33	45.72	45.76	45.73
	GOOD-MIA\GIB	68.23	68.35	68.25	68.53	69.28	70.50	48.20	47.20	47.26
	GOOD-MIA\REx	61.49	61.45	61.50	71.02	70.90	71.30	51.73	51.71	51.73

Table 2: Summary results of Synthetic Data (**Bold**: best; Underline: runner-up).

	Model	ENGB		ES		FR		PTBR		RU		TW	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
TSTS	GCN	57.04	56.99	59.65	59.67	59.07	59.08	58.99	59.07	56.59	56.56	50.05	60.08
	GAT	55.69	55.67	56.81	56.78	56.31	55.32	57.37	57.36	57.33	57.39	56.14	56.18
	SGC	60.27	60.20	59.43	59.47	60.14	60.20	61.37	61.36	58.21	58.27	60.15	60.14
GOOD-MIA	GCN	57.14	57.06	61.06	61.10	63.02	63.15	61.02	61.04	58.88	58.92	61.30	61.27
	GAT	59.97	60.00	59.46	59.50	59.29	59.30	59.52	59.54	58.94	58.96	60.12	60.19
	SGC	**61.29**	**61.31**	**61.56**	**61.58**	**63.74**	**63.77**	**64.80**	**64.92**	**60.38**	**60.36**	**61.56**	**61.33**

Table 3: Attack accuracy and AUC of the GOOD-MIA on the Twitch dataset (**Bold**: best; Underline: runner-up).

5.2 Performance Evaluation

We conduct comprehensive performance verification and ablation experiments on GOOD-MIA.

Distribution Shifts on Synthetic Data. We report the attack scores on the citation network dataset in Tab. 2. We found that when using different GNNs as the backbone, GOOD-MIA consistently showed a significant advantage over corresponding competitors under artificially induced distribution shifts. Also, when the datasets were Cora and PubMed, their performance was close to that of the attack models trained with datasets of the same distribution [Olatunji *et al.*, 2021]. The results indicate that this attack model can effectively capture cross-domain features and their structures when using datasets with different distributions, enabling the model to launch effective attacks.

Distribution Shifts across Domains. Tab. 3 and Tab. 4 demonstrate the attack performance in real-world social network datasets. This kind of dataset has great practical significance because it is rather difficult for us to obtain datasets with the same distribution, while it is relatively easy to acquire datasets with approximate distributions. This kind of dataset is challenging for cross-domain attacks since the nodes in different subgraphs are disconnected. We found GOOD-MIA achieves overall superior performance over competitors. This demonstrates the efficacy of our model in tackling OOD attacks across graphs in different domains.

Overall. For different datasets, the invariance of their features and structures may vary in strength. For example, citation networks may exhibit strong invariance in their structures, making them easier to capture in cross-domain scenarios. Social networks may exhibit weak invariance in their structures, leading to less clear capture and, consequently, poorer attack performance.

5.3 Ablation Study

In this section, we analyze the effectiveness of the three variants:

- **GOOD-MIA ($w/o\,IRM$):** We remove the IRM in the shadow model training objective (Eq. (10)).
- **GOOD-MIA ($w/o\,GIB$):** We remove the GIB in the shadow model training objective (Eq. (10)).
- **GOOD-MIA ($w/o\,REx$):** We remove the REx in the attack model training objective (Eq. (13)), and using the traditional Cross Entropy Loss

We choose GCN as the backbone to evaluate three key modules of GOOD-MIA, namely IRM, GIB, and REx, by removing each module respectively. We report the ablation experiments of artificial data MIA in Tab. 2. Compared with the three variants, the complete model achieves the best attack performance in terms of the total number, indicating that each module is essential for the generalization of the attack model.

Removing the IRM part, when training the shadow model, GOOD-MIA\IRM can only capture features and structures related to downstream tasks through the information bottleneck. It fails to obtain invariant representations of the data when the distributions vary. Merely conducting REx via the attack model cannot yield a satisfactory attack effect.

Target Data \ Shadow Data	Penn						Reed					
	GOOD-MIA			TSTS			GOOD-MIA			TSTS		
	ACC	AUC	Recall	ACC	AUC	Recall	ACC	AUC	Recall	ACC	AUC	Recall
John Hopkins	**57.20**	57.12	**57.20**	53.39	53.42	53.39	57.74	**57.75**	57.73	51.99	51.97	51.99
Amherst	**58.31**	57.83	**58.31**	51.62	51.63	51.62	60.19	**60.29**	57.68	54.56	54.48	54.56
Cornell5	57.52	**57.59**	57.53	53.68	53.62	53.62	56.81	56.78	**56.83**	53.48	53.47	53.48

Table 4: Attack accuracy and AUC of the GOOD-MIA on the Facebook-100 dataset (**Bold**: best).

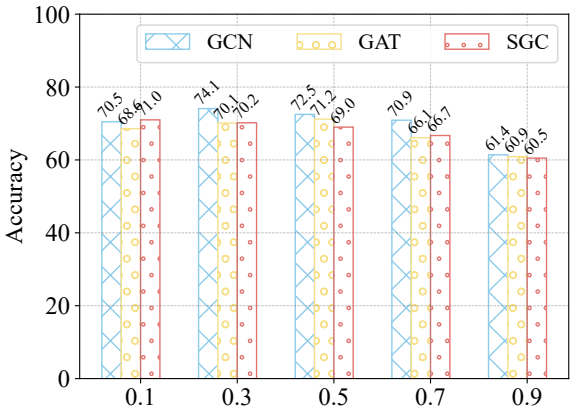

Figure 3: Trade-off parameter α analysis.

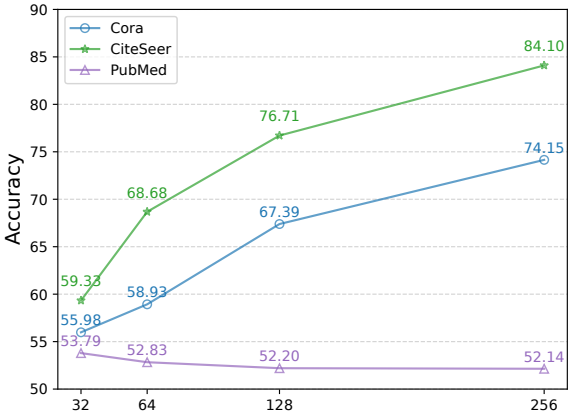

Figure 4: Different numbers of neurons are used in the hidden layer of the shadow model for the MIA.

Omitting the GIB part, GOOD-MIA\GIB can only obtain generalized representations of the data. Although such information may have the same probability under different distributions, it may not be a crucial part in cross-domain attacks.

Without the REx part, although GOOD-MIA\REx can capture invariant representations as well as structures and features related to downstream tasks during the training of the shadow model, simply using an MLP insufficient to achieve effective cross-domain attack. We conclude that only by combining these three can we improve the capability of cross-domain attacks in GOOD-MIA.

5.4 Analysis

Hyperparameter Trade-off Analysis. We conduct a hyperparameter analysis of the trade-off parameter in the shadow model to verify the roles of information bottleneck and invariant learning in GOOD-MIA. The results are shown in Fig. 3, indicating that the sensitivity to attack accuracy when setting α varies between different GNN models. It can be seen from this that the analysis we presented earlier is correct. When the shadow model learns invariant representations using IRM, generalizing to other domains may not be relevant to the downstream tasks. Therefore, it is also necessary to capture the features and structures that are closely related to the downstream tasks. Moreover, when using different GNN models, the attack effect obtained by using different parameters is also different due to different aggregation methods.

In summary, the hyperparameter settings can be further optimized based on the characteristics of different models to improve the overall performance of the attack model.

Different Model Architecture. We further investigate whether the different number of neurons affects the attack performance. We evaluated the target model with 256 neurons in its hidden layer, while the number of neurons in the shadow model varied from 32 to 256. The results are depicted in Fig. 4. We observed that in Cora and CiteSeer, the closer the shadow model approximates the target model, the higher the attack accuracy. PubMed is the opposite. This may be attributed to the lower feature dimensionality of PubMed.

6 Conclusions

In this paper, we propose a novel framework named GOOD-MIA, designed to explore the feasibility of cross-domain membership inference attacks when identically distributed auxiliary datasets are unavailable in real-world scenarios, and to enhance the effectiveness of such cross-domain attacks.

We decompose the overall objectives of GOOD-MIA into shadow model training and attack model training, which have a linear relationship. During the shadow model training phase, invariant features and key graph structures are first captured from the environments of different graph data. Subsequently, attack training sets for different environments are constructed, followed by training the attack model. Finally, through risk extrapolation, the attack model can be generalized to other domains for attacks. Extensive experiments show that GOOD-MIA exhibits excellent attack inference capabilities and domain adaptability.

Acknowledgments

The corresponding author is Xingcheng Fu and Yuecen Wei. This paper was supported by the National Natural Science Foundation of China (Nos. 62162005, 62462007 and U21A20474), National Natural Science Foundation Joint Cultivation Project of Guangxi Normal University (No. 2024PY028), Guangxi Bagui Youth Talent Training Program, Guangxi Collaborative Innovation Center of Multisource Information Integration and Intelligent Processing and the Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education (EBME24-01).

References

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[Beery *et al.*, 2018] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[Boll *et al.*, 2024] Heloísa Oss Boll, Ali Amirahmadi, Mirfarid Musavian Ghazani, Wagner Ourique de Morais, Edison Pignaton de Freitas, Amira Soliman, Farzaneh Etminani, Stefan Byttner, and Mariana Recamonde-Mendoza. Graph neural networks for clinical risk prediction based on electronic health records: A survey. *J. Biomed. Informatics*, 151:104616, 2024.

[Fu *et al.*, 2023] Xingcheng Fu, Yuecen Wei, Qingyun Sun, Haonan Yuan, Jia Wu, Hao Peng, and Jianxin Li. Hyperbolic geometric graph representation learning for hierarchy-imbalance node classification. In *Proceedings of the ACM Web Conference 2023*, pages 460–468, 2023.

[Fu *et al.*, 2024] Xingcheng Fu, Yisen Gao, Yuecen Wei, Qingyun Sun, Hao Peng, Jianxin Li, and Xianxian Li. Hyperbolic geometric latent diffusion model for graph generation. *arXiv preprint arXiv:2405.03188*, 2024.

[Fu *et al.*, 2025a] Xingcheng Fu, Yisen Gao, Beining Yang, Yuxuan Wu, Haodong Qian, Qingyun Sun, and Xianxian Li. Bi-directional multi-scale graph dataset condensation via information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16674–16681, 2025.

[Fu *et al.*, 2025b] Xingcheng Fu, Jian Wang, Yisen Gao, Qingyun Sun, Haonan Yuan, Jianxin Li, and Xianxian Li. Discrete curvature graph information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16666–16673, 2025.

[Hayes *et al.*, 2017] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

[He *et al.*, 2020] Yang He, Shadi Rahimian, Bernt Schiele, and Mario Fritz. Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 519–535. Springer, 2020.

[He *et al.*, 2021] Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. Node-level membership inference attacks against graph neural networks. *arXiv preprint arXiv:2102.05429*, 2021.

[Hintersdorf *et al.*, 2021] Dominik Hintersdorf, Lukas Struppek, and Kristian Kersting. To trust or not to trust prediction scores for membership inference attacks. *arXiv preprint arXiv:2111.09076*, 2021.

[Krueger *et al.*, 2021] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

[Li *et al.*, 2022] Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:11828–11841, 2022.

[Li *et al.*, 2025] Xianxian Li, Zeming Gan, Qiyu Li, Bin Qu, Jinyan Wang, et al. Rethinking the impact of noisy labels in graph classification: A utility and privacy perspective. *Neural Networks*, 182:106919, 2025.

[Liu *et al.*, 2021] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[Liu *et al.*, 2023] Yang Liu, Xiang Ao, Fuli Feng, Yunshan Ma, Kuan Li, Tat-Seng Chua, and Qing He. Flood: A flexible invariant learning framework for out-of-distribution generalization on graphs. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1548–1558, 2023.

[Olatunji *et al.*, 2021] Iyiola E Olatunji, Wolfgang Nejdl, and Megha Khosla. Membership inference attack on graph neural networks. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 11–20. IEEE, 2021.

[Recht *et al.*, 2019] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[Rozemberczki and Sarkar, 2021] Benedek Rozemberczki and Rik Sarkar. Twitch gamers: a dataset for evaluating proximity preserving and structural role-based node embeddings. *arXiv preprint arXiv:2101.03091*, 2021.

[Salem *et al.*, 2018] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

[Sen *et al.*, 2008] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad.

Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

[Sharma *et al.*, 2024] Kartik Sharma, Yeon-Chang Lee, Sivagami Nambi, Aditya Salian, Shlok Shah, Sang-Wook Kim, and Srijan Kumar. A survey of graph neural networks for social recommender systems. *ACM Computing Surveys*, 56(10):1–34, 2024.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[Song and Shmatikov, 2019] Congzheng Song and Vitaly Shmatikov. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206, 2019.

[Traud *et al.*, 2012] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[Tu *et al.*, 2021] Wenxuan Tu, Sihang Zhou, Xinwang Liu, Xifeng Guo, Zhiping Cai, En Zhu, and Jieren Cheng. Deep fusion clustering network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9978–9987, 2021.

[Vapnik, 1991] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[Veličković *et al.*, 2017] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[Wei *et al.*, 2024] Yuecen Wei, Haonan Yuan, Xingcheng Fu, Qingyun Sun, Hao Peng, Xianxian Li, and Chunming Hu. Poincaré differential privacy for hierarchy-aware graph embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9160–9168, 2024.

[Wei *et al.*, 2025] Yuecen Wei, Xingcheng Fu, Lingyun Liu, Qingyun Sun, Hao Peng, and Chunming Hu. Prompt-based unifying inference attack on graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12836–12844, 2025.

[Wu *et al.*, 2019] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. Pmlr, 2019.

[Wu *et al.*, 2020a] Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448, 2020.

[Wu *et al.*, 2020b] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[Wu *et al.*, 2022] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. *arXiv preprint arXiv:2202.02466*, 2022.

[You *et al.*, 2020] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.

[Zhang *et al.*, 2021] Xiao-Meng Zhang, Li Liang, Lin Liu, and Ming-Jing Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12:690049, 2021.

[Zhang *et al.*, 2024a] Guixian Zhang, Shichao Zhang, and Guan Yuan. Bayesian graph local extrema convolution with long-tail strategy for misinformation detection. *ACM Transactions on Knowledge Discovery from Data*, 18(4):1–21, 2024.

[Zhang *et al.*, 2024b] Yi Zhang, Yuying Zhao, Zhaoqing Li, Xueqi Cheng, Yu Wang, Olivera Kotevska, S Yu Philip, and Tyler Derr. A survey on privacy in graph neural networks: Attacks, preservation, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Zhang *et al.*, 2025] Guixian Zhang, Guan Yuan, Debo Cheng, Lin Liu, Jiuyong Li, and Shichao Zhang. Disentangled contrastive learning for fair graph representations. *Neural Networks*, 181:106781, 2025.