Environment-Adaptive Representation Interaction for Privacy-Perturbed Graphs Against Deceptive OOD Attacks

Ju Jia, Renjie Li, Cong Wu, *Member, IEEE*, Yebo Feng, *Member, IEEE*, Siqi Ma, *Member, IEEE*, Lina Wang, *Member, IEEE*, and Robert H. Deng, *Fellow, IEEE*

***Abstract*—Graph neural networks (GNNs) have gained increasing popularity in understanding graph-structured data due to their ability to derive meaningful representations by aggregating complicated topological information. However, privacy operations such as differential privacy mechanisms that inject noise into node features or graph structures to protect sensitive information, and distribution shifts in graph data pose tremendous security risks for the wide application of GNN models. Current researches mainly focus on defending the out-of-distribution (OOD) attacks through robust adversarial training and graph structure purification. Nonetheless, privacy perturbations of graph structures may render OOD attacks more deceptive by obfuscating the distinctiveness of nodes, leading to the failure of existing defense methods. To address these shortcomings, we propose an environment-adaptive representation interaction (EARI) scheme that strengthens the privacy perception of GNNs. Specifically, our scheme leverages the interaction between non-private and private data to enable targeted embedding propagation by the guidance of confidence score feedback. Subsequently, the representation-enriched topological aggregation is implemented to capture more discriminative features by exploiting multi-hop neighborhoods rather than stacked multilayers. Finally, the generalization-enhanced cluster-wise adaptation learning is leveraged to highlight the invariant correlations from nodes across different environments. Extensive experimental results demonstrate that our scheme can enhance the capability of learning representations from privacy-protected graph data, enabling GNNs to effectively defend against deceptive OOD attacks on various graph-structured datasets. Moreover, we reveal that the utilization of interactive topological aggregation can extremely enrich the diversity and guarantee the effectiveness for graph representations.**

***Index Terms*—Privacy-perceptual GNNs, multi-hop topological aggregation, environment-adaptive clustering, interactive representation learning, deceptive OOD attacks.**

I. Introduction

THE powerful representation learning capability of graph neural networks (GNNs) on graph-structured data has led to its wide application in areas such as recommender systems [1], [2], drug discovery [3], traffic prediction [4], [5], *etc*.. However, graph samples in user-related domains typically contain sensitive information. For example, graph-structured financial data may include sensitive account contents and transaction records [6], [7]. Therefore, the regulations like the general data protection regulation (GDPR)[1] underscore the growing importance of privacy protection in the process of data utilization [8]. Unfortunately, the private sensitive attributes in graphs will deteriorate the performance of pre-trained GNN models, which severely hinders their applications in high-risk scenarios [9], [10], [11].

To improve the reliability of GNNs, many studies [12], [13], [14], [15] have explored adaptive weight assignment and dynamic architecture design to extract high-quality embeddings. However, their heavy reliance on graph topology makes them vulnerable to distribution shifts. Adversaries can exploit this by crafting out-of-distribution (OOD) attacks through subtle structural perturbations [16], [17], [18], which mislead predictions during inference. Even minor modifications to graph structures have been shown to significantly impair model performance [16], [19].

Recent efforts have sought to defend GNNs against OOD attacks through robust aggregation [20], [21] and graph structure purification [22], [23]. While effective under structural perturbations, these defenses often rely on identifying anomalies such as added or removed edges. This makes them less reliable under complex distribution shifts, where the boundary between malicious and benign changes becomes blurred. In

Received 28 September 2024; revised 1 June 2025 and 24 September 2025; accepted 14 November 2025. Date of current version 26 November 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62402106 and Grant 62372334, in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK20241272, in part by the Fundamental Research Funds for the Central Universities under Grant 2242025K30025, and in part by the Start-Up Research Fund of Southeast University under Grant RF1028623129. The associate editor coordinating the review of this article and approving it for publication was Prof. Luisa Verdoliva. *(Corresponding author: Ju Jia.)*

Ju Jia and Renjie Li are with the School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China, and also with the Engineering Research Center of Blockchain Application, Supervision and Management, Southeast University, Ministry of Education, Nanjing 211189, China (e-mail: jiaju@seu.edu.cn; lirenjie@seu.edu.cn).

Cong Wu is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: congwu@hku.hk).

Yebo Feng is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: yebo.feng@ntu.edu.sg).

Siqi Ma is with the School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia (e-mail: siqim@uow.edu.au).

Lina Wang is with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: lnwang@whu.edu.cn).

Robert H. Deng is with the School of Information Systems, Singapore Management University, Singapore 188065 (e-mail: robertdeng@smu.edu.sg).

Digital Object Identifier 10.1109/TIFS.2025.3635052

[1]https://gdpr-info.eu/

addition, the privacy noise interferes with the correlations between nodes making OOD attacks more deceptive, which may restrict the utilization of defense methods. This increase in deception is due to privacy perturbations that alter the intrinsic correlation between nodes. When OOD attacks further manipulate the graph, the resulting structural anomalies become more difficult to detect because they may be statistically confused with legitimate privacy noise. As a result, existing defense methods that rely on identifying irregular patterns or purifying anomalous edges may misclassify benign perturbations as attacks and vice versa, which reduces their effectiveness. It is worth noting that in this work, we do not aim to design formal privacy protection mechanisms [24], [25] such as differential privacy. Instead, we employ the term privacy-perceptual to emphasize the ability of models to capture essential patterns from data that inherently includes privacy-sensitive noise.

Specifically, the privacy operations and distribution shifts of graph-structured samples pose substantial challenges for capturing distinctive patterns accurately and adjusting manipulated variables properly in GNNs against deceptive OOD attacks as follows: 1) the privacy noise injected through perturbation operations may induce GNN models to learn superficial surface knowledge rather than essential crucial information; 2) since promising graph representations inevitably rely on robust message propagation and neighborhood aggregation, it is difficult to effectively mitigate the impact of cross-domain distribution intervals by thoroughly purifying graph structures; and 3) the consistent correlations are hard to be fully mined and accurately exploited under the condition of privacy perturbations and distribution shifts, which has a negative influence on the overall performance against deceptive OOD attacks.

To overcome the aforementioned challenges, we propose an environment-adaptive representation interaction (EARI) scheme that learns discriminative representations from privacy-perturbed graph samples and explores intrinsic private patterns, defending against deceptive OOD attacks. Specifically, the targeted embedding propagation is exploited to distinguish between non-private and private data through the guidance of confidence score feedback in an interactive manner. Subsequently, the representation-enriched topological aggregation is achieved by leveraging multi-hop neighborhoods rather than stacking multiple layers to automatically capture more comprehensive features. Ultimately, the adaptive learning at the cluster level attempts to explore and refine the distinctive correlations from the topology level across different environments. Figure 1 illustrates the comparison of our proposed scheme with methods with different privacy-perceptual capabilities under private noise scenarios and deceptive OOD attacks. Our scheme successfully mitigates the interference of noise perturbations on defense effectiveness without compromising the private task performance to avoid distribution shifts. Empirical evaluations show that the proposed EARI outperforms current state-of-the-art methods in comprehensive conditions against deceptive OOD attacks. Notably, our experiments demonstrate that privacy-perceptual GNNs can avoid overdependence on the privacy intensity of graph-structured samples to boost the stability in practical applications. In summary, our contributions are as follows:

- To explore private patterns in graph-structured samples, we introduce the targeted embedding propagation to alleviate noise perturbations by the guidance of confidence score feedback from non-private data to private data.
- To highlight the invariant essence of graph correlations, we design the representation-enriched topological aggregation to facilitate the integration of diversity-rich features through the utilization of multi-hop neighborhoods rather than stacked multilayers.
- To further perceive graph representations across private environments, we construct the generalization-enhanced cluster-wise adaptation learning to approximate unknown domains through information interaction between individuals and clusters.
- To the best of our knowledge, this is the first work to propose an environment-adaptive representation interaction to resist deceptive OOD attacks from privacy operations and distribution shifts on various graph-structured datasets. Comprehensive experimental results demonstrate that our scheme can profoundly perceive the confused distribution shifts and sufficiently reap the biased correction benefits to avoid the risk of adverse effects.

The rest of the paper is structured as follows. Section II reviews the related work. Section III introduces the preliminary knowledge of the proposed scheme. Section IV provides the detailed description of our scheme. Section V presents the experimental evaluation results. Finally, we provide the brief discussion on conclusions and future directions in Section VI.

II. Related Work

A. Deep Graph Representation Learning

Deep graph representation learning aims to learn embedding representations from entire graph-structured data using deep aggregation and propagation techniques to handle downstream-related tasks [26], such as node recognition [27], graph classification [28], link prediction [29], *etc*.. Among the approaches of deep graph representation learning, GNNs have attracted increasing interest in recent years. The classical GNNs, including GCN [30], GAT [31], and GraphSAGE [32], have demonstrated outstanding performance in shallow models by leveraging the concept of neighborhood aggregation. More recently, to improve the computational efficiency, Wu et al. [33] achieved faster computation by simplifying the model of GNN, but it only considered features with maximum hop counts, which limited the ability of feature extraction. In comparison, Bo et al. [12] and Luan et al. [13] enhanced feature learning by applying low-pass and high-pass filters in each graph convolutional layer to extract richer localized information. To facilitate the extraction of targeted graph representations, Chien et al. [14] adaptively learned PageRank weights to achieve the joint optimization of node features and topological structures. Yan et al. [34] extended the adaptive weight learning by introducing the relative degree of nodes compared to neighbors, which can enhance the

model prediction performance. Chen et al. [35] proposed the multi-task learning strategy to improve the performance of decoupled GNNs through the self-supervised objectives. Luo et al. [15] adopted the mixed pseudo-labeling mechanism with the combination of confidence and uncertainty to boost the performance in semi-supervised node classification tasks. However, the graph representation learning without considering the anomalous situations typically assumes that the distribution of the training data is the same as that of the testing data. As a result, the performance of these methods may be limited and susceptible to privacy attitudes or noise perturbations when confronted with deceptive OOD attacks. Therefore, our scheme focuses on learning private patterns and deriving invariant representations across different environments to cope with the interference of privacy noise due to the distribution shifts. Unlike existing methods that are designed for clean or fully observable graphs, our framework introduces a targeted embedding propagation mechanism guided by confidence feedback, which allows more effective utilization of already-perturbed private nodes during representation learning. This enables the model to retain semantic utility under privacy noise and distribution shifts, providing a novel architectural perspective for building GNNs that can operate reliably on perturbed graph data.

B. Topology-Aware Graph Clustering

As GNNs become more and more powerful, this makes graph learning highly desirable and popular. The graph clustering is the classic data mining task in this field, which aims to divide graph nodes into cohesive clusters without manual labeling [36]. Early methods relied solely on restricted node features and simple edge connections. However, as the complexity of graph data increases, the methodologies that incorporate neural networks to capture underlying graph topological features have become mainstream [37], [38]. For instance, Wang et al. [37] introduced the attention networks to weigh the importance of neighboring nodes through embedding learning and graph clustering. Liu et al. [38] improved the clustering performance by dynamically adjusting sample weights based on attributes and structures. Despite the success, these methods still suffer from the over-smoothing problem. Bo et al. [39] and Tu et al. [40] mitigated this issue through the combination of dual self-supervision and structural attribute. Liu et al. [41] improved the clustering algorithms by utilizing the dual information decorrelation mechanism to enhance the discrimination of features. In addition, the optimal clustering is also beneficial for the training of neural networks. Tsitsulin et al. [42] and Liu et al. [43] bridged traditional clustering objectives to strengthen the node representation through self-supervised learning and adversarial optimization. Unlike previous studies that ignore the interplay between local features and cross-environmental variations, our EARI explicitly incorporates environmentally-adaptive clustering based on node correlation to support robust representation transfer. This enables domain-level generalization across noisy private environments. It provides a new perspective on integrating clustering adaptation into graph learning under privacy constraints.

C. Adversarial Graph-Structured Defense

Due to the unique structural characteristics and processing mechanisms of GNNs, they are particularly vulnerable to the variety of malicious attacks, which significantly limits their application in high-risk decision-making systems [44]. The existing methods [22], [23], [45], [46] generally focus on graph sample purification to optimize nodes and links in polluted graph-structured data by detecting anomalies and correcting errors. For instance, Wu et al. [45] employed Jaccard similarity to observe contrastive differences in graph samples by eliminating abnormal links between suspicious nodes and adversarial perturbations. Jin et al. [46] improved the defense by purifying the perturbed graphs based on the characteristics of adversarial samples. Nevertheless, if the original features fail to accurately and deeply describe the graph topology, the structural perturbations can easily exploit the vulnerability of GNNs. To address this issue, Li et al. [23] designed the unsupervised GCN to acquire optimized graph structures, which can enhance the defense robustness without sacrificing computational efficiency. In a more lightweight and model-agnostic manner, Li et al. [22] proposed a universal binary edge mask as a defense patch to eliminate suspicious links. The patch can be applied to any node without requiring access to model internals, making the defense broadly applicable and easy to deploy. Additionally, the methods based on robust aggregation training have been employed to improve the message passing and neighborhood aggregation [16], [47], [48]. For example, Zhu et al. [47] enhanced the model robustness by dynamically adjusting the weights of suspicious samples via the variance-based attention mechanism. Kong et al. [48] iteratively augmented node features with gradient-based adversarial perturbations, which allowed the model to be robust against small fluctuations in the graph data. However, the existing linear neighborhood aggregation easily leads to the excessive smoothing, while the maximal aggregation fails to completely capture the detailed information of neighboring node representations. To overcome these challenges, Wang et al. [49] developed a general non-linear aggregator between maximal and linear aggregation, which can provide excellent non-linearity and complementary sensitivity to reinforce the robustness of the adversarial defense. To further improve generalization under noise, Ennadir et al. [20] injected random noise into hidden representations during training, which enhances robustness while maintaining low computational overhead. In contrast to robust aggregation training, some research [21], [50] focused on improving the stability of graph representation learning to enable graph models to make objective and accurate decisions on distribution-shifted graph data, thus enhancing model defense against adversarial attacks. Jia et al. [21] proposed a causal relationship alignment-based structural rationalization scheme to counter adversarial perturbation attacks on graph data, primarily by implementing interventions to construct invariant causal rationales, thereby enhancing the defense capabilities of graph neural network models. While these approaches have achieved notable results in defending against adversarial perturbations, they frequently fail to take into account the interference of private data on the effectiveness of the defense. Moreover, the enhancement

of such defense tends to sacrifice the performance on clean samples (*i.e.*, non-attacked graphs). In contrast, our scheme combines multi-hop semantic aggregation with adaptive perceptual learning to provide a cohesive framework that can defend against both privacy noise and OOD perturbations. This work pushes the boundaries of privacy-constrained defense strategies against OOD attacks, which has rarely been explored in the previous graph model defense literature.

III. Preliminary

A. Notation

In this work, we primarily focus on the node classification tasks. Accordingly, some necessary definitions employed in this paper are introduced as follows. Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph with private operations (*e.g.*, differential perturbation), where $\mathcal{V} = \{v_1, v_2, v_3, \ldots, v_N\}$ represents the set of nodes. The adjacency matrix of the nodes is defined as $\mathbf{A} \in \{0,1\}^{N\times N}$. The embedding feature matrix of all nodes is denoted as $\mathbf{H} \in \mathbb{R}^{N\times d}$, where d is the feature dimension, N is the total number of nodes, and the node labels are represented by $\mathbf{Y} \in \mathbb{R}^{N\times c}$, where c is the number of label categories. Given a set of labeled nodes $\mathcal{V}_b$, the goal of node classification task is to predict the labels of unlabeled nodes by using the adjacency matrix $\mathbf{A}$ and the node embedding feature matrix $\mathbf{H}$. In practice, both $\mathbf{A}$ and $\mathbf{H}$ are used as inputs to a GNN, where each representation of node is updated by aggregating information from its neighbors as defined by $\mathbf{A}$. This iterative message-passing process generates task-relevant embeddings, and a final classifier (*e.g.*, the softmax layer) is applied to the output node features to predict their corresponding labels.

B. Threat Model

We define deceptive OOD attacks as structural perturbations that intentionally introduce distributional shifts in graph data, leading to generalization errors in inference. These attacks are particularly threatening in practical applications because they are able to mislead the model without changing its parameters. Based on this understanding, we describe the following threat assumptions for various attack scenarios.

1) Attacker's Knowledge: In deceptive OOD attacks, the attacker is assumed to operate under a black-box setting where the model architecture and parameters are unknown and cannot be altered. However, the attacker has full access to the input graph data and labels employed by the target model, which eliminates the need to train a shadow model for black-box approximation. This does not mean that an attacker can influence the training process of the model, just that they can observe the same data that the model employs for training and reasoning. This setup reflects realistic scenarios in privacy-sensitive applications, where data exposure may occur but model internals remain protected.

2) Attacker's Goal: The attackers aim to increase the distribution shifts of graph samples within the limited attack budget, which can mislead the prediction results of target model during the testing phase even with small perturbations. Specifically, given a graph G_t for prediction, attackers may modify it to a perturbed graph G'_t, so that the target model can classify a node v_t labeled Y_{v_t} in G_t as Y'_{v_t} with $Y'_{v_t} \neq Y_{v_t}$. In real-world scenarios, since the target model is usually inaccessible to attackers, we select to allow attackers to execute surrogate attacks through the proxy model while maintaining open access to the testing data. This setup is equivalent to the standard circumvention attack setting. In such applications, model parameters and training data are protected (*e.g.*, via differential privacy), but the inference API or test inputs may be observable or manipulable. In contrast to poisoning attacks, evasion attacks allow an adversary to covertly induce prediction errors without affecting the model training process or internal parameters. Typically, the OOD attack can be classified into direct attacks that modify the target samples and indirect attacks that perturb the adjacent samples. However, in practice applications, the direct attacks usually produce superior results [51], so the direct attacks are selected on testing graphs.

The attacker's goal is to introduce distributional shifts into the input graph while remaining within a limited perturbation budget. These shifts aim to degrade the generalization ability of the model at inference time, leading to incorrect predictions even when the modifications are subtle. Specifically, given a graph G_t at test time, the attacker modifies it to a perturbed graph G'_t, such that a node v_t with ground-truth label Y_{v_t} is misclassified by the target model as $Y'_{v_t} \neq Y_{v_t}$. Since the target model is not accessible, surrogate attacks are executed using a proxy model, under the assumption that the attacker can simulate similar behavior by leveraging the same input data. In addition to direct and indirect manipulation, OOD attacks can be categorized by their degree of restriction:

- **Restricted OOD Attacks:** These attacks operate under explicit constraints, such as bounded perturbation budgets, fixed target areas (*e.g.*, specific node neighborhoods or clusters), or structural constraints (*e.g.*, the maximum degree of variation). This type of attack simulates subtle localized changes. Depending on the mode of operation they can be subdivided into locally directed attacks and budgeted global distribution attacks.
- **Unconstrained OOD Attacks:** These attacks apply perturbations globally or without specific constraints to simulate structural anomalies or graph domain shifts. However, in reality, unconstrained perturbations are easily detected and isolated by defense methods. Therefore, this type of attack is less often considered.

IV. Proposed EARI Scheme

A. An Overview of the Proposed Scheme

In this section, we elaborate on the proposed EARI. The architecture and workflow of our scheme are illustrated in Figure 2. The ultimate goal of our scheme is to guarantee that the model can effectively perform the challenging tasks even if the graph samples with private perturbations are manipulated by adversaries under deceptive OOD attacks. It can be divided into the following three main components. (a) Private attribute-perceived embedding propagation. Both non-private and private data participate in the process of sensitive attribute perception through targeted embedding propagation,

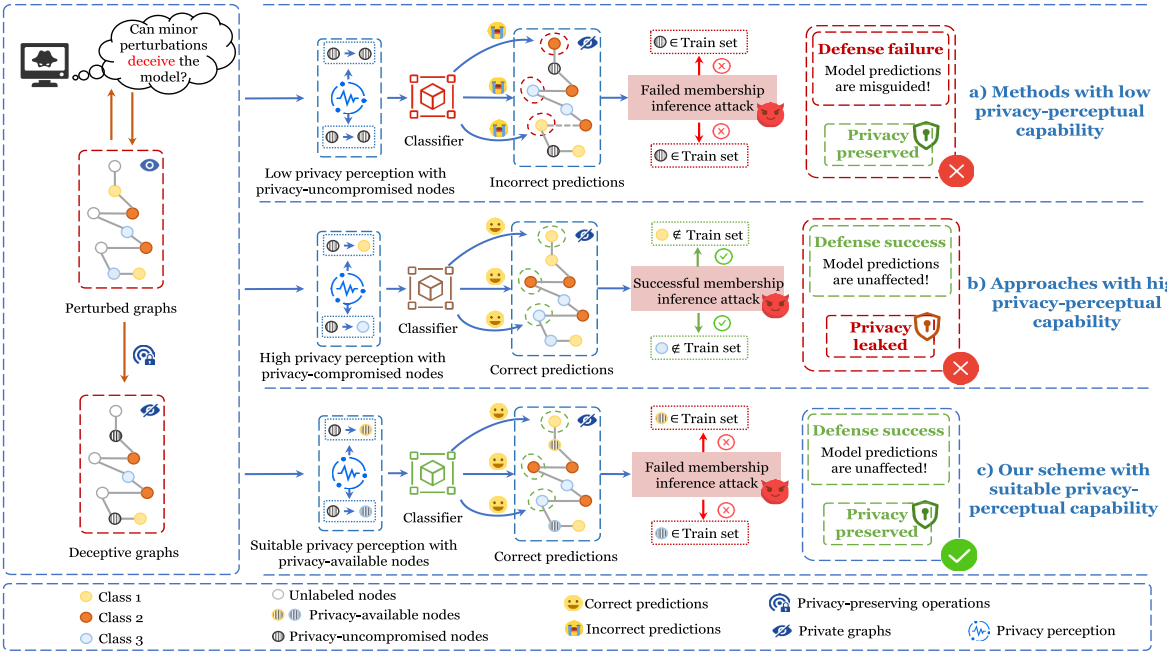

Fig. 1. A comparative illustration of three methods with different privacy-perceptual capabilities under deceptive OOD attacks.

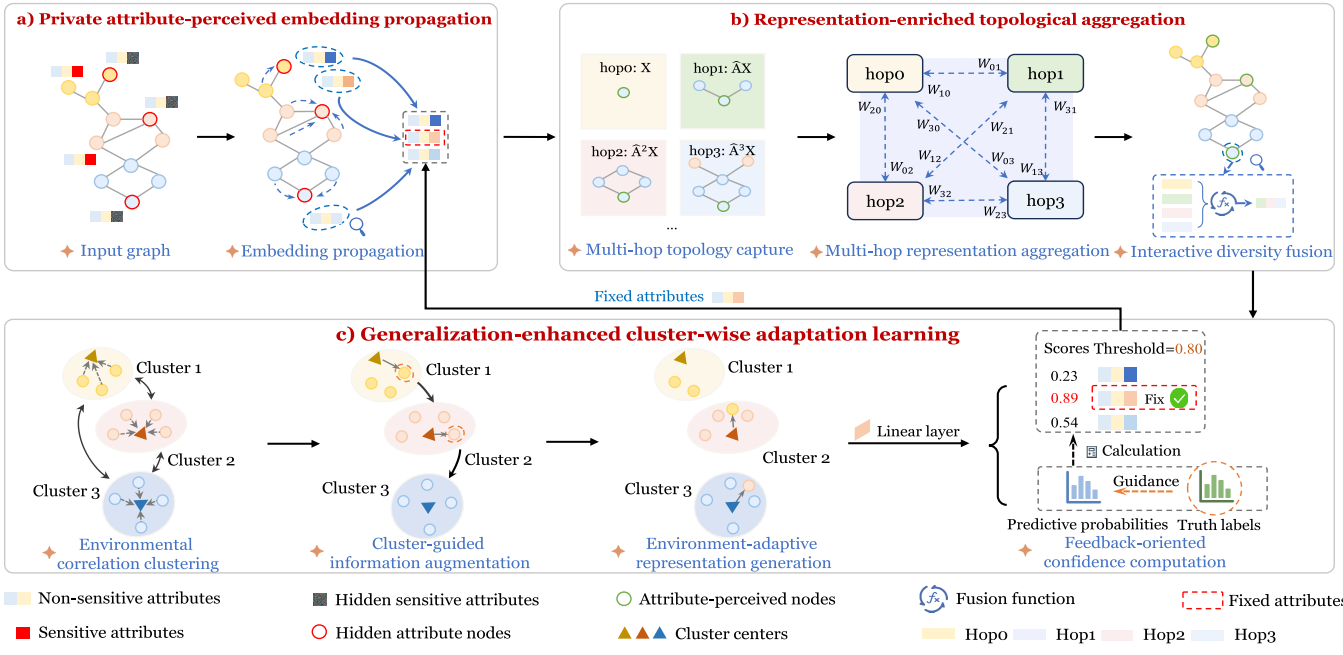

Fig. 2. The overall framework of the proposed EARI consists of three phases to enhance the privacy perception of GNN models against deceptive OOD attacks by exploring intrinsic private patterns and enriching latent semantic representations.

which can be guided by the feedback-oriented confidence computation. (b) Representation-enriched topological aggregation. The multi-hop features captured from embedding propagation graph data are employed to enrich individual characteristics and generate interactive representations through powerful topological aggregation. (c) Generalization-enhanced cluster-wise adaptation learning. This component utilizes continuous information interactions and invariant essential correlations between nodes and clusters to accommodate different environments, which can facilitate the acquisition of environment-adaptive representations.

B. Private Attribute-Perceived Embedding Propagation

Given the large amount of privacy-sensitive information that may be contained in graphs, the straightforward learning from these samples is not conducive to the discovery of underlying clues, which inevitably leads to the utilization of superficial

Algorithm 1 Private Attribute-Perceived Embedding Propagation

Input: Adjacency matrix $\mathbf{A}$, degree matrix $\mathbf{D}$, embedding vectors of nodes $\mathbf{H}$, private node index idx, initial iteration number $iters$;
Output: Post-propagation embedding feature matrix $\widetilde{\mathbf{H}}$;

```
Perform Laplacian normalization on the adjacency matrix A according to Eq. (1);
Classify private embedding matrix H_p and non-private embedding matrix H_o based on private node index idx;
for i to iters do
    Propagate embedding information by H' = ÃH;
    if H_d ≠ 0 and H_d ≠ 1 then
        Constrain the results as 0 or 1 based on Eq. (2);
    end
    Reset feature values of H_o;
    Normalize H through Eq. (4);
end
return H̃;
```

noisy features. In addition, this intricate phenomenon is exacerbated as the intensity of privacy protection increases, which in turn enforces the design of suitable methods to suppress the impact of injected erratic noise (*i.e.*, private perturbations). The private attribute-perceived embedding propagation is exploited to relieve the disturbance of unstable noise. Specifically, the Laplacian normalization of adjacency matrix is represented as follows:

$$\widetilde{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \tag{1}$$

where $\mathbf{A}$ is the adjacency matrix, and $\mathbf{D}$ denotes the degree matrix. We assume that the set of private nodes is recorded as $\mathcal{V}_p$, and the set of non-private nodes is represented as $\mathcal{V}_o$. The embedding matrices are denoted as $\mathbf{H}_p$ and $\mathbf{H}_o$, respectively. The main steps of embedding propagation can be expressed as $\mathbf{H} = \widetilde{\mathbf{A}}\mathbf{H}$.

When the attribute values of features $\mathbf{H}_d$ are represented in binary form, the calculation results need to be constrained:

$$\mathbf{H}_d = \begin{cases} 0 & \mathbf{H}_d \le 0.5 \\ 1 & \mathbf{H}_d > 0.5 \end{cases}. \tag{2}$$

For continuous attribute values, this step is omitted. Subsequently, the feature values of non-private nodes are reset to the original values before executing the embedding propagation, which updates only the embedded features of private nodes. The above steps need to be iteratively performed during the training process, and the formula for γ iterations can be written as:

$$\mathbf{H}^{(\gamma+1)} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A}_{op} & \mathbf{A}_{pp} \end{bmatrix} \mathbf{H}^{(\gamma)}, \tag{3}$$

where $\mathbf{A}_{op}$ and $\mathbf{A}_{pp}$ are the adjacency matrix from non-private nodes to private nodes, and $\mathbf{I}$ is the unit matrix. Since the attribute-perceived embedding propagation algorithm updates the private features by gradually propagating information from neighboring samples, this sufficient interaction reduces the interference between private and non-private nodes during each iteration. Therefore, it is possible to suppose a situation where the absence of non-private nodes around private nodes would affect the convergence speed of the model training.

To avoid this issue, the private attribute-perceived embedding propagation performs several pre-iterations to ensure that each neighbor of nodes contains valuable features at the beginning of the training phase. In this way, even if the private features of nodes are initialized to zero, the convergence speed of the model is still guaranteed after more than 40 executions through embedding propagation [52]. Furthermore, the repeated iterations can lead to the over-amplification of the embedded eigenvalues for the private nodes relative to the other nodes, which makes the imposition of constraints on the eigenvectors in consecutive iterations indispensable. Consequently, it is necessary to normalize the embedding features $\mathbf{H}$, which ensures that the L_2 norm of each row is set to 1, as illustrated by the following formula:

$$\widetilde{\mathbf{H}} = \frac{\mathbf{H}}{\|\mathbf{H}\|_2}, \tag{4}$$

where $\widetilde{\mathbf{H}}$ is the embedding feature matrix after normalization. The overall flow of the private attribute-perceived embedding propagation is shown in Algorithm 1. Such iterative aggregation computations have the significant advantage that each aggregation served as the optimal filter can effectively suppress the noise perturbations generated by the privacy-preserving operations to understand intricate private patterns [53], [54], [55].

C. Representation-Enriched Topological Aggregation

The enhancement of representation learning capability typically requires the utilization of multi-level features. For GNNs, the collection of useful messages from adjacent neighbors is a straightforward and effective method. Unlike traditional GNNs, our scheme concentrates on utilizing multi-hop neighborhoods rather than stacked multilayers to perceive private patterns within a U-hop range. The multi-hop node information is computed based on the obtained embedding features, and U-hop features can be represented as follows:

$$\widetilde{\mathbf{X}} = [\mathbf{X}^0; \mathbf{X}^1; \ldots; \mathbf{X}^U], \quad \mathbf{X}^u = \widetilde{\mathbf{A}}^u\widetilde{\mathbf{H}}, \tag{5}$$

where $\mathbf{X}^u$ is the u-th hop matrix. The processing step does not require any learnable parameters by recalculating $\widetilde{\mathbf{X}}$ when the embedding features are updated. Normally, it needs to be computed only once, which improves the robustness across different datasets through the adaptation of mini-batch training.

Since the calculation of multi-hop representations does not involve learnable parameters, it may limit the expressiveness by relying only on U-hop features. To address this problem, we design the shared linear layer $f(\cdot)$, which performs the trainable linear transformation of the hop features into multi-hop embeddings. Additionally, the sequential relations between multi-hop features facilitate to capture the crucial changes for privacy-sensitive nodes. To enhance the expressiveness, the learnable hop matrix $\mathbf{\Phi}$ is incorporated into the

multi-hop embeddings. The entire transformation process can be represented as:

$$\mathbf{E} = \left[f(\mathbf{X}^0); f(\mathbf{X}^1); \ldots; f(\mathbf{X}^U)\right] + \mathbf{\Phi}, \tag{6}$$

where $\mathbf{E}$ represents the multi-hop embeddings with temporal contextual information. The learnable hop matrix is uniformly concatenated with each transformed embedding to control the weights of features through the depiction of accurate representations across different hops.

Previous studies [35], [52] have shown that there are correlations between the multi-hop features, which to some extent represents the intrinsic characteristics of graphs. Therefore, we replace the traditional message passing with the nonlinear representation interaction to enhance the distinguishability between known and unknown domains. Our scheme is based on the development of leveraging the message propagation to facilitate the interactive learning by aggregating topological knowledge in multi-hop graph representations. Specifically, given a node $v_i \in \mathcal{V}$, the fully connected multi-hop feature graph $\mathbf{G}_i^{\text{hop}}$ is constructed, where each node v'_{iu} corresponds to the u-th hop feature of the original node v_i. To alleviate the vanishing gradient problem, we introduce the residual connections to complete the aggregation process of multi-hop representations:

$$\mathbf{E}_i^l = g(\mathbf{G}_i^{\text{hop}}, \mathbf{E}_i^{l-1}) + \mathbf{E}_i^{l-1}, \quad l = 1, \ldots, L, \tag{7}$$

where $\mathbf{E}_i^l \in \mathbb{R}^{U \times d}$ represents the U-hop features of node v_i after the l-th iteration of multi-hop representation aggregation, with L being the number of executions for topological aggregation and $g(\cdot)$ is an arbitrary GNN model.

Since the constructed hop features are fully connected, the multi-head self-attention [31] is employed to capture the relationships between pairwise graph representations. The calculation can be expressed as follows:

$$\mathbf{S}_i^h = \text{Softmax}\left(\frac{\mathbf{E}_i \mathbf{W}_Q^h (\mathbf{E}_i \mathbf{W}_K^h)^T}{\sqrt{d_H}}\right), \quad d_H = \frac{d}{H}, \tag{8}$$

$$head_i^h = \mathbf{S}_i^h \mathbf{E}_i \mathbf{W}_V^h, \quad h = 1, \ldots, H, \tag{9}$$

$$g(\mathbf{G}_i^{\text{hop}}, \mathbf{E}_i) = \text{Concat}(head_i^1, \ldots, head_i^H)\mathbf{W}_O. \tag{10}$$

For simplicity, the formula is exemplified by a single-layer interaction that the superscript of the hop embedding feature $\mathbf{E}_i$ is not l. The learnable weight matrices are denoted as $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$, and $\mathbf{W}_O$. The d_H represents the feature dimension of each attention head, H is the number of attention heads, and h denotes the h-th attention head.

The $\mathbf{S}_i$ is the weight matrix to capture the interaction behaviors between multi-hop representations of nodes. In this context, the utilization of the multi-head attention mechanism can be considered as the basic GNN (*e.g.*, GAT) for the realization of hop interactions. After the multi-hop representation aggregation, the embedding feature for each node essentially remains in the $U \times d$ dimensional space. To obtain the comprehensive representation for each node, it is necessary to incorporate diverse multi-hop features. Specifically, the primary fusion technique is applied to the multi-hop aggregated representations by averaging the single feature embeddings. The node representation $\mathbf{Z}$ after diversity fusion encompass the commonality of multi-hop features within each node. This aggregation of topological information enhances the recognizability of the nodes against deceptive OOD attacks.

D. Generalization-Enhanced Cluster-Wise Adaptation Learning

Following the implementation of topological aggregation in Section IV-C, the high-quality node representations have been obtained. However, it is evident that these learned representations heavily depend on the topological structure of the surrounding neighbors. Any changes in the topological correlations (*i.e.*, environmental structures) can significantly affect the extraction of representations for nodes. Typically, the characteristics of neighboring nodes can be captured by the cluster information, which is related to the environmental conditions across different views. Therefore, the distribution shifts caused by changes in graph-structured samples can be seen as the impacts on the clustering results to some extent.

To overcome the vulnerability of topological fluctuations, we propose the generalization-enhanced cluster-wise adaptation learning that enables the model to learn invariant features of the data under different environmental structures through the interaction between clusters and nodes. Specifically, the multilayer perceptron (MLP) is designed to compute the cluster assignment matrix $\mathbf{P}$ with the softmax activation function:

$$\mathbf{P} = \text{MLP}(\mathbf{Z}, \Theta_{\text{MLP}}), \tag{11}$$

where Θ_{MLP} denotes the parameters of MLP. Each element p_{ij} of $\mathbf{P}$ represents the probability that the node v_i belongs to the cluster j. The softmax activation function ensures that the values are within the range $[0, 1]$, and the sum of the probability for each node across all clusters equals 1.

To evaluate the effectiveness of the clustering, the cut loss $\mathcal{L}_c$ is introduced, which encourages the assignment of strongly connected nodes to the same cluster:

$$\mathcal{L}_c = -\frac{tr(\mathbf{P}^T \widetilde{\mathbf{A}} \mathbf{P})}{tr(\mathbf{P}^T \mathbf{D} \mathbf{P})}, \tag{12}$$

where $tr(\cdot)$ is the trace of the matrix. The cut loss $\mathcal{L}_c$ ranges from $[-1, 0]$. The smaller values indicate stronger connections between nodes within the same cluster. However, since $\mathcal{L}_c$ is a non-concave function, the direct minimization of $\mathcal{L}_c$ can lead to the cluster collapse that all nodes are assigned to a single cluster. To prevent this, we adopt the orthogonality loss to ensure similar cluster sizes, which can be expressed as:

$$\mathcal{L}_o = \left\| \frac{\mathbf{P}^T \mathbf{P}}{\|\mathbf{P}^T \mathbf{P}\|_F} - \frac{\mathbf{I}_M}{\sqrt{M}} \right\|_F, \tag{13}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The orthogonality loss ranges from $[0, 2]$, and it is minimized to encourage orthogonal cluster assignments and suitable cluster sizes when $\mathbf{P}^T \mathbf{P} = \mathbf{I}_M$.

Overall, the total clustering loss is the combination of these two losses:

$$\mathcal{L}_{clu} = \mathcal{L}_c + \lambda_1 \mathcal{L}_o, \tag{14}$$

where λ_1 is the balancing coefficient. The minimization of the combined loss $\mathcal{L}_{clu}$ can facilitate to obtain meaningful and consistent clusters. Subsequently, the cluster centers $\mathbf{C}$

are calculated by averaging the node representations $\mathbf{Z}$ as $\mathbf{C} = \mathbf{P}^T\mathbf{Z}$. The cluster centers summarize the features of each cluster, while each node is only associated with one cluster. To maintain the high discriminative representation during structural changes, we adopt the clustering-oriented strategy that empowers nodes to adapt to different environments across unknown domains, which can be derived by clustering information. Unlike cluster centers, the cluster information can provide the comprehensive descriptions, which is determined by both the corresponding cluster center and the cluster standard deviation. The standard deviation of the clusters $\sigma(\mathbf{C})$ can be computed as follows:

$$\sigma(\mathbf{C}_m) = \left(\sum_{i=1}^{n} p_{im}(\mathbf{Z}_i - \mathbf{C}_m)^2\right)^{1/2} \tag{15}$$

where $\sigma(\mathbf{C}_m)$ is the standard deviation of the m-th cluster. Subsequently, the node v_i can be transferred to the different environments, such as from cluster m to cluster j, as follows:

$$\mathbf{Z}'_i = \mathbf{C}_j + \sigma(\mathbf{C}_j)\left(\frac{\mathbf{Z}_i - \mathbf{C}_m}{\sigma(\mathbf{C}_m)}\right), \tag{16}$$

where $\mathbf{Z}'_i$ denotes the node representation adapted to the unknown environment. The term $\mathbf{Z}_i - \mathbf{C}_m$ captures the deviation of node v_i from cluster centers, which eliminates the effects of nodes within the original clusters. This deviation can be viewed as the invariant correlations from nodes across different environments.

Since different clusters may have diverse distribution characteristics, the normalization and standardization are necessary to maintain the consistency of invariant correlations across various clusters. Specifically, the regularization operations are achieved by constructing stable topological structures, which reduces the amplitude of range variation from inconsistent distributions. Then, the derived features are rescaled by $\sigma(\mathbf{C}_j)$, which aims to adjust variable distributions across cluster environments. Finally, the scaled features are assigned to the cluster center $\mathbf{C}_j$ of cluster j while preserving the invariant correlations. The above steps can realize the adaptation of representations across diverse environments in graph-structured data. To further enhance the robustness under distribution shifts, The Gaussian perturbations are introduced to generate clusters with slight variations from the original environment [56], [57]. The uncertainty of the cluster center and the standard deviation can be estimated as follows:

$$\theta_\mu = \sigma(\mathbf{C}), \theta_\sigma = \sigma(\sigma(\mathbf{C})), \tag{17}$$

where θ_μ and θ_σ represent the uncertainty estimates for cluster center and standard deviation, respectively, which are applied to adjust the Gaussian noise in various clusters. Therefore, the cluster center and standard deviation can be transformed as follows:

$$\delta(\mathbf{C}_j) = \mathbf{C}_j + \epsilon_\mu\theta_\mu, \epsilon_\mu \sim \mathcal{N}(0,1), \tag{18}$$

$$\beta(\mathbf{C}_j) = \sigma(\mathbf{C}_j) + \epsilon_\sigma\theta_\sigma, \epsilon_\sigma \sim \mathcal{N}(0,1), \tag{19}$$

where $\delta(\mathbf{C}_j)$ and $\beta(\mathbf{C}_j)$ are standard normal random variables. The entire process can be expressed as:

$$\mathbf{Z}'_i = \delta(\mathbf{C}_j) + \beta(\mathbf{C}_j)\left(\frac{\mathbf{Z}_i - \mathbf{C}_m}{\sigma(\mathbf{C}_m)}\right). \tag{20}$$

The introduction of Gaussian noise facilitates the proposed scheme to capture invariant features from diverse environments while avoiding the complexity and uncertainty associated with unstable graph-structured samples. In each iteration, the subset of nodes v_{ij} is randomly selected, and their embeddings $\mathbf{Z}_{ij}$ are replaced with the environment-adaptive paradigms $\mathbf{Z}'_{ij}$ to obtain the updated representations $\mathbf{Z}'$. To ensure the stability during the training, the loss of crucial features is modified to incorporate the interactions among multiple clusters. In this way, the overall loss can be defined as follows:

$$\mathcal{L}_t = \mathcal{L}_{ce} + \lambda_2\mathcal{L}_{clu}, \tag{21}$$

where $\mathcal{L}_{ce}$ denotes the classification loss, and λ_2 is the balancing coefficient of the clustering process. The confidence score of nodes is calculated with the predicted loss to further ensure the feasibility of the privacy-perceptual embedding propagation across different environments. Specifically, the predicted probability of the true category for the private node v_p is considered as the confidence score, which ranges from 0 to 1. In the iteration of sensitive attribute-perception embedding propagation, the privacy nodes with confidence scores above the set threshold remain fixed, while the portions with scores below the threshold are unfixed. To prevent that none of the private nodes reaches the threshold, the dynamic adjustment mechanism is introduced to promptly update the threshold values. If no node reaches the threshold in the current iteration, the threshold is decreased by 5% until some private nodes are fixed in one iteration. In addition, such feedback-oriented dynamic adjustment mechanism guarantees the flexibility of the privacy-perceptual embedding propagation, which is conducive to improving the overall performance.

V. Experiments

In this section, our experimental evaluation is designed to answer the following research questions (RQs):

- **RQ1:** What is the impact of different privacy intensities on the effectiveness of our scheme?
- **RQ2:** How robust is the proposed scheme compared to other methods under deceptive OOD attacks?
- **RQ3:** What is the impact on the defensive capabilities of the methods as the strength of privacy protection increases?
- **RQ4:** Can our scheme maintain satisfactory performance under both transductive and inductive settings?
- **RQ5:** How sensitive is our scheme to the hyperparameters on various datasets under general attack scenarios?

A. Experimental Settings

1) Datasets. We utilize several widely-used publicly available benchmark datasets for the experiments, including Cora, Citeseer, Pubmed[2] [58], DBLPv7, Citationv1, and ACMv9[3] [59], to assess the effectiveness of the proposed scheme. A brief description of the statistics for these datasets can be found in Table I.

[2]https://pytorch-geometric.readthedocs.io/en/latest/index.html
[3]https://www.aminer.cn/citation

TABLE I
STATISTICS OVERVIEW OF GRAPH BENCHMARK DATASETS ADOPTED IN OUR EVALUATION EXPERIMENTS

Datasets	#Nodes	#Edges	#Features	#Classes
Cora	2,708	10,556	1,433	7
Citeseer	3,327	9,104	3,703	6
Pubmed	19,717	88,648	500	3
DBLPv7	5,484	8,130	6,775	5
Citationv1	8,935	15,113	6,775	5
ACMv9	9,360	15,602	6,775	5

TABLE II
CLASSIFICATION ACCURACY (%) RESULTS OF THREE DATASETS UNDER VARYING PRIVACY BUDGETS

Datasets	Privacy Budget			
	$\epsilon = 1$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$
Cora	86.35±0.37	87.45±0.18	88.75±0.74	88.93±0.84
Citeseer	73.27±0.30	75.88±0.13	76.73±0.30	76.75±0.62
Pubmed	82.52±0.17	84.84±0.15	85.37±0.04	85.51±0.07

*2) Attack Models.*The representative methods of OOD attacks are employed in the experiments. We first consider the random perturbation attack (RPA), and then select the state-of-the-art OOD attacks, including projected randomized block coordinate descent (PR-BCD) [16], greedy randomized block coordinate descent (GR-BCD) [16], locally constrained randomized block coordinate descent (LR-BCD) [18]), and partial graph attack (PGA) [17]. Both of these methods employ GCN as the surrogate model and perform evasion attacks on the victim model.

3) Defense Baselines. To evaluate the defense capability against deceptive OOD attacks, the classical GCN [30] is adopted as the baseline method. In addition, we compare the proposed EARI scheme with the following state-of-the-art baseline defense methods robust GCN (RGCN) [47] and MedianGCN [60], which improve the resistance to deceptive OOD attacks by enhancing the message-passing process and designing the aggregation function. In contrast, graph universal adversarial defense (GUARD) [22] and NoisyGNN [20] mitigate OOD perturbations by directly removing malicious edges and appropriately injecting random noise, respectively. We also include two recent OOD-aware graph learning methods in our comparison. ML-GOOD [61] performs multi-level outlier detection by modeling node-level and graph-level inconsistency in feature and structure space, which enables the model to identify and suppress distributional anomalies. NODESAFE [62] is a self-supervised framework that leverages feature augmentation and graph structure perturbation consistency to discriminate out-of-distribution nodes from in-distribution nodes, thereby facilitating effective OOD purification.

B. Evaluation of the Effectiveness for Privacy Perception (RQ1)

To assess the privacy-perceptual capabilities of the proposed scheme, we conduct experiments on three popular datasets (*i.e.*, Cora, Citeseer, and Pubmed). Specifically, differential privacy [53] is applied to protect the sensitive attributes. In this work, we assume that the input graphs have been preprocessed under a differential privacy mechanism. We consider the setting of node-level differential privacy (node-DP), where the privacy of individual nodes (*e.g.*, their features and associated connections) is protected through noise injection (*e.g.*, the Laplace mechanism). This ensures that the presence or absence of any single node does not significantly alter the model output. While our experiments are conducted under node-DP assumptions, the proposed EARI is inherently generalizable to edge-level differential privacy (edge-DP) settings, since its core components (targeted embedding propagation and topological aggregation) are agnostic to the source of perturbation. In edge-DP, where the perturbation occurs on graph structure rather than node features, our strategy still functions by leveraging robust multi-hop information propagation and semantic interaction to counteract structural distortion. Notably, both the training data and the test data are perturbed by the same processing and privacy budget in our setup, which reflects the reality that privatized graphical data are exposed to external analysis only after a uniform privacy-preserving transformation. For simplicity, the protection ratio is set to 50%. Subsequently, we vary the privacy budget ϵ from 1 to 16 and record the predictive accuracy of the model. From the experimental results shown in Table II, we observe that classification accuracy consistently improves as the privacy budget increases, which aligns with the expectation that a higher privacy budget indicates weaker privacy protection. Notably, the accuracy drops by only 4.53% when ϵ is reduced from 16 to 1, even on the Citeseer dataset where the difference in results is the largest. This demonstrates that our proposed scheme possesses excellent privacy-perceptual capabilities. We select the privacy budget $\epsilon = 8$ for the subsequent experimental setups, which maintains accurate results with a relatively small privacy budget.

TABLE III
RESULTS OF CLASSIFICATION ACCURACY (%) UNDER MISMATCHED DP MECHANISMS BETWEEN TRAINING AND TESTING. LAP = LAPLACE PERTURBATION, GAU = GAUSSIAN PERTURBATION

Training	Testing	RGCN	NoisyGNN	Ours
Lap	Lap	71.28±0.21	72.78±0.55	**76.73±0.30**
Gau	Gau	71.66±0.53	72.49±0.42	**76.10±0.58**
Lap	Gau	70.49±0.27	70.25±0.10	**75.83±0.22**
Gau	Lap	70.36±0.44	69.98±0.69	**75.34±0.76**

To further validate the practicality of the proposed scheme, we conduct additional experiments under mismatched differential privacy settings on the Citeseer dataset, where training and testing datasets are perturbed by different differential privacy mechanisms (*i.e.*, Laplace or Gaussian). Table III summarizes the experimental results. Despite performance degradation compared to the matched setting, EARI consistently outperforms baseline defense schemes, which indicates that its environment-adaptive design can be effectively extended to such mismatched privacy-perturbation scenarios.

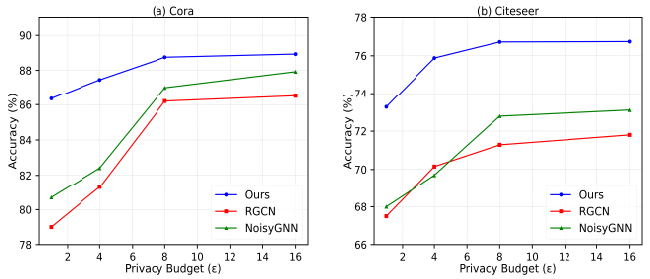

Fig. 3. Experimental evaluation results of model accuracy under varying privacy budgets ϵ on the test set.

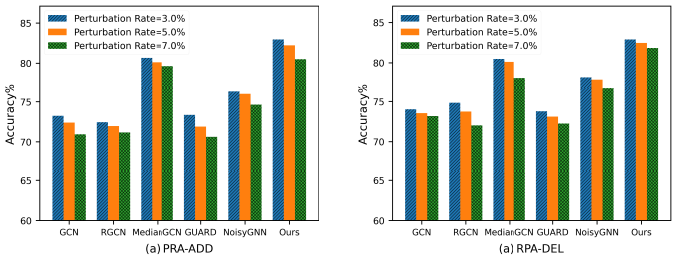

Fig. 4. Experimental evaluation results for defending against the RPA on the DBLPv7 dataset.

In addition to mechanism mismatches, another practical concern is the variation of privacy budgets between training and testing. We fix the training budget at $\epsilon = 8$ and vary the testing budget across $\{1, 4, 8, 16\}$. As shown in Figure 3, performance remains stable when the gap between training and test perturbations is small, but larger discrepancies cause drops. Notably, the accuracy decline becomes more pronounced when test data incorporates stronger noise perturbations ($\epsilon_{\text{train}} < \epsilon_{\text{test}}$), as the heightened noise within test data is more likely to undermine the environmental adaptability of the model during inference. These observations confirm the robustness of EARI in handling privacy intensity mismatches between training and testing environments.

C. Comparative Robustness Under Deceptive OOD Attacks (RQ2)

To evaluate the defensive effectiveness of the proposed scheme, we employ the five representative deceptive OOD attacks outlined in Section 5.2 for the following analyses. The selected OOD attacks include both localized and globally distributed perturbation patterns. Specifically, PR-BCD, GR-BCD, LR-BCD, and PGA are structured directional attacks and belong to locally directed attacks. In contrast, RPA performs random or adaptive modifications throughout the graph and belongs to the budgeted global distribution attack. All perturbations are performed under the same budget constraints, thus ensuring a fair and comprehensive evaluation of model robustness under different OOD threat scenarios.

1) Defense against RPA. We randomly add or delete edges in the DBLPv7 dataset with perturbation rates ranging from 3.0% to 7.0%. These perturbations construct two variants of the attacks (*i.e.*, RPA-ADD and RPA-DEL), which are tested on the perturbed graph by using our proposed scheme

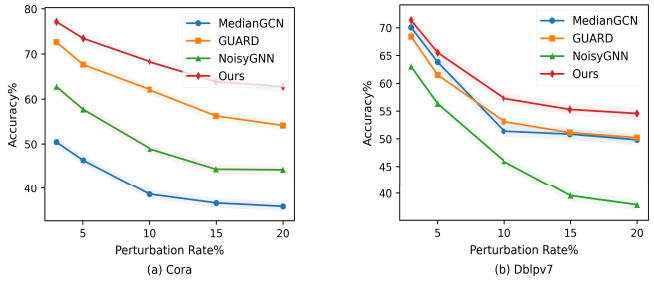

Fig. 5. Experimental evaluation results for defending against the GR-BCD on the Cora and Dblpv7 datasets.

with other defense methods. The results of these experiments are depicted in Figure 4. Under the RPA perturbation, we observe that GCN, GUARD, and RGCN can achieve the similar results. Moreover, the MedianGCN exhibits a more pronounced performance decline when the perturbation rate of RPA-DEL is increased from 5.0% to 7.0%. In contrast, our proposed scheme achieves the highest accuracy among the six defense methods and shows less fluctuation in performance across different perturbations.

2) Defense against PR-BCD. We conduct the PR-BCD attack by various perturbation rates on Citationv1 and Citeseer datasets under distribution shifts. Unlike RPA, the perturbations in PR-BCD attacks are targeted. Consequently, the baseline defense methods exhibit significant prediction accuracy degradation under PR-BCD attack compared to RPA. From Table IV, we notice that the effectiveness of the defense methods declines progressively as perturbation rates increase, which indicates the positive correlations between the intensity of PR-BCD and the difficulty of defense. Nonetheless, our defense scheme demonstrates remarkable robustness under these conditions, while achieving up to 12.52% performance improvement even at the maximum attack intensity. This clearly verifies that the proposed scheme can discriminate the critical properties of nodes even in the presence of PR-BCD perturbations.

To further strengthen the convincing argument that the proposed scheme is robust under deceptive OOD attacks, we additionally choose two node-level OOD detection methods, NODESAFE [62] and ML-GOOD [61] as comparison methods. Specifically, these methods first calculate the OOD score of each node. Subsequently, edges connected to high-scoring nodes are removed to suppress the impact of potential attacks. The results are shown in Table IV. Although both methods are able to moderately reduce the impact of PR-BCD attacks, their performance is still lower than that of our proposed method. We believe it is due to the fact that OOD attacks inherently tend to achieve their attack goals with small modifications, which leads to the fact that OOD detection is not easily aware of these tiny perturbations. And the normal nodes may also be misidentified as OOD nodes due to the presence of privacy noise. In contrast, the proposed EARI does not require explicit detection or pruning operations and is able to maintain stable performance under deceptive OOD attacks.

3) Defense against GR-BCD. We adopt the Cora and Dblpv7 datasets as targets for the GR-BCD attack. The GR-BCD is

TABLE IV

RESULTS OF CLASSIFICATION ACCURACY (%) ON BENCHMARK DATASETS UNDER THE CONDITION OF PR-BCD. THE BEST RESULTS ARE MARKED IN BOLD

Datasets	Perturbation Rate	GCN	RGCN	GUARD	NoisyGNN	NODESAFE	ML-GOOD	Ours
		ACC ▲						
Citationv1	0%	81.32±0.48	82.88±0.24	80.80±0.59	80.69±0.45	79.92±0.31	80.06±0.35	**87.66±0.06**
	3%	73.08±0.38	73.68±0.37	72.56±0.59	69.24±0.16	71.83±0.40	72.11±0.34	**76.43±0.19**
	5%	69.46±0.36	68.94±0.55	69.16±0.37	65.36±0.86	67.84±0.28	68.21±0.51	**71.36±0.18**
	10%	62.53±0.60	62.64±0.16	61.59±0.59	57.42±0.43	59.85±0.42	60.32±0.38	**64.73±0.60**
	15%	56.56±0.56	56.04±0.80	56.86±0.92	52.20±0.29	57.26±0.58	57.04±0.67	**59.88±0.90**
	20%	51.98±0.37	52.91±0.24	52.09±0.28	49.70±0.11	52.66±0.49	52.31±0.45	**57.08±0.74**
Citeseer	0%	69.62±0.44	70.58±0.19	69.87±0.14	71.98±0.19	69.92±0.22	69.14±0.26	**76.14±0.55**
	3%	61.36±0.51	61.16±0.28	61.56±0.42	61.16±1.44	61.84±0.62	62.06±0.35	**63.96±0.88**
	5%	57.66±1.85	58.16±0.99	57.06±0.88	54.75±0.14	55.04±1.02	56.48±0.73	**59.26±1.16**
	10%	41.74±0.24	41.24±0.86	42.24±0.14	41.24±1.23	42.08±0.91	42.36±0.65	**47.85±1.48**
	15%	30.13±1.35	31.13±0.28	29.83±1.39	31.63±0.28	31.84±0.94	31.42±1.06	**39.14±1.26**
	20%	26.23±1.60	24.72±0.28	27.43±1.39	22.72±1.16	26.06±0.85	27.19±1.17	**35.24±1.35**

TABLE V

RESULTS OF CLASSIFICATION ACCURACY (%) AND MACRO-F1 (%) ON BENCHMARK DATASETS UNDER THE CONDITION OF PGA. THE BEST RESULTS ARE MARKED IN BOLD

Datasets	Perturbation Rate	MedianGCN		GUARD		NoisyGNN		Ours	
		ACC ▲	Macro-F1 ▲	ACC ▲	Macro-F1 ▲	ACC ▲	Macro-F1 ▲	ACC ▲	Macro-F1 ▲
Dblpv7	0%	82.48±0.54	81.94±0.52	75.61±0.48	72.44±0.80	78.47±0.39	77.28±0.30	**84.85±0.37**	**84.09±0.54**
	3%	70.99±0.39	70.01±0.52	67.46±0.77	65.32±0.96	68.19±1.35	67.83±1.22	**76.09±0.55**	**74.75±0.51**
	5%	62.41±0.60	60.82±0.52	62.29±1.04	59.89±1.26	57.78±2.37	56.40±1.99	**69.44±0.45**	**67.26±0.33**
	10%	57.18±0.52	56.27±0.59	59.37±0.48	56.80±0.45	53.10±2.84	51.60±2.27	**65.88±1.09**	**63.84±0.58**
	15%	53.83±0.79	53.11±1.08	56.99±1.01	54.22±0.79	48.24±2.46	46.51±2.01	**63.41±1.18**	**61.05±0.98**
	20%	51.89±1.35	51.20±1.90	55.96±0.85	53.48±0.66	45.86±2.13	44.69±1.93	**62.69±1.18**	**60.12±0.76**
ACMv9	0%	78.42±0.53	80.09±0.47	74.75±0.25	76.76±0.45	75.85±0.23	78.06±0.26	**82.52±0.27**	**84.16±0.55**
	3%	67.59±0.57	68.72±0.40	65.85±0.28	66.66±0.53	64.60±0.44	65.80±0.35	**75.00±0.23**	**76.42±0.34**
	5%	60.61±0.45	61.37±0.76	61.40±0.28	62.15±0.53	56.84±0.57	57.80±0.85	**69.23±0.23**	**70.84±0.16**
	10%	56.66±0.13	57.87±0.30	57.94±0.05	58.66±0.40	52.53±1.27	53.61±1.27	**65.70±0.30**	**67.59±0.35**
	15%	53.99±1.04	54.43±1.34	55.63±0.28	56.27±0.74	48.93±1.15	49.94±1.02	**63.99±0.60**	**65.33±0.12**
	20%	53.14±0.61	53.34±0.65	54.81±0.32	55.35±0.69	47.40±1.40	48.19±1.48	**62.46±0.43**	**63.64±0.74**

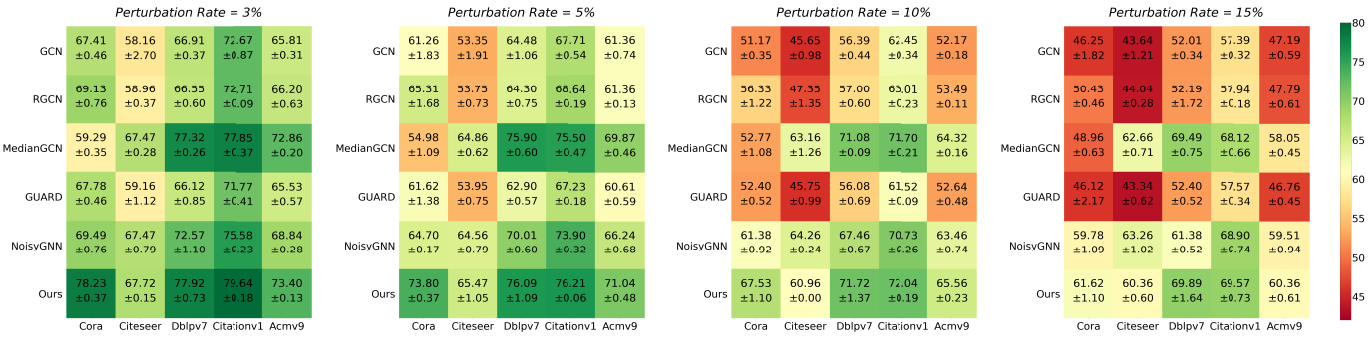

Fig. 6. Experimental evaluation results for defending against the LR-BCD on the Cora, Citeseer, Dblpv7, Citationv1, and ACMv9 datasets.

one of the more challenging OOD attacks, which employs the greedy matching strategy to modify graph structures and mislead classifier decisions. The intensity of this attack can be controlled through the constraint of perturbation ratio. However, as shown in Figure 5, our proposed scheme consistently outperforms the other baseline methods. Interestingly, the state-of-the-art NoisyGNN and MedianGCN are more affected by this attack, which experience a notable degradation in performance as the perturbation rate rises. These findings highlight the significant vulnerability of both NoisyGNN and MedainGCN to GR-BCD as their practical reliability suffers considerably when the attack strength increases.

4) Defense against PGA. The performance of four defense methods with different strategies is evaluated on DBLPv7 and ACMv9 datasets. We adjust the perturbation rate of the PGA from 0% to 20%, and generate the corresponding attack samples for each dataset. In Table V, the accuracy and macro-F1 score are adopted as evaluation metrics for the defense results. It is observed that MedianGCN shows a distinct advantage in defending against PGA by designing the robust aggregation function. Moreover, our proposed scheme further enhances the distinguishability of representations by exploiting multi-hop neighborhoods rather than stacked multilayers, which allows the model to consistently maintain optimal predicted

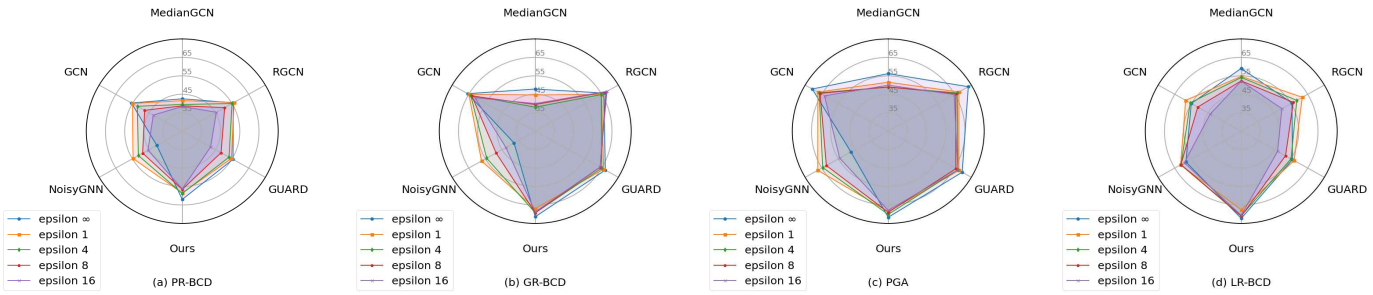

Fig. 7. Experimental evaluation results of defending against four deceptive OOD attacks with different strengths of privacy protection.

results under diverse perturbation rates. This suggests that the enhancement of the representation learning capability facilitates the proposed scheme to maintain predictive stability in the face of PGA.

5) Defense against LR-BCD. Under the condition of evasion attacks, we evaluate the performance of the selected baseline defense methods against the LR-BCD attack on five datasets: Cora, Citeseer, Dblpv7, Citationv1, and ACMv9. To implement the evasion attack, the defense models are trained on clean non-attacked graphs that have only undergone privacy-preserving manipulations. In this context, we conduct evasion attacks based on the gray-box setting (*i.e.*, the attacker knows partial information about the target object) while maximizing the knowledge available to the adversary. Besides manipulating the training data, the attacker can also acquire knowledge about the test data, which increases the difficulty of defense. The LR-BCD attack is employed to measure the classification accuracy of the defense model by generating modified edges with the perturbation rate of 3% to 15% during the testing phase. Figure 6 illustrates the performance across various perturbation rates. We observe that our scheme achieves desirable results compared to other defense methods on perturbed graphs. This demonstrates that our scheme can provide excellent results in defending against LR-BCD attacks by enhancing the inherent distinguishability of sensitive nodes and capturing the invariant characteristics across diverse environments.

D. Analyzing the Impact of Privacy Protection Strength on Defensive Capabilities (RQ3)

1) Defense under common privacy budgets. In this section, the strength of privacy protection is adjusted by changing the privacy budget to investigate the impact on the defensive capability of the model under deceptive OOD attacks. Specifically, as the privacy budget varies within the range of 1 to 16, we compare the defense capabilities of our proposed scheme with baseline defense methods against PR-BCD, GR-BCD, PGA, and LR-BCD attacks on the Cora dataset. Figure 7 reports the performance of these methods in resisting the aforementioned attacks under different strengths of privacy protection. From a global perspective, in the absence of privacy perturbation (*i.e.*, $\epsilon = \infty$), most methods exhibit superior defense effects compared to scenarios with privacy noise. It is surprising to find that the effectiveness of deceptive OOD attacks diminishes as privacy intensity increases, and this effect is most pronounced in the PR-BCD attack. We speculate that excessive perturbations introduced by privacy protection may interfere with the data affected by OOD attacks, which weakens the effectiveness of these attacks. However, the privacy intensity has less impact on PGA than other attacks. This may be due to the fact that the PGA focuses on nodes that are susceptible to small perturbations, which retain some interference effects even when the privacy noise is introduced. Locally, the NoisyGNN is the most affected under PGA, while providing notably weaker defense against the four OOD attacks when $\epsilon = \infty$ compared to the presence of privacy noise. We conjecture that NoisyGNN enhances the robustness to the perturbation by adding noises to the underlying structures, which reduces the interference of privacy protection on the model. These phenomena highlight the strong dependency of current defense methods on privacy intensity. However, the high privacy strength inherently causes serious interference with prediction accuracy. Therefore, the enhancement of the privacy-perceptual capability of defense methods is highly beneficial for improving their practicality and reliability.

In addition, our proposed scheme is observed to be minimally affected by changes in privacy intensity across various attacks. The reason is that our scheme weakens some interference from privacy operations through the private attribute-perceived embedding propagation mechanism, which facilitates the model to perceive private patterns. Furthermore, the multiple interactions enrich the diversity of representations and allow the scheme to maintain stable operations even under deceptive OOD attacks.

2) Impact of smaller privacy budgets. To further investigate the influence of smaller privacy budgets, we extend our experiments beyond the original setting of $\epsilon \in [1, 16]$ and evaluate EARI under more stringent differential privacy conditions (*i.e.*, $\epsilon = \{0.5, 0.3, 0.1, 0.05, 0.01\}$) on the Cora dataset. As shown in Figure 8, The results indicate that model performance gradually declines as ϵ decreases, which aligns with the theoretical expectation that stronger privacy guarantees introduce higher levels of noise. Notably, EARI is still able to maintain a usable level of classification accuracy around $\epsilon \approx 0.1$, whereas performance deteriorates significantly when $\epsilon < 0.1$ due to excessive noise overwhelming the graph features. This observation is consistent with the theoretical understanding that as $\epsilon \to 0$, the DP mechanism approaches complete randomization, rendering the outputs statistically indistinguishable but

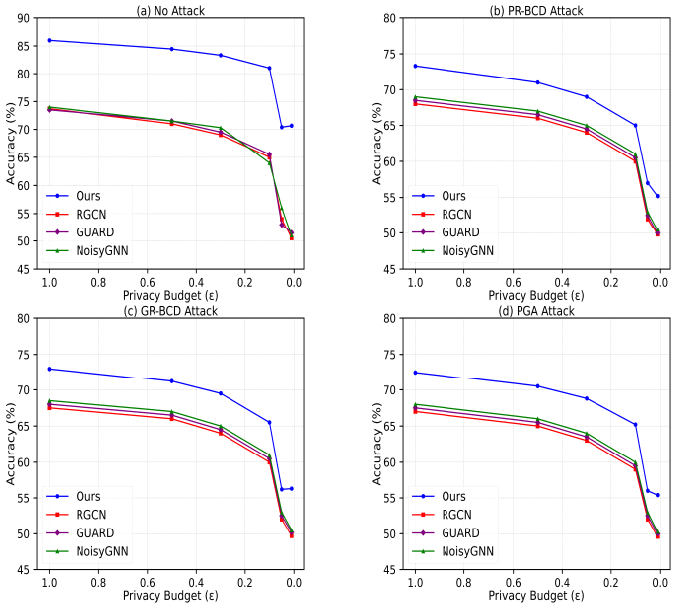

Fig. 8. Experimental evaluation results of model performance under small privacy budgets with and without deceptive OOD attacks.

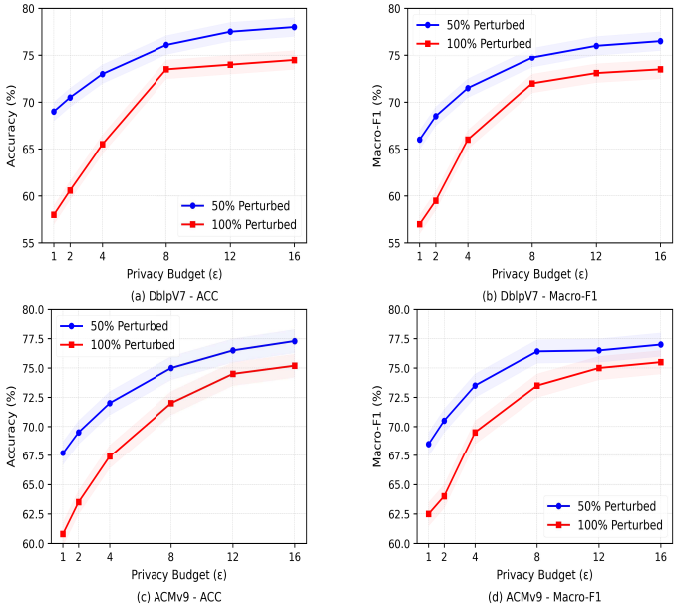

Fig. 9. Experimental evaluation results of model performance under different node privacy perturbation rates.

practically unusable for learning tasks. We further observe that the interaction between privacy perturbations and adversarial attacks depends on whether the attacked node overlaps with the privacy-protected node when OOD attacks (*e.g.*, PR-BCD with perturbation rate of 3%) are introduced at these smaller ϵ settings. If the attacked nodes are covered by DP noise, the attack signal is largely obscured, thus reducing its effectiveness. Conversely, the attack effect may combine with privacy perturbations, which leads to stronger distribution shifts. Despite this uncertainty, the EARI shows robustness in moderate privacy budget settings (*e.g.*, $\epsilon \in [0.5, 1]$). However, the dominant factor shifts to excessive noise under minimal budget constraints (*e.g.*, $\epsilon < 0.1$), severely compromising the result. These findings provide a clearer boundary for the trade-off between privacy and utility, suggesting that extremely small privacy budgets are of limited practical relevance, while EARI remains effective in realistic privacy-preserving scenarios.

3) Effectiveness under full node perturbation. To validate the defense effectiveness of the proposed scheme under full privacy perturbation, we conduct additional experiments on the Dblpv7 and ACMv9 datasets. Unlike the 50% node perturbation coverage utilized in previous settings, we adopt 100% node coverage, where all nodes are perturbed by the Laplace mechanism to ensure differential privacy. Under this condition, the privacy budget ϵ is varied from 1 to 16, and results are compared with those of 50% node perturbation coverage. Experiments are conducted under the 3% perturbation rate of the PGA. The result is shown in Figure 9. As the privacy budget increases (*i.e.*, privacy protection strength decreases), the model performance gradually improves, which is consistent with theoretical expectations. However, the overall performance under 100% coverage is lower than that under 50% perturbation, since excessive perturbation amplifies the impact of privacy noise and reduces the compensatory effect of non-private nodes on private ones. Notably, our scheme still achieves superior results compared to baseline methods (see Table V) at $\epsilon = 8$, demonstrating that EARI retains practical utility and defensive capability even in fully perturbed scenarios. These results indicate that proposed scheme remains robust under full privacy coverage, but optimal performance is typically observed with partial node perturbation (*e.g.*, 50%). This suggests that our scheme can better leverage its advantages in scenarios where only a subset of users choose to disclose their private information.

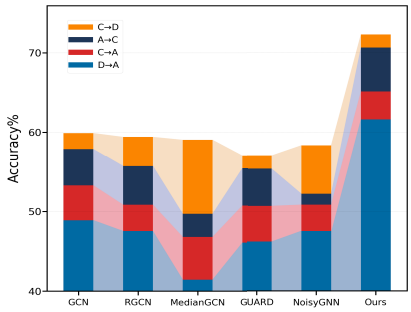

Fig. 10. Experimental evaluation results of performance stability on the Citationv1, Dblpv7 and ACMv9 datasets.

E. Evaluation of the Stability Under Transductive and Inductive Settings (RQ4)

The previous experiments were all under transductive settings, where the model might learn some clean samples during the training stage due to topological connections. However, under the inductive setting, we can ensure that the model does not have access to any information about the test nodes during the training phase by randomly training on one dataset and testing on others. To further validate the overall superiority of our scheme, we compare the proposed EARI with other baseline methods in the inductive setting. Three

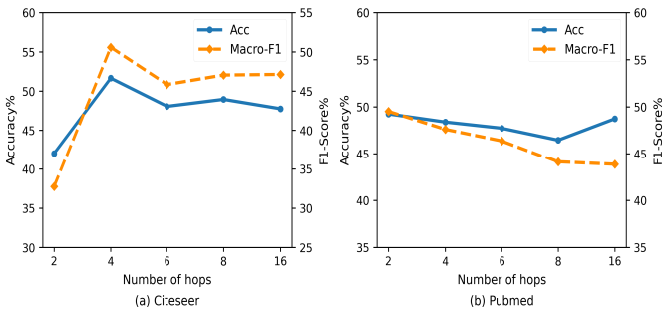

Fig. 11. Experimental evaluation results for hop number on the accuracy/macro-F1 of our scheme.

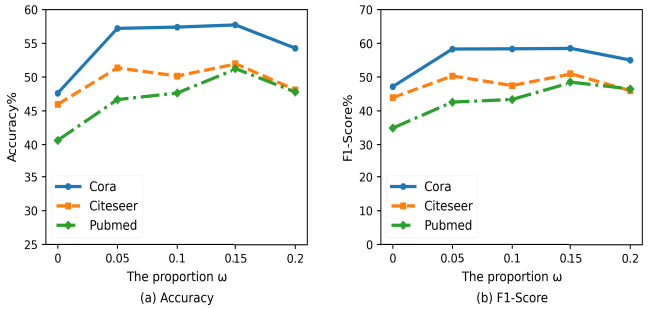

Fig. 12. Experimental evaluation results for environment-adaptive representations on the accuracy/macro-F1 of our scheme.

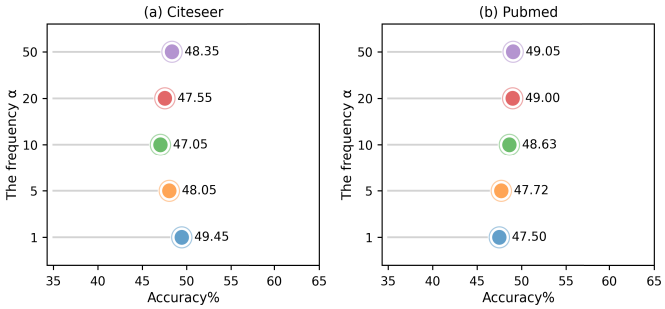

Fig. 13. Experimental evaluation results for representation replacement frequency of our scheme.

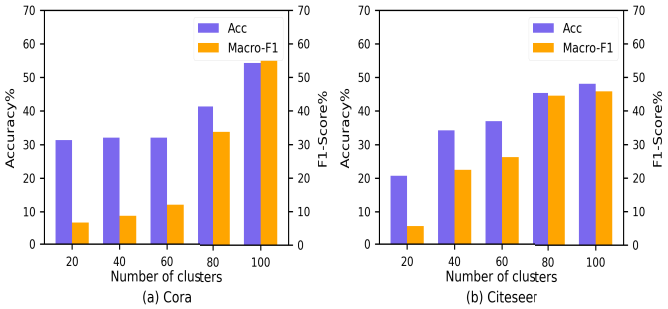

Fig. 14. Experimental evaluation results for cluster numbers on the accuracy/macro-F1 of our scheme.

similar yet distinct datasets are selected: citationv1, dblpv7, and ACMv9, which by nature have different distributions. This selection allows us to perform six tasks: C→D, C→A, D→C, D→A, A→C, and A→D, where C, D, and A represent citationv1, dblpv7, and ACMv9, respectively. Figure 10 shows the results of six defense methods on four tasks (*i.e.*, C→D, A→C, C→A, D→A) under the inductive setting. Among the six defense methods, our proposed scheme consistently demonstrates superior performance across multiple tasks. Specifically, the substantial improvement of 20.7% on the C→D task is achieved by our scheme compared to the best outcome of other defense methods. Surprisingly, the accuracy of EARI outperforms the highest value of other methods by 2.87% even in the D→A task that shows the worst performing of all methods. It is worth noting that while MedianGCN performed well in the other experiments, it performed the worst performance in most of the tasks in this experiment. This suggests that the emphasis of most defense schemes tends to be on preventing OOD attacks while ignoring the performance of the clean non-attacked privacy graphs. Overall, the analysis results above demonstrate the effectiveness of our proposed environment-adaptive representation interaction to ensure competitive performance, which further reveals the importance of enriching latent semantic representations and exploring intrinsic private patterns.

F. Hyper-Parameter Sensitivity Analysis (RQ5)

To explore the impact of different trade-off hyperparameters on the model performance, we further conduct the following hyperparameter experiments on the Cora, Citeseer, and Pubmed datasets under the PR-BCD attack.

1) Effect of Hop Number: In this experiment, the impact of varying the hop number U on the capability of our proposed scheme is investigated under deceptive OOD attacks. We alter the hop number within the range of $\{2, 4, 6, 8, 16\}$, and present the accuracy on different datasets. According to the results shown in Figure 11, the representation-enriched topological aggregation can effectively benefit from multiple hops, which suggests that the representation learning ability of EARI is correlated with the hop number.

2) Effect of Environment-Adaptive Representations: The proportion of environment-adaptive representation is adjusted to investigate the impact on model performance. Ideally, the model can enhance the defense against deceptive OOD attacks by learning various environmental knowledge. Figure 12 shows the significant effectiveness improvement in our proposed scheme as the generation ratio ω increases from 0 to 0.05. Beyond this point, the performance fluctuates with further increases in ω while remaining essentially stable. These observations indicate the effectiveness of the generalization-enhanced cluster-wise adaptation learning can successfully resist the deceptive OOD attacks.

3) Effect of Representation Replacement Frequency: The frequency of representation replacement α controls how frequently the original representation is replaced with environment-adaptive paradigms during the training phase. We vary the value of α unevenly from 1 to 50, and the results are recorded in Figure 13. It is worth noting that the overall performance remains stable, although the changes in α have a slight impact on accuracy. This suggests that our proposed scheme can identify invariant features even when exposed to a relatively small number of different environmental representations.

TABLE VI
INFLUENCE OF DIFFERENT COMPONENTS ON TWO DATASETS. THE BEST RESULTS ARE MARKED IN BOLD

Models	ACMv9		Dblpv7	
	ACC ▲	Macro-F1 ▲	ACC ▲	Macro-F1 ▲
W/O EP	68.02±0.75	66.81±1.30	70.62±0.83	69.02±0.54
W/O TA	66.37±1.24	65.12±1.52	68.25±1.41	66.87±1.65
W/O AL	67.81±0.77	66.43±1.01	69.74±1.10	68.23±0.68
EARI	**70.19±0.63**	**68.93±0.79**	**73.36±1.19**	**71.82±1.02**

4) Effect of Cluster Numbers: The selection of the cluster number M is important for the clustering tasks. We test multiple sets of cluster numbers ranging from 20 to 100, and observe the changes in accuracy and macro-F1 scores across different datasets. As illustrated in Figure 14, the predictive capacity of the model is influenced by the number of clusters. We speculate that the number of clusters directly affects the effectiveness of clustering, which is closely related to the performance of cluster-wise adaptation learning.

G. Ablation Study

To further validate the effectiveness of each component of our proposed EARI scheme, we conducted a specialized ablation study by removing individual modules and evaluated the resulting performance changes. Specifically, we applied the PR-BCD attack on two benchmark datasets (ACMv9 and Dblpv7) with perturbation rates ranging from 3% to 5% to assess robustness and task utility under adversarial conditions. We constructed the following three streamlined variants of our framework:

- **W/O EP**: Disabling the private attribute-perceived embedding propagation module.
- **W/O TA**: Removing the representation-enriched topological aggregation module.
- **W/O AL**: Suppressing the generalization-enhanced cluster-wise adaptation learning module.

The results are demonstrated in Table VI. The removal of any of the three components leads to the significant performance degradation under adversarial perturbation. Notably, the largest decrease in accuracy is caused by removing the TA module, which suggests that representation-rich topological aggregation plays a key role in capturing the underlying structural information and enhancing feature robustness. This experiment emphasizes the indispensability of each design component of EARI in ensuring the stability and effectiveness of the model under deceptive OOD attacks.

VI. CONCLUSION

In this paper, we propose a GNN framework based on environment-adaptive representation interaction (*i.e.*, capturing invariant topological correlations) to learn effectively from privacy-perturbed graph data and enhance the defense performance against deceptive OOD attacks. The following interesting conclusions can be drawn from our research findings: 1) the private attribute-perceived embedding propagation aims to explore private patterns in graph-structured samples, which enables the GNN models to mitigate noise perturbations; 2) the representation-enriched topological aggregation is designed from the perspective of maximizing the utilization of contextual interactions. This design is conducive to distinguishing crucial components from graph structures, which can be employed in other GNN applications to derive representations that are reliant on topological connectivity; 3) this work provides an empirical viewpoint that the interactive topological aggregation across different cluster environments can be utilized to further enrich the diversity and guarantee the effectiveness for graph representations under deceptive OOD attacks; and 4) extensive experiments demonstrate that our proposed scheme exhibits the significant advantage of overall performance (*e.g.*, reliable defensiveness and accurate stability) in deceptive graph-structured scenarios where the privacy operations and distribution shifts are manipulated by the adversarial attacks, and can reap correction benefits through the perception of biased distributions to mitigate the risk of adverse effects. In the near future, we plan to further investigate the topological interaction mechanism between non-private and private data across environments that can construct dependable and secure GNN models under intricate distribution conditions.

REFERENCES

[1] T. Zhou, N. Liu, B. Song, H. Lv, D. Guo, and L. Liu, "RobFL: Robust federated learning via feature center separation and malicious center detection," in *Proc. IEEE 40th Int. Conf. Data Eng. (ICDE)*, May 2024, pp. 926–938.
[2] J. Jia, P. Gao, M. Luo, C. Wu, and J. Guo, "CTEA: Camouflaged topological element attack via causal influence discovery," *Expert Syst. Appl.*, vol. 297, Feb. 2026, Art. no. 129160.
[3] Y. L. Liu et al., "Interpretable chirality-aware graph neural network for quantitative structure activity relationship modeling in drug discovery," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 12, pp. 14356–14364.
[4] Y. Liu, S. Rasouli, M. Wong, T. Feng, and T. Huang, "RT-GCN: Gaussian-based spatiotemporal graph convolutional network for robust traffic prediction," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102078.
[5] J. Jia, R. Li, C. Wu, S. Ma, L. Wang, and R. H. Deng, "SIGFinger: A subtle and interactive GNN fingerprinting scheme via spatial structure inference perturbation," *IEEE Trans. Dependable Secure Comput.*, vol. 22, no. 4, pp. 3629–3646, Jul. 2025.
[6] Q. Zhong et al., "Financial defaulter detection on online credit payment via multi-view attributed heterogeneous information network," in *Proc. ACM Web Conf.*, 2020, pp. 785–795.
[7] X. Pei, X. Deng, S. Tian, J. Liu, and K. Xue, "Privacy-enhanced graph neural network for decentralized local graphs," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 1614–1629, 2024.
[8] I. E. Olatunji, W. Nejdl, and M. Khosla, "Membership inference attack on graph neural networks," in *Proc. 3rd IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, Dec. 2021, pp. 11–20.
[9] Z. Chen, S. Yu, M. Fan, X. Liu, and R. H. Deng, "Privacy-enhancing and robust backdoor defense for federated learning on heterogeneous data," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 693–707, 2024.
[10] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr, "Membership inference attacks from first principles," in *Proc. SP*, 2022, pp. 1897–1914.
[11] Z. Zhang et al., "GraphMI: Extracting private graph data from graph neural networks," in *Proc. Thirtieth Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 3749–3755.
[12] D. Bo, X. Wang, C. Shi, and H. Shen, "Beyond low-frequency information in graph convolutional networks," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 5, pp. 3950–3957.
[13] S. Luan et al., "Revisiting heterophily for graph neural networks," in *Proc. NeurIPS*, 2022, pp. 1362–1375.
[14] E. Chien, J. Peng, P. Li, and O. Milenkovic, "Adaptive universal generalized PageRank graph neural network," in *Proc. ICLR*, 2020, pp. 1–12.

[15] X. Luo et al., "Toward effective semi-supervised node classification with hybrid curriculum pseudo-labeling," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 3, pp. 1–19, Mar. 2024.

[16] S. Geisler, T. S. Schmidt, H. Şirin, D. Zügner, A. Bojchevski, and S. Günnemann, "Robustness of graph neural networks at scale," in *Proc. NeurIPS*, vol. 34, 2021, pp. 7637–7649.

[17] G. Zhu, M. Chen, C. Yuan, and Y. Huang, "Simple and efficient partial graph adversarial attack: A new perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 8, pp. 4245–4259, Aug. 2024.

[18] L. Gosch, S. Geisler, D. Sturm, B. Charpentier, D. Zügner, and S. Günnemann, "Adversarial training for graph neural networks: Pitfalls, solutions, and new directions," in *Proc. NeurIPS*, 2023, pp. 58088–58112.

[19] Z. Chen, S. Wang, A. Fu, Y. Gao, S. Yu, and R. H. Deng, "LinkBreaker: Breaking the backdoor-trigger link in DNNs via neurons consistency check," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2000–2014, 2022.

[20] S. Ennadir, Y. Abbahaddou, J. F. Lutzeyer, M. Vazirgiannis, and H. Boström, "A simple and yet fairly effective defense for graph neural networks," in *Proc. AAAI*, vol. 38, 2024, pp. 21063–21071.

[21] J. Jia, S. Ma, Y. Liu, L. Wang, and R. H. Deng, "A causality-aligned structure rationalization scheme against adversarial biased perturbations for graph neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 59–73, 2024.

[22] J. Li et al., "GUARD: Graph universal adversarial defense," in *Proc. ACM CIKM*, 2023, pp. 1198–1207.

[23] K. Li et al., "Reliable representations make a stronger defender: Unsupervised structure refinement for robust GNN," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2022, pp. 925–935.

[24] I. E. Olatunji, T. Funke, and M. Khosla, "Releasing graph neural networks with differential privacy guarantees," *Trans. Mach. Learn. Res.*, pp. 1–15, 2021.

[25] T. T. Mueller, J. C. Paetzold, C. Prabhakar, D. Usynin, D. Rueckert, and G. Kaissis, "Differentially private graph neural networks for whole-graph classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7308–7318, Jun. 2023.

[26] J. Jia, J. Yu, D. Wu, C. Wu, H. Zhu, and L. Wang, "Prompt as a double-edged sword: A dynamic equilibrium gradient-assigned attack against graph prompt learning," in *Proc. 31st ACM SIGKDD Conf. Knowl. Discovery Data Mining V.2*, Aug. 2025, pp. 1049–1060.

[27] W. Fan et al., "Jointly attacking graph neural network and its explanations," in *Proc. IEEE 39th Int. Conf. Data Eng. (ICDE)*, Apr. 2023, pp. 654–667.

[28] N. Yin et al., "CoCo: A coupled contrastive framework for unsupervised domain adaptive graph classification," in *Proc. Int. Conf. Mach. Learn.*, vol. 202, 2023, pp. 40040–40053.

[29] J. Li et al., "Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking," in *Proc. NeurIPS*, 2023, pp. 3853–3866.

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

[31] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–12.

[32] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. NeurIPS*, vol. 30, 2017, pp. 1024–1034.

[33] F. Wu, A. H. Souza, T. Zhang, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6861–6871.

[34] Y. Yan, M. Hashemi, K. Swersky, Y. Yang, and D. Koutra, "Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2022, pp. 1287–1292.

[35] J. Chen, Z. Li, Y. Zhu, J. Zhang, and J. Pu, "From node interaction to hop interaction: New effective and scalable graph learning paradigm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7876–7885.

[36] Y. Liu et al., "A survey of deep graph clustering: Taxonomy, challenge, application, and open resource," 2022, *arXiv:2211.12875*.

[37] C. Wang, S. Pan, R. Hu, G. Long, J. Jiang, and C. Zhang, "Attributed graph clustering: A deep attentional embedding approach," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3670–3676.

[38] Y. Liu et al., "Hard sample aware network for contrastive deep graph clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 8914–8922.

[39] D. Bo, X. Wang, C. Shi, M. Zhu, E. Lu, and P. Cui, "Structural deep clustering network," in *Proc. Web Conf.*, Apr. 2020, pp. 1400–1410.

[40] W. Tu et al., "Deep fusion clustering network," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 11, pp. 9978–9987.

[41] Y. Liu et al., "Deep graph clustering via dual correlation reduction," in *Proc. AAAI*, 2022, vol. 36, no. 7, pp. 7603–7611.

[42] A. Tsitsulin, J. Palowitch, B. Perozzi, and E. Müller, "Graph clustering with graph neural networks," *JMLR*, vol. 24, no. 127, pp. 1–21, 2023.

[43] Y. Liu et al., "Dink-net: Neural clustering on large graphs," in *Proc. ICML*, 2023, pp. 21794–21812.

[44] X. Wan, K. Xu, X. Liao, Y. Jin, K. Chen, and X. Jin, "Scalable and efficient full-graph GNN training for large graphs," *ACM Manage. Data*, vol. 1, no. 2, pp. 1–23, Jun. 2023.

[45] H. Wu, C. Wang, Y. Tyshetskiy, A. Docherty, K. Lu, and L. Zhu, "Adversarial examples for graph data: Deep insights into attack and defense," in *Proc. Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4816–4823.

[46] W. Jin, Y. Ma, X. Liu, X. Tang, S. Wang, and J. Tang, "Graph structure learning for robust graph neural networks," in *Proc. ACM SIGKDD*, 2020, pp. 66–74.

[47] D. Zhu, Z. Zhang, P. Cui, and W. Zhu, "Robust graph convolutional networks against adversarial attacks," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 1399–1407.

[48] K. Kong et al., "Robust optimization as data augmentation for large-scale graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 60–69.

[49] B. Wang, B. Jiang, J. Tang, and B. Luo, "Generalizing aggregation functions in GNNs: Building high capacity and robust GNNs via nonlinear aggregation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13454–13466, Nov. 2023.

[50] H. Ling, Z. Jiang, Y. Luo, S. Ji, and N. Zou, "Learning fair graph representations via automated data augmentations," in *Proc. ICLR*, 2023, pp. 1–12.

[51] J. Li, T. Xie, L. Chen, F. Xie, X. He, and Z. Zheng, "Adversarial attack on large scale graph," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 82–95, Jan. 2023.

[52] E. Rossi, H. Kenlay, M. I. Gorinova, B. P. Chamberlain, X. Dong, and M. M. Bronstein, "On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features," in *Proc. Learn. graphs Conf.*, 2021, pp. 1–11.

[53] S. Sajadmanesh, A. S. Shamsabadi, A. Bellet, and D. Gática-Pérez, "GAP: Differentially private graph neural networks with aggregation perturbation," in *Proc. USENIX Secur. Symp.*, 2022, pp. 3223–3240.

[54] S. Sajadmanesh and D. Gática-Pérez, "ProGAP: Progressive graph neural networks with differential privacy guarantees," in *Proc. ACM WSDM*, 2024, pp. 596–605.

[55] S. Sajadmanesh and D. Gatica-Perez, "Locally private graph neural networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2021, pp. 2130–2145.

[56] X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan, and L. Duan, "Uncertainty modeling for out-of-distribution generalization," in *Proc. ICLR*, 2022, pp. 1–16.

[57] D. Xia, X. Wang, N. Liu, and C. Shi, "Learning invariant representations of graph neural networks via cluster generalization," in *Proc. NeurIPS*, 2024, pp. 45602–45613.

[58] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," *AI Mag.*, vol. 29, no. 3, pp. 93–106, Sep. 2008.

[59] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 990–998.

[60] L. Chen, J. Li, Q. Peng, Y. Liu, Z. Zheng, and C. Yang, "Understanding structural vulnerability in graph convolutional networks," in *Proc. Thirtieth Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 2249–2255.

[61] T. Cai, Y. Jiang, M. Li, C. Huang, Y. Wang, and Q. Huang, "ML-GOOD: Towards multi-label graph out-of-distribution detection," in *Proc. AAAI*, 2025, vol. 39, no. 15, pp. 15650–15658.

[62] S. Yang, B. Liang, A. Liu, L. Gui, X. Yao, and X. Zhang, "Bounded and uniform energy-based out-of-distribution detection for graphs," in *Proc. ICML*, 2024, pp. 56216–56234.