Construct New Graphs using Information Bottleneck Against Property Inference Attacks

Chenhan Zhang*, Zhiyi Tian*, James J.Q. Yu†, Shui Yu*
*School of Computer Science, University of Technology Sydney, Sydney, Australia.
†Dept. of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China.
Email: {chenhan.zhang, zhiyi.tian}@student.uts.edu.au, yujq3@sustech.edu.cn, shui.yu@uts.edu.au

***Abstract*—Graphs provide a unique representation of real-world data. However, recent studies found that inference attacks can extract private property information of graph data from trained graph neural networks (GNNs), which arouses privacy concerns about graph data, especially in collaborative learning systems where model information is more accessible. While there has been a few research efforts on the property inference attacks against GNNs, how to defend against such attacks has seldom been studied. In this paper, we propose to leverage the information bottleneck (IB) principle to defend against the property inference attacks. Particularly, we involve a threat model, where the attacker can extract graph property from the graph embedding developed by GNNs. To defend against the attacks, we use IB to construct new graph structures from the original graphs. The change in graph structures enables the new graphs to contain less information related to the property information of the original graphs, making it harder for attackers to infer property information of the original graphs from the graph embeddings. Meantime, the IB principle enables task-relevant information to be sufficiently contained in the new graph, enabling GNNs to develop accurate predictions. The experimental results demonstrate the efficacy of the proposed approach in resisting property inference attacks and developing accurate predictions.**

***Index Terms*—Graph-structured data, graph neural networks, inference attacks, information bottleneck.**

I. Introduction

Graphs are ubiquitous in the real world, representing entities and their relationships, such as social networks, e-commerce networks, and traffic networks. However, the privacy of graph data has been a big concern due to the risk of leakage, especially in collaborative machine learning systems such as federated learning and semantic communication [1], [2]. On the one hand, the communication transmission can be eavesdropped and the data is thus theft. On the other hand, the shareable data can be maliciously used. All of these scenarios can cause leakage of private information. Leakage of graph data will result in serious consequences. For example, exposure of a social network of COVID-19 patients naturally goes against the patients as such disease histories belong to their private information [3].

Moreover, recent studies indicated that graph neural networks (GNNs) are vulnerable to *inference attacks* [3]–[5]. Training data samples leave *footprints* on the GNNs, which are recorded by the model gradients or learned embeddings. The attacker can easily trace relevant information of the training graph data using these footprints. It is often assumed that attackers would like to steal graph structures and nodal attributes as they are the fundamental components of a graph [5]. However, some statistical properties of the graph data, such as the number of nodes and the graph density, can also be private. The data curators may not intend to share these properties since they may reveal sensitive information such as business transaction frequency. Also, these properties imply intellectual property since collecting them is laborious. Therefore, the privacy of graph properties is an integral part to the privacy of graph data, which is worthy of in-depth study.

Stealing graph property information from graph embeddings is a realistic assumption — local graph embeddings can be shared to other parties for broad use, which gives access to *man-in-the-middle* [6] attackers. The paradigm of property inference attacks based on graph embeddings can be referred to [4]. The attack model mainly focuses on extracting information from the graph embeddings queried from the GNNs. Inference attacks on graph property are easy to carry out and have a high success rate compared to inference attacks on other targets. Thus, studying the defenses to property inference attacks against GNNs is essential, which has not been given much attention yet [2], [7], [8].

Differential privacy (DP) has been recognized as an effective measure of countering membership inference attacks as this type of attack focuses on the privacy of *individual* records [9]. DP adds controlled noise to target models' gradients or outputs, which can effectively impede the inference. However, graph properties, e.g., graph density and number of edges, are *global*. Previous studies have shown that DP-based defenses are not similarly effective to such property inference attacks [10]. Furthermore, the nature of adding noises makes DP cause inevitable loss on data utility [5], [11]. This situation motivates us to find an effective way to defend against such attacks targeting global properties.

A possible solution is using *compressive privacy*, which compresses the data to juice out the private parts [12]. Information bottleneck (IB) [13] is a key technique of CP, which compresses the data by squeezing out task-irrelevant information while retaining task-relevant information, and it provides a tradeoff between the two parts. This technique drives us to wonder: *How about using the IB principle to squeeze out relevant information about the graph properties but include sufficient predictive information to achieve the privacy-utility tradeoff of graph data?*

In this paper, we propose to leverage the IB principle to defend against the property inference attacks on graph embeddings. Specifically, we leverage IB to construct new graphs, which are predictive yet distorted from the original graph structures. The graph embeddings developed from the new graphs have less information corresponding to the original graph structures, making property inference attackers hard to extract the accurate graph property from them.

The highlights of this paper are summarized below:

- We propose an information bottleneck-based approach to defend against the property inference attacks on graph embeddings. So far as we know, we are among the early works dedicated to countering property inference attacks against GNNs.
- We demonstrate that the information related to the original graph property in the new graph structure is lessened through the information bottleneck. Such a change makes it more difficult for attackers to accurately infer the accurate information about the original graph.
- We conduct comprehensive case studies on three real-world graph-structured datasets. The results show that the proposed approach performs better than conventional approaches regarding the tradeoff between data privacy and utility.

The remainder of the paper is organized as follows. Section II presents the definition of graph learning and information bottleneck. We introduce the threat model in Section III. In Section IV, we elaborate on the proposed approach. Experimental results are shown in Section V. Section VI review the related literature. Section VII draws the conclusion.

II. Preliminaries

A. Graph Neural Networks for Graph Classification

Let $G \in \mathbb{G} = (X, A)$ be a graph with node set $V = \{v_i | i = 1, \ldots, |V|\}$ and edge set $\{E = (v_i, v_j) | i > j; v_i$ and v_j is connected$\}$, where $X \in \mathbb{R}^{|V| \times d}$ is the node feature matrix with d dimension and $A \in \mathbb{R}^{|V| \times |V|}$ is the adjacency matrix. $Y \in \mathbb{Y}$ denotes the label of the graph. Let $\{(G_1, Y_1), \ldots (G_1, Y_1)\}$ be a set of N pairs of graphs. A classifier is to be learned that can correctly map graph data to the corresponding label: $\mathcal{F} : G_i \rightarrow Y_i$.

GNNs learn a representation for each node in the graph by aggregating the non-linear-transformed vectors of neighbor nodes. Let $\mathcal{N}(v)$ be the set of 1-hop neighbor nodes of node v, the canonical aggregation of GNNs is described as:

$$\begin{aligned} \mathbf{h}^l_{\mathcal{N}(v)} &= \text{AGGREGATE}_l \left(\left\{ \mathbf{h}^{l-1}_u, \forall u \in \mathcal{N}(v) \right\} \right), \\ \mathbf{h}^l_v &= \text{UPDATE}_l \left(\mathbf{h}^l_{\mathcal{N}(v)} \right), \end{aligned} \tag{1}$$

where $\mathbf{h}^l_u$ denotes the embedding of node u at layer l. AGGREGATE_l denotes the aggregator function. UPDATE_l represents a non-linear function, e.g., a multilayer perceptron (MLP). Then, GNNs aggregate the embeddings of all nodes in the graph to obtain a whole graph embedding:

$$H_G = \text{POOLING}(\mathbf{h}_v, \forall v \in V). \tag{2}$$

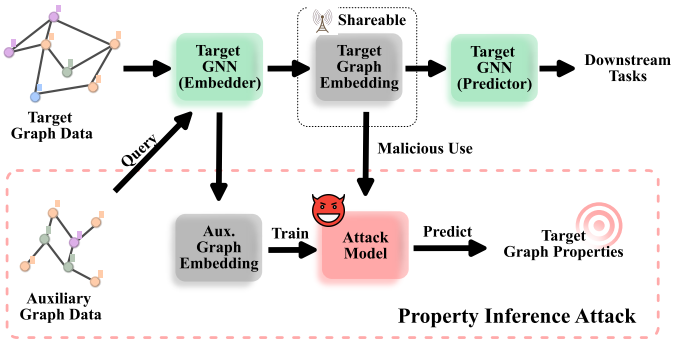

Fig. 1. Threat model: property inference attacks on graph embeddings.

Max pooling and *mean pooling* are two common graph pooling operation. We consider GNNs using mean pooling in this work. Finally, a multi-class classifier is used to predict the label of the graph given the graph embedding H_G.

B. Information Bottleneck

Given the input data X and the label Y, the information bottleneck (IB) Principle seeks to find the latent representation Z, which contains little information of X yet is maximally informative concerning Y. Let $I(X, Z)$ denote the Shannon mutual information (MI) between the input $\boldsymbol{X}$ and the encoded representation Z, and $I(Y, Z)$ denote the Shannon MI between Z and the class label Y. IB principle [13] aims to learn the *minimal suffficient* representation Z:

$$\arg\min_Z -I(Y, Z) + \beta I(X, Z), \tag{3}$$

where β is a Lagrange multiplier to control the compression level of Z.

III. Threat Model

In this paper, we propose a defense approach to property inference attacks against GNNs. In terms of our adversary, the property inference attacker, we generally follow the assumption in [4].

Attacker's Knowledge: We study a grey-box setting. The attacker can obtain the graph embedding by querying the target GNN model with an input graph. All other knowledge, such as training graphs, architectures and parameters of the target GNN model, is not accessible to the attacker.

Attacker's Goal and Capabilities: The attacker aims to extract the property information of the target graph from the graph embedding. Graph embedding-based inference attacks are very realistic since local graph embeddings have been shared with other parties for further graph analysis or learning tasks [14], where data leakage or theft can happen.

We focus on *graph density* inference in this work. To this end, the attacker adopts an attack model $\mathcal{A}$ to input the graph embedding. $\mathcal{A}$ is a MLP that predicts the graph density [4]. The procedure of the attack can be referred to in Figure 1. There also exists an auxiliary dataset $\mathcal{D}_{\text{aux}}$ that the attacker can access to train the attack model. The graphs in the auxiliary dataset are assumed to be from the same distribution as the target graph. However, different from [4] which formed the

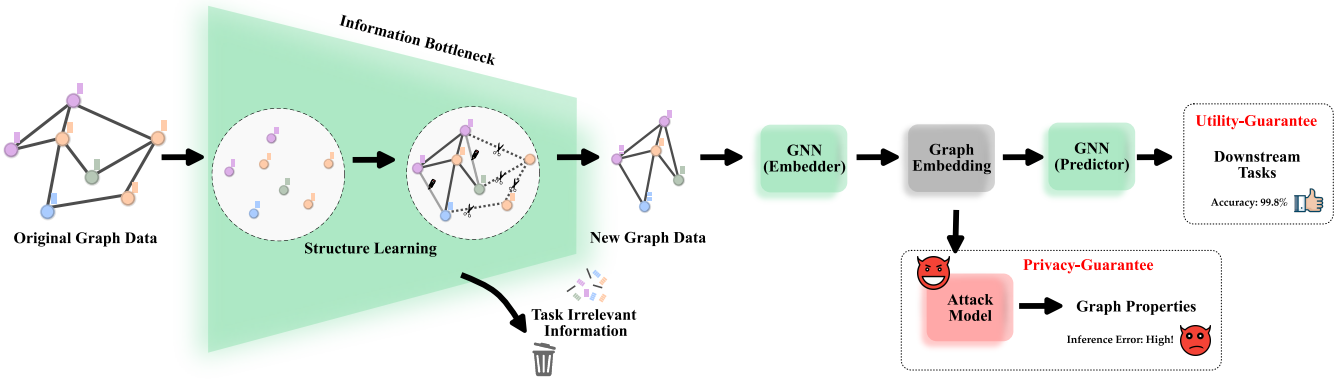

Fig. 2. Schematic of the proposed defense.

property inference into a classification problem, we treat the property inference as a regression problem to evaluate more fine-grained inference results.

Denoting the graph embedding as an intermediate state output by the target GNN $\mathcal{F}$ as $\mathcal{F}_g$, the attacker is optimizing the following attack objective function, that is

$$\arg\min_{\mathcal{A}} \mathop{\mathbb{E}}_{G_{\text{aux}} \in \mathcal{D}_{\text{aux}}} \left[\sum \mathcal{L} \left(\mathcal{A} \left(H_{G_{\text{aux}}} \right), P_{G_{\text{aux}}} \right) \right] \tag{4}$$

where P is the true value of graph density.

IV. Our Information Bottleneck-based Defense

We propose a new defense approach that uses the information bottleneck for graph data to reconstruct the graph structures to mitigate property inference attacks. Figure 2 illustrates the pipeline of our defense approach. We first leverage the IB principle to learn new graphs given the original graphs. Then, we treat the new graphs as the input to train the GNNs. The graph embeddings developed by the GNNs trained on such new graphs will have less information about the original properties of the graphs, which is the core to resisting property inference attacks. We will elaborate on the technical details and analyze the privacy and utility guarantee of the proposed approach in the sequel.

A. Graph Representation Learning based on Information Bottleneck

Graph information bottleneck extends the IB principle to representation learning on graph data [15]–[17]. Given a graph $G \in \mathbb{G} = (X, A)$ and its label Y, the IB-optimized graph is formulated as

$$\arg\min_{G_{\text{IB}}} -I\left(Y, G_{\text{IB}}\right) + \beta I\left(G, G_{\text{IB}}\right), \tag{5}$$

where G_{IB} is constituted of the task-relevant feature matrix X_{IB} and task-relevant adjacency matrix A_{IB}.

Resorting to variational IB [18], we can derive a variational bound which is tractable for Eq. (5), that is

$$\begin{aligned} &I\left(Y, G_{\text{IB}}\right) - \beta I\left(G, G_{\text{IB}}\right) \geq \\ &\quad \frac{1}{N} \sum_{i=1}^{N} \int p\left(G_{\text{IB}} \mid G_i\right) \log q_\phi\left(Y_i \mid G_{\text{IB}}\right) dG_{\text{IB}} \\ &\qquad - \beta \text{KL}\left(p\left(G_{\text{IB},i} \mid G_i\right) \mid r(G_{\text{IB}})\right), \end{aligned} \tag{6}$$

where $\text{KL}(\cdot)$ denotes the Kullback Leibler (KL) divergence, $r(G_{\text{IB}})$ is the variational approximation of $p(G_{\text{IB}})$, and $q_\phi\left(Y \mid G_{\text{IB}}\right)$ and $q_\phi(G_{\text{IB}} \mid G)$ are reparameterized variational approximations to $p\left(Y \mid G_{\text{IB}}\right)$ and $p\left(G_{\text{IB}} \mid G\right)$, respectively.

However, it is hard to estimate $I\left(G, G_{\text{IB}}\right)$ as the irregularity of graph data. Neural network-based mutual information estimation has been demonstrated to be an available solution to this problem. Denoting the neural network-estimated graph representation as Z_{IB}, we have $I(G, G_{\text{IB}}) \geq I(G, Z_{\text{IB}})$ [19], which is in favour of the IB optimization. Particularly, we use the mutual information estimation adopted in [17] to optimize the IB. Consequently, we obtain minimally sufficient Z_{IB} by optimizing this objective, which is less prone to overfitting and thus delivers better performance on downstream tasks. Interested readers can refer to [17] for the technical details.

B. Construct New Graph Structure

We then move forward to how to construct G_{IB} from the results of IB optimization. We leverage the notion of graph auto-encoder (GAE) [20] to construct the new graph structure. We first use MLP to get a latent representation of each node feature by:

$$Z(v) = \text{MLP}(X_v). \tag{7}$$

For any two nodes v and u, we have assignment probability ψ to determine whether the edge (v, u) should be included or not. We consider that two nodes with closing representation are more likely to have an edge. Therefore, we calculate ψ by applying the logistic sigmoid function to the inner product of $Z(v)$ and store the values in the adjacency matrix form, which is formulated as:

$$A_{\text{assign}} = \psi(v, u) = \text{sigmoid}(Z(v)Z(u)^T) \tag{8}$$

Subsequently, we follow [15] by employing Gumbel-softmax to make A_{assign} differentiable from Bernoulli distribution. Finally, we can determine the binary adjacency matrix $A_{\text{IB}} = \{a_{u,v}\}$ by conducting a Bernoulli sampling from A_{assign}. We construct G_{IB} according to A_{IB}. We first identify the largest connected component (LCC) in A_{IB} as the new graph structure. For the node feature matrix X_{IB}, we keep the node feature of the nodes included in the new graph structure and discard others.

So far, new graph structure G_{IB} can be either used for neural mutual estimation to further optimize IB (as described in Section IV-A), or developing the corresponding graph embedding $H_{G_{\mathrm{IB}}}$ for downstream tasks by forwarding G_{IB} to GNN embedders. We consider graph embeddings $H_{G_{\mathrm{IB}}}$ are more privacy-guaranteed than the original ones when exposed to the threat model. We will justify our hypothesis in Section IV-C and our experiments.

C. Privacy and Utility Guarantee by the Information Bottleneck

For the target graph property P_G, the information transmission among P_G, G, and G_{IB} can be described by the Markov chain:

$$P_G \longrightarrow G \longrightarrow G_{\mathrm{IB}} \longrightarrow \hat{P_G}, \tag{9}$$

where $\hat{P_G} = \mathcal{A}(H_{G_{\mathrm{IB}}})$ denotes the graph property predicted by the attacker using $H_{G_{\mathrm{IB}}}$. We can ensure that property inference attackers cannot derive more information from G_{IB} than G from Eq. (11) since 1) G subsumes G_{IB} and 2) G_{IB} is optimized to maximumly squeeze the mutual information with G. According to [19], we know that this assurance also holds for the graph embedding H_G and $H_{G_{\mathrm{IB}}}$ since the amount of information loss from G_{IB} to $H_{G_{\mathrm{IB}}}$ is close to the one from G to H_G when using the same embedder. Eq. (11) also implies that information transmission is diminishing, and we can obtain:

$$I\left(G, P_G\right) \geq I\left(G_{\mathrm{IB}}, P_G\right). \tag{10}$$

Therefore, it is easy to derive that the upper bound of property inference attacks using $H_{G_{\mathrm{IB}}}$ equivalent to the attacks using G. On this basis, we can conclude that, $H_{G_{\mathrm{IB}}}$, which is from G_{IB}, will be much less informative in terms of the property inference.

For the label Y, the information transmission among Y, G, and G_{IB} in Eq. 5 can be described by the Markov chain:

$$Y \longrightarrow G \longrightarrow G_{\mathrm{IB}} \longrightarrow \hat{Y}. \tag{11}$$

We assume that G_{IRR} is the component of G, which is irrelevant to the target Y. We can derive an upper bound for mutual information between G_{IB} and G_{IRR} from [16], that is

$$I\left(G_{\mathrm{IRR}}, G_{\mathrm{IB}}\right) \leq I\left(G, G_{\mathrm{IB}}\right) - I\left(Y, G_{\mathrm{IB}}\right). \tag{12}$$

Consequently, optimizing Eq. (5) amounts to minimizing $I\left(G_{\mathrm{IRR}}, G_{\mathrm{IB}}\right)$, making the optimized G_{IB} with less irrelevant information to target Y.

V. Experimental Evaluation

A. Experiments Preparation

Dataset: In our experiments, we employ three real-world graph-structured social network datasets in terms of graph classification tasks, namely, **IMDB-B**, **IMDB-M**, and **COLLAB** [21]. The statistical information of the three datasets is summarized in Table I. Following [4], the ratios of the training set (for the training of target GNN), auxiliary set (for the training of attack model), testing set (for testing of both target GNN and attack model), and testing sets are 40%, 40%, and 20%, respectively. Additionally, we apply data augmentation to the auxiliary set by adding random edges to ensure sufficient training samples for the attack model.

TABLE I
Statistical summary of graph classification datasets.

Dataset	# Graphs	Avg. Nodes	Avg. Edges	# Classes
IMDB-B	1000	19.77	96.53	2
IMDB-M	1500	13.00	65.94	3
COLLAB	5000	74.49	2457.21	3

Model and Hyperparameters: We incorporate two GNN models in our case studies: Graph Convolution Network (**GCN**) [22] and Graph Attention Network (**GAT**) [23]. We empirically set them to 2-layer and with embedding size 16. Unless other stated, we adopt the following settings. To train the attack model, we set the epoch number to 50 and the mini-batch size to 20. To train target GNN models, we set the epoch number to 200 and the mini-batch size to 100. For the proposed approach, the Lagrange multiplier β is an important hyperparameter to control the distortion level of the new graph structure. We set $\beta = 1 \times 10^{-5}$ as default and conduct a related hyperparameter test later. The learning rates for training GNNs and the attack model are all set to 1×10^{-3}.

Baseline: Since we are among the pioneering work focusing on the defenses against property inference attacks, there are no baselines dedicated to this problem. We denote the proposed approach as **IB**. We first consider the **Original** case, i.e., the one without any defense mechanism. Furthermore, we identify that **differential privacy (DP)** is the most widely-adopted defense strategy against inference attacks in the existing works [4], [5], [14]. Therefore, we mainly compare our proposed approach with DP. Specifically, we apply DP-based noises to the target graph embedding to defend the property inference attacks by $\tilde{H_G} = H_G + \mathrm{Lap}(0, \frac{s}{\epsilon})$, where the noises are from the Laplacian distribution $\mathrm{Lap}(0, \frac{s}{\epsilon})$ with mean 0 and scale $\frac{s}{\epsilon}$. ϵ and S denote the *privacy budget* and *sensitivity*, respectively. We set $s = 1$ with different $\epsilon = \{1, 5, 10\}$ to evaluate the performance with difference scales. In general, smaller values of ϵ provide more privacy preservation and vice versa.

Metrics: To evaluate the resistance to property inference attacks, root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are used as the metrics to evaluate the accuracy of inference. The lower accuracy indicates a better resistance of the approach; otherwise, worse. In particular, according to common practice, MAPE is considered preferable. To measure the data's utility, we consider the graph classification accuracy of the GNN model. Higher classification accuracy indicates a better utility of the learned graph embeddings.

B. Results

1) Resistance to Property Inference Attacks: We first compare the proposed approach and baselines' resistance to property inference attacks. The results are shown in Table

TABLE II
COMPARISON OF PROPERTY INFERENCE ACCURACY.

Inference Accuracy		COLLAB			IMDB-B			IMDB-M		
		RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
GCN	Original	10.26	9.07	13.91	18.92	15.87	13.17	112.91	110.46	29.24
	DP ($\epsilon = 1$)	16.56	15.93	**27.46**	**35.22**	**32.99**	**29.20**	116.92	111.57	30.39
	DP ($\epsilon = 5$)	15.92	15.33	26.24	33.13	29.79	26.36	114.81	110.97	30.27
	DP ($\epsilon = 10$)	15.84	15.69	25.68	32.43	28.87	25.23	114.22	110.96	30.18
	IB (Ours)	**19.19**	**18.91**	26.24	26.77	26.54	25.99	**122.26**	**121.22**	**35.28**
GAT	Original	11.19	8.81	12.54	17.54	14.42	11.91	112.20	110.43	29.97
	DP ($\epsilon = 1$)	17.24	15.81	23.07	**30.28**	28.77	25.61	18.33	16.73	23.96
	DP ($\epsilon = 5$)	17.46	15.73	23.16	26.76	24.10	20.52	18.60	17.73	25.87
	DP ($\epsilon = 10$)	16.37	14.28	20.63	24.63	21.87	18.39	18.07	16.58	25.07
	IB (Ours)	**18.52**	**18.43**	**27.58**	30.24	**29.94**	**27.99**	**122.90**	**121.24**	**35.21**

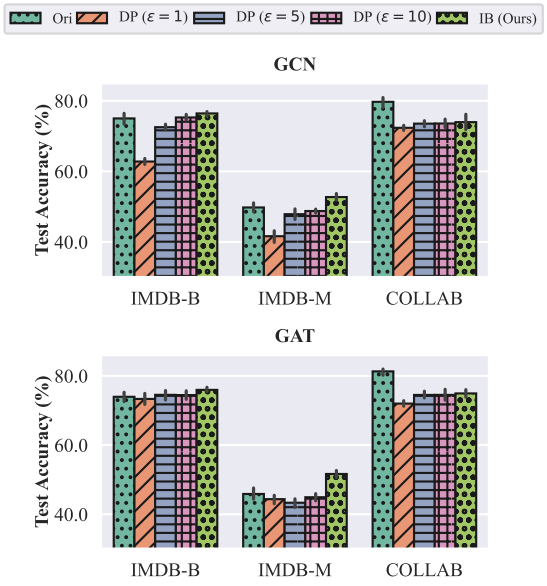

Fig. 3. Comparison of graph classification accuracy.

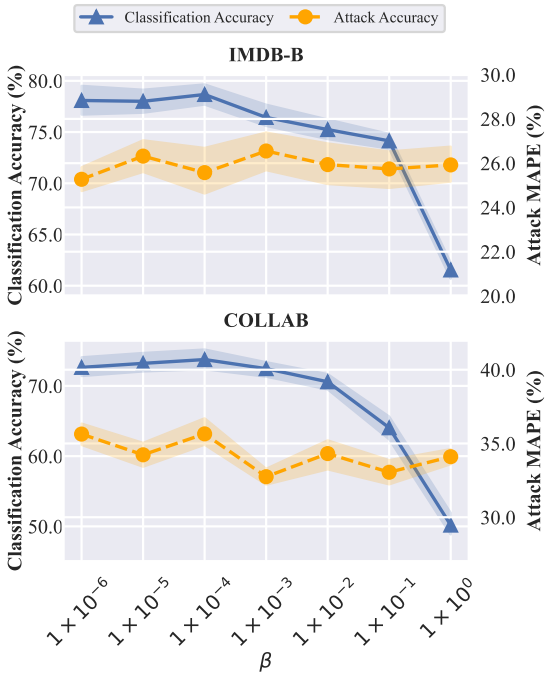

Fig. 4. Sensitivity of β to graph classification accuracy and property inference attack accuracy on IMDB-B and COLLAB datasets with GCN model.

II where the best ones are highlighted in **bold**. Obviously, the proposed approach is effective in resistance to property inference attacks against GNNs. Compared with the original setting where no defense strategy is adopted, the inference accuracy (MAPE) is dropped from 5.24% to 16.08% in different configurations. Additionally, we recognize that the effectiveness of the proposed approach varies under different GNN models and datasets. This is due to that the attack model performs differently on the graph embeddings from these different settings. The proposed approach outperforms DP-based defenses in most cases, even compared with DP with small values of ϵ. Particularly, for those DP-based defenses with good resistance performance, the data utility suffers an obvious loss meantimes, which will be discussed later.

2) Prediction Accuracy on Downstream Tasks: The prediction accuracy developed by the GNN is a significant metric of the data (graph embeddings) utility guarantee provided by the defense approaches. We compare the graph classification accuracy between different baselines as shown in Figure 3. We find that the proposed approach obtains satisfactory classification accuracy — reaches and even outperforms the results of the original setting. It means that the new structure is informative of prediction. Compared with DP-based approaches, especially the ones with small ϵ, the proposed approach shows superiority. As aforementioned, DP-based defenses are successful in protecting privacy in smaller ϵ settings, but at the expense of data utility.

In a nutshell, we can conclude that the IB principle enables the proposed approach to achieve the tradeoff between privacy and data utility to a great extent.

3) Hypereparameter Study on β: Tradeoff between Utility and Privacy: Moreover, we investigate the sentivity of the proposed approach to the Lagrange multiplier β as it is a pivotal hyperparameter for the proposed approach in controlling the distortion of learned graph structures, which is shown in Figure 4. In terms of classification accuracy, we observe that large β usually develops inferior performance. This is due to large β

can lead to over-distorted new graph structures, rendering the loss of predictive information. In terms of the resistance to attacks, we find different result patterns on the two datasets. For IMDB-B, there is a slight uptrend of the attack accuracy with larger β; however, it shows a decline on COLLAB. There may exist some non-trivial influence on the attack model's performance due to the difference in graph size, topology, etc. We will conduct further investigation into this phenomenon in the future.

VI. Related Work

Recent years have witnessed increasing research attention on inference attacks against GNNs [7], including membership inference attacks [24], reconstruction attacks [5], property inference attacks [4], and model extraction attacks [25]. For example, Olatunji *et al.* [3] focused on membership inference attacks and proposed to identify if a node is in training set for GNNs. Zhang *et al.* [4] concentrated on reconstruction attacks and proposed to reconstruct a graph structure from the leaked gradient information. Shen *et al.* [26] systematically categorized the different levels of model extraction attacks according to the attackers' capabilities. Zhang *et al.* [4] recognized the significance of the privacy of graph properties and proposed a threat model of property inference attacks against GNNs. However, the defenses (countermeasures) against these attacks (threats) are not well-studied. While some conventional strategies, such as differential privacy, were involved in this literature, the results indicated the limitation of these conventional approaches in most cases.

VII. Summary and Future Work

In this paper, we propose a novel defense approach to property inference attacks against GNNs. The proposed approach leverages information bottleneck principle to develop new graphs from the original graphs. The information of the original graph are lessened in the new graphs, which makes the inference attackers harder extract accurate property information about the original graph. However, the task-relevant information in the new graph is maximally contained its utility in downstream tasks. The case studies indicate that the proposed approach achieves the tradeoff between privacy preservation and prediction utility of graph data. In the future, we will apply the proposed approach the scenarios of more graph tasks and GNNs to further evaluate the generalizability.

References

[1] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[2] H. Zhang, B. Wu, X. Yuan, S. Pan, H. Tong, and J. Pei, "Trustworthy graph neural networks: Aspects, methods and trends," *arXiv preprint arXiv:2205.07424*, 2022.

[3] I. E. Olatunji, W. Nejdl, and M. Khosla, "Membership inference attack on graph neural networks," in *Proc. International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pp. 11–20, IEEE, 2021.

[4] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang, "Inference attacks against graph neural networks," in *Proc. USENIX Security Symposium*, pp. 1–18, 2022.

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

[6] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE communications surveys & tutorials*, vol. 18, no. 3, pp. 2027–2051, 2016.

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

[8] B. Wu, J. Li, J. Yu, Y. Bian, H. Zhang, C. Chen, C. Hou, G. Fu, L. Chen, T. Xu, *et al.*, "A survey of trustworthy graph learning: Reliability, explainability, and privacy protection," *arXiv preprint arXiv:2205.10014*, 2022.

[9] H. Jiang, J. Pei, D. Yu, J. Yu, B. Gong, and X. Cheng, "Applications of differential privacy in social network analysis: a survey," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[10] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *International Journal of Security and Networks*, vol. 10, no. 3, pp. 137–150, 2015.

[11] Z. Zhang, Q. Liu, Z. Huang, H. Wang, C. Lu, C. Liu, and E. Chen, "Graphmi: Extracting private graph data from graph neural networks," in *Proc. International Joint Conference on Artificial Intelligence* (Z. Zhou, ed.), pp. 3749–3755, 2021.

[12] S.-Y. Kung, "Compressive privacy: From information\/estimation theory to machine learning [lecture notes]," *IEEE Signal Processing Magazine*, vol. 34, no. 1, pp. 94–112, 2017.

[13] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

[14] S. Zhang, H. Yin, T. Chen, Z. Huang, L. Cui, and X. Zhang, "Graph embedding for recommendation against attribute inference attacks," in *Proc. Web Conference 2021*, pp. 3002–3014, 2021.

[15] T. Wu, H. Ren, P. Li, and J. Leskovec, "Graph information bottleneck," *Proc. Advances in Neural Information Processing Systems*, vol. 33, pp. 20437–20448, 2020.

[16] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He, "Recognizing predictive substructures with subgraph information bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[17] Q. Sun, J. Li, H. Peng, J. Wu, X. Fu, C. Ji, and S. Y. Philip, "Graph structure learning with variational information bottleneck," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 36, pp. 4165–4174, 2022.

[18] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. International Conference on Learning Representations*, 2017.

[19] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proc. International conference on machine learning*, pp. 531–540, 2018.

[20] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv:1611.07308*, 2016.

[21] R. Rossi and N. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. AAAI conference on artificial intelligence*, 2015.

[22] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. International Conference on Learning Representations*, 2017.

[23] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. International Conference on Learning Representations*, 2018.

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

[25] B. Wu, X. Yang, S. Pan, and X. Yuan, "Model extraction attacks on graph neural networks: Taxonomy and realisation," in *Proc. ACM on Asia Conference on Computer and Communications Security*, pp. 337–350, 2022.

[26] Y. Shen, X. He, Y. Han, and Y. Zhang, "Model stealing attacks against inductive graph neural networks," in *Proc. IEEE Symposium on Security and Privacy (SP)*, pp. 1175–1192, IEEE, 2022.