Quantifying Privacy Leakage in Graph Embedding

Vasisht Duddu
Univ Lyon, INSA Lyon, Inria, CITI
vduddu@tutamail.com

Antoine Boutet
Univ Lyon, INSA Lyon, Inria, CITI
antoine.boutet@insa-lyon.fr

Virat Shejwalkar
Univ Massachusetts Amherst
vshejwalkar@cs.umass.edu

Abstract

Graph embeddings have been proposed to map graph data to low dimensional space for downstream processing (e.g., node classification or link prediction). With the increasing collection of personal data, graph embeddings can be trained on private and sensitive data. For the first time, we quantify the privacy leakage in graph embeddings through three inference attacks targeting Graph Neural Networks (GNNs). Firstly, we propose a membership inference attack to infer whether a graph node corresponding to individual user's data was member of the model's training or not. We consider a blackbox setting where the adversary exploits the output prediction scores, and a whitebox setting where the adversary has also access to the released node embeddings. This attack provides an accuracy up to 28% (blackbox) 36% (whitebox) beyond random guess by exploiting the distinguishable footprint between train and test data records left by the graph embedding. Secondly, we propose a Graph Reconstruction attack where the adversary aims to reconstruct the target graph given the corresponding graph embeddings. Here, the adversary can reconstruct the graph with more than 80% of accuracy and link inference between two nodes around 30% more confidence than a random guess. Finally, we propose an attribute inference attack where the adversary aims to infer a sensitive attribute. We show that graph embeddings are strongly correlated to node attributes letting the adversary inferring sensitive information (e.g., gender or location). Our analysis indicates serious data privacy risks in graph data processing algorithms and calls for further research to design privacy-preserving embeddings algorithms for graph data.

1 Introduction

Large scale real-world systems are typically modelled in the form of graphs. Consequently, A large number of applications require processing graph data which contains rich relational information between different entities (e.g., online social media, disease outbreaks, recommendation engines, knowledge graphs and navigation systems). Deep Learning and more precisely Convolutional Neural Networks have shown tremendous performance over non-graph data such as images by capturing the spatial relation between pixels of image and extracting features over multiple layers. However, this machine learning scheme has shown its limits for graph data and the learning on such data is still challenging (1). Indeed, the models have to capture the connections in the data while ensuring invariance of graph data representation, even without fixed ordering between the nodes (i.e., the adjacency matrix representing the connections between nodes varies but still results in the same graph). To overcome this limitation, the graph data is passed through embedding algorithms which map the large graphs to lower dimensions which are then used for downstream processing with GNNs. Graph embedding algorithms enable models operating on low dimensional euclidean datasets (i.e., such as images) to graph data by mapping them into a low dimensional embedding. In many applications,

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

such embeddings are released for further processing to save storage cost without considering the privacy implications.

Large graph dataset raises the question of privacy specifically if the algorithms and models are trained with private and potentially sensitive data. Consider a graph capturing the outbreak of a disease where the nodes represent the individuals, medical symptoms as the node features and the edges indicating the disease transmission. Typically, in such datasets a GNN provides state of the art performance for predicting disease for an arbitrary user in the graph (node classification) and determining the future outbreak (link prediction). For such embedding models which do not account privacy, an adversary can however infer the health status of a particular user (node in graph) by identifying whether the user was part of the training data or not. Further, the adversary can potentially reconstruct the sensitive graph input from the low dimensional embeddings. Finally, graph embeddings capture important semantics from the input graph while maintaining the contextual information in the form of preferential connection which can be exploited to infer sensitive attributes about an individual. These three privacy attacks, namely, membership inference, graph reconstruction and attribute inference, are examples of a direct privacy violation of the individual which can further be used without user consent. Further, companies spend enormous resources to annotate the training dataset to achieve state of the art performance and such attacks inferring training data also violates the Intellectual Property.

Prior literature in privacy attacks focus on models trained on non-graph data including text, images and speech to study the vulnerability to membership inference (2; 3), attribute inference (4; 5), property inference (6), model inversion (7) attacks as well as model parameter and hyperparameter stealing attacks (8; 9; 10). While well studied in traditional ML, the privacy risk in graph-based ML models under adversarial setting has not been fully explored and quantified.

2 Contribution and Key Results

In this work, we propose the first comprehensive privacy analysis of Graph Embedding algorithms under different threat models and adversary assumptions. We mainly focus on exploiting publicly released graph embeddings trained with private data, used for different downstream tasks, under various practical attacks which violates the user's data privacy: membership inference, graph reconstruction and attribute inference.

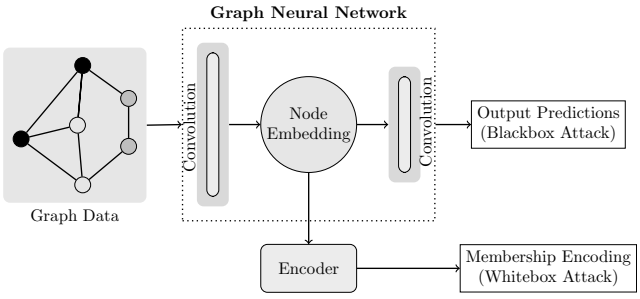
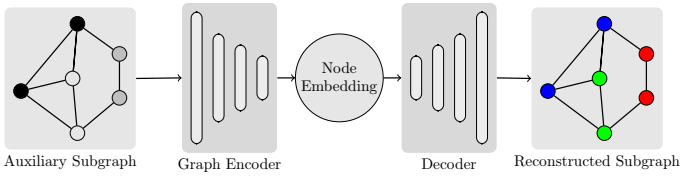


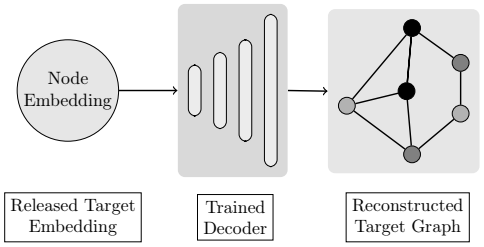
Figure 1: Blackbox and whitebox inference attacks to distinguish members and non-members of the training graph.

First, we evaluate the privacy leakage under membership inference attacks where the gaol of the adversary is to infer whether a given user's node was used in the training graph dataset or not. This is a binary classification problem where the adversary learns the threshold to predict the membership of a user node. Depending on the adversary's knowledge, we consider two settings: blackbox (with and without auxiliary knowledge) and whitebox. As shown Figure 1, to distinguish between members and non-members of the training graph, the blackbox attacks exploit the statistical difference in output predictions while the whitebox attack exploits the intermediate low dimensional embedding. The blackbox setting considers the specific case of downstream node classification task for convolution kernel based graph embedding with neural network. In this setting, we propose two attacks for membership inference: with auxilary knowledge on the data distribution (shadow model attack) and without auxilary knowledge (confidence score attack). Here, we show that the proposed attacks have an inference accuracy of 78%, 63%, and 60% for confidence score attack and 62%,

60%, and 55% for shadow model attack, respectively for three standard benchmarking datasets, i.e., Cora, Citesser and Pubmed dataset. For the whitebox setting, we propose an unsupervised attack for the more generic case of using just the graph embeddings to differentiate whether a given node was part of the training graph or not. We show that an adversary in this setting can predict the training data with a high accuracy (70% on average on the three datasets).



(a) Adversary trains attack model on auxiliary subgraph



(b) Attack model reconstructs target graph

Figure 2: Attack methodology for graph reconstruction from released embeddings.

Second, we propose a novel graph reconstruction attack where the adversary, given access to the node embeddings of a subgraph, trains an encoder-decoder model to reconstruct the target graph from its publicly released embeddings (Figure 2). This attack has serious privacy implications since the adversary reconstructs the input graph dataset which can be potentially sensitive. The proposed attack has high precision: 0.722 for Cora, 0.778 for Citeseer and 0.95 for Pubmed dataset. Moreover, on increasing the adversary's prior knowledge, the attack performance increases significantly. An important privacy implication is link inference, i.e, predicting whether there exists a link between any two nodes in the graph. Through this attack, an adversary infers a link between nodes with 93%, 90% and 57% of accuracy for respectively Cora, Citeseer and Pubmed dataset, compared to the 50% baseline random guess accuracy.

Finally, we propose the attribute inference attack where the adversary tries to infer sensitive attributes for user node in the graph using the released graph embeddings. We consider two state of the art unsupervised random walk based embeddings, Node2Vec (11) and DeepWalk (12), on two real-world social networking datasets: Facebook [1] and LastFM [2], where the adversary aims to infer the user gender and location, respectively. Given access to the embeddings of a subgraph and corresponding sensitive attributes, we model attribute inference as a supervised learning problem. The adversary trains a supervised attack model to predict sensitive hidden attributes for target users given the released publicly available target embeddings. Here, the attack model's F1 score (capturing the balance between precision and recall) on LastFM was as high as 0.65 for DeepWalk and 0.83 for Node2Vec. For Facebook, the F1 score was 0.61 for Node2Vec and 0.59 for DeepWalk.

The extended version of our analysis is available in preprint (13). In addition, the code for all the experiments is made publicly available for easy reproducibility[3].

3 Discussions and Conclusions

This work provides the first comprehensive privacy risk analysis related to graph embedding algorithms trained on sensitive graph data. Specifically, we quantify privacy leakage of three major

[1] http://snap.stanford.edu/data/ego-Facebook.html

[2] http://snap.stanford.edu/data/feather-lastfm-social.html

[3] `https://github.com/vasishtduddu/GraphLeaks`

classes of privacy attacks under practical adversary assumptions and threat models, namely membership inference, graph reconstruction and attribute inference. Our results underlines many privacy risks in graph embeddings and calls for further research to mitigate these privacy threats.

Potential mitigation strategies to lower the privacy risks can be considered. For instance, lowering the precision of the embedding vector for each node by rounding can help to reduce the attack model from learning rich features about the inputs (3; 14). In the proposed attacks, the attacker model is a machine learning algorithm vulnerable to adversarial examples, i.e, imperceptible noise added to the output prediction to force the target model to misclassify. The embeddings can be released with an additional adversarial noise to misclassify the target model while additionally ensuring utility (15; 16). Further, the inference attacks can be modelled within the training process as a minimax adversarial training with joint optimization to minimize the model loss using the graph embeddings (e.g., GNNs) while maximising the adversary's loss on inferring the sensitive inputs (17; 18). Finally, Differential Privacy can provide a theoretical bound on the total privacy leakage from the downstream processing from embeddings on an individual's data point (19; 20). However, the efficacy of these potential mitigations are left for future work.

References

[1] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *arXiv preprint arXiv:1812.08434*, 2018.

[2] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *NDSS*, 2019.

[3] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *SP*, 2017, pp. 3–18.

[4] N. Gong and B. Liu, "You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors," in *USENIX Security*, 2016, pp. 979–995.

[5] 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," *Int. J. Secur. Netw.*, vol. 10, no. 3, pp. 137–150, 2015.

[6] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *CCS*, 2018, pp. 619–633.

[7] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *CCS*, 2015, pp. 1322–1333.

[8] V. Duddu, D. Samanta, D. V. Rao, and V. E. Balas, "Stealing neural networks via timing side channels," 2018.

[9] V. Duddu and D. V. Rao, "Quantifying (hyper) parameter leakage in machine learning," *arXiv:1910.14409*, 2019.

[10] B. Wang and N. Z. Gong, "Stealing hyperparameters in machine learning," in *SP*, 2018, pp. 36–52.

[11] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *KDD*, 2016.

[12] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *KDD*, 2014, pp. 701–710.

[13] V. Duddu, A. Boutet, and V. Shejwalkar, "Quantifying privacy leakage in graph embedding," *Preprint*, 2020.

[14] X. Pan, M. Zhang, S. Ji, and M. Yang, "Privacy risks of general-purpose language models," in *SP*, 2020.

[15] J. Jia and N. Z. Gong, "Attriguard: A practical defense against attribute inference attacks via adversarial machine learning," in *USENIX Security*, 2018, pp. 513–529.

[16] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *CCS*, 2019, pp. 259–274.

[17] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *CCS*, 2018, pp. 634–646.

[18] C. Song and A. Raghunathan, "Information leakage in embedding models," 2020.

[19] L. J. Xuan-Son Vu, Son N. Tran, "dpugc: Learn differentially private representation for user generated contents," in *CICLing*, 2019.

[20] D. Xu, S. Yuan, X. Wu, and H. Phan, "Dpne: Differentially private network embedding," 2018, pp. 235–246.