International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data

K.Venkatesh[1] *M.shanmukha anjaneya raju* [2] *C.Giri Babu* [3] *A. Akhilesh* [4]

1, 2, 3, 4 students, Department of Electronics and Communication, Santhiram Engineering College,A.P, India.

ABSTRACT

Recent advances in protecting node privacy on graph data and attacking graph neural networks (GNNs) gain much attention. The eye does not bring these two essential tasks together yet. Imagine an adversary can utilize the powerful GNNs to inferusers' private labels in a social network. How can we adversarial defend against such privacy attacks while maintaining the utility of perturbed graphs? In this work, we propose a novel research task, adversarial defenses against GNN-based privacy attacks, and present a graph perturbation-based approach, NetFense, to achieve the goal. NetFense can simultaneously keep graph data unnoticed ability (i.e., having limited changes on the graph structure), maintain the prediction confidence of targeted label classification (i.e., preserving data utility), and reduce the prediction confidence of private label classification (i.e., protecting the privacy of nodes). Experiments conducted on ingle- and multiple-target perturbations using three real graph data exhibit that the perturbed graphs by NetFense can effectively maintain data utility (i.e., model unnoticed ability) on targeted label classification and significantly decrease the prediction confidence of private label classification (i.e., privacy protection). Extensive studies also bring several insights, such as the flexibility of NetFense, preserving local neighbourhoods in data unnoticed ability, and better privacy protection for high-degree nodes.

1.INTRODUCTION

GRAPH data, such as citation networks, social networks, and knowledge networks, are attracting much attention in real applications. Graphs can depict not only node features but their relationships. With the development of deep learning, graph neural networks (GNNs) are currently the most popular paradigm to learn and represent nodes in a graph. GNN encodes the patterns from node features, aggregates the representations of neighbors based on their edge connections, and generates effective embeddings for downstream tasks, such as node classification, link prediction, and community detection. Typical GNN models include graph convolution-based semi-supervised learning, and generating relational features by incorporating input features with a columnar network. In addition, graph attention is developed to estimate the contribution of incident edges. The theory of information aggregation in GNNs has also discussed to enhance the representation ability. Powerful GNNs make us concern about the disclosure of private information if the adversary considers private labels (e.g., gender and age) as the label. For example, while online social networks, such as Facebook, Twitter, and LinkedIn allow users to do privacy controls, partial data still have leakage crisis if users do not actively enable the privacy settings or agree with the access for external apps. As the adversary has partial data, GNNs can be trained to infer and acquire private information. For example, GNNs can be used to detect the visited location via user-generated texts on Twitter, and to predict age and gender using users An elaboration of privacy-protected graph perturbation. Left: we expect a method to perturb the given graph by removing an edge and adding a new one such that two requirements are satisfied: (1) The prediction confidence (yaxis) on private labels (i.e., square and circle) is lowered down, i.e., decreasing the risk of leaking privacy. (2) The prediction confidence on targeted labels (i.e., light green and yellow) is maintained, i.e., keeping the data utility. Right: The proposed NetFense model can achieve such two requirements, compared to clean and perturbed data generated by Netteck. e-commerce records. Differential privacy (DP) is a typical approach to add noise into an algorithm so that the risk of leaking private data can be lowered down. DPN and PPGD devise shallow DP-based embedding models to decrease performances of link prediction and node classification. However, since the original graph data cannot be influenced, such two models still lead to a high potential of risk exposure of private information. We use Fig. 1 (left) to elaborate the idea. Through a well-devised defense model, the graph is perturbed by removing one edge and adding another one. We expect that the new graph misleads the inference on private labels by decreasing the prediction confidence while keeping the data utility on target labels by maintaining the prediction confidence. In fact, an adversary can train a model based on the public-available data of some users' profiles in online social platforms, such as Twitter and Instagram. Not all of the users seriously care about their privacy. Hence, personal attributes and connections can be exposed due to two factors. First, some users may be not aware of privacy leaking when they choose to set some fields public. Second, some users do not care whether their private information is obtained by other people but are eager to promote themselves and maximize the visibility of themselves by proving full personal data. Data collected from such kinds of users allow the attackers to train the attack model. Therefore, we aim to find and fix the weaker parts from the data that can cause the risk of privacy exposure inferred by the attack model, and also to maintain the data utility. Then, the attackers would see the privacy-preserved data and cannot disclose the private labels of users by using the attack model. Nettack is the most relevant study. A gradient-based attack model is developed to perturb node features and graph structure so that the performance of a task (e.g., node classification) is significantly reduced. However, Nettack cannot work for privacy protection in two aspects. First, when the targeted private label is binary, the adversary can reverse the misclassified labels to obtain the true value if she knows there is some protection. Second, while Nettack can be used to defend against privacy attacks by decreasing the performance, it does not guarantee the utility of the perturbed data on inferring non-private labels. As shown in Fig. 1 (right) conducted on real graph data, Nettack leads to misclassification on the private label, but fails to maintain the prediction confidence on the target label. Note that the y-axis is the classification margin, indicating the difference of prediction probabilities between the ground truth and the 2-nd probable label. It can be also regarded as the prediction confidence of a model. Negative values mean higher potential to be identified as the 2-nd probable label. In this paper, we propose a novel problem of adversarial defense against privacy attack on graph data. Given a graph, in which each node is associated with a feature vector, a targeted label (e.g., topic or category), and a private label (e.g., gender or age), our goal is to perturb the graph structure by adding and removing edges such that the privacy is protected and the data utility is maintained at the same time. To be specific, we aim at lowering down the prediction confidence on the private label to prevent privacy from being inferred by the adversary's GNNs, and simultaneously maintaining the prediction confidence on the targeted label to keep the data utility under GNNs when the data is released. This task can be treated as a kind of privacy defense, i.e., defending the model attack that performs learning to infer private labels. We create Table 1 to highlight the key differences between model attack (i.e., Nettack) and our proposed privacy defense (i.e., NetFense) on graph data. First, since the model attack is performed by the adversary and the privacy defense is conducted by the data owner, their scope of accessible data is different. Second, as mentioned above, our problem is to tackle two tasks at the same time (i.e., Summary of differences between model attack and privacy defense on graph data in terms of who is doing the attack/defense (WHO), accessible data (AD), strategy (STG), perturbation objective (PO), non-noticeable perturbation (NP), number of tackled task (#Task), and number of concerned targets (#Trg). “Pred-Acc” and “Pred-Confi” are prediction accuracy and confidence. “on” is maintenance.

2.METHODOLOGY

In this paper, we propose a novel problem of adversarial defense against privacy attack on graph data. Given a graph, in which each node is associated with a feature vector, a targeted label (e.g., topic or category), and a private label (e.g., gender or age), our goal is to perturb the graph structure by adding and removing edges such that the privacy is protected and the data utility is maintained at the same time. To be specific, we aim at lowering down the prediction confidence on the private label to prevent privacy from being inferred by the adversary's GNNs, and simultaneously maintaining the prediction confidence on the targeted label to keep the data utility under GNNs when the data is released. This task can be treated as a kind of privacy defense, i.e., defending the model attack that performs learning to infer private labels. We create Table 1 to highlight the key differences between model attack (i.e., Nettack and our proposed privacy defense (i.e., NetFense) on graph data. First, since the model attack is performed by the adversary and the privacy defense is conducted by the data owner, their scope of accessible data is different. Second, as mentioned above, our problem is to tackle two tasks at the same time (i.e., fool the model on private labels and keep data utility), but model attack deals with only fooling the model on targeted labels. Third, in the context of privacy protection, decreasing the prediction accuracy on private labels cannot prevent them from being inferred if the private label is binary. The adversary can reverse the prediction results if she knows the existence of defense mechanism. Therefore, we reduce the prediction confidence as close as possible to 0:5. Fourth, while model attack tends to make one or fewer nodes misclassified on targeted labels, privacy defense is expected to shield the private labels of more nodes from being accurately inferred. Last, both model attack and privacy defense need to ensure the perturbation on graph data is unnoticeable. Privacy defense further requires to achieve model unnoticed ability, which is maintaining the performance of target label prediction using the perturbed graph under the same model (i.e., equivalent to maintain data utility). It is quite challenging to have a defense model that meets all of these requirements at the same time.

To tackle the proposed privacy defense problem, we propose an adversarial method, NetFense, based on the adversarial model attack. NetFense consists of three phases, including candidate selection, influence with GNNs, and combinatorial optimization. The first phase ensures the perturbed graph is unnoticeable while the second and third phases ensure both privacy preservation and data utility (i.e., model unnoticeable) of the perturbed graph.

3. WORKING PRINCIPLE

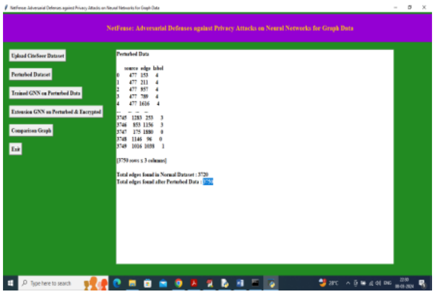
In propose paper author is providing security to data just by adding and removing some dummy or fake edges but the ID of the authors are clearly visible to attackers so as extension we are encrypting and perturbing above author ID's so attacker should not understand ID also. After encrypting accuracy may be high or low but privacy will be more.

Modules:

To implement this project we have designed following modules

1) Upload CiteSeer Dataset: using this module will upload CITESEER data to application and then record size of dataset
2) Perturbed Dataset: this module will perturbed data by adding and removing edges and after adding and removing edges we can see the change size of dataset
3) Trained GNN on Perturbed Data: perturbed data will be input to GNN algorithm to train a model and this model will be applied on test data to calculate accuracy. Wrong predicted percentage will be consider as privacy margin for the dataset. The more records the GNN predicted wrongly the more data security will be achieved
4) Extension GNN on Perturbed & Encrypted: perturbed & encrypted data will be input to GNN algorithm to train a model and this model will be applied on test data to calculate accuracy. Wrong predicted percentage will be consider as privacy margin for the dataset. The more records the GNN predicted wrongly the more data security will be achieved
5) Comparison Graph: will plot algorithm training loss graph between propose and extension algorithm. The lower the loss the better is the algorithm

4.RESULTS


NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data
NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data
Upload CiteSeer Dataset
Perturbed Dataset
Trained GNN on Perturbed Data
Extension GNN on Perturbed & Encrypted
Comparison Graph
Exit
Type here to search

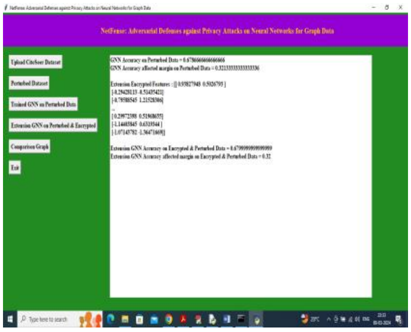
```
Perturbed Data

      source  edge  label
0        477   153      4
1        477   211      4
2        477   957      4
3        477   789      4
4        477  1616      4
...      ...   ...    ...
3745    1283   253      3
3746     853  1156      3
3747     175  1880      0
3748    1146    96      0
3749    1016  1038      1

[3750 rows x 3 columns]

Total edges found in Normal Dataset : 3720
Total edges found after Perturbed Data : 3750
```

In above screen can see now dataset having source node, connecting edge node name and label and after perturbing edges size increase to 3750 from 3720 and now click on ‘Trained GNN on Perturbed Data’ button to trained GNN algorithm on perturbed data and get below output

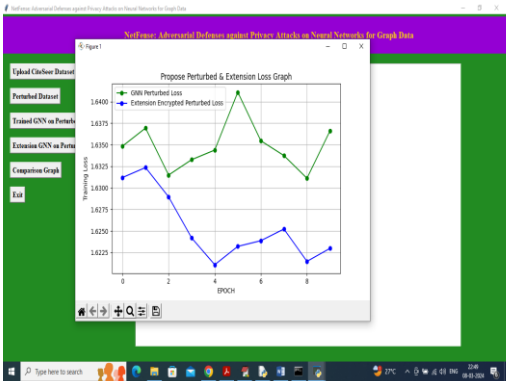

NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data
NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data
Upload CiteSeer Dataset
Perturbed Dataset
Trained GNN on Perturbed Data
Extension GNN on Perturbed & Encrypted
Comparison Graph
Exit
Type here to search

```
GNN Accuracy on Perturbed Data = 0.6786666666666666
GNN Accuracy affected margin on Perturbed Data = 0.32133333333333336

Extension Encrypted Features : [[ 0.93827948  0.5026793 ]
 [-0.29428113 -0.51435421]
 [-0.79588545  1.21528306]
 ...
 [ 0.29972388  0.51968655]
 [-1.14483845  0.6319344 ]
 [ 1.07143782 -1.36471669]]

Extension GNN Accuracy on Encrypted & Perturbed Data = 0.6799999999999999
Extension GNN Accuracy affected margin on Encrypted & Perturbed Data = 0.32
```

In above screen extension accuracy is increased in points and data is having encrypted points so data will be more secured and in above screen can see some encrypted values. So with extension technique by adding encrypted training we can provide more security or privacy to data. Now click on 'Comparison Graph' link to get below graph


NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data
NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data
Figure 1
Upload CiteSeer Dataset
Perturbed Dataset
Trained GNN on Perturb
Extension GNN on Pertu
Comparison Graph
Exit
Propose Perturbed & Extension Loss Graph
GNN Perturbed Loss
Extension Encrypted Perturbed Loss
1.6400
1.6375
1.6350
1.6325
1.6300
1.6275
1.6250
1.6225
Training Loss
0
2
4
6
8
EPOCH
Type here to search
27°C
ENG
22:49
08-03-2024

In above graph x-axis represents training epochs and y-axis represents training loss and then green line represents 'GNN Perturbed Loss' and blue line represents 'Extension Encrypted Perturbed Loss'. In both algorithms can see Extension got less loss so it will provide more security or privacy to data.

6.CONCLUSION

This paper presents a novel research task: adversarial defences against privacy attack via graph neural networks under the setting of semi-supervised learning. We analyze and compare the differences between the proposed problem and model attacks on graph data, and realize the perturbed graphs should keep data unnoticed ability, maintain model unnoticed ability (i.e., data utility), and achieve privacy protection at the same time. We develop an adversarial approach, NetFense, and empirically find that the graphs perturbed by NetFense can simultaneously lead to the least change of local graph structures, maintain the performance of targeted label classification, and lower down the prediction confidence of private label classification. We also exhibit that perturbing edges brings more damage in misclassifying private labels than perturbing node features. In addition, the promising performance of the proposed NetFense lies in not only single-target perturbations, but also multi-target perturbations that cannot be well done by model attack methods such as Nettack. The evaluation results also deliver that moderate edge disturbance can influence the graph structure to avoid the leakage of privacy via GNNs and alleviate the destruction of graph data. Besides, we also offer the analysis of hyper parameters and perturbation factors that are highly related to the performance. We believe the insights found in this study can encourage future work to further investigate how to devise a privacy-preserved graph neural networks, and to study the correlation between the leakage of multiple private labels and attributed graphs.

7.REFERENCES

[1] Ghazaleh Beigi, Ahmadreza Mosallanezhad, Ruocheng Guo, Hamidreza Alvari, Alexander Nou, and Huan Liu. Privacy-aware recommendation with private-attribute protection using adversarial learning. In Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM ’20, pages 34–42, 2020.

[2] Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. Enhancing robustness of machine learning systems via data transformations. In 52nd Annual Conference on Information Sciences and Systems (CISS), pages 1–5, 2018.

[3] Aleksandar Bojchevski, Johannes Klicpera, Bryan Perozzi, Martin Blais, Amol Kapoor, Michal Lukasik, and Stephan Gunnemann. ¨ Is pagerank all you need for scalable graph neural networks? In Proceedings of the 15th international workshop on mining and leaning with graphs, 2019.

[4] Zhipeng Cai, Zaobo He, Xin Guan, and Yingshu Li. Collective data-sanitization for preventing sensitive information inference attacks in social networks. IEEE Transactions on Dependable and Secure Computing, 15(4):577–590, 2016.

[5] Liang Chen, Jintang Li, Jiaying Peng, Tao Xie, Zengxu Cao, Kun Xu, Xiangnan He, and Zibin Zheng. A survey of adversarial learning on graphs. arXiv preprint arXiv:2003.05730, 2020.

[6] Weijian Chen, Yulong Gu, Zhaochun Ren, Xiangnan He, Hongtao Xie, Tong Guo, Dawei Yin, and Yongdong Zhang. Semi-supervised user profiling with heterogeneous graph attention networks. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, pages 2116–2122, 2019.

[7] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. SIAM review, 51(4):661– 703, 2009.

[8] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In Proceedings of the 35th International Conference on Machine Learning, pages 1115–1124, 2018.

[9] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 196–204, 2018.

[10] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference, pages 265–284. Springer, 2006.

[11] Negin Entezari, Saba A. Al-Sayouri, Amirali Darvishzadeh, and Evangelos E. Papalexakis. All you need is low (rank): Defending against adversarial attacks on graphs. In Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM), 2020.

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In 3rd International Conference on Learning Representations, ICLR, 2015.

[13] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In 5th International Conference on Learning Representations, ICLR, 2017.