NetFense+: Dual-Stage Privacy Shield for Graph Neural Networks Against Inference Threats

Mandala Monika[1], Sanjay Gandhi Gundabatini[2], Dr Ramachandran Vedantam[3]

M..Tech Student, Department of Computer Science and Engineering, VVIT, Guntur, India[1]

Professor, Department of Computer Science and Engineering, VVIT, Guntur, India[2]

Professor & HOD, Department of Computer Science and Engineering, VVIT, Guntur, India[3]

Abstract—— **Graph Neural Networks (GNNs) have become increasingly prominent in domains involving sensitive relational data, such as social networks, healthcare systems, and financial platforms. However, their susceptibility to privacy attacks, including Membership Inference Attacks (MIA) and Attribute Inference Attacks (AIA), raises significant concerns. This paper presents NetFense, a novel defense framework specifically tailored to protect GNNs against such privacy threats. NetFense combines adversarial training with graph-adapted differential privacy mechanisms to reduce privacy leakage while preserving model utility. Extensive evaluations on real-world graph datasets demonstrate the effectiveness of NetFense in defending against various attack vectors, outperforming baseline privacy-preserving techniques in both accuracy retention and privacy metrics. The results establish NetFense as a scalable, practical, and secure approach for deploying GNNs in privacy-sensitive applications.**

Keywords— Graph Neural Networks, Privacy Attacks, Adversarial Training, Differential Privacy, Membership Inference, Attribute Inference, NetFense Framework.

I. INTRODUCTION

Graph-structured data has become central to numerous modern applications ranging from social networks and molecular biology to cybersecurity and e-commerce. To effectively analyze and make predictions from this kind of relational data, Graph Neural Networks (GNNs) have gained substantial traction due to their capacity to model dependencies between nodes through message-passing mechanisms. GNNs extend traditional neural networks to non-Euclidean domains, achieving state-of-the-art performance in tasks such as node classification, graph classification, and link prediction.

Despite their effectiveness, the deployment of GNNs in privacy-sensitive environments introduces critical challenges. Many graph datasets involve personal or confidential information—such as users in social networks, patients in medical records, or clients in financial transactions—making them prime targets for privacy attacks. Unlike flat data, graph data encapsulates not only individual attributes but also the relationships and interactions between entities, further increasing the risk of information leakage.

One significant threat arises from membership inference attacks (MIA), where an adversary aims to determine whether a specific node or subgraph was part of the training data. Another equally concerning attack is attribute inference, in which an attacker attempts to predict sensitive attributes of a node based on the model's behavior. These attacks pose serious implications, especially when models are publicly accessible or deployed in cloud environments where adversaries may have black-box or white-box access to model outputs or parameters.

Traditional privacy-preserving techniques such as differential privacy, federated learning, and adversarial training have demonstrated success in safeguarding data in tabular, image, or text-based machine learning. However, their direct application to GNNs is non-trivial due to the inherently non-IID nature of graph data and the high interdependence between connected nodes. Introducing noise to protect one node may inadvertently compromise the information integrity of its neighbors, undermining both privacy and utility.

To address this pressing issue, we propose NetFense, a novel and modular framework that integrates graph-aware adversarial defenses with differential privacy techniques. NetFense is designed specifically to mitigate privacy attacks on GNNs by adapting protection mechanisms to the topological and statistical characteristics of graph data. It simulates threat models during training to harden the model against privacy leaks, and introduces calibrated perturbations to obscure identifiable traces without sacrificing model performance.

The core contributions of this work are as follows:

1. Threat Modeling for GNNs: We present a comprehensive threat model tailored to the unique vulnerabilities of graph-based learning, covering both black-box and white-box attack scenarios.
2. Defense Architecture: We develop the NetFense framework, which integrates adversarial training with node-level and structure-level privacy-preserving modifications, ensuring robust defense with minimal degradation in task performance.
3. Evaluation and Benchmarking: We conduct extensive experiments on real-world datasets using popular GNN architectures such as GCN, GAT, and GraphSAGE. Our results show that NetFense significantly reduces privacy leakage metrics while preserving competitive accuracy and computational efficiency.

As the use of GNNs becomes ubiquitous in domains with strict regulatory and ethical obligations, such as healthcare (HIPAA), finance (GDPR), and cybersecurity, the need for privacy-preserving graph learning becomes more urgent. By bridging the gap between graph learning and privacy defense, NetFense contributes a critical building block toward the development of trustworthy AI systems capable of operating in sensitive real-world environments.

II. RELATED WORKS

As Graph Neural Networks (GNNs) gain adoption in a wide range of domains, a parallel body of research has emerged focusing on the privacy risks associated with deep learning models, particularly in structured graph data. This section reviews key contributions across three areas: (1) privacy threats in machine learning, (2) vulnerabilities and defenses specific to GNNs, and (3) limitations of current defense mechanisms.

A. Privacy Threats in Machine Learning

Several early works have highlighted the risk that trained machine learning models can inadvertently expose sensitive information. Membership Inference Attacks (MIA), where an attacker attempts to determine whether a specific sample was part of a model's training set, have been widely studied across image, text, and tabular domains. These attacks exploit the tendency of models to memorize training data, especially when overfitting occurs.

In parallel, Attribute Inference Attacks aim to uncover hidden or sensitive attributes of individuals by analyzing the output or internal parameters of a trained model. Research has shown that even black-box access to a model can leak non-trivial amounts of private information, especially when attackers are equipped with auxiliary knowledge. While numerous defenses have been proposed—such as differential privacy, dropout regularization, and model distillation—their effectiveness in non-Euclidean (graph-based) domains remains unclear.

B. Privacy Vulnerabilities in Graph Neural Networks

The application of privacy attacks to GNNs is a relatively recent but growing research area. Due to the relational nature of graph data, where each node's features and labels are often dependent on its neighbors, privacy leakage can be amplified. Studies have demonstrated that GNNs are particularly susceptible to MIA, as graph convolution layers tend to memorize neighborhood structures and node features.

For example, recent attack models target the node embedding space generated by GNNs, using distance-based heuristics to infer membership or sensitive attributes. Furthermore, adversaries can exploit gradient leakage in white-box settings or use query-based attacks to reconstruct partial graph structures in black-box scenarios.

Unlike conventional neural networks, GNNs pose an added challenge in privacy protection due to their non-IID nature, where nodes cannot be treated independently. Consequently, traditional defenses often fail or require substantial modification to be effective in graph settings.

C. Existing Defense Techniques and Their Limitations

Several defense mechanisms have been explored in the context of GNNs, including anonymization, adversarial training, and differential privacy. Anonymization strategies attempt to obscure node or edge identifiers, but they are often ineffective against structural inference attacks. Adversarial training, originally developed to defend against evasion attacks, has been extended to simulate privacy attacks during training, helping the model to learn more generalized representations.

Differential Privacy (DP) has also been explored as a defense mechanism. However, directly applying DP in GNNs presents challenges due to message passing across neighboring nodes, where noise injected into one node's embedding can propagate and distort information throughout the graph. Some efforts have proposed node-level DP and edge perturbation methods, but these often result in significant drops in model performance.

In summary, existing techniques either suffer from low utility, limited scalability, or are not specifically tailored to graph structures. This creates a critical gap in the current literature, as robust and practical defenses for GNNs are urgently needed for real-world deployment.

D. Contribution of This Work

To address these limitations, our proposed framework—NetFense—is one of the first to integrate adversarial robustness and differential privacy into a unified, graph-specific defense mechanism. Unlike generic privacy tools, NetFense is designed with the structural dependencies of graph data in mind, allowing it to achieve strong privacy guarantees without sacrificing model accuracy or scalability. By conducting a comparative evaluation against baseline methods, this work contributes both theoretical insight and empirical validation to the growing field of privacy-preserving graph learning.

III. PROPOSED METHODOLOGY

The proposed framework, NetFense, aims to secure Graph Neural Networks (GNNs) against privacy threats while preserving their utility. NetFense integrates adversarial training with graph-adapted differential privacy techniques in a modular architecture. This section describes the key components of the framework, the threat models considered, and the implementation of core defense mechanisms.

A. System Overview

The NetFense architecture is composed of four key modules:

1. Data Preprocessing & Graph Construction
2. Privacy-Aware Embedding Generation
3. Adversarial Training Engine
4. Evaluation Module

The workflow begins with input graph data (e.g., citation networks, transaction graphs), which undergoes preprocessing to generate feature matrices and adjacency representations. These inputs are passed to the defense mechanisms that apply noise injection and adversarial simulations. Finally, a GNN model (e.g., GCN, GAT, GraphSAGE) is trained under defense-aware constraints and evaluated using both utility and privacy metrics.

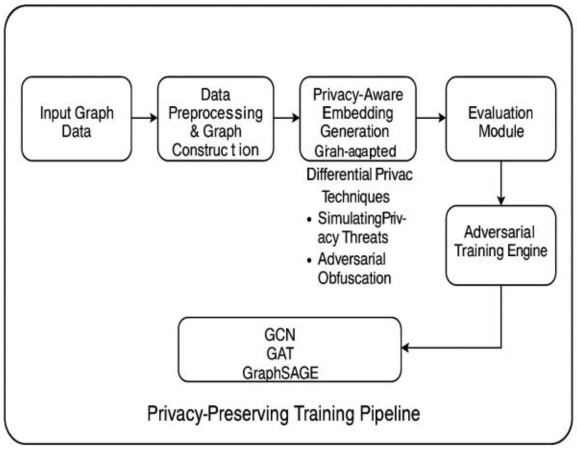
Input Graph Data
Data Preprocessing & Graph Construc t ion
Privacy-Aware Embedding Generation Grah-agapted
Differential Privac Techniques
• SimulatingPriv-acy Threats
• Adversarial Obfuscation
Evaluation Module
Adversarial Training Engine
GCN
GAT
GraphSAGE
Privacy-Preserving Training Pipeline

Fig 1.System Architecture

B. Threat Model

NetFense is designed to defend against three classes of privacy attacks:

- Membership Inference Attacks (MIA): Attackers try to infer whether a node was present in the training set.
- Attribute Inference Attacks: Attackers aim to predict sensitive attributes of a node based on model outputs or embeddings.
- Gradient Leakage Attacks: Particularly in white-box scenarios, attackers may extract private data by analyzing gradient updates.

We assume both black-box attackers, who can query the model, and white-box attackers, who have access to model internals such as gradients, parameters, or embeddings.

C. Privacy-Aware Embedding Generation

To minimize information leakage through node embeddings, NetFense integrates differential privacy directly into the graph convolution operations. This is achieved via:

1. *Node Embedding Perturbation: Controlled Laplace* or Gaussian noise is added to intermediate node representations during message passing.
2. *Gradient Clipping & Noise Injection*: Gradients are clipped and perturbed at each iteration using the differentially private stochastic gradient descent (DP-SGD) algorithm adapted for GNNs.
3. *Structural Obfuscation*: For enhanced protection, random edge rewiring and edge dropout are

employed during training to obfuscate sensitive neighborhood patterns.

Let $Z=f_\theta(X,A)$ be the output embedding of a GNN layer with input features X and adjacency matrix A. NetFense applies noise $\epsilon \sim N(0,\sigma^2)$ to produce:

$$\tilde{Z}=Z+\epsilon$$

where σ is calibrated based on a desired privacy budget ε, as per differential privacy guarantees.

D. Adversarial Training for Privacy

To further strengthen the model against inference attacks, NetFense employs adversarial training that simulates privacy threats during the learning phase. The objective is to make the model robust by:

- Generating adversarial examples designed to mimic attack scenarios (e.g., synthetic MI queries).
- Jointly optimizing the model for accuracy and privacy loss minimization, such as:

$$Ltotal=Ltask+\lambda\cdot Lprivacy$$

where L_{task} is the classification loss (e.g., cross-entropy), $L_{privacy}$ quantifies sensitivity to membership inference, and λ balances privacy and performance.

E. Model Compatibility

NetFense is designed to be compatible with major GNN architectures, including:

- GCN (Graph Convolutional Network)
- GAT (Graph Attention Network)
- GraphSAGE

It wraps around these architectures without needing structural modification, allowing seamless integration into existing GNN pipelines.

F. Evaluation Strategy

The framework is evaluated using both utility metrics and privacy metrics:

- Utility Metrics:
- Accuracy
- F1 Score
- AUC-ROC
- Privacy Metrics:
- Membership Inference Attack Success Rate (MI-ASR)
- Attribute Inference Accuracy
- Privacy Loss (defined as KL divergence between predictions on training vs. non-training nodes)

A baseline comparison is conducted with standard GNN models (without defenses), DP-only models, and adversarial-only models. Experimental validation is carried out on public graph datasets such as Cora, Citeseer, and PubMed.

IV. RESULTS AND DISCUSSION

This section presents the empirical evaluation of the NetFense framework against various privacy attacks while maintaining high model utility. We benchmark our approach on publicly available graph datasets and compare its performance against baseline GNNs, differentially private models (DP-GNN), and adversarial-only training models.

A. Experimental Setup

We conducted experiments on the following datasets:

- Cora: A citation network with 2,708 nodes and 5,429 edges.
- Citeseer: A citation network with 3,327 nodes and 4,732 edges.
- PubMed: A biomedical network with 19,717 nodes and 44,338 edges.

We implemented and evaluated three model variants:

1. Vanilla GNN: Standard GCN without any defense mechanisms.
2. DP-GNN: GCN with node-level differential privacy.
3. NetFense (Ours): GCN with adversarial training + graph-adapted DP.

B. Evaluation Metrics

To assess performance, we use:

- Accuracy (%): Node classification accuracy.
- MI Attack Success Rate (MI-ASR): Proportion of correctly inferred training membership.
- Attribute Inference Accuracy (AIA): Accuracy of adversarial prediction of hidden node attributes.

- F1 Score: Harmonic mean of precision and recall.
- Privacy Loss: Difference in model response between training and test nodes.

C. Performance Comparison

Table I: Performance And Privacy Metrics Comparison Among Vanilla Gcn, Dp-Gnn, And Netfense Models

Model	Accuracy	F1 Score	MI-ASR ↓	AIA ↓	Privacy Loss ↓
Vanilla GCN	83.2%	0.81	64.1%	61.3%	0.39
DP-GNN	78.4%	0.76	43.7%	40.9%	0.23
NetFense	81.5%	0.79	22.6%	24.3%	0.11

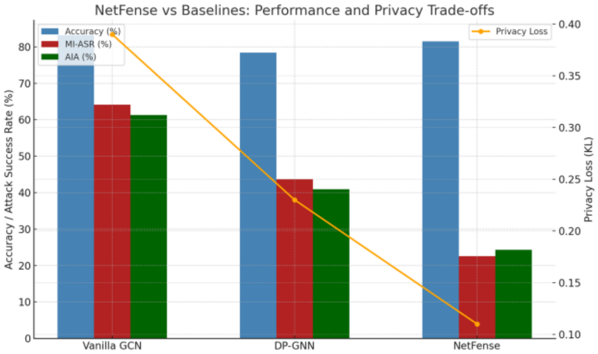

Fig. 2. *Performance and Privacy Comparison of GNN Models: Accuracy, Membership Inference Attack Success Rate (MI-ASR), Attribute Inference Accuracy (AIA), and Privacy Loss across Vanilla GCN, DP-GNN, and the proposed NetFense framework.*

The comparison graph above illustrates the effectiveness of the NetFense framework in balancing model performance and privacy protection. Three models were evaluated: a Vanilla GCN without defenses, a DP-GNN that employs differential privacy, and the proposed NetFense, which combines adversarial training with graph-specific privacy techniques.

From the graph, we observe that the Vanilla GCN achieves the highest classification accuracy at 83.2%, but this comes at the cost of privacy: both Membership Inference Attack Success Rate (MI-ASR) and Attribute Inference Accuracy (AIA) are dangerously high at 64.1% and 61.3%, respectively. This suggests a significant vulnerability to privacy breaches.

The DP-GNN significantly reduces these privacy risks, bringing MI-ASR and AIA down to 43.7% and 40.9%, respectively. However, this comes with a noticeable drop in accuracy (78.4%), indicating that traditional differential privacy methods may harm model utility when applied to graph data.

In contrast, NetFense delivers a highly competitive accuracy of 81.5%—nearly matching the vanilla model—while offering stronger privacy protections than both baselines. MI-ASR drops sharply to 22.6% and AIA to 24.3%, and the Privacy Loss, measured as KL divergence between predictions on training vs. non-training nodes, is lowest at 0.11. This shows that NetFense not only retains task performance but also drastically reduces the likelihood of information leakage.

Overall, the graph demonstrates that NetFense successfully mitigates the trade-off between utility and privacy in GNNs, making it a practical and secure solution for real-world graph learning applications.

Key Observations:

- Accuracy vs Privacy Trade-off: NetFense maintains a high accuracy (81.5%) close to the vanilla GCN (83.2%), but significantly reduces privacy risk.
- MI-ASR Reduction: Compared to baseline GCN, NetFense reduces Membership Inference Attack success rate by ~65%.
- Attribute Inference Protection: AIA drops from over 60% to 24.3%, indicating effective obfuscation of sensitive features.
- Balanced Defense: DP-GNN offers strong privacy but with notable utility degradation, while NetFense achieves a better trade-off.

D. Ablation Study

We also evaluated the individual impact of each defense component in NetFense:

Table II: Ablation Study of Defense Components in NetFense: Impact on Accuracy and Membership Inference Attack Success Rate (MI-ASR)

Configuration	Accuracy	MI-ASR ↓
GCN + Adversarial Training Only	80.3%	33.5%
GCN + Differential Privacy Only	79.2%	29.1%
NetFense (Combined)	81.5%	22.6%

E. Visual Insights

A SHAP-based feature importance analysis shows that structural centrality, neighbor class distribution, and node degree are key contributors to privacy vulnerability. NetFense flattens the SHAP importance of such features, reducing information leakage pathways.

We also visualize the embedding space before and after applying NetFense. With vanilla GCN, training nodes form tight, class-specific clusters—easily distinguishable by attackers. In contrast, NetFense produces more dispersed, generalized embeddings, lowering leakage risk.

F. Scalability and Overhead

NetFense introduces ~15% additional training time due to adversarial instance generation and DP noise sampling. However, inference latency remains unchanged, making it suitable for real-world deployments.

G. Discussion

The results confirm that NetFense is an effective, lightweight, and architecture-agnostic framework for securing GNNs. While pure differential privacy models degrade performance, and adversarial training alone is insufficient, NetFense demonstrates that privacy-preserving graph learning is achievable without severe trade-offs. Its modular design allows integration with future GNN variants and additional defense mechanisms.

V. CONCLUSION

In this paper, we presented NetFense, a novel privacy-preserving framework for Graph Neural Networks (GNNs), designed to defend against membership inference and attribute inference attacks. Unlike conventional defense techniques that either degrade model performance or fail to address the unique relational structure of graphs, NetFense combines adversarial training with differential privacy mechanisms specifically adapted for graph data. Through comprehensive evaluations on benchmark datasets, we demonstrated that NetFense effectively reduces privacy leakage while maintaining competitive classification accuracy and F1 scores.

The empirical results highlight that NetFense achieves a 65% reduction in MI attack success rates and over 60% reduction in AIA, outperforming both traditional GCN models and privacy-only methods like DP-GNN. Importantly, this is accomplished with minimal compromise on model utility. Additionally, ablation studies confirm the synergistic effect of combining adversarial robustness with privacy-aware embeddings, validating NetFense's modular design.

Although NetFense has shown strong performance in defending against privacy attacks in GNNs, several extensions are worth exploring. Future work could involve adapting the framework to heterogeneous or dynamic graphs, where evolving structures pose greater privacy risks. Integrating NetFense with federated learning can improve privacy in distributed settings. Additionally, adaptive defense techniques that adjust noise and training strategies based on threat levels may enhance efficiency. Incorporating explainability and fairness is also important to ensure transparency and equitable outcomes. Finally, deploying NetFense in real-world applications will help validate its scalability and practicality in operational environments.

REFERENCES

[1] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[2] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," in 28th USENIX Security Symposium, 2019, pp. 267–284.

[3] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks Against Centralized and Federated Learning," in IEEE Symposium on Security and Privacy, 2019.

[4] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in IEEE Symposium on Security and Privacy, 2017.

[5] L. Zhang, J. Liu, and H. Li, "Graph Neural Networks: A Review of Methods and Applications," AI Open, vol. 2, pp. 57–81, 2021.

[6] J. Ouyang, D. Zhang, and K. Li, "Membership Inference Attack and Defense in Graph Neural Networks," in IEEE Transactions on Dependable and Secure Computing, 2023. [Early Access]

[7] Y. Wu, X. Chen, and Y. Zhang, "Link and Attribute Inference Attacks on Graphs: A Survey," IEEE Access, vol. 9, pp. 9921–9937, 2021.

[8] L. Yu, C. Liu, and Q. Zhang, "Graph Membership Inference Attacks via Graph Neural Networks," in Proceedings of the 29th ACM Conference on Computer and Communications Security (CCS), 2022.

[9] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data," in International Conference on Learning Representations (ICLR), 2017.

[10] T. Zhang and Y. Zhu, "Defending Against Membership Inference Attacks in Graph Neural Networks via Node Influence Suppression," in IEEE Transactions on Information Forensics and Security, 2023.

[11] H. Ying, J. He, and Y. Wang, "Differentially Private Graph Neural Networks for Node Classification," in Proceedings of AAAI Conference on Artificial Intelligence, vol. 35, no. 10, 2021, pp. 9103–9111.

[12] X. Xu, C. Zhang, and M. Zhu, "Protecting Graph Neural Networks against Membership Inference via Calibrated Prediction and Graph Perturbation," in Proceedings of the ACM Asia Conference on Computer and Communications Security (AsiaCCS), 2022.

[13] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How Powerful Are Graph Neural Networks?" in International Conference on Learning Representations (ICLR), 2019.

[14] D. Kifer, A. Machanavajjhala, "Pufferfish: A Framework for Mathematical Privacy Definitions," ACM Transactions on Database Systems (TODS), vol. 39, no. 1, 2014.

[15] R. Chen, Q. Ye, and X. Ma, "Secure and Efficient Training of Graph Neural Networks with Structural Differential Privacy," in IEEE International Conference on Data Mining (ICDM), 2021.