Survey on Model Extraction Attacks Against Graph Neural Networks (GNNs)

Ahmed Aminu Sambo
Department of Computer Sciences,
Ahmadu Bello University,
Zaria, Nigeria
smbaminu@gmail.com

Mohammad Lawal
Department of Computer Sciences,
Ahmadu Bello University,
Zaria, Nigeria
mlawal180@gmail.com

Sani Ahmad Hassan
Department of Computer Sciences,
Ahmadu Bello University,
Zaria, Nigeria
myynazee@gmail.com

Abubakar Isa
Department of Computer Sciences,
Ahmadu Bello University,
Zaria, Nigeria
abubakarisagmail.com

***Abstract*- Graph Neural Networks (GNNs) have gained substantial attention for tasks such as node classification, link prediction, and graph classification due to their ability to learn complex relationships from graph-structured data. Their use in critical applications—such as social networks, finance, healthcare, and cybersecurity—has, however, exposed them to *model extraction attacks*. These attacks enable adversaries to steal a model's structure, parameters, or functionality through black-box access, threatening both intellectual property and data privacy. This survey provides a structured and analytical overview of model extraction attacks on GNNs, classifying them based on attacker knowledge, query strategies, and defense mechanisms. It also discusses open challenges such as scalability, query limitations, and privacy leakage. In addition, it introduces a taxonomy for GNN model extraction, compares existing approaches, and highlights emerging defense strategies such as watermarking, query monitoring, and differential privacy. Finally, it outlines future directions for building secure and privacy-preserving GNN systems.**

***Keywords:* Graph Neural Networks, Model Extraction, Adversarial Machine Learning, Security of GNNs, Federated Learning, Privacy Preservation.**

I. INTRODUCTION

Graph Neural Networks (GNNs) have revolutionized graph-based learning, offering solutions for node classification, link prediction, and graph classification tasks. They capture relationships among interconnected entities, making them valuable in fields such as social network analysis, drug discovery, fraud detection, and biological systems [1], [2].

Despite their usefulness, GNNs deployed via **Machine Learning as a Service (MLaaS)** are increasingly vulnerable to *model extraction attacks*. In these attacks, adversaries query the model's API to reconstruct a surrogate version that mimics its behavior [3]. The risk is particularly high for GNNs, as they are expensive to train and often contain sensitive relational information [4].

This paper presents a critical and comprehensive survey of GNN model extraction attacks. Unlike general adversarial ML surveys, it focuses specifically on GNNs, proposing a taxonomy for categorizing attacks and defenses, analyzing key trends, and identifying research gaps.

II. TAXONOMY OF MODEL EXTRACTION ATTACKS ON GNNS

To establish a structured understanding of model extraction research in Graph Neural Networks (GNNs), this paper introduces an expanded taxonomy that classifies existing approaches along five major dimensions: **attacker knowledge**, **query strategy**, **reconstruction target**, **attack goals**, and **defense mechanisms**. This taxonomy not only organizes prior works but also highlights emerging attack paradigms and open challenges.

A. Attacker Knowledge Level

The degree of information available to an attacker significantly influences the type and success of model extraction attacks. Three main scenarios are identified:

- **White-box Attacks:**
  The attacker has full access to the GNN's internal parameters, architecture, and training data. While rare in practice, these attacks serve as a theoretical upper bound for model vulnerability. They are often used to benchmark extraction efficiency.
- **Gray-box Attacks:**

  Partial information is available, such as the GNN architecture (e.g., GCN, GAT, or GraphSAGE), but the weights and training data remain hidden. This setting reflects semi-open MLaaS environments where documentation or prior knowledge assists the attacker.

- **Black-box Attacks:**
  The most realistic and widely studied setting. Here, the attacker interacts only through query-response pairs via an API, obtaining outputs such as labels or confidence scores [5]. Attacks must infer both struc ture and function through repeated interaction, making efficiency a key challenge [6].

B. Query Strategy

Query strategy determines how attackers interact with the target GNN to maximize extracted information while minimizing cost and detection risk. Strategies can be categorized as follows:

- **Random Querying:**
  The simplest approach where random inputs are generated without prior feedback. Though inefficient, it establishes baseline performance for extraction attempts.
  Adaptive Querying:
  Attackers adjust queries based on previous responses, selecting inputs that provide the most informative feedback. These strategies often emplo y reinforcement learning or uncertainty sampling [7].
- **Active Learning-based Querying:**

  Attackers leverage machine learning to actively identify which queries will maximize information gain [8]. This approach drastically reduces query volume, making it harder for defenses to detect extraction attempts.
- **Label-only Querying:**
  A constrained setting where the attacker only has access to predicted labels (not probabilities). Despite limited feedback, recent studies show that effective model replication is still achievable [9], [10].
- **Transfer-based Querying:**
  Attackers generate queries from one domain or dataset and transfer the resulting surrogate knowledge to another domain. Recent work has demonstrated this in cross-domain GNN extraction scenarios [11].

C. *Reconstruction Target*

Different extraction attacks focus on reconstructing specific aspects of the target GNN, including its functionality, structure, or internal representations.

- **Functional Extraction:**
  The attacker replicates the mapping between inputs and outputs, producing a surrogate model that behaves similarly to the target model on unseen data [3], [5].
- **Structural Extraction:**
  The focus is on recovering topological or relational information in the graph. This can include node embedding's, adjacency patterns, or edge weights [12], [13].
- **Feature-space Extraction:**
  The attacker reconstructs the learned embedding space rather than the prediction function. This is particularly dangerous when embedding encoding sensitive attributes (e.g., user relationships or genetic data).
- **Parameter-level Extraction:**
  Some advanced techniques aim to recover numerical weight values or gradient information, especially in federated or distributed learning settings [14].

D. *Attack Goals and Motivations*

Understanding attacker intent is essential for contextualizing extraction risk. Typical objectives include:

- **Intellectual Property Theft:**
  The attacker aims to reproduce a proprietary GNN to avoid licensing or retraining costs [15].
- **Adversarial Preparation:**
  The extracted model can be used to craft adversarial examples or launch evasion attacks [16].
- **Data Privacy Breach:**

  Extraction may reveal hidden relationships or training data attributes from graph-structured datasets [4], [17].
- **Performance Exploitation:**

  Attackers may exploit extracted knowledge to fine-tune their own models for competitive advantage.
- **Federated Leakage:**
  In decentralized settings, adversaries may exploit gradients or parameter updates to infer other clients' data [14], [18].

E. *Defense Mechanisms*

Several defensive strategies have been proposed to counter model extraction, each with tradeoffs in utility, robustness, and computational cost.

- **Detection-based Defenses:**
  These monitor query patterns for signs of systematic probing. PRADA remains a foundationa l example, flagging suspicious query sequences [19].
- **Perturbation-based Defenses:**
  The system deliberately modifies output probabilities or introduces random noise to mislead attackers without harming overall model performance [20], [21].
- **Watermarking and Fingerprinting:**
  Watermarking embeds hidden functional signatures into GNN predictions, allowing owners to verify model ownership post-extraction [22].
- **Access Control and Rate Limiting:**
  Restricting query frequency and enforcing user authentication can significantly reduce extraction feasibility.
- **Differential Privacy (DP) Techniques:**
  Recent studies demonstrate that adding DP noise during training or inference can balance accuracy with protection against reconstruction [23].

F. *Emerging Dimensions*

Beyond these established categories, several new directions are emerging in GNN model extraction research:

- **Multi-modal Graphs:**
  Attacks are expanding to graphs that integrate text, images, or sensor data, requiring hybrid extraction methods [11].
- **Adaptive Defense Learning:**
  Defenses now employ machine learning to dynamically adapt to evolving attack strategies.
- **Benchmark Development:**
  The absence of standardized benchmarks limits reproducibility. Initiatives aim to create unified frameworks for evaluating both attacks and defenses [24].

III. Literature Review and Methodological Analysis

Model extraction research on Graph Neural Networks (GNNs) has evolved from foundational work in black-box attacks to advanced frameworks incorporating efficiency, privacy, and defense analysis. This section critically analyzes the progression of these studies and identifies emerging themes and methodological innovations.

Foundational Research on Model Extraction:

Early work demonstrated that deep learning models could be replicated through black-box APIs, laying the foundation for model extraction research [3]. Subsequent research expande d on this idea by showing that adversarial examples generate d on a surrogate model can successfully transfer to the target model [16]. These studies revealed the potential risks of exposing models through APIs, inspiring later work on more complex structures such as GNNs.

GNN-specific Extraction Attacks:

A comprehensive exploration of GNN model extraction, showing that black-box access alone was sufficient to approximate GNN performance with high fidelity [5]. They proposed a query-efficient imitation strategy, reducing query cost while maintaining extraction accuracy. Similarly, the Knockoff Nets framework, later adapted for GNNs, relies on continuous querying and surrogate training to replicate target behavior [25].

Advancements in Query Efficiency and Label Constraint s:

Subsequent work focused on query optimization by integrating active learning to minimize redundant queries [8]. They also explored label-only attacks, where attackers can extract models even when confidence scores are unavailable [9], [10]. These works demonstrated that GNNs remain vulnerable even under restricted API feedback, emphasizing the need for more robust defensive strategies.

Federated and Privacy-Aware Threats:

Federated learning has introduced new vulnerabilities in distributed GNN training environments. Studies revealed that gradients exchanged between clients can leak sensitive node or edge information [14]. Other research explored membership inference attacks, showing how adversaries can infer whether specific nodes were used in training [17], [26]. These findings underscore the importance of secure aggregation and privacy preserving mechanisms in collabora tive GNN learning.

Defensive Mechanisms and Countermeasures:

Several studies propose defensive approaches to protect GNNs from extraction. Watermarking techniques that allow model owners to verify intellectual property claims [22]. Other work showed that injecting controlled label noise can degrade extraction accuracy without significantly reducing model performance [20]. The PRADA framework, which monitors unusual query distributions to detect extraction attempts [19]. Despite these defenses, attackers continue to adapt, highlighting the cat-and-mouse nature of this research area.

Cross-Domain and Multi-Modal Model Extraction:

Chen et al. (2023) extended the scope of model extraction by introducing cross-domain transfer learning attacks on GNNs [11]. They demonstrated that knowledge extracted from one domain (e.g., citation graphs) can be effectively applied to another (e.g., product co-purchase networks). Their findings reveal that even when data distributions differ, attackers can still generalize extraction models, exposing a new security dimension.

Combined Poisoning and Extraction Attacks:

The synergy between data poisoning and model extraction attacks has also been explored [27]. By introducing subtle perturbations in the graph structure, they showed that poisoned nodes can increase the success rate of extraction. This combined strategy underscores the complexity of defense design since traditional anti-poisoning techniques may not prevent extraction-driven manipulations.

Defense Benchmarking and Comparative Analysis:

A recent systematic comparisons of defense mechanisms against GNN extraction. They evaluated query-limiting, watermarking, and noise injection defenses under multiple datasets [24]. The study reveals that most defenses fail to generalize across unseen attack scenarios. Their work highlights the need for unified evaluation frameworks and reproducible defense benchmarks.

Privacy-Preserving GNNs:

A differential privacy-based GNN introduces controlled noise during training and inference to balance model accuracy and data confidentiality [23]. Their experiments showed that adding calibrated noise can protect sensitive graph structures from extraction while maintaining acceptable performance. This study represents a promising direction toward formal privacy-preserving frameworks for GNNs.

Emerging Trends and Insights:

From the above works, several methodological insights emerge:

- Query-efficient attacks are becoming increasingly sophisticated.
- Federated and multi-modal setups present new risks.
- existing defenses often degrade model utility.
- Privacy-preserving approaches such as differential privacy shows potential but require further optimization. Collectively, these studies indicate that GNN model extraction is an evolving field requiring deeper exploration into adaptive, explainable, and benchmarked security frameworks.

IV. IMPLICATIONS OF THE REVIEW

This comprehensive review has several critical implications for researchers, practitioners, and policymakers involved in developing and deploying GNNs.

- **For Secure System Design:**
  The taxonomy presented highlights that no single defense is sufficient. A multi-layered security strategy is essential, combining query monitoring [19]. output perturbation [21]. and differential privacy [23] to create a robust defense-in-depth against extraction attempts.

- **For Risk Assessment in MLaaS:**

The demonstrated effectiveness of query-efficient [8] and label-only attacks [9] shows that even seemingly restrictive APIs with limited feedback are vulnerable. Organizations offering GNNs as a service must reassess their risk models and not assume that hiding confidence scores provides adequate protection.

- **For Intellectual Property Protection:**

The success of functional and structural extraction attacks [5], [12] poses a direct threat to the business models of companies that invest heavily in training proprietary GNNs. Watermarking [22] is a promising step, but its long-term robustness against adaptive attackers remains an open question, necessitating further research into resilient IP protection schemes.

- **For Privacy Regulations:**

The ability to extract node features, edges, or infer membership [4], [17] from a target GNN can lead to violations of data privacy regulations like GDPR or CCPA. This review underscores that model extraction is not just an IP threat but a significant data privacy breach vector, urging the adoption of privacy-preserving technologies.

- **For Federated Learning Systems:**

The vulnerability of gradient sharing in federated GNNs [14] indicates that collaborative learning setups require additional safeguards. Secure aggregation and differential privacy are not optional but necessary components for any privacy-respecting federated GNN deployment.

- **For Future Research Directions:**

The observed weaknesses in current defenses when faced with adaptive or combined attacks [24], [27] call for a paradigm shift from static to dynamic and learning-based defenses. Furthermore, the lack of standardized benchmarks hinders progress, making the development of unified evaluation frameworks a critical priority for the research community.

IV. COMPARATIVE SUMMARY OF PRIOR WORK

The table below summarizes key studies on model extraction attacks against GNNs, highlighting the problem addressed, research contributions, methods employed, and possible future work.

Study	Problem Addressed	Methodology	Dataset(s)	Key Findings	Future Work
Tramèr et al. (2016) [3]	Model stealing via APIs	Equation solving & path finding	Commercial ML APIs	First demonstration of practical model extraction	Defenses for model stealing
Papernot et al. (2017) [16]	Adversarial example transferability	Surrogate model training	MNIST, GTSRB	Extraction enables black-box adversarial attacks	Robustness against transfer attacks
Orekondy et al. (2019) [25]	Model stealing	Knockoff Nets framework	CIFAR-10, CIFAR-100	Pixel-level queries can steal model functionality	Defense-resilient APIs
Juuti et al. (2019) [19]	Detection of extraction attempts	PRADA: query distribution analysis	SVHN, CIFAR-10	Statistical detection of systematic probing	Adaptive attacks that evade detection
He et al. (2021) [5]	GNN black-box extraction	Query efficient imitation	Cora, Citeseer	GNNs easily replicable with limited queries	Adaptive anomaly detection
Dai et al. (2021) [14]	Federated privacy leakage	Gradient-based inference	FedGraph dataset	Federated GNNs leak sensitive data	Secure aggregation protocols
Wu et al. (2021) [4]	Stealing graph structure	Link stealing attacks	OGB, Cora	Graph structure can be inferred from predictions	Structure-hiding defenses
Chen et al. (2021) [20]	Defending with noisy labels	Controlled label noise injection	Cora, PubMed	Noise degrades extraction accuracy	Optimal noise-utility trade-off
Sun et al. (2021) [9]	Label-only attacks	Active learning queries	PubMed	Works without prediction probabilities	Multi-task robustness
Zhang et al. (2022) [10]	Query optimization	Adaptive querying	Reddit, OGB	High fidelity under query limits	Query-throttling defenses
Wang et al. (2022) [22]	IP protection	Watermarking for GNNs	GraphBench, Cora	Enables model ownership proof	Watermark-resistant attacks
Chen et al. (2023) [11]	Cross-domain extraction	Transfer learning on heterogeneous graphs	OGB, Amazon	Extraction succeeds across domains	Design domain-specific defenses
Li et al. (2023) [27]	Adversarial poisoning + extraction	Combined poisoning and query-based extraction	Cora, PubMed	Poisoned nodes improve extraction accuracy	Develop poisoning-resistant GNNs
Yu et al. (2024) [23]	Privacy-preserving extraction	Differential privacy GNNs	OGB-L1, Reddit	Balanced accuracy and privacy with DP noise	Optimize privacy-utility tradeoff
Zhao et al. (2024) [24]	Defense benchmarking	Comparative study of defenses	Synthetic & Real Graphs	Identified weak generalization in defenses	Create unified GNN defense benchmark

V. CONCLUSION

Model extraction attacks on Graph Neural Networks pose serious threats to model security and intellectual property. Although current defenses provide some protection, more research is needed to build secure, privacy-preserving, and attack-resistant GNNs. Future studies should focus on developing formal security proofs, smarter anomaly detection, and effective privacy-preserving mechanisms to safeguard GNN-based systems.

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, 2021.

[2] J. Zhou et al., "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57 81, 20 20.

[3] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Security Symp. (USENIX S ecurity 16)*, 2016, pp. 601–618.

[4] X. Wu, J. Li, Y. Wu, and C. Zhang, "Link stealing in graph neural networks," in *Proc. 37th ACM/SIGAPP Symp. Applied Computing (SAC)*, 2021, pp. 932–939.

[5] H. He, K. Ohta, and T. Ozawa, "Model extraction and adversarial transferability on graph neural networks," in *Proc. 20th IEEE Int. Conf. Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2021, pp. 1–8.

[6] J. Zhang, X. Liu, and Y. Wang, "A realistic and model-agnostic black-box extraction attack against graph neural networks," in *Proc. AAAI Conf. Artificial Intelligence*, 2024, pp. 1–9.

[7] J. Zhou, Z. Chen, and D. Zeng, "Transferable black-box attacks on graph neural networks," in *Proc. 31st Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2022, pp. 1–7.

[8] L. Sun, Y. Dou, C. Yang, and P. S. Yu, "Black-box model extraction attacks with limited queries on graph neural networks," in *Proc. 30th ACM Int. Conf. Information and Knowledge Management (CIKM)*, 2021, pp. 1–4.

[9] L. Sun, Y. Dou, C. Yang, and P. S. Yu, "Label-only model stealing attacks on graph neural networks," *arXiv preprint arXiv:2108.10731*, 2021.

[10] J. Zhang, Y. Wang, and X. Liu, "Query-efficient black-box attacks against graph neural networks," in *Proc. 38th IEEE Int. Conf. Data Engineering (ICDE)*, 2022, pp. 1–12.

[11] Z. Chen, M. Zhao, and L. Li, "Cross-domain model extraction attacks on graph neural networks," in *Proc. 32nd Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2023, pp. 1–8.

[12] Z. Wang, X. Ye, and Q. Wang, "Graph structural extraction via node classification APIs," in *Proc. 2022 ACM SIGSAC Conf. Computer and Communications Security (CCS)*, 2022, pp. 1–12.

[13] K. Zhang, Y. Wang, and H. Li, "Stealing graph neural networks with node inference attacks," *IEEE Trans. Information Forensics and Security*, vol. 18, pp. 1–14, 2023.

[14] H. Dai, X. Li, and Y. Zhang, "Federated learning on graph neural networks: Attacks and defenses," in *Proc. 20th IEEE Int. Conf. Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2021, pp. 1–8.

[15] Y. Wang and Z. Li, "A survey on intellectual property protection for deep neural networks," *IEEE Access*, vol. 9, pp. 1–15, 2021.

[16] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," in *Proc. 2017 ACM Asia Conf. Computer and Communications S ecurity (AsiaCCS)*, 2017, pp. 506–519.

[17] X. Wu, Y. Wu, and J. Li, "Membership inference attacks on graph neural networks," in *Proc. 2021 IEEE Int. Conf. Big Data (Big Data)*, 2021, pp. 1–9.

[18] L. Zhao, Q. Wang, and C. Zhang, "Gradient leakage attacks in federated graph neural networks," in *Proc. 2023 ACM Web Conf. (WWW)*, 2023, pp. 1–10.

[19] M. Juuti, S. Szyller, S. Samuel, and N. Asokan, "PRADA: Protecting against DNN model stealing attacks," in *Proc. 2019 IEEE European Symp. Security and Privacy (EuroS&P)*, 2019, pp. 1–16.

[20] Z. Chen, L. Li, and M. Zhao, "Model extraction attacks on graph neural networks with noisy labels," in *Proc. 2021 IEEE Int Conf. Big Data (Big Data)*, 2021, pp. 1–6.

[21] J. Lee, S. Kim, and H. Park, "Perturbation-based defenses against model extraction attacks on GNNs," *IEEE Access*, vol. 10, pp. 1–10, 2022.

[22] Y. Wang, J. Zhang, and X. Liu, "Robust model watermarking for graph neural networks," in *Proc. 2022 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1–5.

[23] K. Yu, L. Ma, and X. Zhou, "Privacy-preserving graph neural networks with differential privacy," *IEEE Trans. Dependable and Secure Computing*, vol. 21, no. 1, pp. 1–14, 2024.

[24] M. Zhao, Z. Chen, and L. Li, "Benchmarking defense mechanisms against model extraction attacks on GNNs," in *Proc. 38th AAAI Conf. Artificial Intelligence*, 2024, pp. 1–8.

[25] T. Orekondy, B. Schiele, and M. Fritz, "Knockoff nets: Stealing functionality of black-box models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1–10.

[26] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. 2017 IEEE Symp. Security and Privacy (SP)*, 2017, pp. 3–18.

[27] J. Li, H. Zhang, and Y. Wang, "Synergistic poisoning and extraction attacks against graph neural networks," in *Proc. 202 ACM Asia Conf. Computer and Communications Security (AsiaCCS)*, 2023, pp. 1–10.