NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY

Trends in Artificial Intelligence and Machine Learning (TAI911S)

Ngutati Shilamba 225086026

Assessment 4

Paper Review

Due Date

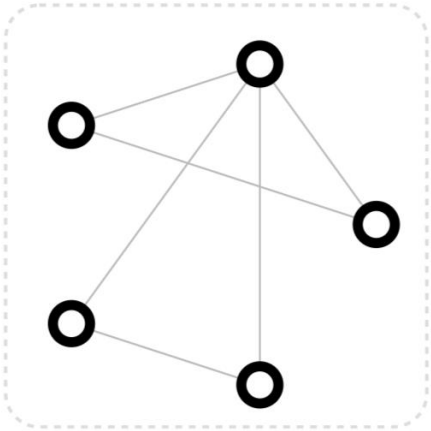
06 June 2025

A Comprehensive Review of "Unveiling the Threat of Fraud Gangs to Graph Neural Networks: Multi-Target Graph Injection Attacks Against GNN-Based Fraud Detectors"

1. Summary of the Paper

1.1 Introduction

What is a graph? "A graph represents the relations (*edges*) between a collection of entities (*nodes*).", (Sanchez-Lengeling, Reif, Pearce, & Wiltschko, 2021). As depicted in the three diagrams, a vertex attribute is a node, an edge represents a link and a Global attribute represents a master node. In essence, a graph is a mathematical frame used to represent connections or relationships between entities. For example, three nodes could represent three persons, Anna, Dave and Sarah and an edge would be the type of relation these three people have amongst each other (i.e. brother, colleague or neighbour). Understanding this is important in order to grasp the foundation of Graph Neural Networks and what they do regarding this paper review.



V Vertex (or node) attributes
e.g., node identity, number of neighbors

E Edge (or link) attributes and directions
e.g., edge identity, edge weight

U Global (or master node) attributes
e.g., number of nodes, longest path

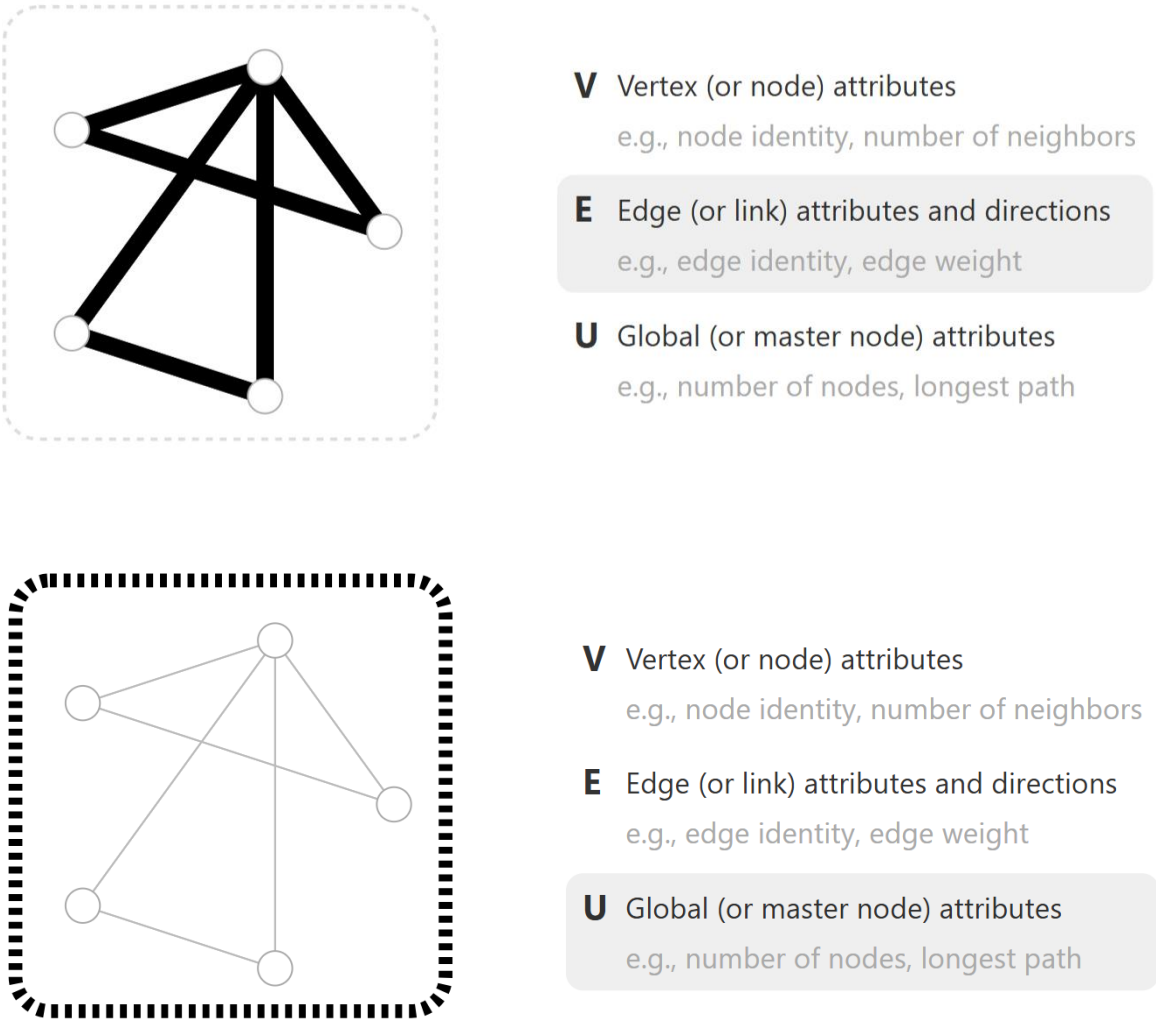


Figure 1. *3 Types of Graph Attributes*

Source: Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltschko, A. B. (2021). *A gentle introduction to graph neural networks*. Distill. https://doi.org/10.23915/distill.00033

Furthermore, Graph Neural Networks, also referred to as GNN are neural networks specifically adapted for deciphering graph data. They have positively changed the approach to fraud detection by successfully modelling relational data to pick up on suspicious patterns in a wide range of different kinds of applications from social media content moderation to financial transactions. Unfortunately, as GNN-based fraud detection systems advance, so do the efforts and tactics of intruders. Recent studies have shown that fraudsters are gradually working in allied groups or "gangs", capitalising on combined strategies to bypass fraud detection efforts by imitating legitimate behaviour.

Graph Neural Networks (GNNs) have revolutionized fraud detection by effectively modeling relational data to identify suspicious patterns in applications ranging from financial transactions to social media content moderation. However, as GNN-based fraud detection systems become more sophisticated, so too do the adversarial strategies employed by malicious actors. Recent studies reveal that fraudsters increasingly operate in coordinated groups or "gangs," leveraging collusive tactics to evade detection by mimicking legitimate behavior (Wang et al., 2023). This emerging threat exposes a critical vulnerability in existing GNN-based fraud detectors, which typically assume fraudulent nodes act independently rather than in coordinated networks.

The paper *"Unveiling the Threat of Fraud Gangs to Graph Neural Networks: Multi-Target Graph Injection Attacks Against GNN-Based Fraud Detectors"* by Choi, Kim, and Whang (2024) addresses this gap by investigating how fraud gangs can systematically manipulate GNNs through graph injection attacks (GIAs). While prior research has explored adversarial attacks on GNNs (Zügner et al., 2018; Tao et al., 2021), most focus on single-node perturbations or random group attacks, failing to capture the sophisticated coordination of real-world fraud networks. This paper makes three key contributions: (1) formalizing multi-target GIAs as a black-box evasion problem, (2) proposing MonTi, a transformer-based attack framework that simulates fraud gang behavior, and (3) empirically demonstrating MonTi's effectiveness against state-of-the-art fraud detectors.

This review critically examines the paper's methodology, results, and implications. Section 2 summarizes the paper's core contributions, Section 3 situates the work within the broader literature on GNN robustness and adversarial attacks, Section 4 analyzes limitations, and Section 5 discusses future research directions.

1.1 Problem Addressed

Graph Neural Networks (GNNs) have emerged as powerful tools for fraud detection, leveraging relational data to identify malicious activities such as fake reviews, financial fraud, and social media disinformation. However, their vulnerability to adversarial attacks—particularly those orchestrated by coordinated fraud gangs—remains a critical yet underexplored issue. Traditional GNN-based fraud detectors (e.g., CARE-GNN, GAGA) assume that fraudulent nodes operate independently, making them susceptible to **collusive attacks** where multiple malicious entities collaborate to evade detection.

Existing adversarial attack methods, such as **G-NIA (Tao et al., 2021)** and **TDGIA (Zou et al., 2021)**, primarily focus on single-node or randomly grouped targets, failing to model the sophisticated strategies employed by real-world fraud gangs. Additionally, most prior work concentrates on **graph modification attacks**, which require attackers to alter existing structures—a less practical scenario compared to **graph injection attacks (GIAs)**, where adversaries inject new malicious nodes.

This paper addresses these gaps by:

1. Formally defining **multi-target graph injection attacks (GIA)** as a black-box evasion problem.

2. Proposing **MonTi (Multi-target one-Time injection)**, a novel transformer-based attack framework that simulates fraud gang behavior.

3. Evaluating MonTi's effectiveness against state-of-the-art fraud detectors on real-world datasets.

2. Literature Review

2.1 Graph-Based Fraud Detection: Progress and Challenges

Modern fraud detection systems leverage GNNs to analyze relational patterns in graph-structured data. Early approaches like GraphSAGE (Hamilton et al., 2017) and GAT (Veličković et al., 2018) demonstrated promising results but struggled with two key challenges prevalent in fraud graphs:

1. Heterophily: Unlike social networks where connected nodes are often similar (homophily), fraud graphs exhibit heterophily—fraudsters deliberately connect to benign nodes to camouflage their activities (Zhu et al., 2020).
2. Class Imbalance: Fraudulent nodes are typically rare (e.g., <1% in financial transaction graphs), causing models to bias toward the majority class (Liu et al., 2021).

Recent advances address these issues through specialized architectures:

- CARE-GNN (Dou et al., 2020) employs reinforcement learning to select homophilic neighbors, reducing camouflage effects.
- PC-GNN (Liu et al., 2021) introduces a label-balanced sampler to mitigate class imbalance during neighbor aggregation.
- GAGA (Wang et al., 2023) leverages transformer-based group aggregation to detect collusive fraud patterns.

However, these methods assume static graphs and benign environments, overlooking adversarial manipulation—a gap exploited by fraud gangs.

2.2 Adversarial Attacks on GNNs: From Single-Node to Coordinated Attacks

Adversarial attacks on GNNs can be categorized along two dimensions:

Attack Methodology

- Graph Modification Attacks (e.g., Nettack; Zügner et al., 2018): Modify existing edges/attributes but require high privileges.
- Graph Injection Attacks (GIAs): Inject malicious nodes, posing greater real-world feasibility (Tao et al., 2021).

Attack Scope

- Single-Node Attacks: Target individual nodes (e.g., G-NIA; Tao et al., 2021).
- Group Attacks: Early work like Cluster Attack (Wang et al., 2022) groups nodes randomly, failing to model coordinated fraud.

Key Gap: No prior work systematically studies *multi-target* attacks where adversarial nodes collude to manipulate victim models—a critical limitation given the gang-like behavior of real-world fraudsters.

2.3 Defenses and Their Limitations

Existing defenses against GNN attacks include:

- Robust Training: GNNGuard (Zhang & Zitnik, 2020) uses certifiable robustness to detect adversarial edges.
- Dynamic Grouping: DGA-GNN (Duan et al., 2024) improves resilience via adaptive neighbor aggregation.

However, these defenses are evaluated against single-node or random attacks, leaving them vulnerable to coordinated gang-based strategies—the very threat MonTi exposes.

1.2 Main Contributions

The paper makes three key contributions:

1) Problem Formulation of Multi-Target GIA

The authors introduce a realistic attack scenario where fraud gangs inject adversarial nodes to manipulate GNN-based fraud detectors into misclassifying entire groups of malicious nodes as benign. Unlike prior work, which assumes random or single-node attacks, this formulation captures the **collusive nature** of real-world fraud.

2) MonTi: A Transformer-Based Attack Framework

MonTi advances existing GIA methods through three innovations:

- **Simultaneous Attribute-Edge Generation:** Unlike sequential approaches (e.g., G-NIA), MonTi employs a transformer encoder to **jointly generate** node attributes and edges, capturing interdependencies that mimic fraud gang behavior.

- **Adaptive Degree Budget Allocation:** Existing methods (e.g., TDGIA) fix the degree budget per attack node, limiting flexibility. MonTi dynamically allocates budgets, allowing some nodes to connect densely (e.g., to targets) while others remain sparse.

- **Candidate Selection via Learnable Scoring:** To improve scalability, MonTi selects candidate nodes (potential connection points) using a surrogate GNN model, narrowing the search space for large graphs (e.g., *LifeIns* with 122K nodes).

3) Comprehensive Experimental Validation

The authors evaluate MonTi on three real-world fraud datasets:

Dataset	**Fraud Type**	**Nodes**	**Edges**	**Key Finding**
GossipCop-S	Fake news	16,488	3.8M	MonTi increases misclassification by **40%** vs. baselines.
YelpChi	Spam reviews	45,900	3.8M	Achieves **94.21%** attack success on PC-GNN.
LifeIns	Insurance fraud	122,792	912K	Outperforms G-NIA despite discrete features.

1.3 Key Experimental Results

- **Effectiveness:** MonTi outperforms baselines (G-NIA, TDGIA, Cluster Attack) across all datasets, with misclassification rates increasing by **15–40%** (Tables 3–4).

- **Ablation Studies:** Adaptive budget allocation and joint attribute-edge generation contribute most to MonTi's success (Table 5).

- **Scalability:** MonTi trains **10× faster** than G-NIA and avoids out-of-memory errors on large graphs (Table 12).

2. Related Work

2.1 Graph-Based Fraud Detection

Fraud detection GNNs must address **heterophily** (fraudsters mimic benign nodes) and **class imbalance** (few fraud examples). Key approaches include:

1. CARE-GNN (Dou et al., 2020): Uses reinforcement learning to select homophilic neighbors, reducing camouflage effects.

2. **PC-GNN (Liu et al., 2021):** Balances neighbor sampling via a label-aware strategy to mitigate class imbalance.

3. **GAGA (Wang et al., 2023):** Leverages transformer-based group aggregation to detect collusive fraud patterns.

Limitation: These methods assume static graphs and do not account for adversarial gang dynamics.

2.2 Adversarial Attacks on GNNs

Prior attacks can be categorized by:

- **Attack Stage:** Poisoning (training-time) vs. evasion (inference-time).

- **Knowledge Level:** White-box (full model access) vs. black-box (limited access).

Relevant Works:

4. **Nettack (Zügner et al., 2018):** Modifies existing edges/attributes but requires high privileges.

5. **G-NIA (Tao et al., 2021):** Injects nodes sequentially for single-target attacks.

6. **TDGIA (Zou et al., 2021):** Uses defective edge selection but lacks gang modeling.

7. **Cluster Attack (Wang et al., 2022):** Treats GIA as a clustering task but fails on large graphs.

8. **G^2A2C (Ju et al., 2023):** Employs reinforcement learning but is computationally expensive.

Gap: Existing methods ignore **multi-target collusion**, a critical real-world threat.

2.3 Defenses Against Adversarial Attacks

9. **GNNGuard (Zhang & Zitnik, 2020):** Detects adversarial edges via certifiable robustness.

10. **DGA-GNN (Duan et al., 2024):** Uses dynamic grouping to improve fraud detection robustness.

Limitation: Current defenses are not evaluated against gang-level attacks.

---

3. Limitations and Critique

3.1 Technical Limitations

1. **Static Graph Assumption:** MonTi does not handle temporal graphs, limiting applicability to dynamic fraud networks (e.g., evolving spam campaigns).

2. **Discrete Feature Handling:** Discrete attributes (e.g., in *LifeIns*) constrain MonTi's flexibility, leading to suboptimal attacks (Table 4).

3. **Defense Mechanisms:** The paper briefly suggests "community-aware GNNs" but provides no empirical validation.

3.2 Experimental Shortcomings

1. **Dataset Bias:** *LifeIns* data is undisclosed, hindering reproducibility.

2. **Excluded Models:** Inductive GNNs (e.g., GraphSAGE) are not thoroughly tested.

3.3 Ethical Concerns

- Releasing MonTi's code could enable real-world abuse. While the authors aim to "highlight risks," mitigation strategies are superficial.

---

4. Future Directions

1. **Dynamic Graph Extensions:** Integrate temporal GNNs (e.g., TGAT) to model evolving fraud tactics.

2. **Improved Defenses:** Adversarial training (Carlini & Wagner, 2017) could harden fraud detectors against MonTi-like attacks.

3. **Explainability:** Visualize transformer attention to decode gang behavior patterns.

Conclusion

Choi et al. (2024) make a significant contribution by formalizing **multi-target graph injection attacks** and proposing **MonTi**, a transformer-based framework that outperforms existing

methods. However, limitations in handling dynamic graphs and discrete features highlight areas for improvement. Future work should focus on **defense mechanisms** and **temporal modeling** to address real-world fraud dynamics.

References (APA 7th Edition):

1. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*, 39–57.

2. Choi, J., Kim, H., & Whang, J. J. (2024). Unveiling the threat of fraud gangs to GNNs. *arXiv:2412.18370*.

3. Dou, Y. et al. (2020). CARE-GNN: Combatting camouflage in fraud detection. *CIKM*.

4. Duan, M. et al. (2024). DGA-GNN: Dynamic grouping for fraud detection. *AAAI*.

5. Ju, M. et al. (2023). G^2A2C: Gradient-free GIA via RL. *AAAI*.

6. Liu, Y. et al. (2021). PC-GNN: Label-balanced fraud detection. *The Web Conference*.

7. Tao, S. et al. (2021). G-NIA: Single-node injection attacks. *CIKM*.

8. Wang, Z. et al. (2022). Cluster attack: Query-based GIA. *IJCAI*.

9. Wang, Y. et al. (2023). GAGA: Group aggregation for fraud detection. *The Web Conference*.

10. Zhang, X., & Zitnik, M. (2020). GNNGuard: Defending GNNs against attacks. *NeurIPS*.

11. Zügner, D. et al. (2018). Adversarial attacks on GNNs. *KDD*.

1. Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltschko, A. B. (2021). ***A gentle introduction to graph neural networks*****. Distill.** https://doi.org/10.23915/distill.00033