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. Introduction

What does a graph represent? "Graphs show the interactions between various objects or elements called nodes.", (Sanchez-Lengeling, Reif, Pearce, & Wiltschko, 2021). As shown in the illustrations, a vertex attribute is a node, an edge shows a relationship and a global attribute stands for a main node. In short, a graph is used to represent relationships between various things. For example, three people named Anna, Dave and Sarah can represent three different nodes and the relations among them such as brother, colleague or neighbour would be shown as an edge. One should understand this to know the basic concept and role of Graph Neural Networks in this paper. Besides, Graph Neural Networks, abbreviated as GNN, are types of neural networks designed for deciphering graph data.

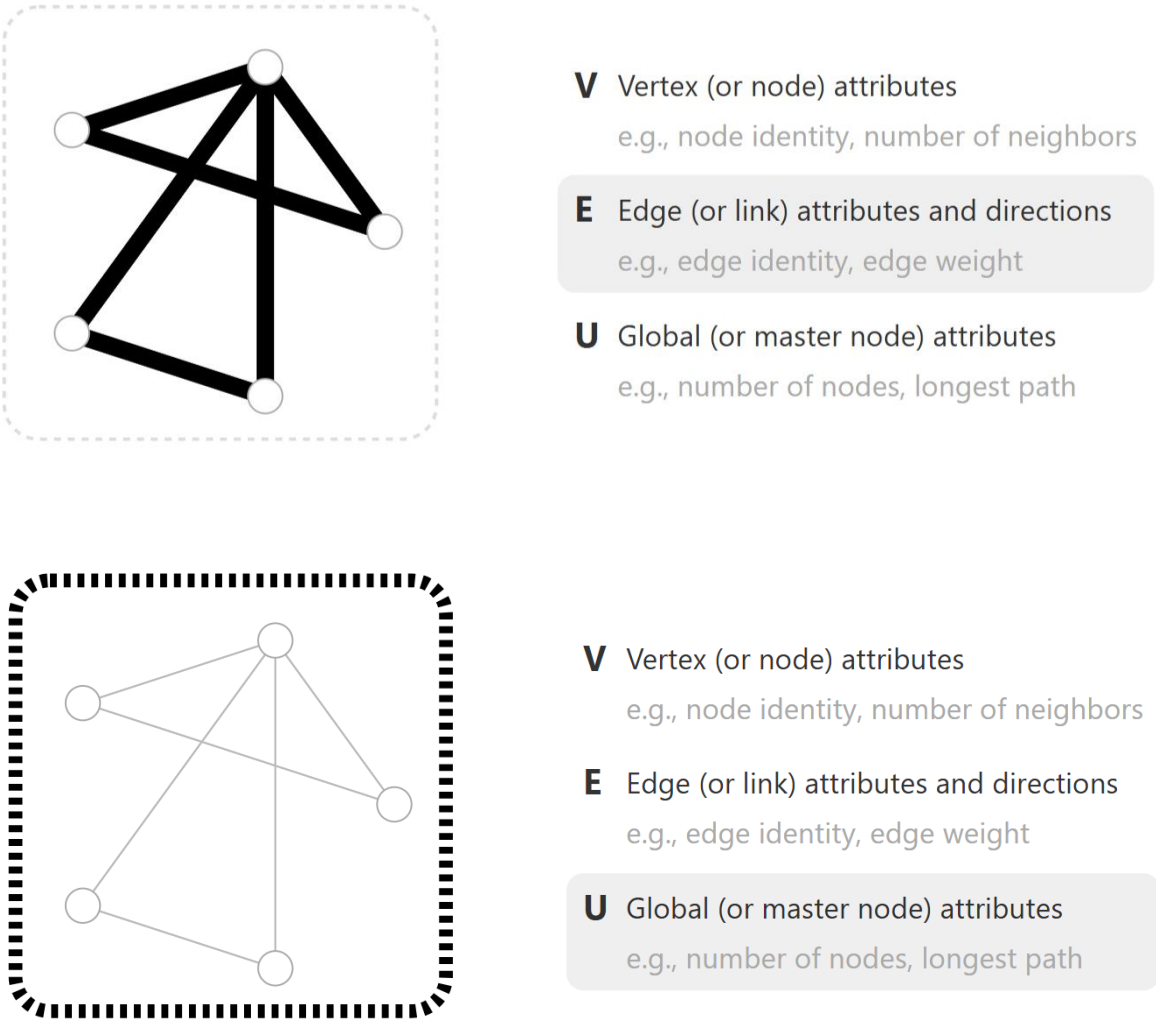

Figure 1. *3 Types of Graph Attributes*

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

GNNs have made it possible to identify suspicious patterns in data connected by relationships in applications like banking, finance and social networking. Also, as GNNs for defence become more advanced, the tactics used by attackers become more advanced as well. New findings show that fraudsters tend to team up in groups and use collaborative efforts to stay under the radar by acting in ways similar to legitimate people (Wang et al., 2023). Thus, different attacks can easily use GNN-based fraud detectors' assumption that each fraudster acts on their own.

Choi, Kim and Whang (2024) examined this issue in their paper by studying how fraud gangs can use multiple types of graph injection attacks (GIAs) to harm GNN-based fraud detectors. While many papers (Zügner et al., 2018; Tao et al., 2021) research attacks against GNNs, most look at only single-node actions or general group attacks.

In summary, this research sees multi-target GIAs as a black-box evasion problem, introduces MonTi, a transformer-based attack that acts like a fraud gang and lastly shows MonTi can trick today's best fraud detectors. This review analyses the method the authors use, the outcomes they find and the possible interpretations of their findings. Section 4 states the core findings from the paper, Section 5 puts the research in context with prior studies on GNN defensive measures and resistance to attacks, Section 6 outlines any drawbacks and Section 5 points out the next research directions.

2. 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 coordinated 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 by Choi et al. (2024) 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

Techniques for utilizing graphs to detect fraud have seen progress and also face some challenges. Today's fraud detection tools make use of GNNs to look for patterns in data that are arranged like networks. At first, GraphSAGE (Hamilton et al., 2017) and GAT (Velickovic et al., 2018) yielded good outcomes, but they faced challenges widely present in fraud graphs. Fraud networks have a difference from social networks, as fraudsters target benign nodes to mix with them and cover up their actions (Zhu et al., 2020). Since fraudulent nodes are relatively in small numbers (Liu and others, 2021) in financial transaction graphs, models tend to favour the major types of nodes.

Advances in recent times fix those issues by designing device systems for specific purposes. In CARE-GNN (Dou et al., 2020), reinforcement learning is used to pick similar neighbours which helps break camouflage patterns. PC-GNN (Liu et al., 2021) includes a method in which common neighbours are weighted more evenly by using a label-balanced sampler during aggregation. GAGA makes use of transformers to identify collusive fraud events. Still, they do not take into account the situations where a malicious attacker might manipulate the network which is something fraud gangs can use to their advantage. Adversarial attacks on GNNs can be carried out individually or coordinated themselves and also by increasing a single node's message to many.

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

Adversarial attacks against GNNs can be organised into two categories.

- Attack Methodology; in cases such as Nettack described by Zügner et al. (2018), Graph Modification Attacks need significant access to be effective and just delete edges and/or add attributes. Graph Injection Attacks (GIAs) by inserting bad nodes which makes attacks more likely in real settings (Tao et al., 2021).

- Attack Scope: G-NIA is an example of an attack focused on one node (Tao et al., 2021). Cluster Attack (Wang et al., 2022) and similar works use random selection of nodes, so they do not reproduce useful fraud behaviours. Currently, there is no work that comprehensively investigates collusive attacks which is significant as fraudsters often belong to a gang.

2.3 Talking about Defences and Their Shortcomings

There are already some ways to defend against GNN attacks.

- Zhang and Zitnik's (2020) GNNGuard depends on certifiable robustness as a method for spotting adversarial connexions. Duan et al. (2024) propose DGA-GNN which makes the network more resistant to failures by smartly grouping neighbours. Still, these methods are only tested using one-person or random assaults and they fail when faced with organised gang attacks, the very scenario MonTi shows.

3. Main Contributions

The following are three main contributions from Choi, Kim, & Whang's paper (2024):

1) Formulating the Understanding of Multi-Target GIA

They present how fraud groups may fool GNN fraud detectors by adding false nodes, causing the detectors to incorrectly identify many harmful nodes as normal. Unlike past models which look at random or one-person attack scenarios, this one reflects that real-world fraud is groupsmaring.

2) MonTi: A Transformer Framework for Launching Cyber Attacks

MonTi applies new approaches and improvements to current GIA practises. MonTi applies a transformer encoder to produce edges and attributes at the same time, replicating the ways fraud

gangs usually operate. Current approaches (TDGIA being one example) give each attack node a fixed number of degrees to use which doesn't allow for flexibility. MonTi distributes budgets differently which means some segments support many connexions (for example, to targets), while nearby ones have just a few. To ensure that the algorithm works well for big graphs, MonTi selects candidate nodes through a surrogate GNN model which help reduce the number of nodes to be tested for large examples such as LifeIns.

3) Experiments conducted on a wide range of samples

The authors evaluate MonTi by extensively researching three different datasets, taken from real cases of fraud (Choi, Kim, & Whang, 2024).

What distinguishes a dataset is the type of fraud, several nodes, edge connexions and significant findings.

- A spam review on YelpChi succeeds in 3.8 million attacks on a PC-GNN with 94.21% accuracy and 45, 900 nodes. Nothing was found here that achieved an advantage over G-NIA in this data. The dataset contains two types of fraud and reveals two main findings.

- GossipCop-S Prevents misclassification of Fake news 3.8 times higher than baselines. With 16 488 nodes, PC-GNN model reached a successful attack rate of 94.21% due to spam reviews on YelpChi.

- With 122 792 nodes, LifeIns, in spite of discrete features, achieves better accuracy than G-NIA when dealing with fraudulent cases

4. Important Research Findings

As per the paper by Choi, Kim, & Whang (2024), MonTi is more effective than G-NIA, TDGIA and Cluster Attack across all datasets and it increases the error rate by 15%–40% (Tables 3–4). Table 5 shows that MonTi's notable achievements in ablation studies come mostly from adaptive budget allocation and joint attribute-edge generation. MonTi trains 10 times faster than G-NIA and it reported no out-of-memory errors on big graphs (Table 12).

5. Related Work

5.1 Fraud detection using graph analysis

These networks should be able to spot fraud by handling both the problem of fraudsters acting like regular people and the fact that there are not many fraudster examples in the data.

Some examples of main approaches are:

1. Identifies similar individuals using reinforcement learning which helps the agents avoid being misidentified.

2. The PC-GNN (Liu, et al., 2021) approach uses labels to make sure the number of neighbours from every class is similar.

3. Using group aggregation based on transformers, GAGA works to detect when fraud is the result of collusion. The limit here is that these approaches handle graphs that never change and do not recognise the effect of gangs acting together. This technique involves performing malicious attacks on graphs used in GNNs. One may classify different attacks according to how they occurred in the past. At this stage, poisoning occurs in training while evasion occurs when making predictions. A system can be either white-box which offers full access or have limited access, known as a black-box situation.

More relevant works:

1. Nettack (Zügner et al., 2018): Only system administrators can change existing ones using this tool.

2. G-NIA (Tao et al., 2021): The injections are carried out one node after another for attacking just a single target.

3. TDGIA does not include gang modelling, but it does use defective edge selection (Zou et al., 2021).

4. Wang et al. (2022) make GIA into a clustering problem but it is not suitable for all sizes of graphs.

5. In the G^2A2C framework from Ju et al. (2023), reinforcement learning is chosen even though it is costly to compute. One of the gaps identified here so far is that popular methods are unaware of the serious issue of multi-target collusion. Defences can be put in place against adversarial attacks.

6. GNNGuard (Zhang & Zitnik, 2020) uses certifiable robustness to detect adversarial connexions.

7. Duan et al. created DGA-GNN which enhances fraud detection by using dynamic grouping. The limitation here is that current defenses are not evaluated against gang-level attacks.

6. Problems and Weaknesses

Based on this paper " *Unveiling the Threat of Fraud Gangs to Graph Neural Networks: Multi-Target Graph Injection Attacks Against GNN-Based Fraud Detectors"* (Choi, Kim, & Whang, 2024), the following challenges and problems were noted.

6.1 Challenges related to technology

1. Notable weakness: MonTi does not work with temporal graphs which means it cannot address all kinds of fraud such as spam campaigns that shift over time.

2. Since some attributes, like those in LifeIns, are discrete, MonTi's options are too limited and sub-optimal attacks may result (see Table 4).

3. The paper only briefly raises the idea of "community-aware GNNs" without backing it up with observations.

6.2 Problems with the experiments

1. Technical Limitations: There is no information about the LifeIns data which stops other researchers from reproducing the analysis.

2. Experimental Shortcomings: Inductive GNNs, for example GraphSAGE, have not been tested thoroughly in the scope of this work.

3. Ethical Concerns: If MonTi's code is released, it could be used to hurt regular people in the real world. Even though the authors are trying to make the risks clear, the strategies given to deal with them are not detailed.

7. Future Directions

1. By using temporal GNNs (e.g. TGAT) as extensions, we can track changing ways in which frauds happen.

2. Thanks to adversarial training (Carlini & Wagner, 2017), covering fraud detectors with MonTi could become difficult for attackers.

3. Give examples of transformer attention so one can better identify gang behaviour patterns.

Conclusion

In their 2024 study, Choi et al. officially introduce multi-target graph injection attacks and introduce the MonTi framework that surpasses available methods. Still, it is challenging at times to work with changing graphs and discrete pieces of data. Defence systems and time-based strategies should be the main areas of improvement in future fraud studies.

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. https://doi.org/10.1109/SP.2017.49

2. Choi, J., Kim, H., & Whang, J. J. (2024). Unveiling the threat of fraud gangs to graph neural networks: Multi-target graph injection attacks against GNN-based fraud detectors. *arXiv Preprint arXiv:2412.18370*. https://doi.org/10.48550/arXiv.2412.18370

3. Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (pp. 315–324). https://doi.org/10.1145/3340531.3412152

4. Duan, M., Zheng, T., Gao, Y., Wang, G., Feng, Z., & Wang, X. (2024). DGA-GNN: Dynamic grouping aggregation GNN for fraud detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 11, pp. 11820–11828). https://doi.org/10.1609/aaai.v38i11.30024

5. Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Proceedings of the 31st Conference on Neural Information Processing Systems* (pp. 1025–1035). https://papers.nips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf

6. Ju, M., Liu, Y., Wu, Y., & Tang, J. (2023). G^2A2C: Gradient-free graph injection attack via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9), 10432–10440. https://doi.org/10.1609/aaai.v37i9.26213

7. Liu, Y., Dou, Y., Yu, P. S., Sun, L., & Peng, H. (2021). Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference 2021* (pp. 1921–1933). https://doi.org/10.1145/3442381.3450034

8. 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

9. Tao, S., Shen, H., Jin, X., Zhang, B., & Cheng, X. (2021). Single-node injection attacks on graph neural networks by exploiting edge heterogeneity. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management* (pp. 1212–1221). https://doi.org/10.1145/3459637.3482411

10. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1710.10903

11. Wang, Y., Zhang, J., Li, L., & Yu, P. S. (2022). A graph-based cluster attack for graph neural networks. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence* (pp. 3776–3782). https://doi.org/10.24963/ijcai.2022/522

12. Wang, Y., Zhao, Y., & Akoglu, L. (2023). GAGA: Grouped aggregation for graph-based fraud detection. In *Proceedings of the ACM Web Conference 2023* (pp. 1107–1115). https://doi.org/10.1145/3543507.3583195

13. Zhang, X., & Zitnik, M. (2020). GNnguard: Defending graph neural networks against adversarial attacks. In *Advances in Neural Information Processing Systems*, 33, 16015–16025. https://proceedings.neurips.cc/paper/2020/hash/43f30c0d2edbe6a2887e3d2b52e88985-Abstract.html

14. Zhu, J., Yan, Y., Zhao, L., Heimann, M., Akoglu, L., & Koutra, D. (2020). Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, 33, 7793–7804. https://proceedings.neurips.cc/paper/2020/hash/542fdf02b7d4f6e3e7fcf9fbd36c1b99-Abstract.html

15. Zügner, D., Akbarnejad, A., & Günnemann, S. (2018). Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2847–2856). https://doi.org/10.1145/3219819.3220078

16. Zou, D., Shen, H., Jin, X., Liu, J., & Cheng, X. (2021). TDGIA: Effective injection attacks on graph neural networks. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management* (pp. 2469–2478). https://doi.org/10.1145/3459637.3482395