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

Jinhyeok Choi, **Heehyeon Kim**, and **Joyce Jiyoung Whang***

* Corresponding Author

School of Computing, KAIST

The 39th AAAI Conference on Artificial Intelligence (AAAI 2025)

▼ BDILab ▼ GitHub

KAIST BDILab BIG DATA INTELLIGENCE

Main Contributions

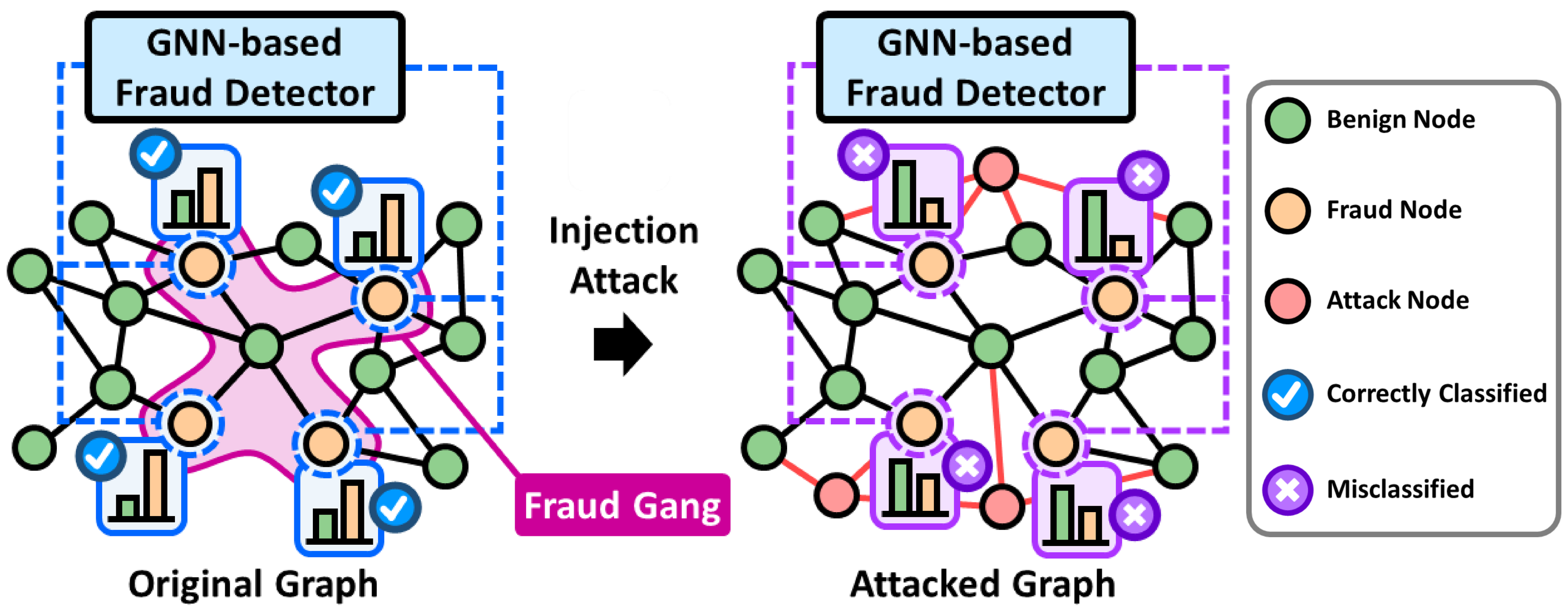
- Investigate **adversarial attacks on GNN-based fraud detectors** by **fraud gangs**
  - First study on graph injection attacks for **multiple target nodes** organized by **groups** based on **metadata or relations** in real-world graphs.
- Propose **M**ulti-target **on**e-**Ti**me graph injection attack model (**MonTi**)
  - Allocate **adaptive degree budgets** and inject all attack nodes **at once**.
  - Capture **interdependencies** between **node attributes and edges**.
- MonTi outperforms state-of-the-art graph injection attack methods in **both multi- and single target settings** on real-world graphs.

GNN-Based Fraud Detection and Fraud Gangs

- **Fraud Detection with GNNs**
  - **Interactions of fraudsters** can be effectively modeled with **graphs**.
    - **Nodes** represent **distinct entities** such as news, reviews, and claims.
    - **Edges** represent **relationships between entities**.
- **Fraud Gangs with Collusive Patterns**
  - Frauds are increasingly **organized into gangs or groups** to carry out fraudulent activities **more effectively with reduced risk**.
  - e.g., Fraudsters can spread misinformation by using **multiple fake accounts**.

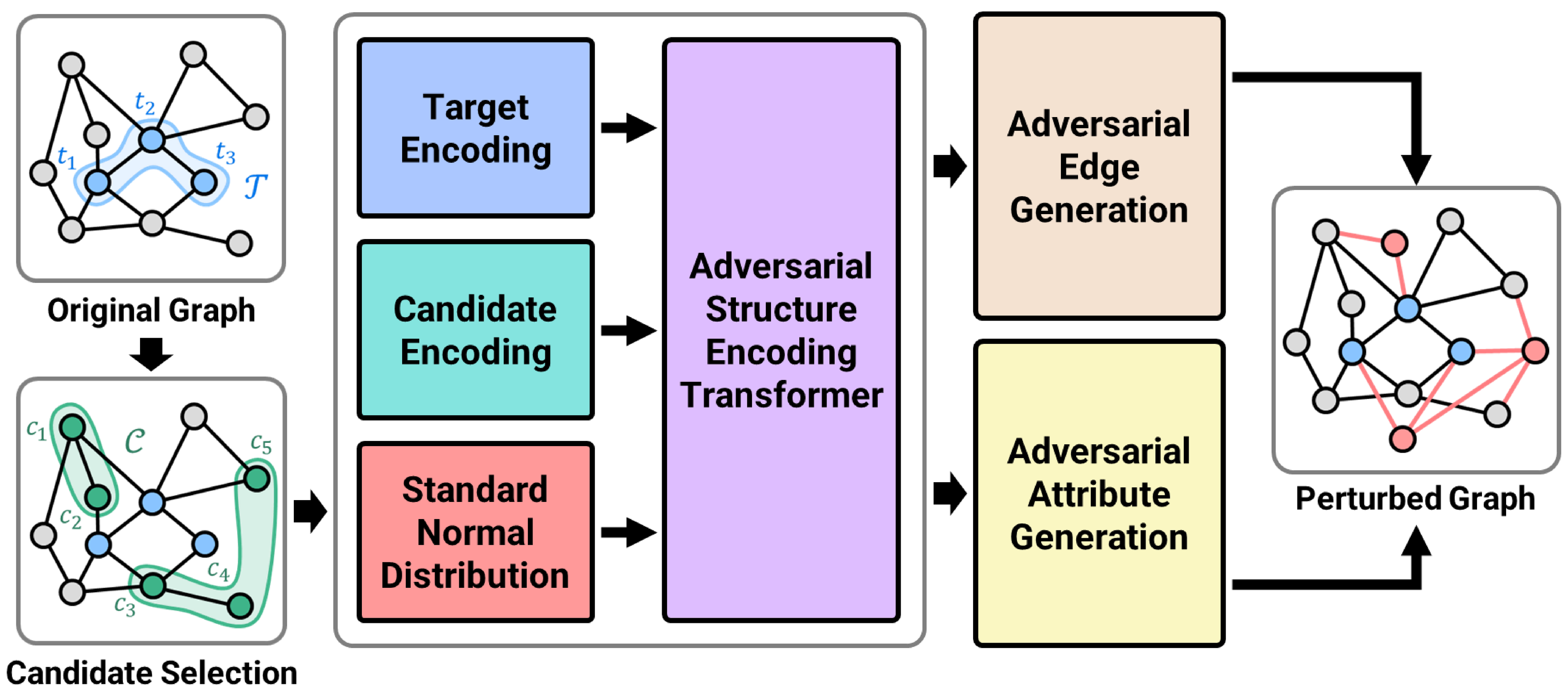
Attack Scenarios: Multi-Target Graph Injection Attack

- **Adversarial Attacks against GNN-Based Fraud Detectors**
  - Design the attack scenarios where **fraud gangs attack GNN-based fraud detectors** to make them **misclassify the fraud nodes as benign**.



- **Black-Box Graph Injection Evasion Attack**
  - **A feasible approach** that does not require access to modify existing structures.
  - The attacker can access only **the original graph**, **partial labels**, and **a surrogate model**, and the attack occurs during **the inference phase**.
- **Limitations of Existing Graph Injection Attack Methods**
  - Inject attack nodes sequentially, **fixing the graph structure at each step**.
  - **Sequentially** generate **attributes and edges** of attack nodes.

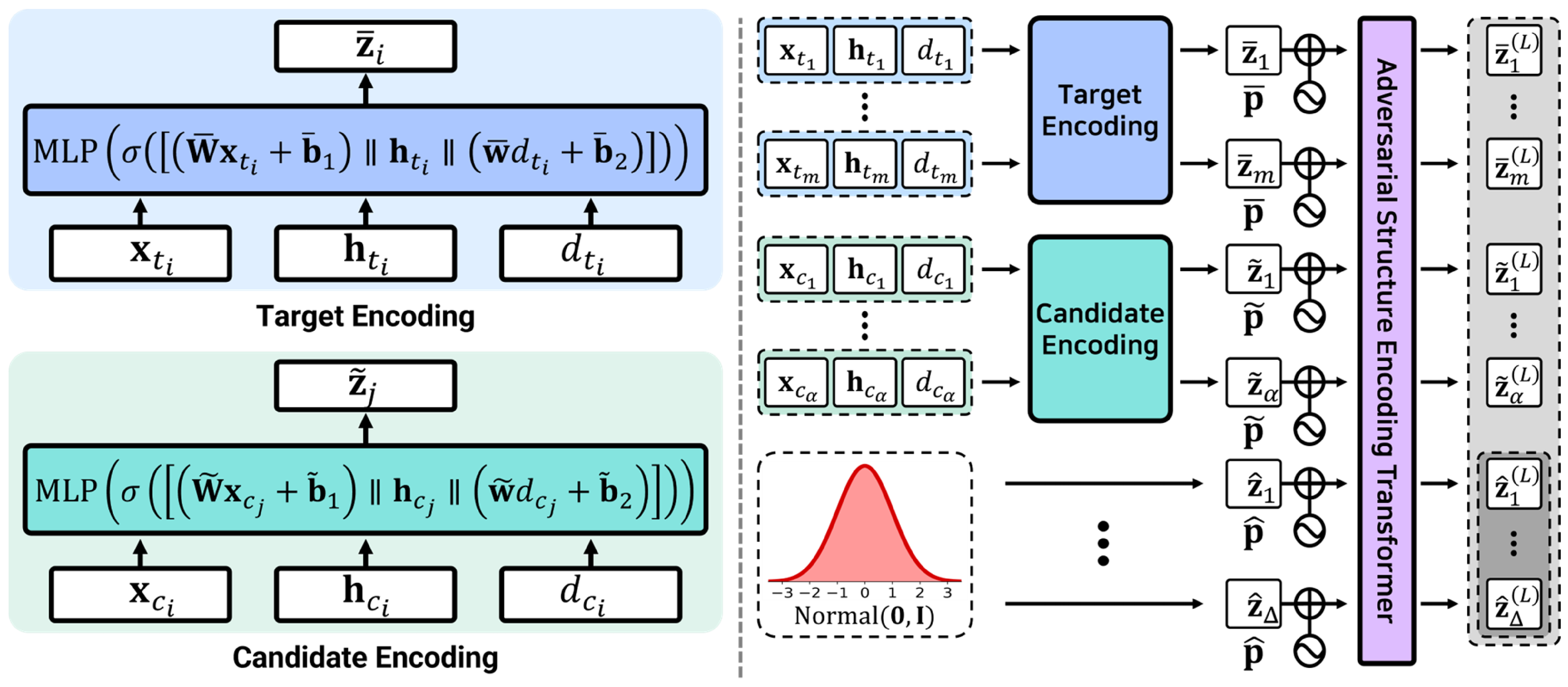
MonTi: Multi-Target One-Time Graph Injection Attack Model



- **Candidate Selection with Learnable Scoring Function**
  - Select candidate nodes to **narrow the search space** with **scoring function**.
- **Adversarial Structure Encoding to handle Interdependencies**
  - Capture **interdependencies** among **all nodes** involved in the attack.
- **One-Time Graph Injection with Intermediate Representations**
  - Generate **attributes and edges of attack nodes** at once.

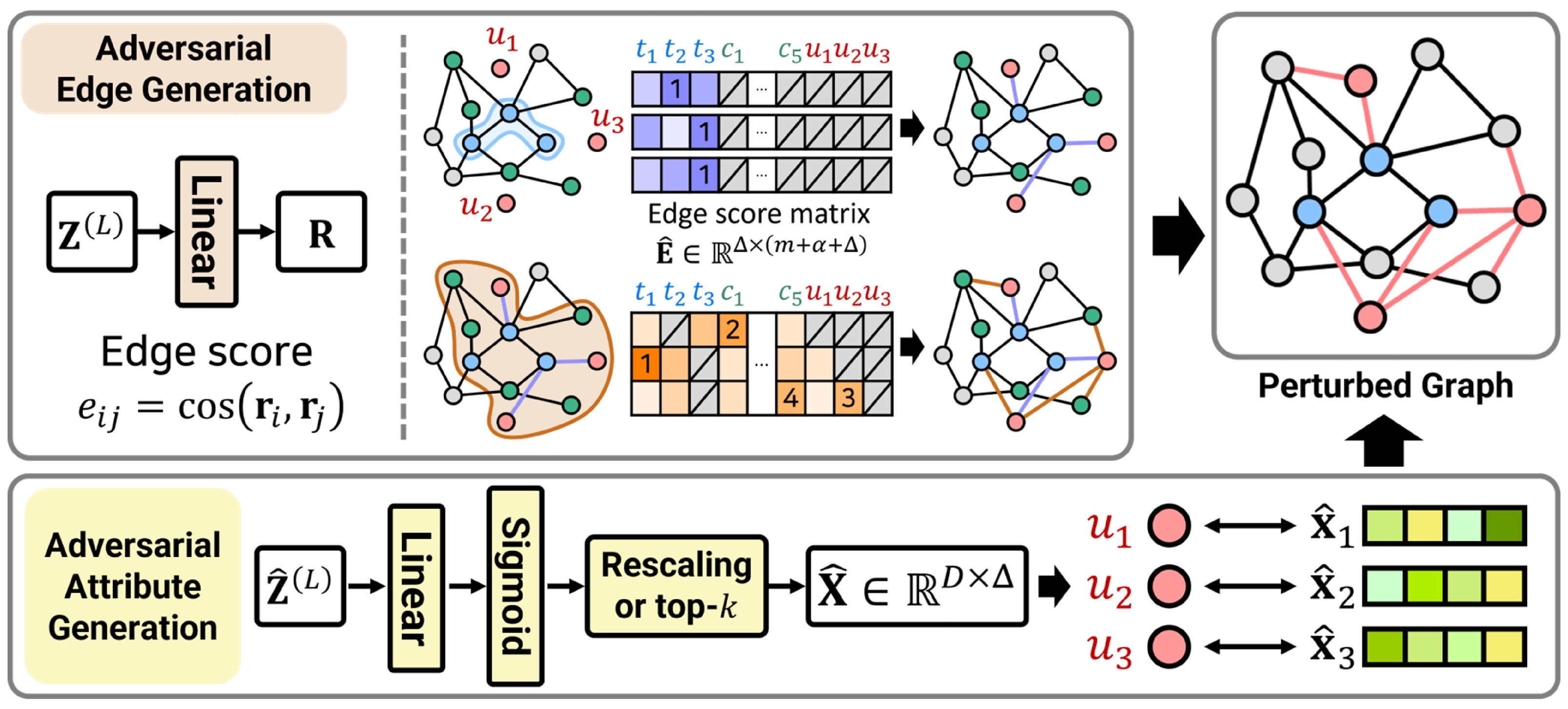
Adversarial Structure Encoding

- **Adversarial Structure Encoding Transformer**
  - Compute the intermediate representations for attribute and edge generation using **raw attributes**, **representations**, and **degree information** as input.



One-Time Graph Injection

- **Adversarial Attribute and Edge Generation**
  - Generate **edges** by projecting the representation into **the edge score space**.
  - Generate **attributes** with **rescaling** or **top-k selection** based on attribute type.



Experiments

- **Surrogate/Victim Models:** GCN, GraphSAGE, GAT, CARE-GNN, PC-GNN, GAGA
- **Attack Baselines:** G-NIA(CIKM'21), TDGIA(KDD'21), Cluster Attack(IJCAI'22), G²A2C(AAAI'23)
- **Evaluation Metric:** Average misclassification rates (%) of target sets
- **Multi-Target Attack Performance on Real-World Fraud Graphs**
  - **When GCN is the Surrogate Model**

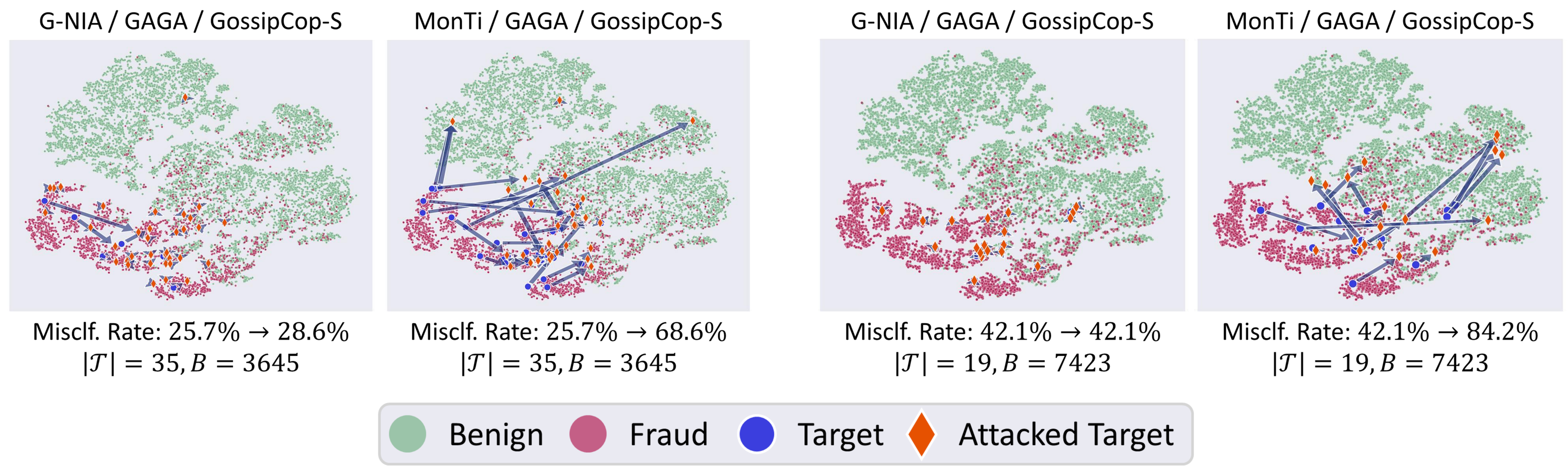
		CARE-GNN	PC-GNN	GAGA
GossipCop-S	Clean	48.02	55.62	21.68
	Best-baseline	60.67	66.25	25.76
	MonTi	**88.40**	**89.36**	**41.34**
YelpChi	Clean	29.79	59.13	28.00
	Best-baseline	34.81	63.57	28.83
	MonTi	**55.59**	**97.21**	**29.63**
LifeIns	Clean	16.42	16.17	15.68
	Best-baseline	18.34	**20.08**	23.38
	MonTi	**18.63**	19.78	**27.25**

  - **Where the Types of Surrogate and Victim Models are the Same**

		GCN	GraphSAGE	GAT	CARE-GNN	PC-GNN	GAGA
GossipCop-S	Clean	46.70	26.04	11.29	48.02	55.62	21.68
	Best-baseline	75.12	67.70	63.21	59.96	62.60	25.69
	MonTi	**92.60**	**97.05**	**94.30**	**90.15**	**90.12**	**46.94**
YelpChi	Clean	87.14	43.81	35.15	29.79	59.13	28.00
	Best-baseline	90.93	64.56	55.51	**32.45**	63.18	31.08
	MonTi	**92.23**	**65.31**	**93.27**	31.92	**69.93**	**37.66**
LifeIns	Clean	27.72	13.70	16.75	16.42	16.17	15.68
	Best-baseline	83.28	37.80	96.60	18.05	17.90	16.87
	MonTi	**99.47**	**60.97**	**100.00**	**26.80**	**20.64**	**35.03**

Qualitative Analysis

- **Effects of the Size of Fraud Gangs**
  - MonTi significantly **shifts the representations** from the fraud to the benign area.



Misclf. Rate: 25.7% → 28.6%, $|\mathcal{T}| = 35, B = 3645$ | Misclf. Rate: 25.7% → 68.6%, $|\mathcal{T}| = 35, B = 3645$ | Misclf. Rate: 42.1% → 42.1%, $|\mathcal{T}| = 19, B = 7423$ | Misclf. Rate: 42.1% → 84.2%, $|\mathcal{T}| = 19, B = 7423$