Graph Mining Course Project Proposal

Submission Date: December 12, 2025
Course: Graph Mining 4041
Instructor: Dr. Zeinab Maleki

- **Student Name(s):** Mojtaba Mollaei, Rasoul Salehi, Seyed Ali Amiri
- **Student ID(s):** 40131383, 4012593, 40118389
- **Email(s):** m.mollaei@ec.ac.ir, r.salehi@ec.ac.ir,

Comparative Re-Implementation of Two GNN-Based Anomaly Detectors on the Bitcoin Blockchain

Abstract

This project reproduces and evaluates two recent anomaly-detection models designed for blockchain transaction graphs:

1. **MDST-GNN** (Chen et al., 2025), a multi-distance spatial-temporal GNN that models how transaction patterns evolve across time steps, and
2. **SGAT-BC** (Chang et al., 2025), a subtree-attention GAT combined with a bagging ensemble to address severe class imbalance.

Using only the **Elliptic dataset** (Elliptic, 2025) a temporal Bitcoin transaction graph with licit, illicit, and many unknown nodes, we re-implement both models from scratch, replicate their preprocessing pipelines, and run controlled experiments. The goal is to compare detection performance, robustness to imbalance, and behavior under varying temporal and structural conditions. Expected outcomes include reproducible code, clear head-to-head model comparisons, and practical insight into which architecture is better suited to different constraints in blockchain anomaly detection.

Problem and Motivation

Blockchain transaction graphs evolve across discrete time steps and contain complex structural patterns that make them well-suited for graph mining methods. The Elliptic dataset exemplifies these challenges: although it provides rich node-level features for more than 200K Bitcoin transactions, only a small fraction of nodes are labeled illicit, resulting in extreme label imbalance. Since each transaction is a node and the task is to infer its class from features and connectivity, the problem reduces to node-level classification where the quality of learned node embeddings is crucial. Prior studies, such as Asiri and Somasundaram (2025) and Chang et al. (2025), show that structural indicators including degree-based centrality and multi-hop neighborhood context, often correlate with fraudulent behavior, underscoring the need for models that can integrate both feature and topological signals.

At the same time, illicit activity in blockchain networks often unfolds through sequences of related transactions across time, creating dependencies that purely static GNNs fail to capture. Temporal sparsity and evolving structural patterns require models that combine spatial reasoning with temporal modeling, as in MDST-GNN (Chen et al., 2025). Meanwhile, the dominance of licit and unknown nodes introduces substantial imbalance problems that can degrade generalization unless mitigated by sampling or ensemble techniques like those used in SGAT-BC. Understanding how these two modeling philosophies (spatial-temporal representation learning versus imbalance-focused structural attention) behave under a unified evaluation on Elliptic is therefore both practically important for blockchain forensics and intellectually valuable for graph-based anomaly detection research.

Objectives

- **Objective 1:** Implement the MDST-GNN architecture exactly as described in the 2025 paper.
- **Objective 2:** Implement the SGAT-BC model (subtree attention + GAT + bagging + CatBoost).
- **Objective 3:** Compare and analyze both models on Elliptic to determine which performs better under different structural and temporal conditions.

Related Work

Chen et al. (2025) introduce MDST-GNN, which integrates multi-distance spatial aggregation with temporal modeling to capture both local and global dependencies in evolving transaction graphs. Chang et al. (2025) propose SGAT-BC, combining subtree attention, graph attention, and a bagging ensemble to combat severe label imbalance especially relevant for Elliptic. Our project directly re-implements both models and evaluates them on the same dataset, enabling a clean, head-to-head comparison of temporal modeling vs. imbalance-focused design.

Proposed Methodology

Dataset(s)

We use only the Elliptic Dataset, a Bitcoin transaction graph with 203,769 nodes across 49 time steps. Nodes are labeled as licit, illicit, or unknown. Following the preprocessing approaches from SGAT-BC and MDST-GNN, we keep only the labeled nodes for supervised training and evaluation, while unknown nodes are excluded from the loss. Node features follow the paper settings: SGAT-BC uses the full 166-dimensional feature vector, while MDST-GNN uses the 42-dimensional subset. All numeric features are standardized.

Because the dataset is temporal, we apply the standard chronological split used in the MDST-GNN paper (70% train, 15% validation, 15% test), ensuring that the models are evaluated on future time steps. This preprocessing keeps the setup consistent across both implementations and maintains the real-world temporal structure of illicit activity.

Techniques and Algorithms

SGAT-BC

SGAT-BC follows a three-stage pipeline structured inside a **Bagging ensemble**.

1. **Training-set sampling:** From the labeled Elliptic data, the method draws *k* bootstrap subsets to counter the severe class imbalance. Each subset is used to train an independent base model.
2. **Base-model training:** Each subset is fed into the **SGAT model**, which combines multi-hop **subtree attention** with a GAT-based message-passing layer to capture local structural patterns and emphasize important subtrees around each transaction node.
3. **Meta-model fusion:** The predictions of the *k* base SGAT models are concatenated into a *k*-dimensional representation. This vector becomes the input to a **CatBoost classifier**, which serves as the meta-model and produces the final node-level classification.

This design allows the model to stabilize predictions under extreme imbalance and noise by exploiting multi-view structural attention and ensemble diversity.

MDST-GNN

MDST-GNN is a spatial–temporal architecture tailored for blockchain anomaly detection. It integrates four key components:

1. **Multi-Distance Graph Convolutions (MD-GCN):** Captures spatial dependencies across multiple hop distances, enabling the model to incorporate both immediate and long-range transaction relationships.
2. **Spatial–Temporal Feature Extractor:** Models the temporal evolution across the 49 Elliptic time steps using temporal convolutions or attention, allowing the network to detect evolving illicit behavior patterns.
3. **Adaptive Fusion Mechanism:** Dynamically merges spatial and temporal features, giving the model flexibility to prioritize whichever signal (structure or time) is more informative at each step.
4. **Self-Supervised Component:** Adds auxiliary objectives that improve the quality of learned embeddings and help generalization under limited labeled illicit data.

Overall, MDST-GNN addresses the core challenges of blockchain anomaly detection i.e. temporal sparsity, multi-hop structural dependencies, and label scarcity by combining multi-distance message passing with adaptive temporal modeling and representation learning.

Tools and Frameworks:

- PyTorch Geometric for GNN implementations.
- Scikit-learn for evaluation metrics.
- Pandas and NumPy for data preprocessing.

Evaluation Plan

- **Metrics:**
  Macro F1, Macro Recall (primary due to imbalance), AUC, and G-Mean (as used by SGAT-BC).
- **Baselines:**
  - GCN, GAT, GraphSAGE for reference.

Challenges and Resources

- Main challenge: **Elliptic is extremely imbalanced** i.e. illicit ≈ 2%, licit ≈ 21%, unknown ≈ 77%. Unknowns cannot be used as standard labels but still influence graph structure. This hurts both training stability and evaluation reliability. Mitigation: careful masking, class-weighted losses, and ensemble methods.
- Secondary challenge: incomplete hyperparameter details in the papers; we will match them as closely as possible and document assumptions.

References

1. Chang, Z.; Cai, Y.; Liu, X.F.; Xie, Z.; Liu, Y.; Zhan, Q. Anomalous Node Detection in Blockchain Networks Based on Graph Neural Networks. Sensors 2025, 25, 1. https://doi.org/10.3390/s25010001
2. Chen, S., Liu, Y., Zhang, Q., Shao, Z. and Wang, Z. (2025), Multi-Distance Spatial-Temporal Graph Neural Network for Anomaly Detection in Blockchain Transactions. Adv. Intell. Syst., 7: 2400898. https://doi.org/10.1002/aisy.202400898
3. Elliptic Dataset: https://www.elliptic.co/dataset/

4. Asiri, A., Somasundaram, K. Graph convolution network for fraud detection in bitcoin transactions. Sci Rep 15, 11076 (2025). https://doi.org/10.1038/s41598-025-95672-w