



A Novel Graph-Based Framework for Cryptocurrency Fraud Detection

Rongyu Yang*

Qingdao WeiMing school, Qingdao City, 266555, China

*Corresponding author's e-mail: 358829199@qq.com

Abstract. Cryptocurrencies have been widely applied in industries such as finance and physical trade, with Bitcoin and Ethereum becoming the mainstream cryptocurrencies. However, cryptocurrency transactions face numerous security issues, such as money laundering, Ponzi schemes, and high-investment-plan scams. As a product of blockchain, cryptocurrency transactions possess the characteristics of anonymity and immutability. While this anonymity protects the privacy of the parties involved in a transaction, it significantly increases the difficulty for security agencies and government institutions to monitor and regulate these transactions. Although existing methods for fraud detection achieve high accuracy, they often lack interpretability and fail to help security teams identify the entities linked to fraudulent transactions or offer insights into similar fraudulent patterns. To address these issues and improve the interpretability of fraud detection, we propose an innovative graph-based framework. By gathering multi-dimensional data, we can provide a more detailed and holistic view of each transaction record. We develop a heterogeneous graph to model transaction entities, their related transaction records, and transaction flows. Using graph fusion and reasoning techniques, this model aids in analyzing market and entity behaviors and supports heuristic exploration by experts. Finally, a pre-trained Graph Neural Network (GNN) is utilized to quickly pinpoint fraudulent entities and their associated transactions within the graph.

Keywords: Cryptocurrency, Blockchain, Data Mining, Fraud Detection

1 Introduction

Since the introduction of Bitcoin in 2009 [1], digital currencies have gained significant attention, with over 9,500 cryptocurrencies and a market capitalization exceeding \$2.33 trillion as of April 2024 [2]. Blockchain's decentralized nature, while offering transparency and immutability, has also facilitated illegal activities such as money laundering and fraud. For instance, Ether fraud profits surged from \$17 million in 2017 to \$36 million in 2018 [3], and over 1,800 Ponzi schemes were reported on bitcointalk.org between 2011-2016 [4]. To combat this, machine learning methods like Decision Trees, Random Forests, and deep learning techniques such as CNN and LSTM [5]–[15] have been employed for fraud detection. However, these methods

lack interpretability, hindering the identification of entities behind fraudulent activities. This paper proposes a novel graph-based fraud detection framework, utilizing a heterogeneous graph to represent entities, transactions, and flows. By applying graph fusion and reasoning, combined with a pre-trained GNN model, we aim to enhance fraud detection interpretability and swiftly identify fraudulent entities [16]–[20].

2 Background

In this section, we first summarize recent researches on digital currency fraud detection from the past 1 to 4 years, covering detection methods based on traditional machine learning, deep learning, and graph-based representations. Following this, the motivations and goals of the new framework are proposed, addressing the limitations of existing graph-based fraud detection methods.

2.1 Related Work

Machine Learning-Based. To address the limitations of fixed detection rules and lack of flexibility, which are unable to cope with complex and dynamically changing fraud patterns, NOOR et al. [5] proposed a machine learning-based approach. They employed the ADASYN-TL data balancing technique to solve the problem of imbalanced positive and negative samples in the training data. Three hyperparameter optimization methods were then used, followed by stacking ensemble learning and SHapley Additive exPlanation (SHAP) methods to detect and interpret Bitcoin fraud records. Similarly, Sharma et al. [8] used three machine learning models—AdaBoost, Random Forest (RF), and XGBoost—for real-time fraud detection. For Ethereum, Neogi et al. [6] used Decision Trees, Random Forest, CatBoost, and XGBoost to detect fraud. Rabia et al. proposed a method based on Light Gradient Boosting Machine (LGBM) for accurately detecting fraudulent activities in Ethereum smart contract transactions. Balakrishnan et al. [9] introduced a machine learning approach based on Isolation Forest, which is well-suited to handle class imbalance or sparse instances in datasets, enabling high-precision fraud transaction detection.

Deep Learning-Based. Hu et al. [11] proposed a fraud detection method based on the BERT model, which can handle highly repetitive, skewed distributions, and heterogeneous Ethereum transactions. Zhao et al. [12] introduced a Long Short-Term Memory (A-LSTM) method for detecting whether transaction records are part of Ponzi schemes or other online scams, enabling fine-grained multi-class classification of fraud types. Umer et al. [13] proposed an Ensemble Deep Learning-Based approach that combines CNN and LSTM for fraud detection in cryptocurrency transactions. Hu et al. [14] also designed a GRU network with an attention mechanism to detect fraudulent smart contracts. Krishnan et al. [15] employed a unique fraud detection method based on social media information, using autoencoders to learn dynamic strategies related to cryptocurrency price fluctuations, allowing for simultaneous prediction of cryptocurrency prices and potential scams.

Graph-Based. Ding et al. [16] constructed a directed multi-graph with attribution edges to represent cryptocurrency accounts and related transaction information. They designed a method combining message passing and attention mechanisms to identify fraudulent account nodes in the graph. Milner et al. [17] designed a unique knowledge representation approach that integrates relationships between advertisement addresses, transactions, and smart contracts on the graph. Fraudulent smart contract transactions are detected through similarity-based empirical knowledge. Kang et al. [18] proposed a discrete prototype graph convolutional autoencoder method to detect phishing fraud on transaction graphs, enhancing the ability to identify subtle patterns that may be concealed through separated representation learning. Singh et al. [19] introduced a GNN-based approach that utilizes adversarial learning to capture fraud patterns with long-term temporal dependencies, improving the model's long-term stability in fraud detection. Jia et al. [20] aimed to leverage the potential synergistic effects of semantic information and similarity patterns in fraud detection by constructing a transaction attribute similarity graph and an account interaction graph. These graphs capture transaction similarities and network structural information, respectively, and fraud detection on Ethereum is performed using a joint language model.

2.2 Motivations and Goals

Motivations. Traditional machine learning methods suffer from several drawbacks: (1) low accuracy. (2) the methods are not general and the domain is too restrictive: only some known fraudulent activities can be detected. (3) poor interpretability. (4) the algorithms rely on centralized datasets to train the models, and the native machine needs to collect and store all the data. This increases the computational load while greatly increasing the risk of privacy leakage of transaction characterization data [21]. Unruly elements may utilize traceability techniques to obtain the real identity of the user through the stolen data [22]–[25]. (5) Cannot better mine anomalies based on local features, global features and attribute features of the network structure. (6) To construct the whole transaction network, a large amount of external network information needs to be used and the model is complex. All of these place extremely high demands on computer hardware, computing power and human resources. In addition, the accuracy is unsatisfactory and the model scalability is low. For existing unsupervised learning methods, they generally use clustering methods K-means clustering has the ability to group instances together, but lacks the prowess of detecting outliers. While LOF is popular for outlier detection, it does not scale well in large datasets with computational time.

However, most of the existing graph-based fraud detection methods are aimed at detecting the transaction records of digital currencies, and few studies have been conducted to construct user address profiles based on transaction addresses, or analyze the pattern similarity of different transaction records to infer the correlation between different address identities. This is due to the anonymity of digital currencies, with many addresses themselves often not directly linked to real identity information, mak-

ing it difficult to determine who a particular user is. In addition, address reuse and sharing and obfuscation techniques increase the difficulty of identity inference.

Goals. Based on the motivations mentioned above, we propose the following three goals.

- More clearly reflect the transaction behavior presented by different addresses and the transaction flow between different addresses.

Proposed Method: Construct a address-centralized knowl-edge graph of digital currency transaction record to portray the trading behavior of different user entities.

- Determine whether different transaction addresses are owned by the same entity user and whether different fraudulent transaction records point to the same entity user.

Proposed Method: Analyze the correlation between address and transaction records, and extract the transaction patterns of different addresses.

- High precision, good interpretability fraud detection method.

Proposed Method: Design a graph embedding technique, which can extract the local and global features of the graph structure, and try various graph models to accurately detect digital currency fraud records.

3 Framework Details

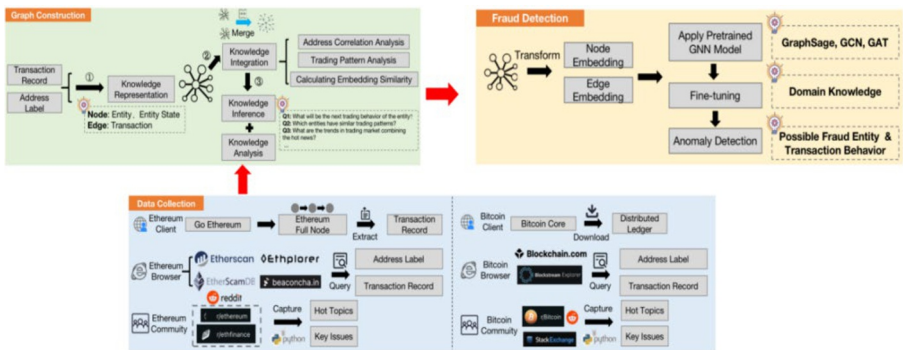


Fig. 1. The graph-based framework for cryptocurrency transaction analysis and fraud detection.

As shown in Figure 1, the methodology framework consists of three modules. In the first module, diverse transaction data, address labels, and key issues are collected from various sources. The second module constructs a user-centralized graph model to represent transaction behaviors. Based on this model, typical characteristics of user transactions and overall features of the digital currency market are analyzed. Finally, in the third module, deep learning models are applied to the constructed knowledge graph to detect fraudulent transactions.

3.1 Data Collection

Compared with previous research, user-centered model building requires richer information about entities. However, the anonymization of user information by digital currencies greatly increases the difficulty of building user profiles. In order to solve this problem, we plan to collect multi-dimensional information of bitcoin and Ethereum from the Internet, including the transaction records of these digital currencies and the address labels, address IDs or address tokens of the user entities, which include the time of the transaction, the entity associated with the transaction, the amount of the transaction, the transaction contract, and other important information. Specifically, we will take different means to collect receipts for each of the two digital currencies, Bitcoin and Ethereum.

Bitcoin: Bitcoin's transaction data is collected from multiple channels: Bitcoin clients and browsers. The Bitcoin client connects directly to the network, storing the full Bitcoin blockchain. We plan to obtain transaction data from Bitcoin Core, which maintains the entire transaction history. Additionally, the Bitcoin Browser (Blockchain.com, Blockstream Explorer) allows querying transaction data, addresses, and labels.

Ethereum: Ethereum's blockchain offers extensive data, accessed through the Go Ethereum client and various browsers like Etherscan and Ethplorer. These tools provide transaction records and address mappings. We also plan to leverage datasets like EtherScamDB and Kaggle's top open-source datasets to enrich the data.

Data Cleaning and Organization: After gathering the data, we will take the following three steps to filter and process each record: (1) Merge data from different sources, expanding feature dimensions. (2) Clean data through cross-validation and prioritization of authoritative sources. (3) Organize data by timestamps and aggregate it by time blocks (e.g., hourly, daily).

Real-time news and trending topics in the cryptocurrency space will help analyze market trends. User identities can be difficult to infer due to the anonymity of digital currencies, so we plan to collect public records from social media or compliance programs (e.g., Bitcoin addresses disclosed on forums). We will gather news from top Bitcoin and Ethereum communities (r/Bitcoin, r/Ethereum, r/ethfinance) and extract key topics and data metrics.

3.2 Graph Construction

Unlike previous studies, which mainly focus on detecting fraud or money laundering, our goal is to construct a user-centric model that infers correlations between different transactions, identifying similar behavioral patterns that suggest the same user. Previous graph-based studies focused on fraud detection but did not effectively capture the relationships between transactions made by the same user. We aim to enhance these models by adding inference and analysis operations for more valuable insights. Our graph construction involves three steps: knowledge representation, knowledge fusion, and knowledge inference and analysis.

Knowledge Representation: We represent transaction records and address labels in a heterogeneous graph. User identities are nodes labeled “User” and transactions are represented as nodes labeled “Transaction”. Each transaction contains attributes such as the transaction address, amount, currency type (BTC/ETH), and time. Relational edges (“Flow”) connect transactions based on transaction flows.

Knowledge Fusion: This step merges transactions with high similarities into a single user node. We aim to merge transactions based on address correlation, transaction patterns, and embedding vector similarity: (1) Address Correlation: We analyze the co-occurrence of addresses in transactions. If two transactions involve similar addresses, they are considered similar. This is measured using a common neighbor algorithm within a neighborhood of 10 or fewer transaction nodes. (2) Transaction Pattern Analysis: We examine the input/output patterns (e.g., number of addresses, transaction amounts) and the transaction frequency/timestamp patterns to identify similar transactions. (3) Embedding Vector Similarity: We apply Graph Convolutional Networks (GCN) or Graph Neural Networks (GNN) to compute node embeddings and measure similarity between transactions based on graph representation learning. Two similarity thresholds are used: one for merging subgraphs and another for determining the similarity between transaction flows.

Knowledge Inference and Analysis: After constructing the graph, we add temporal features to capture trends. We use Long Short-Term Memory (LSTM) networks to predict future market trends based on the temporal characteristics of transaction data. Additionally, knowledge analysis identifies subtle similarities between transactions not merged by the second threshold, indicating potential links between entities (e.g., belonging to the same institution or sharing social attributes).

3.3 Fraud Detection

Fraudulent transactions are a significant issue in digital currencies. By constructing transaction graphs and applying deep learning models, we can detect hidden patterns and anomalous behaviors. Specifically, we design the following three steps for detecting fraud transaction.

Graph Conversion to Trainable Representation: (1) Node and Edge Encoding: Encode nodes (e.g., addresses, transactions) and edges (e.g., transaction relationships) into vector forms using embedding techniques. (2) Adjacency Matrix Construction: Build an adjacency matrix to represent node connectivity and graph structure. (3) Feature Engineering: Add extra features like node degree and transaction statistics to help the model better understand the graph's content.

Graph Model Application: (1) Graph Model Structure: Select an appropriate deep learning model (e.g., GCN, GraphSage, GAT) to process the transaction graph and learn embeddings. (2) Learning Embeddings: Feed the graph data into the model to learn node and edge embeddings that reflect the relationships between transaction.

Model Fine-Tuning: (1) Data Augmentation: Generate more fraudulent transaction samples to improve the model's ability to detect fraud patterns. (2) Anomaly Detection Objective: Train the model using labeled fraudulent and normal transactions

to fine-tune it for fraud detection. (3) Parameter Tuning: Optimize model hyperparameters (e.g., learning rate, regularization) to improve performance.

Fraud Detection: (1) Threshold Setting: After training, set a threshold to classify transaction records as normal or fraudulent based on their embeddings. (2) Post-Processing: Use techniques like clustering and time series analysis to further refine fraud detection accuracy.

4 Conclusion

Due to the anonymity of cryptocurrency transactions, it has become much more difficult for regulatory agencies to track fraudulent activities. To assist security personnel in analyzing the identities of entities involved in anonymous transactions, detecting transaction behaviors and patterns associated with these entities, and identifying fraudulent transactions, this paper proposes a novel graph-based framework for cryptocurrency transaction analysis and fraud detection. Multi-dimensional information collection can provide a more comprehensive view of entity profiling and the motivations behind transaction behaviors. By representing this information as a knowledge graph, graph fusion and graph reasoning techniques can be applied to analyze transaction patterns and market trends. Finally, the pre-trained Graph Neural Network (GNN) is fine-tuned to better adapt to the fraud detection task in cryptocurrency transactions.

References

1. Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.
2. "coinmarketcap," accessed: Dec. 19, 2024. [online], 2024. <https://coinmarketcap.com/>.
3. Mauro Conti, E Sandeep Kumar, Chhagan Lal, and Sushmita Ruj. A survey on security and privacy issues of bitcoin. *IEEE communications surveys & tutorials*, 20(4):3416–3452, 2018.
4. Marie Vasek and Tyler Moore. Analyzing the bitcoin ponzi scheme ecosystem. In *Financial Cryptography and Data Security: FC 2018 International Workshops, BITCOIN, VOTING, and WTSC, Nieuwpoort, Curacao, March 2, 2018, Revised Selected Papers 22*, pages 101–112. Springer, 2019.
5. Noor Nayyer, Nadeem Javaid, Mariam Akbar, Abdulaziz Aldegeishem, Nabil Alrajeh, and Mohsin Jamil. A new framework for fraud detection in bitcoin transactions through ensemble stacking model in smart cities. *IEEE Access*, 2023.
6. Anandarupa Neogi, Disha Mukhopadhyay, Anubhav Jaiswal, Ankush Kumar, and Bitan Misra. Fraud detection in ethereum transactions: A machine learning approach. In *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, pages 1–7. IEEE, 2024.
7. Rabia Musheer Aziz, Mohammed Farhan Baluch, Sarthak Patel, and Abdul Hamid Ganie. Lgbm: a machine learning approach for ethereum fraud detection. *International Journal of Information Technology*, 14(7):3321–3331, 2022.
8. Anshika Sharma and Himanshi Babbar. Machine learning-driven detection and prevention of cryptocurrency fraud. In *2023 International Conference on Research Meth-*

- odologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), pages 1–5. IEEE, 2023.
9. D Balakrishnan, Umasree Mariappan, Sharan Sagar Seerla, Pyrapu Varun Tej, Yerroju Eswar Vani, and Shaik Fazi. Isolation forest based fraud detection for cryptocurrency transaction. In 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS), pages 1–6. IEEE, 2023.
 10. Mohammad Javad Rajaei and Qusay H Mahmoud. A survey on pump and dump detection in the cryptocurrency market using machine learning. *Future Internet*, 15(8):267, 2023.
 11. Sihao Hu, Zhen Zhang, Bingqiao Luo, Shengliang Lu, Bingsheng He, and Ling Liu. Bert4eth: A pre-trained transformer for ethereum fraud detection. In *Proceedings of the ACM Web Conference 2023*, pages 2189–2197, 2023.
 12. Puyang Zhao, Wei Tian, Lefu Xiao, Xinhui Liu, and Jingjin Wu. An attention-based long short-term memory framework for detection of bitcoin scams. In 2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS), pages 21–26. IEEE, 2022.
 13. Qasim Umer, Jian-Wei Li, Muhammad Rehan Ashraf, Rab Nawaz Bashir, and Hamid Ghous. Ensemble deep learning based prediction of fraudulent cryptocurrency transactions. *IEEE Access*, 2023.
 14. Huiwen Hu, Qianlan Bai, and Yuedong Xu. Scsguard: Deep scam detection for ethereum smart contracts. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE, 2022.
 15. Lakshmi P Krishnan, Iman Vakilinia, Sandeep Reddivari, and Sanjay Ahuja. Analyzing cryptocurrency social media for price forecasting and scam detection. In 2024 12th International Symposium on Digital Forensics and Security (ISDFS), pages 1–6. IEEE, 2024.
 16. Zhihao Ding, Jieming Shi, Qing Li, and Jiannong Cao. Effective illicit account detection on large cryptocurrency multigraphs. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 457–466, 2024.
 17. Helen Milner, Redowan Mahmud, Mahbuba Afrin, Sashowta G Sid-dhartha, Sajib Mistry, and Aneesh Krishna. On-graph machine learning-based fraud detection in ethereum cryptocurrency transactions. In 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pages 1279–1285. IEEE, 2023.
 18. Junha Kang and Seok-Jun Buu. Graph anomaly detection with disentangled prototypical autoencoder for phishing scam detection in cryptocurrency transactions. *IEEE Access*, 2024.
 19. Aditya Singh, Anubhav Gupta, Hardik Wadhwa, Siddhartha Asthana, and Ankur Aro-ra. Temporal debiasing using adversarial loss based gnn architecture for crypto fraud detection. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 391–396. IEEE, 2021.
 20. Yifan Jia, Yanbin Wang, Jianguo Sun, Yiwei Liu, Zhang Sheng, and Ye Tian. Ethereum fraud detection via joint transaction language model and graph representation learning. *arXiv preprint arXiv:2409.07494*, 2024.
 21. Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

22. Meng Shen, Yiting Liu, Liehuang Zhu, Xiaojiang Du, and Jiankun Hu. Fine-grained webpage fingerprinting using only packet length information of encrypted traffic. *IEEE Transactions on Information Forensics and Security*, 16:2046–2059, 2020.
23. Alex Biryukov, Dmitry Khovratovich, and Ivan Pustogarov. Deanonimi-sation of clients in bitcoin p2p network. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 15–29, 2014.
24. Florian Tramèr, Dan Boneh, and Kenny Paterson. Remote {Side-Channel} attacks on anonymous transactions. In *29th USENIX security symposium (USENIX security 20)*, pages 2739–2756, 2020.
25. Meng Shen, Jinpeng Zhang, Liehuang Zhu, Ke Xu, and Xiaojiang Du. Accurate decentralized application identification via encrypted traffic analysis using graph neural networks. *IEEE Transactions on Information Forensics and Security*, 16:2367–2380, 2021.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

