FRAUD DETECTION AND ANTI-MONEY LAUNDERING APPLYING MACHINE LEARNING TECHNIQUES IN CRYPTOCURRENCY TRANSACTIONAL GRAPHS

Ana P. Martins and Miguel A. Brito
Centro Algoritmi
University of Minho, Portugal

ABSTRACT

Cryptocurrencies have advantages such as lower costs, efficiency, and security, but are vulnerable to fraud due to a lack of controls and anonymity. Criminals use virtual currencies for quick, anonymous transactions. Robust measures are needed to prevent illegal activities like money laundering. Machine learning (ML) and graph analysis can help detect fraud in the cryptocurrency market, despite criminals mimicking normal transactions. This study aims to use cutting-edge technologies like ML and graph learning to find fraudulent patterns in cryptocurrency transactions.

KEYWORDS

AML, Cryptocurrencies, Graphs, Machine Learning, Fraud Detection

1. INTRODUCTION

Cryptocurrencies are digital currencies issued on blockchain technology with advantages over traditional monetary systems. However, cryptocurrency fraud is a risk due to the lack of central regulation and controls, and markets are susceptible to fraudulent schemes like Ponzi, phishing, and money laundering. Effective anomaly detection methods and anti-money laundering (AML) controls are needed to prevent illegal cryptocurrency activities. ML algorithms are a promising area of research in AML.

This article explores using blockchain data and deep learning, specifically neural networks, for detecting cryptocurrency fraud. Combining deep learning with graph analysis is essential for detecting fraudulent transactions and extracting hidden patterns. The study aims to prevent illegal transactions and comply with industry standards.

The remainder of the paper is organized as follows. Section 2 presents the theme of AML and the role of ML, providing a clear understanding of the topic. Section 3, it is introduced the concepts of graph learning and explained its importance in this domain. Section 4 incorporates a literature review of ML and graph learning applied to AML. Finally, the main conclusions are discussed in Section 5.

2. ANTI-MONEY LAUNDERING

Effective AML systems are necessary for organizations to reduce the risk of financial losses from money laundering. AML involves identifying suspicious activities through customer transaction analysis using software, databases, and analytical tools (Labanca *et al.,* 2022). Cryptocurrencies and blockchain technology present challenges in implementing traditional AML practices due to their decentralized and transparent nature.

Updated technologies and methods are necessary to comply with regulations and effectively detect new money laundering techniques. ML is crucial in AML to prevent and detect financial crimes.

Graph learning improves AML solutions by analyzing financial data, identifying patterns, and increasing accuracy and efficiency (Han *et al*., 2020). This technology overcomes traditional AML system limitations and helps detect and prevent money laundering activities involving cryptocurrencies (Vassallo *et al*., 2021).

3. TRANSACTIONAL NETWORKS

Analyzing cryptocurrency transactions using graph analysis and ML techniques can efficiently detect illegal financial activities. This method saves time by eliminating the need for extensive feature engineering and provides a better understanding of the interactions within the cryptocurrency ecosystem. Graph Anomaly Detection (GAD) is a graph learning technique that detects anomalies in high-dimensional graph data using deep learning methods. It involves using graph representations and algorithms to recognize patterns.

Graph-based methods, an extension of deep learning, can uncover hidden patterns in graph data that traditional ML cannot (Xia *et al.,* 2021). Transaction network graphs are frequently used, representing nodes and connections in a vectorial space. UTXO blockchain address, transaction, and user networks are analyzed for link prediction, node classification, and forensic investigation.

Graph embedding techniques such as graph2vec and node2vec are popular for learning latent representations of vertices on large-scale networks, effectively reducing the transaction network data's dimensionality (Lin *et al.,* 2020).

Graph neural networks (GNNs) process graph data and extract graph representations for classification, recommendation, and clustering. GCN is a commonly used GNN technique that outputs node embeddings using a learnable kernel on local graph neighborhoods (Alarab *et al*., 2020b).

4. LITERATURE REVIEW

This section reviews the use of ML algorithms for detecting fraud and money laundering and suggests the application of graph learning and analysis in this area, using concepts from the previous section for support and clarity.

4.1 Machine Learning for Anti-Money Laundering

Chen *et al.* (2018) studied how ML techniques are used for detecting suspicious transactions in banking. Effective data prep and ML models like SVM, neural networks, and Bayesian networks are used for accurate detection, scalability, and fast performance. Choithani *et al.* (2022) found that SVM, LSTM, and Artificial Neural Networks were the most effective AI and ML algorithms for predicting cryptocurrency behavior for risk management in banking. Kamišalić *et al.* (2021) reviewed studies on using blockchain and data mining to identify fraud and anomalies. They found that Gradient Boosting, SVM, and Random Forest (RF) were popular methods, but Deep Learning with Neural Networks and LSTM is becoming more prevalent, especially with their graph learning approach.

The Elliptic dataset, which was introduced by Weber *et al. (*2019), is a significant and extensively labeled dataset available for any cryptocurrency. It comprises a graph network of Bitcoin transactions and has gained widespread recognition as a valuable resource for research in AML. RF was found to be better than GCN by Weber *et al.* (2019), but graph-based methods had potential. Alarab *et al.* (2020a) used ensemble learning and showed better results than Weber *et al*.'s method.

Ostapowicz and Żbikowski (2019) and Fawaz Ibrahim *et al.* (2021) proposed ML models to identify fraudulent accounts on Ethereum, with RF having the best performance in recall and false-positive rate. However, Ostapowicz and Żbikowski (2019) cautioned that the recall of 85% obtained may not be adequate for a real-world anti-fraud system.

Supervised methods like RF show promising results but require large, labeled datasets, which can be costly and time-consuming to create (Labanca *et al.,* 2022; Song *et al.,* 2023). Labeling historical transactions to detect money laundering is particularly difficult due to the low percentage of labeled illicit transactions. Unsupervised methods and active learning can be effective alternatives when labeled transactions are scarce. Recent studies have shown that unsupervised and active learning can perform similarly to supervised methods with only a few hundred labeled transactions. An active learning system that includes an unsupervised model can identify both known and unknown anomalous patterns, making it effective in detecting money laundering in capital markets transactions. The Isolation Forest algorithm has been found to be the most effective for this task. (Lorenz *et al.,* 2020; Labanca *et al.,* 2022)

Zhang and Trubey (2019) used ML and sampling to detect money laundering in financial transactions, concluding that artificial neural networks were the most effective and addressing the problem of unbalanced data.

As blockchain data grows, finding information and detecting money laundering becomes challenging due to its rarity and resulting data imbalance (Chen *et al.,* 2018; Zhang and Trubey, 2019). Researchers suggest analyzing trade-offs between recall and False Negative Rate (FNR) to achieve high accuracy. Various methods have been proposed, such as clustering, semi-supervised techniques, and heuristic user address clustering. For example, Monamo (2017) used k-means clustering and supervised classification models to detect fraudulent activity, while Yuan *et al.* (2018) employed a heuristic user address clustering method followed by verification using the Gaussian Mixture Model. Chen *et al.* (2018) suggested clustering or semi-supervised techniques to address data imbalance.

The literature suggests that traditional ML algorithms such as RF have been utilized in detecting fraud in finance and crypto transactions, but they have drawbacks such as costly manual labeling, difficulty in detecting rare events, and challenges in balancing recall and FPR.

4.2 Graph Learning and Analysis for Detecting Money Laundering

Motamed and Bahrak (2019) analyzed five cryptocurrencies, focusing on Bitcoin and Ethereum. They devised algorithms to extract transaction features and used graphs to detect anomalous transactions. Their findings showed that the address and transaction graphs for Bitcoin and Ethereum's money flow transaction graphs were particularly helpful.

GCNs, DeepWalk, and node2vec encode graph structures into dense representations to group similar nodes based on their neighborhoods. However, they don't account for temporal information, which is crucial for some applications.

Alarab *et al.* (2020) enhanced node classification on transaction graphs by merging GCN with linear layers for superior performance compared to GCN alone. Alarab and Prakoonwit (2022) achieved 98% accuracy and 80% f1-score in detecting illicit transactions. They improved previous studies by combining LSTM and GCN models and using active learning to label unlabeled data, considering the temporal aspect of transactions.

Singh *et al.* (2021) used a GNN-based adversarial loss architecture to reduce temporal bias in the elliptic dataset, leading to better performance. Pocher *et al.* (2022) found that GCN outperformed baseline methods over GAT. Geng *et al.* (2022) proposed an enhanced version of GCN, a graph attention mechanism, with the potential for detecting illegal activities in blockchain networks, demonstrating the usefulness of graph learning in addressing blockchain system security.

Caglayan and Bahtiyar (2022) proposed a graph embedding algorithm called node2vec to detect money laundering in financial transactions. It proved to be superior to other methods for detecting money laundering in a banking transaction dataset. Lopes *et al.* (2022) also found node2vec effective in identifying criminal and non-criminal relationships in criminal networks when combined with predictive ML methods. The studies show that node2vec has the potential to be a valuable tool for analyzing financial and criminal transactions. Further, Zhou *et al.* (2021) used node2vec to classify and predict financial fraud by representing the topological features of a financial network graph as low-dimensional dense vectors. This approach improved the efficiency of financial fraud detection. These findings demonstrate the potential of using graph embedding and deep learning for Internet financial fraud detection.

Li *et al.* (2022) proposed TA-Struc2Vec, a novel method to identify fraudulent users in financial transaction networks. It uses ML to classify users by considering both structural and transaction amount homogeneity in the network. This approach was found to be more effective than previous methods and has the potential to improve fraud detection in financial transaction networks.

The literature review highlights the significance of graph representation techniques in detecting fraudulent transactions, including financial and criminal activities. Graph-based approaches such as GCN, DeepWalk, node2vec, and their variants (e.g., GNN-based adversarial loss architecture, GAT mechanism) have shown promise in improving fraud detection accuracy. Combining graph representation with deep learning and incorporating temporal and weighted information can further enhance the performance of fraud detection models. These findings have implications for future research and demonstrate the potential of graph representation in addressing security concerns in financial and criminal transactions.

5. DISCUSSION

ML algorithms struggle with complex and interconnected graph data. Representation learning automatically extracts features, and node embedding maps nodes to a vector, characterizing their features. Node2vec generates expressive node embeddings using biased random walks. GNNs are specialized deep-learning techniques that work on graph data, incorporating information from neighboring nodes and individual node features. GCNs are effective in identifying patterns of illicit activity, detecting suspicious transactions, and analyzing blockchain networks in cryptocurrency transactions.

Previous studies found that combining GCN with node embeddings yielded the best results for RF algorithm. However, recent research suggests that GCN alone may be sufficient. This raises the question of whether embeddings play a crucial role in achieving superior performance. We propose to investigate this by testing different scenarios on the Elliptic Dataset to establish a benchmark for evaluating results.

We aim to understand the impact of different features, embeddings, and graph information on model performance and establish the optimal way to integrate traditional supervised models with deep learning techniques for improved performance in analyzing complex graph data.

6. CONCLUSION

The article discusses the challenges of using ML algorithms for complex graph data and how representation learning and GNNs can address these issues. Specifically, the focus is on the use of GNNs, such as GCNs for identifying patterns of illicit activity, detecting suspicious transactions, and analyzing blockchain networks in cryptocurrency transactions. The article proposes to investigate the impact of different features, embeddings, and graph information on model performance and establish the optimal way to integrate traditional supervised models with deep learning techniques for improved performance in analyzing complex graph data.

ACKNOWLEDGEMENT

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

REFERENCES

Alarab, I. and Prakoonwit, S. (2022) 'Graph-Based LSTM for Anti-money Laundering: Experimenting Temporal Graph Convolutional Network with Bitcoin Data', *Neural Processing Letters* [Preprint]. Available at: https://doi.org/10.1007/s11063-022-10904-8

Alarab, I., Prakoonwit, S. and Nacer, M.I. (2020a) 'Comparative Analysis Using Supervised Learning Methods for Anti-Money Laundering in Bitcoin', in *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp. 11–17. Available at: https://doi.org/10.1145/3409073.3409078

Alarab, I., Prakoonwit, S. and Nacer, M.I. (2020b) 'Competence of Graph Convolutional Networks for Anti-Money Laundering in Bitcoin Blockchain'. Available at: https://doi.org/10.1145/3409073.3409080

Caglayan, M. and Bahtiyar, S. (2022) 'Money Laundering Detection with Node2Vec', *Gazi University Journal of Science*, 35(3), pp. 854–873. Available at: https://doi.org/10.35378/GUJS.854725

Chen, Z. *et al.* (2018) 'Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review', *Knowl Inf Syst*, 57, pp. 245–285. Available at: https://doi.org/10.1007/s10115-017-1144-z

Choithani, T. *et al.* (2022) 'A Comprehensive Study of Artificial Intelligence and Cybersecurity on Bitcoin, Crypto Currency and Banking System', *Annals of Data Science*, pp. 1–33. Available at: https://doi.org/10.1007/S40745-022-00433-5/TABLES/2

Fawaz Ibrahim, R., Mohammad Elian, A. and Ababneh, M. (2021) 'Illicit Account Detection in the Ethereum Blockchain Using Machine Learning'. Available at: https://doi.org/10.1109/ICIT52682.2021.9491653

Geng, Z. *et al.* (2022) 'Novel blockchain transaction provenance model with graph attention mechanism', *Expert Systems with Applications*, 209. Available at: https://doi.org/10.1016/J.ESWA.2022.118411

Han, J. *et al.* (2020) 'Artificial intelligence for anti-money laundering: a review and extension', *Digital Finance*, 2(3–4), pp. 211–239. Available at: https://doi.org/10.1007/s42521-020-00023-1

Kamišalić, A., Kramberger, R. and Fister, I. (2021) 'Synergy of Blockchain Technology and Data Mining Techniques for Anomaly Detection', *Applied Sciences 2021, Vol. 11, Page 7987*, 11(17), p. 7987. Available at: https://doi.org/10.3390/APP11177987

Labanca, D. *et al.* (2022) 'Amaretto: An Active Learning Framework for Money Laundering Detection', *IEEE Access*, 10, pp. 41720–41739. Available at: https://doi.org/10.1109/ACCESS.2022.3167699

Li, R. *et al.* (2022) 'Internet Financial Fraud Detection Based on Graph Learning; Internet Financial Fraud Detection Based on Graph Learning', *IEEE Transactions on Computational Social Systems*, PP. Available at: https://doi.org/10.1109/TCSS.2022.3189368

Lin, D. *et al.* (2020) 'T-EDGE: Temporal WEighted MultiDiGraph Embedding for Ethereum Transaction Network Analysis', *Frontiers in Physics*, 8. Available at: https://doi.org/10.3389/FPHY.2020.00204

Lopes, D.D. *et al.* (2022) 'Machine learning partners in criminal networks', *Scientific Reports 2022 12:1*, 12(1), pp. 1–9. Available at: https://doi.org/10.1038/s41598-022-20025-w

Lorenz, J. *et al.* (2020) 'Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity'. Available at: https://doi.org/10.1145/3383455

Monamo, P.M., Marivate, V. and Twala, B. (2017) 'A Multifaceted Approach to Bitcoin Fraud Detection: Global and Local Outliers', pp. 188–194. Available at: https://doi.org/10.1109/ICMLA.2016.0039

Motamed, A.P. and Bahrak, B. (2019) 'Quantitative analysis of cryptocurrencies transaction graph', *Applied Network Science*, 4(1). Available at: https://doi.org/10.1007/S41109-019-0249-6

Ostapowicz, M. and Żbikowski, K. (2019) 'Detecting Fraudulent Accounts on Blockchain: A Supervised Approach', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11881 LNCS, pp. 18–31. Available at: https://doi.org/10.48550/arxiv.1908.07886

Pocher, N. *et al.* (2022) 'Detecting Anomalous Cryptocurrency Transactions: an AML/CFT Application of Machine Learning-based Forensics'.

Sabry, F. *et al.* (2020) 'Cryptocurrencies and artificial intelligence: Challenges and opportunities', *IEEE Access*, 8, pp. 175840–175858. Available at: https://doi.org/10.1109/ACCESS.2020.3025211

Singh, A. *et al.* (2021) 'Temporal Debiasing using Adversarial Loss based GNN architecture for Crypto Fraud Detection', in *Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021*. Institute of Electrical and Electronics Engineers Inc., pp. 391–396. Available at: https://doi.org/10.1109/ICMLA52953.2021.00067

Song, W. *et al.* (2023) *Blockchain Data Analysis from the Perspective of Complex Networks: Overview*. Available at: https://www.webofscience.com

Vassallo, D., Vella, V. and Ellul, J. (2021) 'Application of Gradient Boosting Algorithms for Anti-money Laundering in Cryptocurrencies', *SN Computer Science*, 2, p. 143. Available at: https://doi.org/10.1007/s42979-021-00558-z.

Weber, M. *et al.* (2019) *Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics*.

Xia, F. *et al.* (2021) 'Graph Learning: A Survey', *IEEE Transactions on Artificial Intelligence*, 2(02), pp. 109–127. Available at: https://doi.org/10.1109/TAI.2021.3076021

Yuan, Y., Member, S. and Wang, F.-Y. (2018) 'Blockchain and Cryptocurrencies: Model, Techniques, and Applications; Blockchain and Cryptocurrencies: Model, Techniques, and Applications', *SYSTEMS*, 48(9). Available at: https://doi.org/10.1109/TSMC.2018.2854904

Zhang, Y. and Trubey, P. (2019) 'Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection', *Computational Economics*, 54(3), pp. 1043–1063. Available at: https://doi.org/10.1007/S10614-018-9864-Z

Zhou, H. *et al.* (2021) 'Internet Financial Fraud Detection Based on a Distributed Big Data Approach with Node2vec', *IEEE Access*, 9, pp. 43378–43386. Available at: https://doi.org/10.1109/ACCESS.2021.3062467