Explain First, Trust Later: LLM-Augmented Explanations for Graph-Based Crypto Anomaly Detection

Adriana Watson
School of Engineering Technology
Purdue University
West Lafayette, United States
watso213@purdue.edu

***Abstract*—The decentralized finance (DeFi) community has grown rapidly in recent years, pushed forward by cryptocurrency enthusiasts interested in the vast untapped potential of new markets. The surge in popularity of cryptocurrency has ushered in a new era of financial crime [1]. Unfortunately, the novelty of the technology makes the task of catching and prosecuting offenders particularly challenging [2]. Thus, it is necessary to implement automated detection tools related to policies to address the growing criminality in the cryptocurrency realm.**

***Index Terms*—Cryptocurrency, Fraud Detection, Anomaly Detection, Graph Neural Networks, Explainable Artificial Intelligence (XAI), Large Language Models, Human-in-the-Loop, Blockchain Forensics**

I. Introduction

As the popularity of cryptocurrency and other decentralized finance mediums have grown, so too have concerns surrounding the rising trend of cryptocurrency attacks. These attacks are constantly evolving, rendering existing fraud detection systems unreliable. With the threat of cryptocurrency attacks becoming increasingly minacious, the need for novel solutions to identify cryptocurrency fraud has become apparent.

The emergence of machine learning technology has promised a new wave of solutions to conventionally complex tasks, including fraud detection. With ever-evolving banking methods and increasingly clever attacks, fraud detection is an ideal application for machine learning strategies.

Although comparable solutions exist, such as JPMorgan's fraud detection mechanisms and advanced credit card fraud detection tools, these systems are hindered by the black-box nature of their models. Furthermore, because anomalies tend to be sparse in training datasets, anomaly detection algorithms (which perform fraud detection) commonly fall victim to model imbalances. These two challenges combined make the regulation of fraud detection formidable, as models tend to accuse innocent parties more frequently and fail to explain why the accusation was made.

Thus, a pipeline that includes a graph-based fraud detection model, explainable artificial intelligence (XAI) algorithm, human-in-the-loop audit, and regulator emerges as a strong solution. Using a graph-based model addresses gaps in existing research, which commonly only utilizes relational databases from major cryptocurrency coins. Implementing an XAI model helps interpret model bias and removes the ethical concerns surrounding the use of black box models. The integration of a human-in-the-loop further protects innocent users. Finally, placing regulation stage only at the end of the process allows the system itself to be regulation-agnostic as it can be easily adapted to address any rules.

II. Literature Review

Existing literature reveals that the growing range of cryptocurrency attacks make regulation particularly difficult [3]. Furthermore, as cryptocurrency is still a relatively new development, existing fraud detection mechanisms are incapable of addressing the complex nature of DeFi fraud [4]. Thus far, efforts to address cryptocurrency fraud have emerged from one of three distinct areas: industry use cases, regulatory challenges, or academic contributions.

A. Industry Use Cases

JPMorgan has been a key player in recent efforts to combat cryptocurrency fraud. Although CEO Jamie Dimon has expressed a clear distaste for the medium that he once described as "hyped-up fraud" [5], the company been a public proponent of blockchain technologies [6].

This support is most clear in their use of AI cryptocurrency fraud detection tools that use a combination of neural networks, deep learning, natural language processing, reinforcement learning, and computer vision to catch fraudulent behavior as well as detect both traditional and DeFi fraud. The use of AI-powered mechanisms has saved JPMorgan $250 million annually and enabled the company to more effectively respond to customer reports [7].

Despite the great success of the program, JPMorgan has suffered greatly from the black-box nature of their models. Due to data imbalances, anomaly detection models have a tendency to yield high false positive rates, resulting in the targeting of innocent customers. The company also faces continuous challenges integrating new systems into legacy platforms. Despite these problems, JPMorgan continues to

dominate the market space, demonstrating the clear need for advanced fraud detection investments in the private sector [7].

B. Regulatory Challenges

In sharp contrast to the novelty of private industry's anti-fraud efforts, public enforcement of fraud policy has remained antiquated.

Companies investigated by the Market Integrity and Major Frauds Unit (MIMF) under the Department of Justice Criminal Division, were most commonly prosecuted for violations of lesser crimes than the ones committed, as no regulations exist to specifically address the cryptocurrency space. For example, BitConnect, a company that carried out a $2.4 billion Ponzi scheme was ultimately charged with wire fraud, operating an unlicensed money transmitting business, and conspiracy [8]. Despite the $2.4 billion lost by users, only $17 million was paid out to the victims of the company, likely in part due to these insufficient charges [9]. Similarly, Forsage, a company that defrauded investors out of $340 million was only charged with two counts of conspiracy to commit wire fraud [10]. At the time of writing, the victims of the scheme have not been compensated.

A similar theme plays out in cases brought by the U.S. Securities and Exchange Commission (SEC). Of the cases presented by the SEC that were actually prosecuted for fraud, the most common charge was under securities fraud statues [11]. While the definition of this charge (a misrepresentation of securities offerings) fits the crime to an extent, Ponzi schemes, rug-pulls, and other cryptocurrency attacks are much more devastating. Given the glaring lack of legislation to adequately mete out justice for these types of violations, the solution will likely be found by the more nimble private sector. Emerging academic research may hold the key to bridging the gap between the profit-driven private sector and technologically lagging public sector. This link is critical for both advancing cryptocurrency fraud detection and informing legislation.

C. Academic Contributions

There are many ongoing academic research efforts that can guide the search for cryptocurrency fraud solutions both, directly and indirectly.

Machine learning (ML) has grown to play a critical role in the identification and prevention of similar fraudulent activities. One such application is that of credit card fraud detection. While traditional methods for credit card fraud detection have been in place for nearly as long as credit cards, the growing interest in ML has naturally led to its application to the field. As credit card fraud detection is similar in both industry and nature to cryptocurrency fraud detection, the problems and solutions addressed by ML approaches can likely be transferred between the two. For example, researchers have proposed the use of Federated Learning (FL) and hybrid ML models to address data privacy concerns and data imbalances, respectively [12].

Beyond applying known application methods to new problems, there is a growing body of research surrounding cryptocurrency-specific approaches to fraud detection. Similar to the research surrounding credit card fraud detection, supervised and semi-supervised models combined with hybrid learning approaches tend to produce the most accurate fraud detection models [13]. Other researchers, however, have begun to propose novel ML models that more effectively address the graphical, interconnected nature of blockchain transaction data [14].

Many challenges surrounding ML-based fraud detection, particularly for cryptocurrency, have yet to be addressed. Notably, the balance of privacy and security poses a unique challenge [15]. As privacy is a major draw for cryptocurrency users, this attribute must be maintained. Unfortunately, security naturally challenges privacy, thus a technological paradox is formed. Related research has proposed FL as a solution to privacy concerns, as training data is decentralized, however, FL itself has inherent security loopholes that have yet to be addressed [16].

While ML has been leveraged to perpetrate cryptocurrency scams, promising research indicates that advanced algorithms can be part of a solution [17]. As blockchain data is best presented as a graphical dataset (a network of nodes and edges storing information in both the instance itself and the connection between instances), traditional data analytics techniques have often fallen short. Machine learning models, however, excel in this domain making them part of a very strong solution. Furthermore, developments in ML have proposed encouraging solutions to problems with false positive rates and real-time detection [18].

Existing research surrounding fraud detection for blockchain transactions and the application of XAI strategies largely neglects the graphical nature of cryptocurrency transactions and focuses only on major coins [19]. While the existing research is certainly valuable, much of the fraudulent activity occurs away from established coins such as Bitcoin and Ethereum. Additionally, identifying and explaining fraudulent behavior from a graphical perspective better encompasses the nature of the task.

As much of the academic literature implies, the best solution is one that reaches across all three of the above sectors so that all stakeholders are integrated effectively.

III. Proposed System Architecture

As shown in Figure 1, a solution that addresses many of the previously mentioned risks requires a system of checks and balances at each stage of the cryptocurrency transaction and detection process. The proposed model is trained on graph-based transaction data to address limitations in existing research, enabling the model to detect fraud occurring from any coin. Additionally, an XAI model is run in tandem with the fraud detection model to explain the instances of fraud detected. This addresses the model imbalance risk by forcing the model to justify decisions. A human-in-the-loop audit enhance by LLM generated explanations is performed on instances that are marked as fraud as a final step before the

case, with the evidence generated by the proposed process, is taken to regulators.

The multi-stage solution proposed ensures that each instance of the fraud detected process is verified and documented. This integrates a natural audit mechanisms for regulators into the solution. Additionally, the solution is legislatively agnostic, thus it could be adapted to conform to regulations from any area or multiple areas at once. The solution could also be expanded to integrate a Retrieval-Augmented Generation (RAG) model between the XAI and human reviewer phase to cross-compare transaction behavior with existing legislative policy documents.

IV. Implementation and Results

The research presented provides a working solution for all non-human components discussed above. The solution was executed in three primary stages: (1) the graph-based anomaly detection, (2) the XAI and LLM explanations, and (3) the interactive user interface (UI). It builds on existing research by Hasan et al. by adding the LLM explanation layer and user interference to improve the output literature [19].

A. Anomaly Detection

To meet the graph-based cryptocurrency database criteria, the Elliptic++ Dataset was used [20]. The dataset, which contains node and edge files, was used to create a graphical database of cryptocurrency transactions. Nodes in the dataset, which represented wallets, contained information such as the number of transactions, transaction totals, and bitcoin sent and received. Edges, which represented transactions, contained information such as mean, median, and maximum bitcoin sent through the edge. The graph was then used to train an unsupervised Graphical Neural Network (GNN) to detect anomalies. The result of this process was an updated graph database that included a binary anomaly designation as well as the trained GNN model.

B. XAI and LLM Explanations

The GNN model was then used to train a GraphLIME model so that explanations for the anomaly designation could be generated. The GraphLIME model could then be used to assign weights to features which could provide insight into which node features lead the model to mark it as anomalous.

After the GraphLIME model had been trained, this output combined with the original node features was sent to an OpenAI model to generate human-readable explanations. The prompt, shown below, included context for the task, variables where the node features and GraphLIME weights would be added, and few shot prompting.

```
    You are a financial crime analyst specializing
    in cryptocurrency fraud.
A graph-based anomaly detection model has flagged
    the following wallet as suspicious.

Your task is to analyze both:
1. The top features that *influenced the model's
    decision* (from GraphLIME), and
2. The actual transaction statistics of the wallet.

**Note:** The feature importance scores do NOT
    reflect actual values - they only indicate how
    strongly each feature contributed to the anomaly
     detection.

---
**Example 1**
**Feature Importances (from GraphLIME):**
- btc_sent_total: 9.812e-01
- degree: 3.442e-02
- btc_received_total: 0.000e+00

**Actual Node Data:**
- btc_sent_total: 0.0
- btc_received_total: 45.1
- degree: 24

**Interpretation**:
Even though the model heavily weighted `
    btc_sent_total`, the actual value is 0 -
    indicating the node received a lot of funds but
     hasn t sent anything. This may suggest
    hoarding behavior, commonly seen in Ponzi
    schemes.

---
**Example 2**
**Feature Importances (from GraphLIME):**
- transacted_w_address_mean: 7.623e-01
- degree: 1.124e-01
- btc_sent_total: 4.132e-03

**Actual Node Data:**
- btc_sent_total: 90.55
- btc_received_total: 21.38
- total_txs: 27
- transacted_w_address_mean: 1.0

**Interpretation**:
The model flagged the wallet based on consistent
    interactions with unique addresses (`
    transacted_w_address_mean`), but the node also
    exhibits significant sending behavior. This
    could suggest transaction structuring or
    layering in a money laundering pattern.

---
Now analyze this real case:

**Node ID**: {node_id}

**Features that most influenced the anomaly model (
    importance scores only):**
{formatted_weights}

**Actual Node Values:**
{formatted_data}

---
Your tasks:
1. Explain the suspicious behavior based on these
    two views.
2. If appropriate, classify it using known crypto
    fraud types: {fraud_types}
3. If the behavior appears normal, say so explicitly
    .
```

Listing 1: LLM Input Script

The input for the prompt included the most important weights from GraphLIME as well as the primary node features. A sample of the information automatically added to the

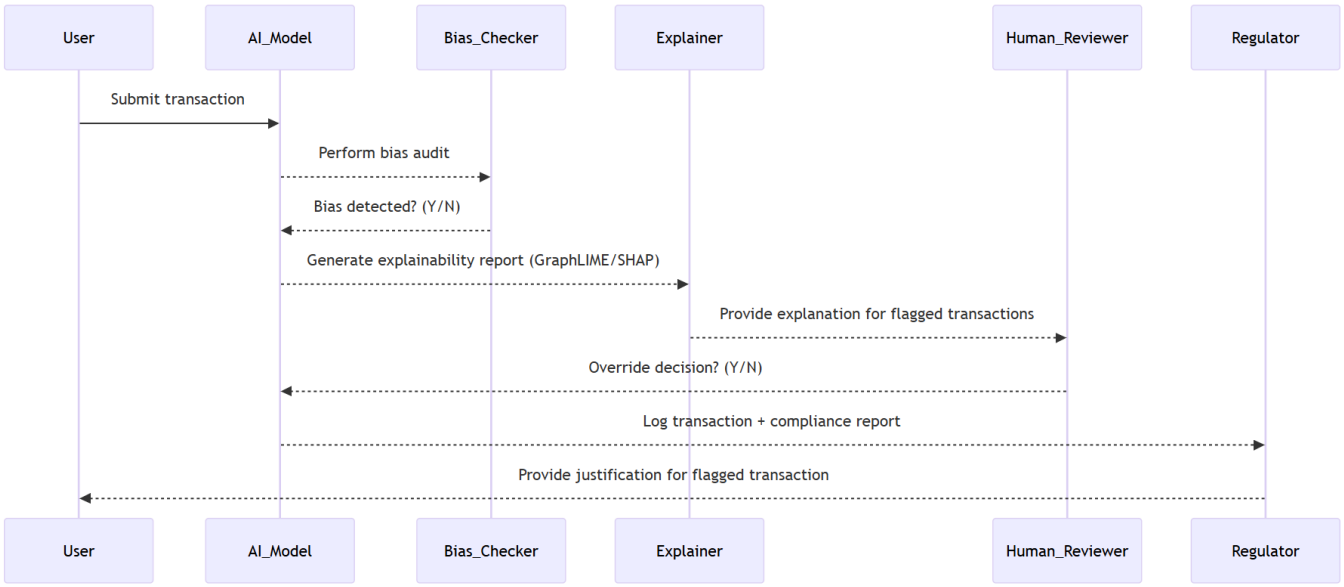

Fig. 1: A Sequence Diagram of the Proposed Solution

prompt is shown below.

```
LLM input:
Node ID: 1EQPoYt9DAnpTrAYjTBRCSD5bj5e1an4tF
Type: unknown
Class Label: N/A
Time Step: N/A
Lifetime (blocks): 0.0

Top contributing features from GraphLIME:

- degree: 9.941e-01
- btc_received_median: 9.941e-01
- btc_sent_total: 0.000e+00

Additional node statistics:

- total_txs: 2.0
- btc_received_total: 5159.84
- btc_sent_total: 5159.84
- num_txs_as_sender: 1.0
- num_txs_as_receiver: 0.0
- btc_transacted_total: 10319.7
- fees_total: 0.0013
- degree: 5
```

Listing 2: Sample Input

Based on the prompt and information provided, the LLM created a three part summary of the information including a determination on what kind of fraud may have been detected as well as a judgment regarding how fair the anomaly flagging seemed.

```
LLM says:

1. Based on the feature importances and actual node
    values provided, the model flagged the wallet as
     suspicious primarily due to the high degree
    centrality and the high median amount of Bitcoin
     received. Additionally, the wallet has sent out
     the same amount of Bitcoin it received, which
    could indicate a quick turnover of funds. The
    low number of total transactions (only 2) and
    the fact that the wallet has only been a sender
    in one transaction also raise suspicions. This
    behavior could be indicative of a money
    laundering scheme where funds are quickly moved
    through the wallet to obfuscate their origin.
2. This behavior aligns more closely with money
    laundering than any of the specific crypto fraud
     types listed. Money laundering involves the
    process of making illegally obtained funds
    appear legitimate by moving them through a
    complex sequence of banking transfers or
    commercial transactions. In this case, the high
    volume of funds being quickly moved through the
    wallet could be an attempt to disguise the
    origin of the funds.
3. Given the information provided, the behavior does
     appear abnormal and warrants further
    investigation to determine the true nature of
    the transactions and the wallet's involvement in
     potentially illicit activities.
```

Listing 3: Sample Output

C. Interactive UI

Finally, a small version of the database combined with the generated explanations was used to create an interactive UI that would allow users to visually see the interactions between wallets (nodes) and transactions (edges). A sample of the UI is shown in 2 as well as in appendix VI-B.

D. Discussion of Results

The final output of the system clearly identifies anomalies and provides believable explanations for the reasoning behind the label. While scaling the solution presents a clear challenge, as the size of graph based data in particular tends to be exponential in nature, the presented pipeline can be implemented in a variety of settings to add explainability to black box ML solutions.

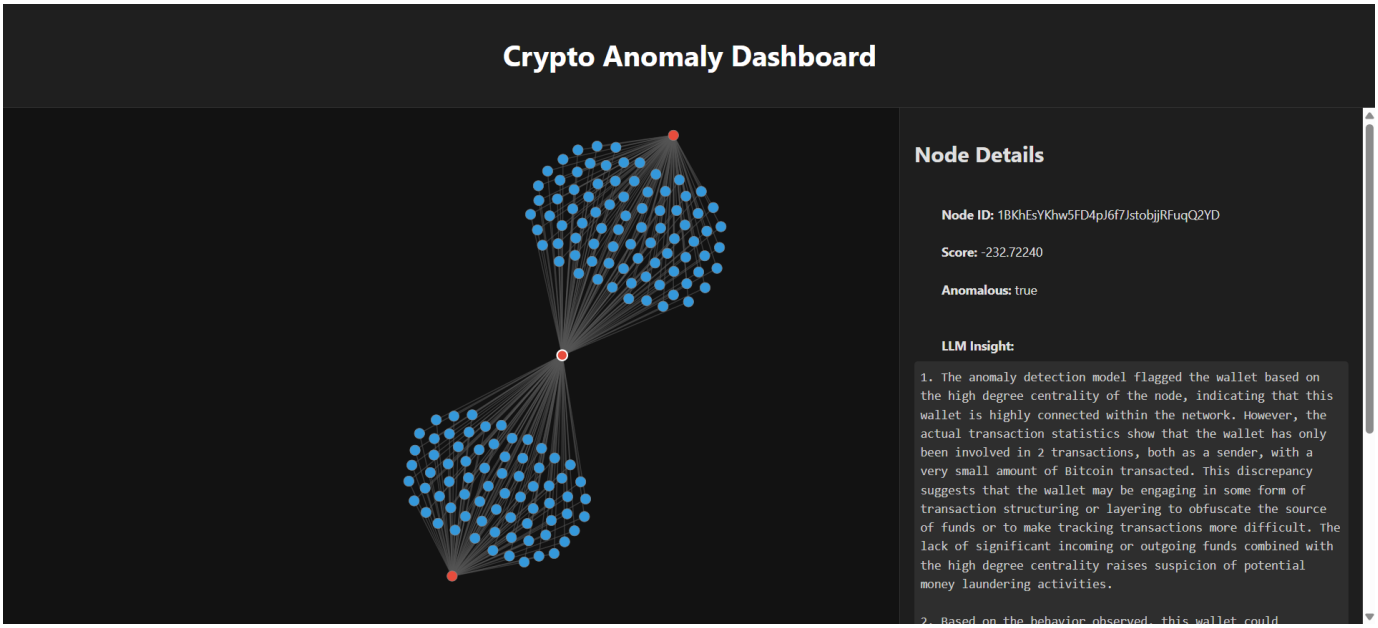

Fig. 2: A Sample of the Dashboard with the LLM Explanation and k=1

V. Future Work

Improvements that could further extend the project include:

- Generating LLM insights for non-anomalous nodes could provide further clarity into regular versus irregular behavior.
- Adding data from other blockchain transactions (the dataset used only included Bitcoin transactions) would add a layer of complexity and improve the range of applications.
- Connecting the insights with a more RAG-like system to more carefully define fraud types would improve the LLM insights.

References

[1] Federal Bereau of Investigation, "Cryptocurrency Fraud Report," Tech. Rep., 2023. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2023_IC3CryptocurrencyReport.pdf

[2] E. Wessan and P. Pillari, "Problems with Rulemaking by District Court Enforcement Action: the SECs Improper Cryptocurrency Regulation Eric Wessan & Phil Pillari," Aug. 2024. [Online]. Available: https://journals.law.harvard.edu/jlpp/problems-with-rulemaking-by-district-court-enforcement-action-the-secs-improper-cryptocurrency-regulation-eric-wessan-phil-pillari/

[3] A. Trozze, J. Kamps, E. A. Akartuna, F. J. Hetzel, B. Kleinberg, T. Davies, and S. D. Johnson, "Cryptocurrencies and future financial crime," *Crime Science*, vol. 11, no. 1, p. 1, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8730302/

[4] C. Leuprecht, C. Jenkins, and R. Hamilton, "Virtual money laundering: policy implications of the proliferation in the illicit use of cryptocurrency," *Journal of Financial Crime*, vol. 30, no. 4, pp. 1036–1054, Sep. 2022, publisher: Emerald Publishing Limited. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/jfc-07-2022-0161/full/html

[5] M. Sigalos, "Jamie Dimon says he's done talking about bitcoin: I don't care," Jan. 2024, section: Davos WEF. [Online]. Available: https://www.cnbc.com/2024/01/17/jamie-dimon-says-hes-done-talking-about-bitcoin-i-dont-care.html

[6] K. Balevic, "Jamie Dimon says bitcoin is the crypto of choice for 'sex traffickers, money launderers, ransomware'," Jan. 2025. [Online]. Available: https://www.businessinsider.com/jamie-dimon-bitcoin-criticism-crypto-criminals-choice-2025-1

[7] K. Tulsi, "Transforming Financial Services: The Impact of AI on JP Morgan Chases Operational Efficiency and Decision-Making," *International Journal of Scientific Research and Engineering Trends*, vol. 10, no. 1, pp. 207–213, Feb. 2024. [Online]. Available: https://ijsret.com/2024/02/12/transforming-financial-services-the-impact-of-ai-on-jp-morgan-chases-operational-efficiency-and-decision-making/

[8] "United States v. Kumbhani," Aug. 2021. [Online]. Available: https://www.justice.gov/jmd/media/1236066/dl?inline

[9] Internal Revenue Service, "Victims of BitConnect scheme to receive more than $17 million to compensate for losses," Internal Revenue Service, Press Release, Jan. 2023. [Online]. Available: https://www.irs.gov/compliance/criminal-investigation/victims-of-bitconnect-scheme-to-receive-more-than-17-million-to-compensate-for-losses

[10] "United States v. Okhotnikov," Feb. 2023. [Online]. Available: https://www.justice.gov/criminal/criminal-fraud/file/1570081/dl?inline

[11] U.S. Securities and Exchange Commision, "Crypto Assets," U.S. Securities and Exchange Commision, Tech. Rep., Feb. 2025. [Online]. Available: https://www.sec.gov/securities-topics/crypto-assets

[12] R. Bin Sulaiman, V. Schetinin, and P. Sant, "Review of Machine Learning Approach on Credit Card Fraud Detection," *Human-Centric Intelligent Systems*, vol. 2, no. 1, pp. 55–68, Jun. 2022. [Online]. Available: https://doi.org/10.1007/s44230-022-00004-0

[13] M. Bhowmik, T. Sai Siri Chandana, and B. Rudra, "Comparative Study of Machine Learning Algorithms for Fraud Detection in Blockchain," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Apr. 2021, pp. 539–541. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9418470?casa_token=pwE8EgS8vccAAAAA:O2WTLocNLx5wxrTFS8iO4nMFmvkV_69lFCQ-GtpeX1XvZl0G3dNo4V-8v-rYygA3UE3E3V41eQ

[14] R. M. Aziz, M. F. Baluch, S. Patel, and A. H. Ganie, "LGBM: a machine learning approach for Ethereum fraud detection," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3321–3331, Dec. 2022. [Online]. Available: https://doi.org/10.1007/s41870-022-00864-6

[15] A. A. Ahmed and O. O. Alabi, "Secure and Scalable Blockchain-Based Federated Learning for Cryptocurrency Fraud Detection: A Systematic Review," *IEEE Access*, vol. 12, no. 2024, Jul. 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10599372

[16] J. Zhang, H. Zhu, F. Wang, J. Zhao, Q. Xu, and H. Li, "Security and Privacy Threats to Federated Learning:

Issues, Methods, and Challenges," *Security and Communication Networks*, vol. 2022, no. 1, p. 2886795, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/2886795. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/2886795

[17] T. Ashfaq, R. Khalid, A. S. Yahaya, S. Aslam, A. T. Azar, S. Alsafari, and I. A. Hameed, "A Machine Learning and Blockchain Based Efficient Fraud Detection Mechanism," *Sensors*, vol. 22, no. 19, p. 7162, Jan. 2022, number: 19 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/19/7162

[18] G. Kathareios, A. Anghel, A. Mate, R. Clauberg, and M. Gusat, "Catch It If You Can: Real-Time Network Anomaly Detection with Low False Alarm Rates," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2017, pp. 924–929. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8260756?casa_token=95VRGfgVTaQAAAAA:S6ffvAl9yKKpl_C8thRUrRL_C3sevGQ4259P46MG5HzMX9SmEDvFZbwstbLaL5x65bpXN-fL1Q

[19] M. Hasan, M. S. Rahman, H. Janicke, and I. H. Sarker, "Detecting anomalies in blockchain transactions using machine learning classifiers and explainability analysis," *Blockchain: Research and Applications*, vol. 5, no. 3, p. 100207, Sep. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2096720924000204

[20] Y. Elmougy and L. Liu, "Demystifying Fraudulent Transactions and Illicit Nodes in the Bitcoin Network for Financial Forensics," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '23. New York, NY, USA: Association for Computing Machinery, Aug. 2023, pp. 3979–3990. [Online]. Available: https://dl.acm.org/doi/10.1145/3580305.3599803

VI. Appendix

A. GPT Usage

A GPT was used to generate mermaid graphs based on the following prompts (note: some formatting which was initially done in a markdown platform, was altered to address LaTeXrendering issues):

```
Generate a Mermaid `sequenceDiagram` that visualizes the interaction between different stakeholders in an AI-based transaction monitoring system. The diagram should include the following participants and interactions:

Participants:
- `User`: The entity submitting a transaction
- `AI_Model`: The automated system analyzing the transaction
- `Bias_Checker`: A module that audits for bias
- `Explainer`: A system generating explainability reports (GraphLIME/SHAP)
- `Human_Reviewer`: A reviewer who can override AI decisions
- `Regulator`: A compliance entity receiving logs and providing justifications

Interactions:
1. `User` > (`Submit transaction`) > `AI_Model`
2. `AI_Model` > (`Perform bias audit`) > `Bias_Checker`
3. `Bias_Checker` > (`Bias detected? (Y/N)`) > `AI_Model`
4. `AI_Model` > (`Generate explainability report (GraphLIME/SHAP)`) > `Explainer`
5. `Explainer` > (`Provide explanation for flagged transactions`) > `Human_Reviewer`
6. `Human_Reviewer` > (`Override decision? (Y/N)`) > `AI_Model`
7. `AI_Model` > (`Log transaction + compliance report`) > `Regulator`
8. `Regulator` > (`Provide justification for flagged transaction`) > `User`

Ensure the output follows correct Mermaid syntax with appropriate message direction (`->>`, `-->>`) and clear labeling. The sequence should accurately reflect the decision-making and compliance process in transaction monitoring.
```

This conversation can be found here: https://chatgpt.com/share/67c8c692-aa9c-8001-9d74-35598d9c3a5f

B. Dashboard Examples

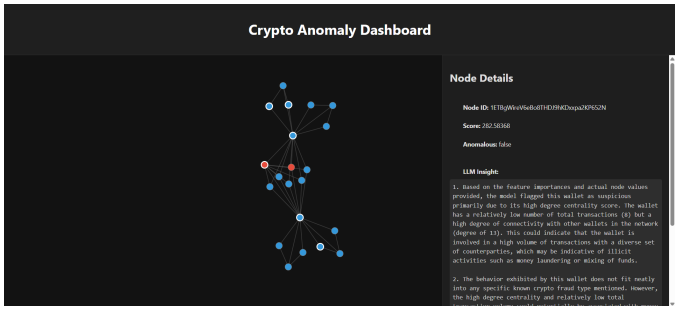

Fig. 3: A Sample of the Dashboard with the LLM Explanation and k=1

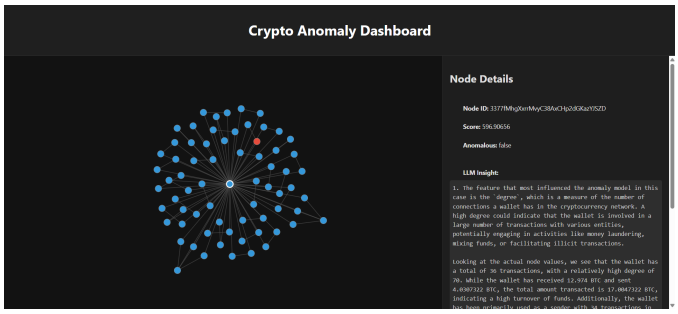

Fig. 4: A Sample of the Dashboard with the LLM Explanation and k=1

C. Code Availability

The codebase, including the anomaly detection model, GraphLIME explanations, and the dashboard interface, is available at: https://github.com/awatson246/crypto-anomaly-detection-policy