UNSUPERVISED ANOMALY DETECTION IN BITCOIN TRANSACTIONS

MACHINE LEARNING ALGORITHMS PROJECT
B. TECH IT SEMESTER VI

Group Members:
A024 Ira Malik
A032 Krisha Garg
A034 Ksama Arora

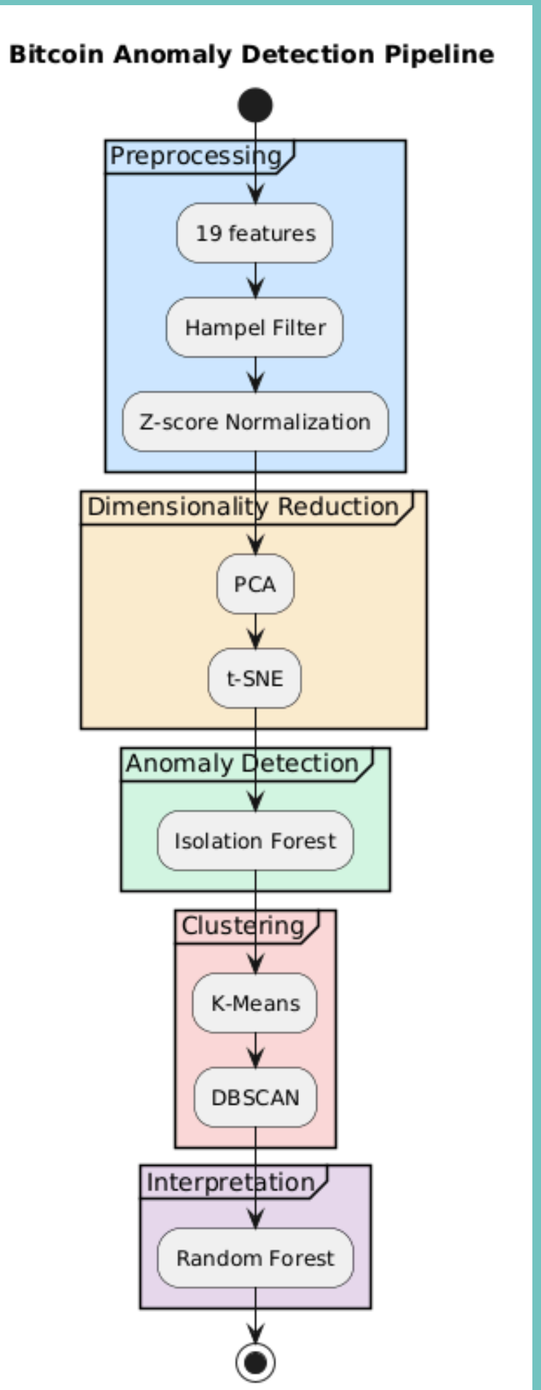
INTRODUCTION

- Bitcoin is currently the most widely used cryptocurrency, facilitating millions of daily transactions in a peer-to-peer, decentralized system. It provides both risk and innovation through its immutability, anonymity, and open ledger.

- Money laundering, double spending, ransomware activities, and sudden spikes in transaction volume are a few instances of unusual activity on the Bitcoin network that can gravely impact user trust and market integrity.

OUR PROPOSED SOLUTION

To identify anomalies in Bitcoin transactions, our paper proposes an end-to-end methodology combining statistical filtering, unsupervised learning, and feature importance analysis. Our approach is towards applying the Hampel filter as preprocessing , clustering for identifying behaviour clusters, and dimensionality reduction to visualize and interpret patterns. Through silhouette scores, we compare the performance of algorithms and utilize a Random Forest model to expose the most salient features .

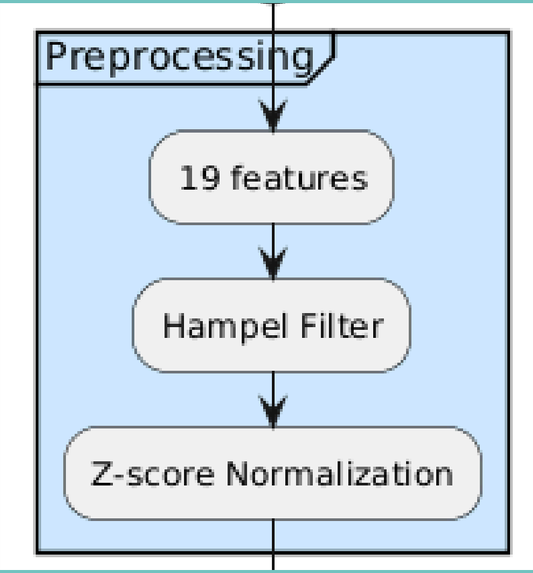


PROPOSED MODEL

- The dataset used in this study is a high-dimensional daily time-series collection of Bitcoin transaction metrics, including over 700 engineered blockchain indicators aggregated from open-source blockchain analytics platforms.

- Network activity features such as transaction volume, average fees paid, mining difficulty, market mood, and address-level behaviors are all represented through these indicators.

PREPROCESSING



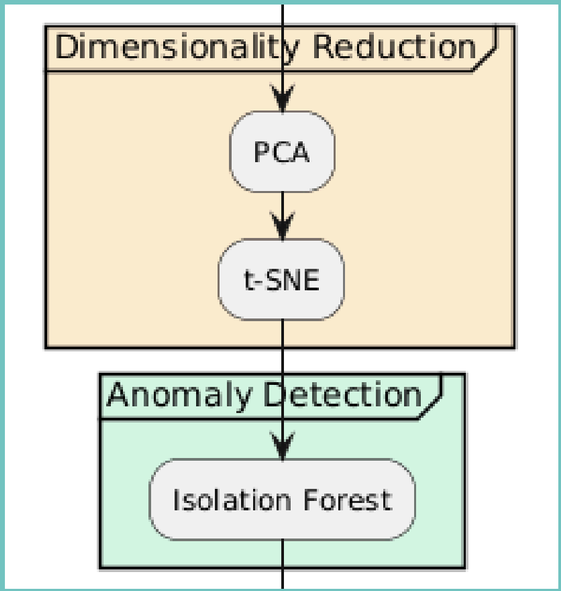
1. Robust Outlier Correction using Hampel Filter

- The Hampel filter was applied across all feature columns to mitigate local noise and outliers caused by sudden market fluctuations or irregular mining activity. By using median and MAD in a moving window, it preserves global trends while adjusting local anomalies, making it ideal for non-Gaussian blockchain data.

2. Feature Normalization with Z-score Scaling

- Post-filtering, all 19 features were standardized using Scikit-learn's StandardScaler. This ensures equal contribution of each feature to distance-based models (e.g., Isolation Forest, K-Means), and is crucial for dimensionality reduction techniques like PCA and t-SNE which are sensitive to scale differences.

DIMENSIONALITY REDUCTION & ANAMOLY DETECTION



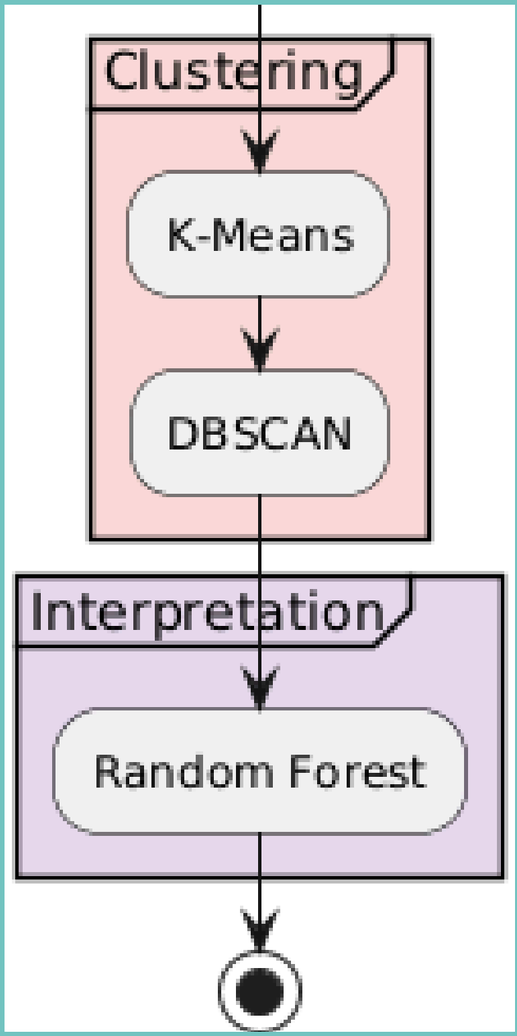
3. Dimensionality Reduction with PCA & t-SNE

- PCA: Used to reduce dimensionality linearly by maximizing variance; after filtering, the first two components explained 48% of variance, revealing improved cluster separation.
- t-SNE: Captured complex nonlinear patterns and visually confirmed clustering and outliers, validating the effectiveness of the filtering stage.

4. Anomaly Detection via Isolation Forest

Isolation Forest flagged the top 5% most anomalous transactions using random feature space partitioning. It is ideal for high-dimensional, unlabeled data like blockchain metrics and significantly enhanced the clarity of downstream clustering and visualization.

CLUSTERING & FEATURE IMPORTANCE



5. Clustering with K-Means and DBSCAN

- K-Means: Applied with k=3, chosen heuristically and validated by silhouette scores.
- DBSCAN: Detected clusters of arbitrary shapes and automatically marked sparse areas as noise, suitable for identifying irregular behavior in dense transaction networks.

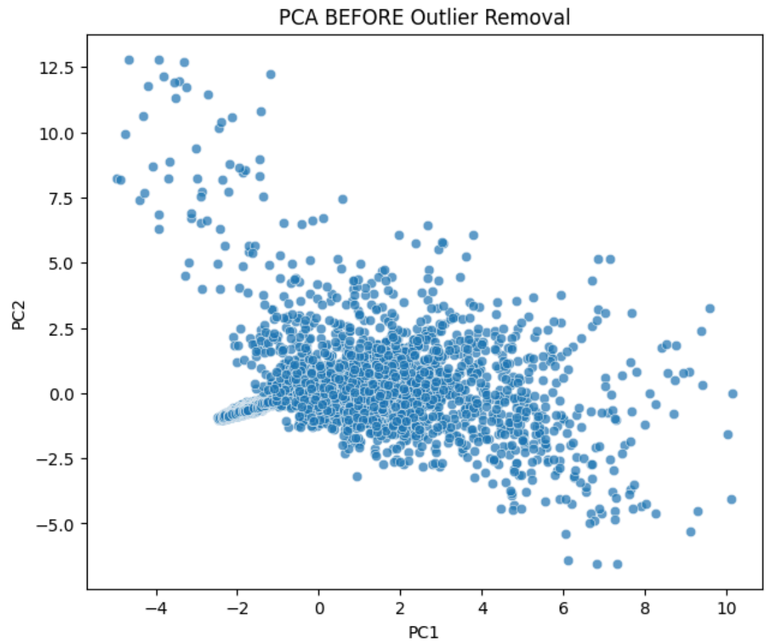
6. Feature Importance through Random Forest

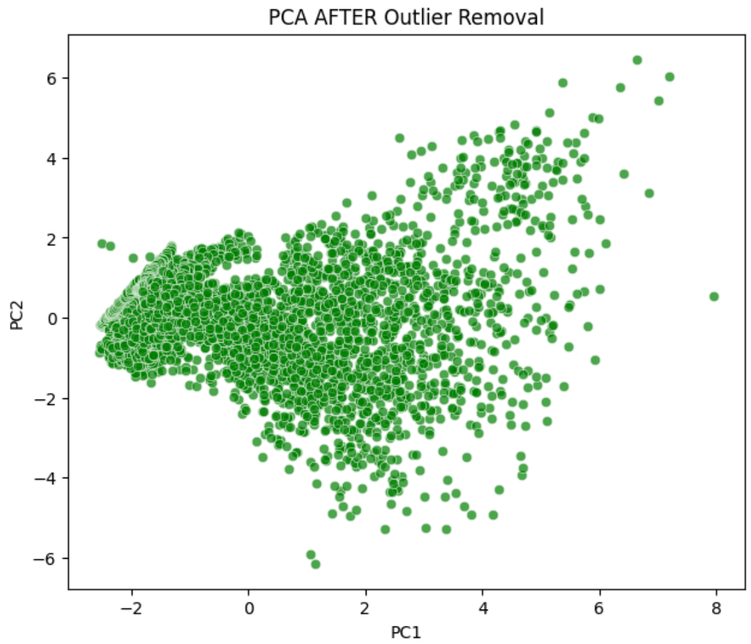
A Random Forest Classifier trained on K-Means cluster labels revealed the most influential features using Gini impurity reduction. Key attributes like difficulty90mom, transactionvalue90momUSD, and sentinusd90momUSD emerged as critical economic indicators for identifying behavioral anomalies.

RESULTS & DISCUSSION

1. Improved Data Structure through Outlier Removal

Outlier detection via Isolation Forest enhanced the internal structure of the dataset, as seen in PCA results—explained variance by the first two components increased to 48%. This suggests improved signal-to-noise ratio, aiding in the effectiveness of subsequent dimensionality reduction and clustering techniques (Figure 1).

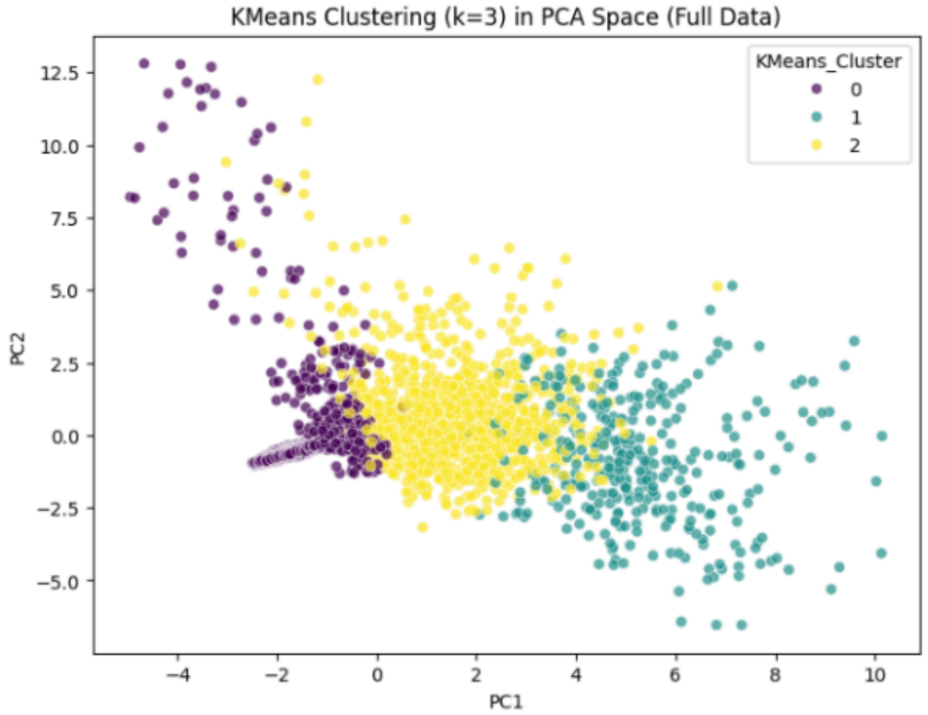


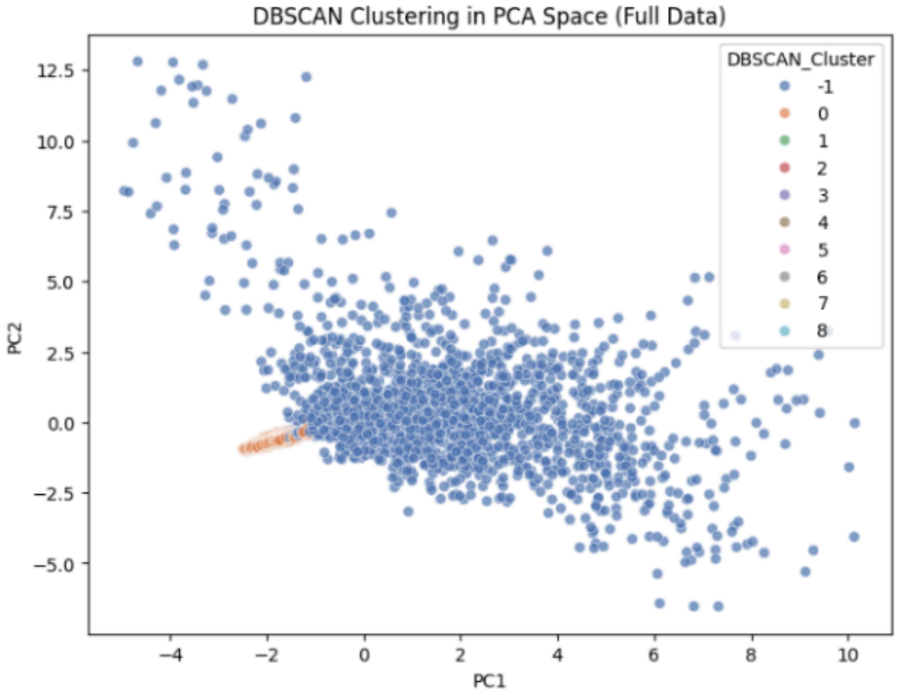


RESULTS & DISCUSSION

2. Clustering Performance Evaluation: K-Means vs DBSCAN

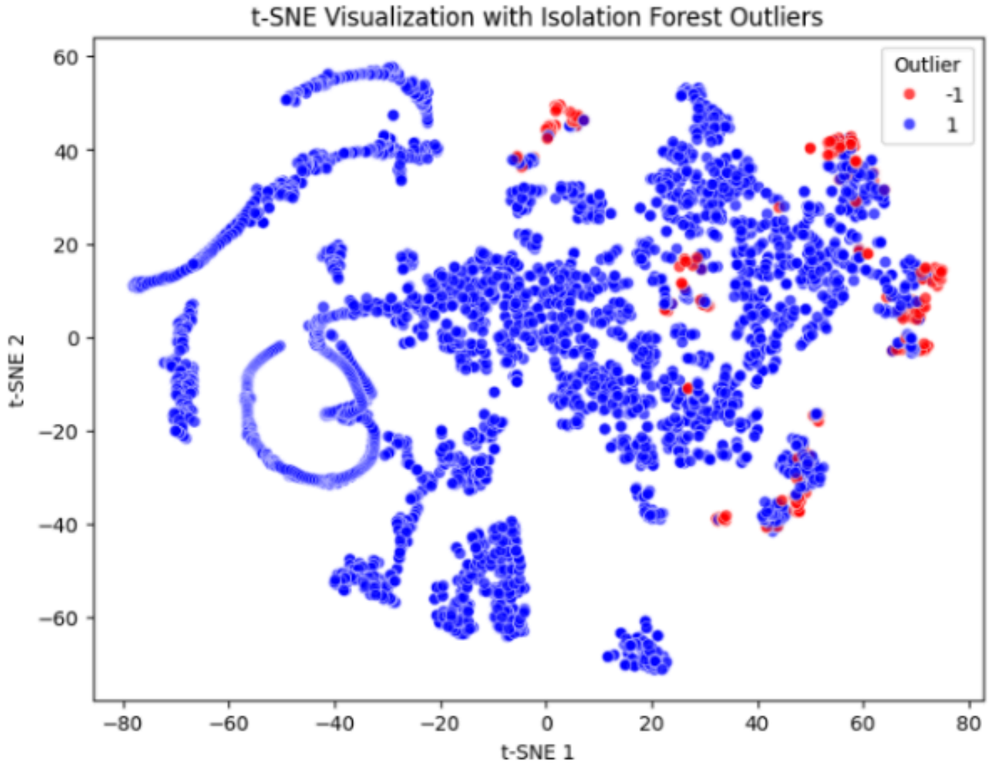
- K-Means (k=3): Achieved a silhouette score of 0.412, showing moderate compactness and better-defined clusters post-outlier removal.
- DBSCAN: Silhouette score of 0.367, but effectively identified irregularly shaped clusters and noise, especially overlapping outliers detected by Isolation Forest. However, it was sensitive to parameter tuning (eps, min_samples).





RESULTS & DISCUSSION

3. Non-Linear Pattern Discovery with t-SNE
t-SNE visualizations revealed distinct and well-separated clusters that matched K-Means and DBSCAN groupings. It also flagged several anomalies consistent with those detected by Isolation Forest. Unlike PCA, which preserves global variance, t-SNE preserved local structure, giving better insight into Bitcoin transaction behavior.



RESULTS & DISCUSSION

4. Feature Importance via Random Forest Analysis

Using K-Means cluster labels, a Random Forest classifier was trained to identify key discriminative features, which included:

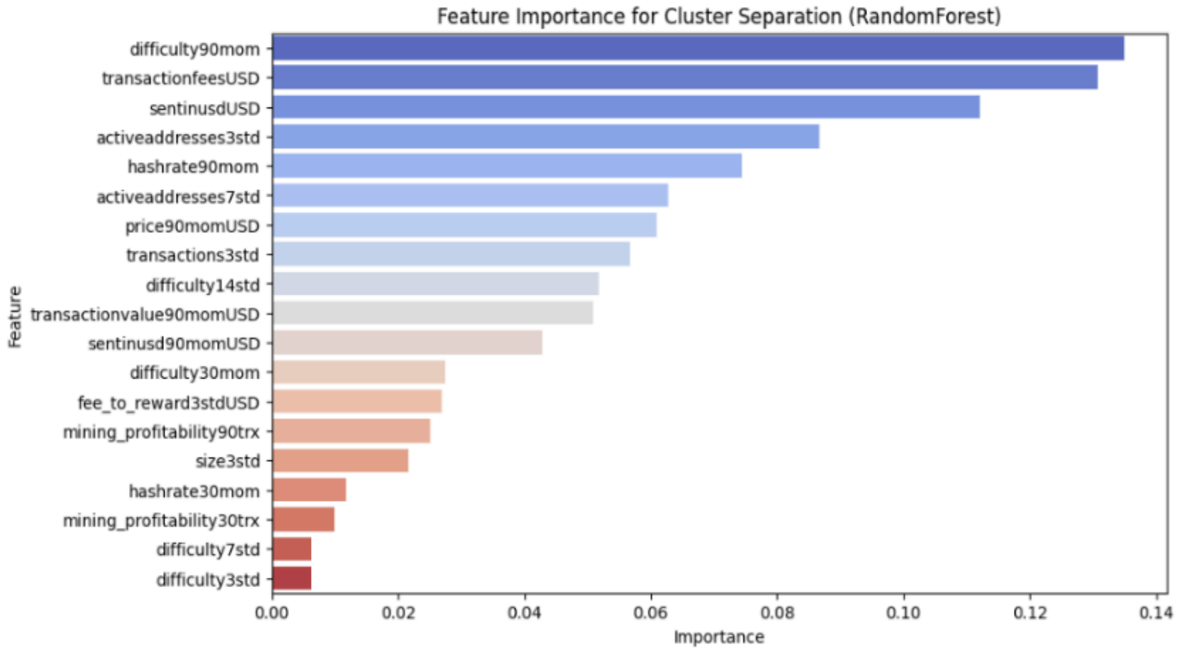


Figure 5: Feature importance plot showing the top contributors to KMeans cluster formation based on the Random Forest classifier.

CONCLUSION

- This study proposes an unsupervised learning pipeline for detecting anomalies in Bitcoin transaction data.
- Using the Hampel filter, PCA, t-SNE, Isolation Forest, and clustering methods (K-Means and DBSCAN), we effectively filtered noise, revealed clear cluster patterns, and identified abnormal transactions without relying on labeled data.
- Feature importance analysis highlighted key behavioral and economic indicators influencing transaction anomalies.
- The framework is scalable, interpretable, and well-suited for real-world blockchain monitoring, with future scope for real-time detection, alerting systems, and federated learning integration.

REFERENCES

[1] B. Bharathi, A. S. Begum, and G. Arulprakash, "Anomaly Detection in Blockchain Transactions: A Comparative Study of Isolation Forest, K-Means Clustering and LSTM Models," Int. J. Comput. Sci. Netw. Secur. (IJCSNS), vol. 23, no. 6, pp. 58–66, 2023.
[2] S. Siddamsetti, C. Tejaswi, and P. Maddula, "Anomaly Detection in Blockchain Using Machine Learning," J. Electr. Syst., vol. 20, no. 3, pp. 619–634, 2024.
[3] V. Dhanawat, "Anomaly Detection in Financial Transactions using Machine Learning and Blockchain Technology," Int. J. Bus. Manag. Vis. (IJBMV), vol. 5, no. 1, pp. 34–39, Jan.–Jun. 2022.
[4] Y. Hu et al., "Characterizing and Detecting Money Laundering Activities on the Bitcoin Network," arXiv preprint arXiv:1912.12060, 2019.
[5] X. Liu and P. Zhang, "A Scan Statistics Based Suspicious Transactions Detection Model for AML in Financial Institutions," in 2010 Int. Conf. Multimedia Communications, pp. 210–213.
[6] D. Ron and A. Shamir, "Quantitative Analysis of the Full Bitcoin Transaction Graph," in Financial Cryptography and Data Security, 2013.
[7] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in 2008 Eighth IEEE Int. Conf. Data Mining, pp. 413–422.