A Model for Detecting Cryptocurrency Transactions with Discernible Purpose

Hyochang Baek, Junhyoung Oh, Chang Yeon Kim, Kyungho Lee
Institute of Cyber Security and Privacy (ICSP)
Korea University
Seoul, Republic of Korea
{hcbaek90, ohjun02, zx0614, kevinlee}@korea.ac.kr

Abstract— The perpetration of financial fraud progresses parallel with the innovation in the field of finance. Consequently, the emergence of the blockchain technology has also manifested financial transaction obfuscation through the use of de-anonymization of the blockchain technology. This study identifies the suspicious transaction from Binance, an open-source cryptocurrency, through the means of defining and detecting the cryptocurrency wallets. By drawing the metadata of 38,526 wallets from etherscan.io, this study investigates the transactions with discernible purpose. This study performed an unsupervised learning expectation maximization (EM) algorithm to cluster the data set. Based on the features engineered from the unsupervised learning, we performed an anomaly detection using Random Forest (RF). In this study, we offered an insight into labeling the cryptocurrency wallets by providing a model for detecting the cryptocurrency with anomalous transactions. We advocate that labeling the wallets with discernible transactions may help financial institutions, private sectors, financial intelligence, and government agencies identify and detect the transactions with illicit activities.

Keywords—Cryptocurrency; Blockchain; Ethereum; Anti-Money Laundering; Smart Contract; Machine Learning

I. INTRODUCTION

Cryptocurrencies are built atop a novel blockchain technology that can be shared by a group of non-trusted parties without a central administration such as SQL or NoSQL databases [1]. With the exception of adversaries wielding large fractions of the computational resources, the miners run distributed consensus utilizing the hash and block principle [2][3]. Each block contains a timestamp that allows it to be traced back to its previous block. The decentralized digital currency has gained momentum, as it promised a peer-to-peer payment that ensures the integrity and anonymity. With the introduction to Bitcoin by Satoshi Nakamoto's white paper in 2008 [3], modified designs of the Bitcoin have introduced altcoins such as Ethereum. In comparison to the traditional financial contracts, Ethereum offers a rich programming language for writing 'smart contracts' that ensures low legal and transaction costs [4].

The World Economic Forum reported that 58% of the survey respondents expected 10% of the global domestic product to be stored using Blockchain Technology by the year 2025 [5]. The perpetration of financial fraud progresses parallel with the innovation in the field of finance, and as the volume of financial transaction data has increased intelligent agencies across different nations are able to only counter fraction of the money laundering activities. Based on the footprint left on the public Bitcoin blockchain, the market for ransomware has a minimum worth of USD 12,768,536 [6]. United Nations Office on Drugs and Crime estimates the annually laundered money between 2-5% of the global GDP, or USD 800 billion to 2 trillion [7].

Decentralized electronic cash systems is a currency that does not discriminate based on the citizenship or location [8]. The decentralized system aims to protect the blocks from being tampered, deleted, and revised. Hence, past studies have been interested in the anonymity of the network, weakness in the protocols, and topological properties [9] [1].

Our study aims to investigate the cryptocurrency wallets with anomalous transactions or features to detect fraudulent and suspicious parties. In this paper, by drawing the metadata of 38,526 wallets from etherscan.io, the paper uses anomaly detection-based approach and random forest to help identify the anomalous wallets from Binance. In Section II, we briefly discuss the prior studies conducted on illicit actives and cryptocurrency. In Section III, we iterate upon our methodology and we present our results in the penultimate section. Finally, we conclude our paper by covering the implications and limitations of this study.

II. RELATED WORKS

A. Money Laundering

Prior studies have investigated money laundering patterns through the use of histogram analysis [10], and Meiklejohn et al. [11] use a heuristic clustering to group Bitcoin wallet to classify the operators. Moreover, Huang et al. [12] have tracked over USD $16 million in ransomware operations during a span of two years. Thus, attracting illicit markets and criminal organizations such as Silk Road 2.0 and Decentralized Autonomous Crime Organization (DACO). Thus, the de-anonymization of blockchain transactions manifests financial transaction obfuscation.

B. Cryptocurrency

Prior to the introduction of cryptocurrencies such as Bitcoin and Ethereum, our economy relied on the fiat-based currencies which are often regulated through the government bodies [13]. Yet, with the collapse of the gold-based currencies, the new incarnation of the electronic currency has created a cryptocurrency that utilizes smart contracts that allow for rich programming language [4].

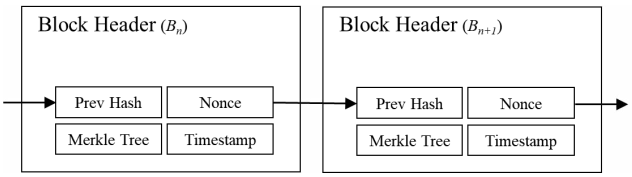

Fig. 1. Schema view of a blockchain within cryptocurrency

Blockchain differs from the traditional financial contracts as it avoids trusted central intermediaries while relying on the on-chain costs. Each block contains a timestamp that allows it to be traced back to its previous block, while using data encryption and public key infrastructure (PKI), as shown in Fig 1.

C. Anomaly Detection Approach

Digital forensic examiners often deploy signature-based detection and anomaly-based detection to identify various threats. The signature-based detection relies on human experts to disassemble the files, but it suffers from detecting unknown code and malware [14]. As such, this study focuses on the anomaly-based detection utilizing application programming interface (API) for the feature extraction method. Yet, due to the incipient nature of cryptocurrency, there has been little to no study that investigates the machine learning (ML) patterns for AML.

D. Ethereum Platform

Ethereum is a blockchain platform that utilizes Ethereum Virtual Machine (EVM). The cryptocurrency has the second largest cryptocurrency transactions after Bitcoin with a total turnover of 7 billion USD on a daily basis [15]. The rapid growth of tokens and decentralized application (DApp) have led to various implementation of smart contracts [16][17]. In our study, we selected Binance as our data collection target. Also, we gathered Ethereum wallets from ethersan.io through the API.

III. Methodoloy

A. Unsupervised Learning

In this study, we aimed to identify and detect the anomalous wallets from Binance. Hence, a total of 38,526 wallets were crawled from etherscan.io. We have selected the cryptocurrency wallets with the most trading volumes and was able to distinguish the Binance wallets into five different types: Binance_1, Binance_2, Binance_3, Binance_4, and $\text{Binance}_{\text{Token}}$ wallets. The data included wallets with a total of 10,041 transactions and each wallet contained at least a single transaction. The wallets encapsulate the Block Number, Time Stamp, Hash, Nonce, Transaction Index.

Although various clustering methods can be useful to organize the data collections, we did not assume any prior knowledge and deemed the massive data to be difficult to assimilate. As such, adopted the expectation maximization (EM) algorithm for Gaussian Mixture Model. By using the k-means algorithm, we clustered the data and defined the weight of the features. The Gaussian mixture model was used to maximize the likelihood function for parameters including the means and covariance of components and coefficients, as shown in the equation (1).

$$\gamma_j(x) = \frac{\pi_k N(x|\mu_k,\sum k)}{\sum_{j=1}^{K} \pi_j N(x|\mu j,\sum j)} \quad (1)$$

B. Supervised Learning

The random forest classifier, an ensemble-based classifier, relies on a randomly selected subset of samples and variables. Supervised classifiers are able to target a specific set of classes from the training samples in the unclassified data. The architecture of the random forest model in Fig. 2. Illustrates how the input data set can select and rank the subset of data.

In order to identify and recognize the anomalous wallets, we set a defined parameter from the features extracted from the EM Algorithm. We then assigned the classifiers to help collect and perform the classification. With the ensemble-based classification, there is a potential for overfitting the model. For this reason, we selected the RF-based approach to help sort the problem.

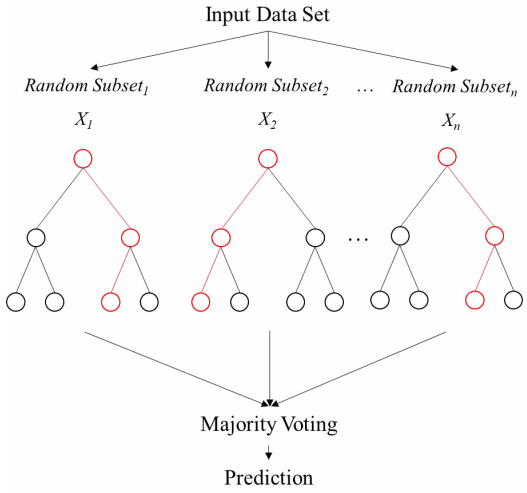

Fig. 2. Architecture of the random forest model

C. Model Verification

After classifying the data sets, the predictive models were used. The predictive model is a 2×2 table that consists of four components: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Table 1. depicts the prediction error of false positives (FP) and false negatives (FN) [18]. Based on equation (2), we were able to measure the *Accuracy* of the TP components. The *Precision* is defined as equation (3) as the FP and TP are the two measures to measure the precision. Equation (4) defines how the TP cases are identified. Moreoever, the *Recall* identifies the proportion of TP cases. The F-measure can be defined as equation (5). And the proportion of the TN and FN defines the True Negative Rate, as shown in equation (6).

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn} \quad (2)$$

$$\text{Precision} = \frac{tp}{tp+fp} \quad (3)$$

$$\text{Recall} = \frac{tp}{tp+fn} \quad (4)$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision+recall} \quad (5)$$

$$\text{True Negative Rate} = \frac{tn}{tn+fp} \quad (6)$$

D. Proposed Solution

The overview of the anomaly detection model for cryptocurrency in our study consists of six steps, as depicted in Fig. 3. As mentioned above, the Python API collects the data of the cryptocurrency. Then the preprocessing step and feature extraction is performed, which allows the unsupervised learning to cluster the data set. The identified clusters are classified into different wallets and then labeled accordingly. Finally, the RF is performed to distinguish the anomaly and then verified.

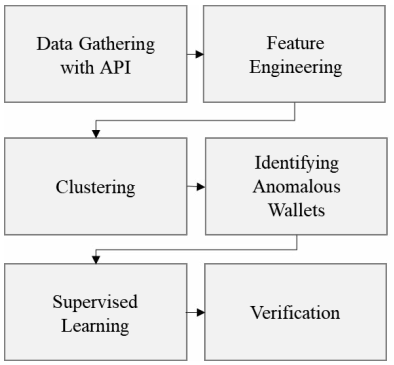

Fig. 3. Overview of the anomaly detection system for cryptocurrency

IV. Experiments

A. Data Description

Our study performed an anomaly detection to identify and detect anomalous transactions within the cryptocurrencies. Each wallet consists of the instance number, index, and address of the wallet (i.e. 0x0a0b8d2f61bffc4ec50273343b9812d4f15a7f8e). We were able to extract nine features from these data. The nine features are as following:

- *Feature$_1$ (valueave1)*: The average value of the difference between the amounts of both the previous and subsequent transaction.
- *Feature$_2$ (valueV1)*: The variance value from Feature$_1$.
- *Feature$_3$ (adIn)*: The standard deviation derived from the deposited amount to the wallet.
- *Feature$_4$ (adOut)*: The standard deviation from the withdrawn amount from the wallet.
- *Feature$_5$ (totalValueave)*: The average value of the transactions from the wallet.
- *Feature$_6$ (totalValueStd)*: The standard deviation of the corresponding wallet's transaction costs.
- *Feature$_7$ (delayCV)*: The difference between the previous and following transactions. 2
- *Feature$_8$ (timeCV)*: The coefficient variance of the time stamp's transactions.
- *Feature$_9$ (inoutGap)*: The account balance of the wallet.

B. Results

Our study performed an anomaly detection for the wallets to identify and detect anomalous transactions. Using the Scikit-learn tool to cluster the wallets. Based on equation (1), we were able to create seven separate clusters: Cluster A, Cluster B, Cluster C, Cluster D, Cluster E, Cluster F, and Cluster G. The data set were clustered into seven separate sets, as shown in Table I. From the seven sets, Cluster A, B, C, D, F, and G consisted of transactions on the same cryptocurrency with equivalent value. Yet, the accounts in Cluster F contained transactions with discernible purpose, as the wallets such as 0x593605e8f33342c43f716cbf6ca486155c435c9d made numerous anonymous transaction to numerous Bittrex accounts.

TABLE I. The Number of Wallets Distinguised By the Wallets

Types of Cluster	*Number of Wallets*
Cluster A	15,694
Cluster B	4,326
Cluster C	1,352
Cluster D	454
Cluster E	3,059
Cluster F	876
Cluster G	12,765

We then performed a binary classification using Random Forest (RF) to categorize the wallets. The RF is a way to create multiple decision trees with various subsamples of a dataset and use their averages to improve the accuracy of the prediction. Using the RF method we performed a binary classification with the range of [0, 1] to categorize the data. The wallets with the label '1' indicated anomalous wallets.

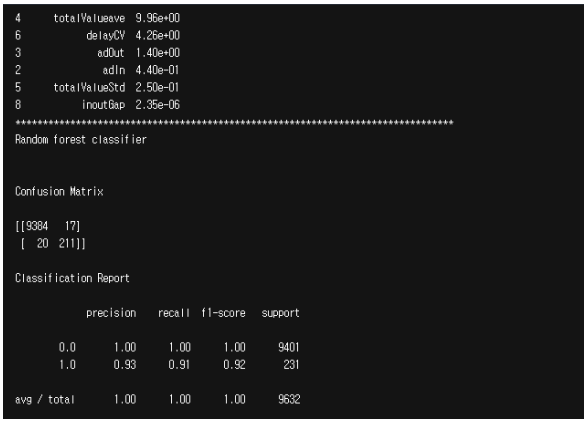

Fig. 4. Result of the predictive model utilizing Pycharm.

Fig. 4. shows the result of compiling Python machine learning program using PyCharm. The map learning algorithm used Scikit-learn's RandomForest Classifier. We used 20,000 wallets out of 38,526 wallets as learning data, and we performed model performance tests with 9,632 wallets. As you can see from Confusion Matrix, there are 9,384 correctly detected normal and 211 abnormal detected more than 9,632 purses. The number of false positives and True negatives was 17 and 20, respectively. Below are the results of calculating Precision, Recall, and F1-Score values.

TABLE II. PREDICTIVE MODEL FOR THE ETHEREUM TRANSACTION

Predictive Results / Classification	***Actual result***	
	Normal	***Suspicious wallet***
Normal	9384	17
Suspicious Wallets	20	211

Based on the confusion matrix we calculated the precision, recall, and F1 scores. The precision ratio represents the ratio between the normal and suspicious wallet. The recall is a percentage of the predicted normal and suspicious wallets. The F1 score is the average of both the wallet. The prediction accuracy of the normal wallet and suspicious wallet of the F1 score is 0.96.

TABLE III. PERFORMANCE OF THE RANDOM FOREST

Label	***Precision***	***Recall***	***F1 score***	***Support***
Normal wallet	0.99	0.99	0.99	9401
Suspicious wallet	0.91	0.93	0.92	231
Ave / Total	0.96	0.96	0.96	9632

The results from the two classifiers can be constructed through the 2 x 2 table that cross-tabulates. As shown in Fig. 4., dotted line from the Receiver-Operating Characteristic (ROC) plot for the random forest traces the sensitivity, the true positive rate, and specificity, the false positive. Since the ROC accuracy can be measured with the maximum value of 1, our results was able to achieve a high detection rate. Moreover, the plot indicates the high accuracy of the model.

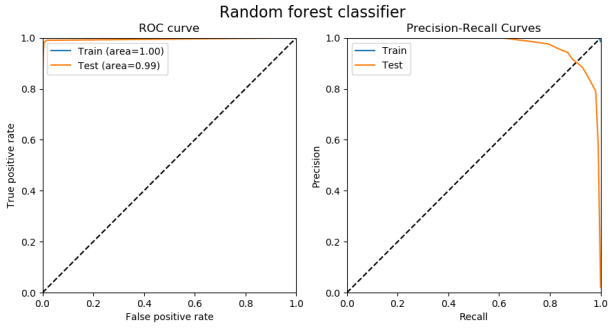

Fig. 5. ROC curve and Precision Recall Curves

V. CONCLUSION

The potential and future of cryptocurrencies are expected to reshape the digital assets and financial industries. Based on the blockchain technology, cryptocurrency ensures the integrity and anonymity as the technologies relies on the blockchain technology. The potential and benefits of the low transaction costs and absence of a central intermediaries have attracted a lot of attention during the past decade. In spite of the inherent benefits of the cryptocurrencies, they are susceptible to illicit actives such as money laundering. While past studies have focused on the anonymity of the network, protocols, and topology of the blockchain, there has been a lack of study on the detecting suspicious transactions

Our study exposes the wallets with discernible transactions through the means of supervised and unsupervised machine learning. By using the Scikit-learn algorithms, our study classified the wallets into seven separate clusters. While a high number of clusters were grouped into clusters with similar purpose. A single cluster consists of a wallet with anomalous transactions. In this study, a model using unsupervised learning was used to cluster the collected database. The EM algorithm was used to engineer the features of the wallets, and we labeled the wallets through the RF-classifiers.

Utilizing the EM algorithm for the Gaussian mix model, we were able to engineer nine features of each of the wallets. The features were engineered from the encapsulated data within each blocks. Based on the features that were engineered, we used the RF approach to label the suspicious wallets with high precision. In this study, we offered an insight into detecting the cryptocurrency wallets with anomalous transactions. We advocate that labeling the wallets with discernible transactions may help financial institutions, private sectors, financial intelligence, and government agencies trace and identify the transaction used for illicit activities.

ACKNOWLEDGEMENT

This research was supported by the Institute for Information &communications Technology Planning &Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01853, Machine Learning based Intelligent Malware Analysis Platform)

REFERENCES

[1] Nazri Abdullah, Anne Hakansson, and Esmiralda Moradian. Blockchainbased approach to enhance big data authentication in distributed envi-ronment. In2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), pages 887–892. IEEE, 2017.

[2] Maria Rona L Perez, Bobby Gerardo, and Ruji Medina. Modified sha256 for securing online transactions based on blockchain mechanism. In2018 IEEE 10th International Conference on Humanoid, Nanotechnol-ogy, Information Technology, Communication and Control, Environmentand Management (HNICEM), pages 1–5. IEEE, 2018.

[3] Satoshi Nakamoto et al. Bitcoin: A peer-to-peer electronic cash system.2008.

[4] Kevin Delmolino, Mitchell Arnett, Ahmed Kosba, Andrew Miller, andElaine Shi. Step by step towards creating a safe smart contract: Lessonsand insights from a cryptocurrency lab. InInternational Conferenceon Financial Cryptography and Data Security, pages 79–94. Springer,2016.

[5] Matthias Mettler. Blockchain technology in healthcare: The revolutionstarts here. In2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), pages 1–3. IEEE,2016.

[6] Masarah Paquet-Clouston, Bernhard Haslhofer, and Benoit Dupont.Ransomware payments in the bitcoin ecosystem.arXiv preprintarXiv:1804.04080, 2018.

[7] United Nations Office on Drugs and Crime (UNODC). Money-laundering and globalization, 2018

[8] Dorit Ron and Adi Shamir. Quantitative analysis of the full bitcoin trans-action graph. InInternational Conference on Financial Cryptographyand Data Security, pages 6–24. Springer, 2013.

[9] Ahmed Kosba, Andrew Miller, Elaine Shi, Zikai Wen, and CharalamposPapamanthou.Hawk: The blockchain model of cryptography andprivacy-preserving smart contracts. In2016 IEEE symposium on securityand privacy (SP), pages 839–858. IEEE, 2016.

[10] Jason Kingdon. Ai fights money laundering.IEEE Intelligent Systems,19(3):87–89, 2004.

[11] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko,Damon McCoy, Geoffrey M Voelker, and Stefan Savage.A fistfulof bitcoins: characterizing payments among men with no names. InProceedings of the 2013 conference on Internet measurement conference,pages 127–140. ACM, 2013.

[12] Danny Yuxing Huang, Maxwell Matthaios Aliapoulios, Vector Guo Li,Luca Invernizzi, Elie Bursztein, Kylie McRoberts, Jonathan Levin, KirillLevchenko, Alex C Snoeren, and Damon McCoy. Tracking ransomwareend-to-end. In2018 IEEE Symposium on Security and Privacy (SP),pages 618–631. IEEE, 2018.

[13] David Yermack. Is bitcoin a real currency? an economic appraisal. InHandbook of digital currency, pages 31–43. Elsevier, 2015.

[14] Mamoun Alazab, Sitalakshmi Venkataraman, and Paul Watters. Towardsunderstanding malware behaviour by the extraction of api calls. In2010Second Cybercrime and Trustworthy Computing Workshop, pages 52–59. IEEE, 2010.

[15] coinmarketcap. Top 100 cryptocurrencies by market capitalization,2019

[16] italik Buterin et al. A next-generation smart contract and decentralized application platform. white paper, 2014.

[17] Wesley Egbertsen, Gerdinand Hardeman, Maarten van den Hoven, Gert van der Kolk, and Arthur van Rijsewijk. Replacing paper contracts with ethereum smart contracts, 2016.

[18] Alan H Fielding and John F Bell. A review of methods for the as-sessment of prediction errors in conservation presence/absence models.Environmental conservation, 24(1):38–49, 1997.