Behavioral classification of Bitcoin addresses based on transaction history

Shayan Hamidi Dehshali*, Mehdi Modarressi†
*College of Computer and Electrical Engineering, University of Tehran
shayanhamidi@ut.ac.ir
†College of Computer and Electrical Engineering, University of Tehran
modarressi@ut.ac.ir

***Abstract*—User anonymity was one of the main motivations of the developers of Bitcoin. The developers achieved anonymity by assigning an address that has a specified format and reveals no clue about its owner. Although the transaction history of an address cannot expose the actual user behind the address, it can disclose hints about the behavioral role of the user. These behavioral roles could be centralized exchange, cyber-security service, darknet market, gambler, mining pool, peer-to-peer financial service and tumbler.**

In this paper, we intend to classify a variety of Bitcoin addresses belonging to the stated wallets types. First, we gathered addresses and their types from wallet explorer websites and similar works. Then, we defined features from the existing transaction history of an address, collected from a live full node peer. Later, we inserted the bulk of data in a PostgreSQL database management system to retrieve data efficiently. We scripted queries to calculate primary features from database. Furthermore, we added secondary features derived from primary features using Pandas python library.

At last, we applied different models, such as Decision Tree, Random Forest, KNN and XGBoost, on our training and test sets. Our results show the best model, regarding the evaluative metrics, was XGBoost with weighted average f1-score of 98.7%.

***Index Terms*—Bitcoin, Blockchain, Anonymity, Classifications, Machine learning, Behavioral pattern extraction**

I. Introduction

Traditional banking systems are centralized. In other words, a third party functioning as a central watcher, controls all the monetary affairs, user authentication and user balances. As traditional banking systems are not transparent, these systems can be manipulated. Bitcoin, introduced in 2008 by Nakamoto, provides transparency, anonymity, and security that proves itself advantageous to the traditional banking systems [1]. Bitcoin is a peer to peer network with a decentralized and distributed structure. In this network, transactions are made with the help of cryptographic mechanisms, for instance, digital signatures [2]. The history of the transactions is kept among a distributed and public ledger. Since the security mechanisms of Bitcoin utilize hash functions [3] and blockchain structure [4], forging transactions or repudiating them is impossible. Moreover, all of the users could validate the transactions. In this network, each user forms a transaction with his/her public and private key, and computes an address from the public key that must be attached to the transaction. Also, it is possible for each individual to own multiple addresses, since there is not any central authenticator. Thus, we can assert existence of anonymity. However, with the transparency of the transaction belonging to an address, we can argue that the privacy of Bitcoin users is flawed.

User anonymity was one of the main motivations of the developers of Bitcoin. Even though this property is advantageous in many ways, it has challenged the security of our society. Many illegal markets, for example, gun shops, drugs, smugglers, and also money laundering individuals chose this platform in order to stay anonymous. Moreover, gambling and betting have become prevalent among Bitcoin users. Although matching a Bitcoin address to its actual user is impractical, It is possible to extract behavioral patterns from the transparent transactions on the blockchain and reveal the role of the actual user behind an address. This feature enables criminal activity detection and their possible prevention.

In order to detect felony of an address, its activities should be tracked. Referring to web pages or messages that an address is appeared on or finding its IP address are some methods for revealing true identity behind the address.

There were several attempts to trace Bitcoin addresses and network. In [5] and [6], Bitcoin Network Graph was introduced which consists of each address as node and any transaction relation as edge. In [7], several machine learning models, with XGBoost as the most accurate one, were trained on the Bitcoin Network Graph. As building the Bitcoin Network Graph is costly, and also not so time efficient, testing or training any model on the graph is slow and resource consuming. Therefore, we propose a method on feature preparation to provide data for our model **at real time**. Moreover, we gathered a data-set which consists of 13 classes, the name of the classes are presented in Section III, while in [6], which is the nearest work to ours, provided only 10 classes of strongly labeled addresses. In other words, our model covers a vaster chunk of addresses of the Bitcoin.

Our main objective is classifying Bitcoin addresses into five main services of the Bitcoin network. Machine learning methods are applied, since they are the most common techniques of classification. Features are extracted from the blockchain data, and evaluation metrics are calculated for each learned model to choose the best model.

In Section II, we take a brief look at the necessary background for this research. In Section III, we elaborate our method that consists of data gathering, feature extraction, and

applying different models on the data. In Section IV, we present the results of the models, and we compare them based on evaluative metrics.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, p. 21260, 2008.

[2] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM*, vol. 21, no. 2, pp. 120–126, 1978.

[3] F. Wang, Y. Chen, R. Wang, A. O. Francis, B. Emmanuel, W. Zheng, and J. Chen, "An experimental investigation into the hash functions used in blockchains," *IEEE Transactions on Engineering Management*, vol. 67, no. 4, pp. 1404–1424, 2019.

[4] S. Haber and W. S. Stornetta, "How to time-stamp a digital document," in *Conference on the Theory and Application of Cryptography*. Springer, 1990, pp. 437–455.

[5] I. Alqassem, I. Rahwan, and D. Svetinovic, "The anti-social system properties: Bitcoin network data analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 1, pp. 21–31, 2018.

[6] B. Tao, I. W.-H. Ho, and H.-N. Dai, "Complex network analysis of the bitcoin blockchain network," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

[7] Y. Xiang, Y. Lei, D. Bao, W. Ren, T. Li, Q. Yang, W. Liu, T. Zhu, and K.-K. R. Choo, "Babd: A bitcoin address behavior dataset for pattern analysis."