Poster:Deanonymization of Bitcoin transactions based on network traffic analysis with semi-supervised learning

Chuanyong Tian
Beijing University of Posts and Telecommunications
tiancy4real@bupt.edu.cn

Ruisheng Shi✉
Beijing University of Posts and Telecommunications
shiruisheng@bupt.edu.cn

Lina Lan
Beijing University of Posts and Telecommunications
lanlina@bupt.edu.cn

Zhiyuan Peng
Beijing University of Posts and Telecommunications
2018213646@bupt.edu.cn

Yulian Ge
Beijing University of Posts and Telecommunications
geyl@bupt.edu.cn

Yuxuan Liang
Beijing University of Posts and Telecommunications
dzqlyx@bupt.edu.cn

Peng Liu
The Pennsylvania State University
pxl20@psu.edu

***Abstract*—Many cybercriminals have adopted Bitcoin as a channel to receive ransoms or launder money. This cryptocurrency poses challenges to law enforcement, as it is decentralized and largely unregulated. Bitcoin network traffic analysis is an important approach for cryptocurrency transaction deanonymization and many methods are proposed by researchers. However, the existing methods suffer from low precision and high cost. In this paper, we propose NTSSL , an efficient transaction deanonymization approach using network traffic analysis with semi-supervised learning. The core idea of this method is to generate pseudo-labels using several unsupervised learning algorithms to achieve comparable performance with lower costs. In addition, we propose a clustering method for transactions originating from the same bitcoin node, and synergistically analyze the clustering results with the NTSSL's deanonymization results to further improve the performance of our deanonymization method.**

Keywords—Bitcoin network, network traffic analysis, semi-supervised learning, transaction deanonymization

I. INTRODUCTION

Blockchain technology has received widespread attention due to its decentralization, anonymity, tamper resistance, and traceability. Bitcoin provides criminals with opportunities for money laundering, black market trading and other illicit activities[1]. The research on deanonymization mechanism of illicit cryptocurrency transactions is necessary.

Network layer deanonymization technology [2]-[4] means using the traffic transmitted in the Bitcoin network to analyze the broadcast path and feature of transactions in the network, so as to trace the IP of the node that generated the transaction. The existing technology has a lower precision and is often affected by computing and storage resources[4]. In addition, there exist some strategies to protect the anonymity of transactions at the network layer, such as Diffusion[5] and Dandelion++[6].

In this paper, we analyze two kinds of traffic collection methods in the Bitcoin network from both active and passive angles. We evaluate the performance of existing deanonymization method based on unsupervised learning[4]. We find that the precision of this method is not ideal after experiments. Therefore, we try to solve this problem by proposing a new deanonymization method based on semi-supervised learning and improving model performance by combining transaction clustering information on transaction layer.

In this paper, we make the following contributions:

- We propose a transaction deanonymization algorithm NTSSL based on semi-supervised learning. This method can identify the transaction created by the node itself from the transaction broadcast by the node, so as to associate the transaction with the IP address of the originating wallet node, thereby realizing the deanonymization of the transaction. The performance of this method is significantly better than the existing ones based on unsupervised learning.
- We propose a transaction clustering method based on transaction layer data analysis, clustering transactions that may originate from the same Bitcoin node within a certain period of time. By designing a cross-layer collaborative analysis tracing method that combines transaction clustering results and node-originated transaction identification results, the deanonymization effect is further improved. Our claims have been preliminarily validated in real-world network environments.

II. OVERVIEW OF DEANONYMIZE

A. Traffic collection method

We have considered two traffic collection methods. From the perspective of passive traffic capture, we consider the node traffic collection method based autonomous systems (AS)[7]. Taking advantage of the network topology of ASes, we can capture all the bitcoin traffic that naturally passes through the AS. From the perspective of active traffic capture, we consider the probe-based traffic collection method to capture the active communication traffic with the target node by using the probe node to actively connect to the target node.

B. NTSSL:node-originated transaction identification algorithm based on semi-supervised learning

We design a transaction deanonymization method based on semi-supervised learning, NTSSL. The entire NTSSL algorithm process can be roughly divided into four stages, as shown in **Figure.1**.

Stage 1: Labeling the suspected transactions originated by the target node. We apply three unsupervised learning algorithms including Isolation Forest, Auto Encoder, and One-class SVM to the dataset. We take the intersection of the results of the respective algorithms, and label the data in the intersection as positive sample, that is, node-originated transactions.

Stage 2: Pseudo label labeling. We regard the pseudo-labels marked in the previous stage as the known labels of the data and use them to label the remaining samples in the training set. We use unsupervised representation learning to calculate the abnormal score of each transaction. In order to minimize possible false labeling, we believe that transaction samples that satisfy all of the following conditions can be labeled as positive samples '1', and the rest are labeled as negative samples '0': (1) The anomaly score is not lower than the lowest known positive sample anomaly score in **Stage 1**; (2) The Send_inv_n value is not lower than the lowest Send_inv_n value of the known positive samples in **Stage 1**; (3) The Recv_getdata_n value is not lower than the lowest known Recv_getdata_n value for positive samples in **Stage 1**.

Stage 3: Expanding the feature vector. We only select the abnormal score of Isolation Forest as a new feature value to add to the original feature vector, getting the final feature vector: [Send_inv_n, Recv_getdata_n, Recv_getdata_n/ Send_inv_n, Score].

Stage 4: Model training and predicting. We use XGBoost to obtain node-originated transaction identification result.

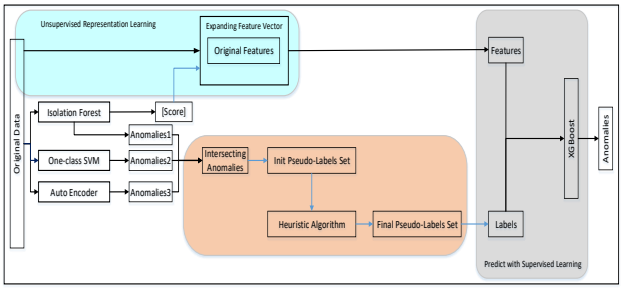
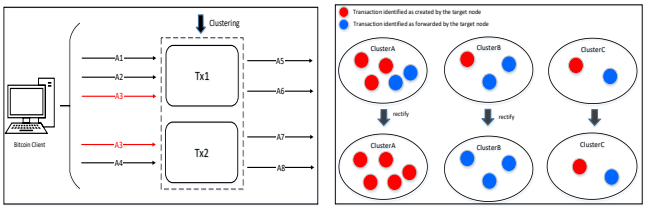


Fig. 1. Algorithm process of NTSSL

C. *Cross-layer collaborative analysis*

1) *Transaction clustering*

We propose a clustering method for transactions originating from the same node: in a relatively short period of time, multiple transactions involving the same address or addresses in the same cluster may originate from the same node, and the transactions in each cluster originate from the same Bitcoin node. The clustering process is shown in **Figure 2(a)**, the Tx1 and Tx2 are regarded as originating from the same Bitcoin node.



（a）Transaction clustering (b) Collaborative analysis

Fig. 2. Cross-layer collaborative analysis

2) *Collaborative analysis*

We try to use result of transaction clustering to find possible false positives or false negatives in the results of NTSSL: for transactions identified by the NTSSL as node-originated, we will observe the cluster where the transaction is located, if the majority of transactions in this cluster are identified as node-originated, we will consider the remaining transactions identified as node-forwarded to be false negatives and rectify them as node-originated. We perform the same operation on these transactions identified as forwarded. In this way, the recall of the algorithm is improved, the false positive rate is reduced, and the analysis process is shown in **Figure 2(b)**.

III. Preliminary Experiment

We make the performance evaluation and comparison of baseline, NTSSL and cross-layer collaborative analysis when intercepting different proportions of node connections in our preliminary experiment. Even for the weakest attacker who can intercept 25 % of the target node's connections, our cross-layer collaborative analysis method achieves F1 score of 0.45, the NTSSL can reach 0.41, and the method [4] which is chosen as our baseline can only reach 0.27 F1 score. Compared with our baseline, our cross-layer collaborative deanonymization method can achieve 50~70% performance improvement when we choose F1-score as a metric.

IV. Conclusion

In this work, we implement a complete network-layer transaction deanonymization. We first give an analysis of the traffic collection method, and propose a deanonymization method NTSSL based on semi-supervised learning. Preliminary evaluation results illustrate that our proposed NTSSL method outperforms existing deanonymization methods based on unsupervised learning. Moreover, our proposed cross-layer collaborative deanonymization method can further improve the performance of NTSSL.

Acknowledgment

This work was supported by the Beijing Natural Science Foundation under Grant M21037， the National Key Research and Development Plan in China (2022YFB2702405), 2022 Industrial Internet Public Service Platform - Industrial Internet Oriented Virtual Currency Mining Governance Public Service Platform Project by the Ministry of Industry and Information Technology of PRC，Major Research and Application Project for the Supervision Platform of Virtual Currency Mining Behavior by the Ministry of Education of PRC.

References

[1] Chainalysis:The 2022 Crypto Crime Report [EB/OL].https://blockbr.com.br/wp-content/uploads/2022/06/2022-crypto-crime-report.pdf

[2] Biryukov A, Khovratovich D, Pustogarov I. Deanonymisation of clients in Bitcoin P2P network[C]//ACM Conference on Computer and Communications Security (CCS). ACM, 2014.

[3] Biryukov A, Tikhomirov S. Deanonymization and linkability of cryptocurrency transactions based on network analysis[C]//2019 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2019: 172-184.

[4] Apostolaki M, Maire C, Vanbever L. PERIMETER: A Network-Layer Attack on the Anonymity of Cryptocurrencies[C]//Proceedings of the 25th International Conference on Financial Cryptography and Data Security (FC'21). 2021.

[5] Diffusion.[EB/OL].https://github.com/Bitcoin/Bitcoin/blob/da4cbb7927497ca3261c1504c3b85dd3f5800673/src/net_processing.cpp#L3813

[6] Fanti G, Venkatakrishnan S B, Bakshi S, et al. Dandelion++ lightweight cryptocurrency networking with formal anonymity guarantees[J]. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2018, 2(2): 1-35.

[7] CAIDA's ranking of Autonomous Systems (AS).[EB/OL]. https://asrank.caida.org/