Dragnet: A Method for Tagging Bitcoin Addresses of Exchanges

Zhen Li [1], Yi Zheng [2], Qi Li [1], Ming Wu [1], and Kunbin Peng [1]

[1]Affiliation not available
[2]Qtum Chain Foundation

October 30, 2023

Abstract

Currently, there are hundreds of Bitcoin exchanges on the market, so choosing a reliable exchange is a critical issue for users. We know that the amount of Bitcoin holdings is an essential indicator for evaluating an exchange, but people have very few ways to access this information. Besides, many reports indicate that the trading volumes of most Bitcoin exchanges do not match their real situations, and the fake volume has become an unspoken rule of the whole industry. It causes the public to doubt the actual amount of Bitcoin owned by each exchange.

To solve the problem of information asymmetry between users and exchanges, we propose a method for tagging Bitcoin addresses of exchanges. Through vertical, forward, and backward address mining, the method can utilize only one or several addresses of an exchange to find out all its addresses and distinguish different address types: deposit wallet, hot wallet, and cold wallet. Then the balance and transfers of the exchange can be further obtained through these addresses, helping users understand the real Bitcoin holdings of the exchange. Several experiments are conducted to evaluate the effectiveness of the proposed Bitcoin address tagging method.

Our method has very little dependence on off-chain information. Only one address is needed for each exchange as a seed to find out all the other addresses. Such a seed address can be easily obtained by depositing some Bitcoin into the exchange or withdrawing some from it, which makes our method feasible for all exchanges.

Dragnet: A Method for Tagging Bitcoin Addresses of Exchanges

Zhen Li*, Yi Zheng†, Qi Li*, Ming Wu† and Kunbin Peng‡
*Chain Info Research, zhen@chain.info
†Qtum Chain Foundation, zhengyi@qtum.info
‡Heilongjiang University of China, 20164698@s.hlju.edu.cn

***Abstract*—Currently, there are hundreds of Bitcoin exchanges on the market, so choosing a reliable exchange is a critical issue for users. We know that the amount of Bitcoin holdings is an essential indicator for evaluating an exchange, but people have very few ways to access this information. Besides, many reports indicate that the trading volumes of most Bitcoin exchanges do not match their real situations, and the fake volume has become an unspoken rule of the whole industry. It causes the public to doubt the actual amount of Bitcoin owned by each exchange.**

To solve the problem of information asymmetry between users and exchanges, we propose a method for tagging Bitcoin addresses of exchanges. Through vertical, forward, and backward address mining, the method can utilize only one or several addresses of an exchange to find out all its addresses and distinguish different address types: deposit wallet, hot wallet, and cold wallet. Then the balance and transfers of the exchange can be further obtained through these addresses, helping users understand the real Bitcoin holdings of the exchange. Several experiments are conducted to evaluate the effectiveness of the proposed Bitcoin address tagging method.

Our method has very little dependence on off-chain information. Only one address is needed for each exchange as a seed to find out all the other addresses. Such a seed address can be easily obtained by depositing some Bitcoin into the exchange or withdrawing some from it, which makes our method feasible for all exchanges.

***Index Terms*—Bitcoin, address, mining, tagging, clustering, exchange**

I. Introduction

Bitcoin exchanges are the platform for users to trade Bitcoin. Currently, there are hundreds of exchanges on the market, so choosing a secure and reliable exchange is a critical issue for users. We know that the amount of Bitcoin holdings is an essential indicator for evaluating an exchange, but people have very few ways to find this information. In April 2019, the Blockchain Transparency Institute (BTI) released a report [1] indicating that 17 of the CoinMarketCap (CMC) top 25 exchanges are found to be over 99%+ fake with many greater than 99.5% fake volumes, including 35 of the top 50 adjusted volume rankings. It means that many exchanges not only hide their Bitcoin holdings but also highly exaggerate their trading volumes to gain the favor of users.

A. Our Contributions

To solve the problem of information asymmetry between users and exchanges, we propose a method for tagging exchange addresses, to better understand the Bitcoin holdings of exchanges. First, based on the transaction structure of Bitcoin, an address mining algorithm is introduced to mine all possible addresses of an exchange from only one or several exchange addresses. Three models are included in the algorithm to establish relationships between addresses in a transaction: vertical, forward, and backward mining. Then, according to the Bitcoin storage and transfer pattern of exchanges, the possible addresses are filtered and classified into three types: deposit wallet, hot wallet, and cold wallet. On this basis, we can further get the balance and transfers of the exchange. Finally, several experiments are conducted to evaluate the effectiveness of the proposed Bitcoin address tagging method.

Our method has very little dependence on the off-chain information, that is, information from the Internet and other places outside the blockchain. Only one address is needed for each exchange as a seed to find out all the other addresses. Such a seed address can be easily obtained by depositing some Bitcoin into the exchange or withdrawing some from it, which makes our method feasible for all exchanges.

B. Related Work

The address tagging problem is solved by address mining and classification in this paper, whereas other existing solutions would like to use tag collection and address clustering to solve it. This is because the address tags of exchanges have a specific collection method (by deposit and withdrawal) and behavior patterns (deposit, cold, and hot wallets), which can be utilized to solve the problem in a better way.

1) Tag Collection: The most common approach [2] [3] [4] to collect address tags is crawling Bitcoin-related websites, such as Bitcointalk, Twitter, and Reddit. Some entities would like to use specific prefixes for their addresses. For example, SatoshiDICE uses “1dice” and LuckyBit uses “1Lucky” [5] prefixes. Therefore, the prefix is another way for tagging addresses [4].

2) Address Clustering: Some works [4] [6] [7] [8] use clustering algorithms to solve the problem of address tagging. All addresses in a cluster are considered to be controlled by the same entity and thus share the same tag. They rely on the interaction of addresses to measure the similarity between them. Afterward, algorithms such as k-means and DBSCAN are utilized for clustering.

II. Address Mining

This section introduces an algorithm for mining all relevant addresses from a single address. It leverages three models to associate input and output addresses in a transaction: vertical, forward, and backward mining. The algorithm is used to find out all possible addresses of an exchange as a rough result that needs to be filtered afterwards.

A. Preliminaries

Bitcoin transactions are based on the UTXO (Unspent Transaction Output) model. Each transaction consists of one or several inputs as well as outputs. There are two parts inside an input or output: address and amount. An input must point to a UTXO, which is an output from a previous transaction and not spent before. The sum of input amounts minus the sum of output amounts equals the transaction fee.

Fig. 1 shows an example of the Bitcoin transaction, where 1 BTC is transferred from the sender to the recipient. The sender uses three UTXOs as inputs and builds two outputs: 1 BTC for the recipient and 0.020956 BTC as the change. The fee for this transaction is 0.001779 BTC.

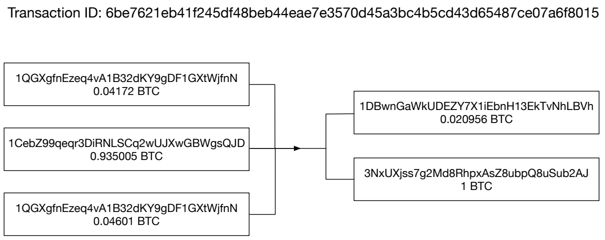

Transaction ID: 6be7621eb41f245df48beb44eae7e3570d45a3bc4b5cd43d65487ce07a6f8015
1QGXgfnEzeq4vA1B32dKY9gDF1GXtWjfnN
0.04172 BTC
1CebZ99qeqr3DiRNLSCq2wUJXwGBWgsQJD
0.935005 BTC
1QGXgfnEzeq4vA1B32dKY9gDF1GXtWjfnN
0.04601 BTC
1DBwnGaWkUDEZY7X1iEbnH13EkTvNhLBVh
0.020956 BTC
3NxUXjss7g2Md8RhpxAsZ8ubpQ8uSub2AJ
1 BTC

Fig. 1. An example of the Bitcoin transaction.

B. Vertical Mining

Heuristic 1: For a transaction, if one of the input addresses belongs to an exchange, the rest belong to the same exchange.

Typically, a Bitcoin transaction is sent by a single user, so the input addresses are all from the sender. Although someone would use the CoinJoin method [9] to combine UTXOs from multiple senders into a single transaction to make it more challenging to determine the relationship between input and output addresses, we detect this method has not been adopted by the exchange so far.

C. Forward Mining

Heuristic 2: For a transaction from an exchange, if it has two outputs while one output amount is an integer and the other is a decimal, the latter output belongs to the same exchange as the input.

This theorem describes a typical payment-and-change transaction as Fig. 1. Since integers are easy to remember and communicate, many payments have integers as their amounts. For instance, many withdrawals are integers since they are easy to type into the form on the website or mobile app of the exchange. In this situation, the change address can be effectively associated with the exchange.

D. Backward Mining

Heuristic 3: For a transaction transferring numerous Bitcoin to an exchange, if all inputs come from the same address, this address also belongs to the exchange.

Such transactions mainly occur during exchanges reorganizing their wallets. Exchanges sometimes move their funds to new addresses for security reasons. They would transfer a large amount of Bitcoin and often need to combine UTXOs from the same address. Backward mining can effectively find those addresses newly created during this process.

E. Mining Process

With the models above, a BFS (Breadth-First Search) algorithm can be used to perform the mining process, as described in Algorithm 1. The search depth is controlled by the parameter m, and setting it to 3 or 4 is enough for most exchanges. If m is too large, many addresses irrelevant to the exchange would be included in the output A.

Algorithm 1: The process of address mining based on BFS

```
Input: An exchange address $a_0$
Output: Possible address set for the exchange
        $A = \{a_0, \dots, a_n\}$
Initialize $A = \{a_0\}$, a queue $Q = [a_0]$ and an empty
  address set $T = \{\}$
for $i = 1$ to $m$ do
    if $Q.empty()$ then
        break
    while $a = Q.pop()$ do
        foreach $t$ in $a.transactions()$ do
            $T.insert(VerticalMining(t))$
            $T.insert(ForwardMining(t))$
            $T.insert(BackwardMining(t))$
    foreach $a$ in $T$ do
        if $A.insert(a)$ then
            $Q.push(a)$
    $T.clear()$
return $A$
```

III. Address Classification

In this section, we first introduce different types of exchange addresses. Then, several classifiers are trained for filtering and classifying the possible addresses from the mining output.

A. Address Types

For most exchanges, addresses can be classified into three types: deposit wallet, cold wallet, and hot wallet. Fig. 2 shows the interaction among them.

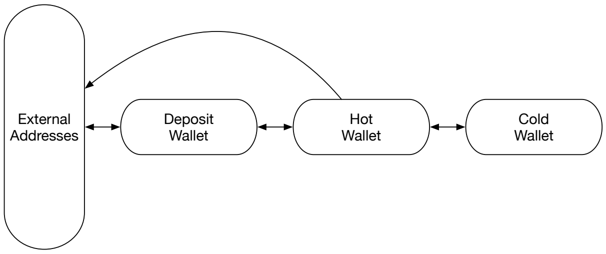


Fig. 2. The interaction between different types of addresses, where the arrow stands for the transfer of Bitcoin.

1) Deposit Wallet: More than 99.9% of the exchange addresses belong to the deposit wallet. It consists of the deposit address for each user, through which the user can deposit Bitcoin from external addresses to the exchange. Bitcoin in the deposit wallet would be transferred to two places: external addresses (when withdrawing Bitcoin) and the hot wallet.

2) Hot Wallet: The hot wallet is responsible for the transfers of Bitcoin. It is involved in almost all transactions created by the exchange as an input or an output. There are only 10 to 30 addresses in the hot wallet, but the number of transactions is enormous. This feature can be used to distinguish the hot wallet from other addresses effectively.

3) Cold Wallet: The exchange stores most of its Bitcoin in the cold wallet. The cold wallet needs to stay offline for the sake of security. It only interacts with the hot wallet. For the cold wallet, the number of transactions is small, but the amount of a single transaction is relatively large (usually more than 100 BTC).

B. Training Classifiers

Addresses in different wallets have different features. These features can be used to build classifiers for distinguishing address types and filtering out addresses that do not belong to the exchange. We list all these features in Table I. Besides, for each wallet, we have selected some sample addresses to observe the values of these features. They are also shown in the table below.

The feature value here refers to the value after normalization. We use a variant of the min-max normalization to rescale the range of values to $[0, 1]$. The formula for the normalization is given as:

$$x' = \frac{x - x_{min,e}}{x_{max,e} - x_{min,e}}$$

where x is the original value, and x' is the normalized value. e is the exchange of the current address. $x_{max,e}$ and $x_{min,e}$ stand for the max and min value of x for all possible addresses in exchange e. The reason for this design is that the range of values differs for each exchange. We need to reduce the impact of such differences on the classifier.

For the hot and cold wallets, classifiers can be built by predefined decision trees with part of the features as well as their thresholds. This is because, on the one hand, there are not many samples for them. On the other hand, we observe that they can be easily distinguished by some features, such as the total transaction count and the balance.

TABLE I
ADDRESS FEATURES FOR BUILDING CLASSIFIERS

Feature Description	**Feature Value**		
	Deposit	**Hot**	**Cold**
UTXO count	Low	High	Mid
Balance	Low	Mid	High
Total Received	Low	High	High
Total Sent	Low	High	High
Total transaction count	Low	High	Low
Average transaction count per block	Low	High	Low
Average transaction interval	High	Low	High
Total transaction amount	Low	High	Mid
Average transaction amount per block	Low	High	Mid
Average transaction amount	Low	High	Mid
Total input address count	Low	High	Low
Input address count per transaction	Low	High	Low
Input address count per block	Low	High	Low
Total output address count	Mid	High	Low
Output address count per transaction	Mid	High	Low
Output address count per block	Mid	High	Low
Total received from the hot wallet	Low	High	High
Total sent to the hot wallet	Low	High	High
Total received from the cold wallet	Low	High	High
Total sent to the cold wallet	Low	High	High

For the deposit wallet, we can use manually selected samples to train a classifier based on any machine learning model, such as the Logistic Regression model. The output of the Logistic Regression model is expressed by:

$$y = f\left(\mathbf{W}^T\mathbf{x}\right) = f\left(\sum_{i=0}^{n-1} w_i x_i + w_n\right)$$

where

$$f(z) = \frac{1}{1 + e^{-z}}$$

Here $\mathbf{x} = (x_0, \ldots, x_{n-1}, 1)$ represents n feature values and $\mathbf{W} = (w_0, \ldots, w_{n-1}, w_n)$ are $n + 1$ model parameters obtained through the training process. If an address has a $y > 0.5$, the address is classified as the type of deposit wallet. More details about the Logistic Regression model and other machine learning models are beyond the scope of this paper. All that matters for the purposes here is that they can be used to find out the address that belongs to the deposit wallet and filter out the remaining addresses.

IV. EXPERIMENTAL EVALUATION

We have successfully tagged the addresses of 10 famous Bitcoin exchanges. This section first introduces the experimental setup. Then the tagging results of two exchanges are presented as examples to evaluate our method. At last, we show the balances and transfers of all these exchanges through a table.

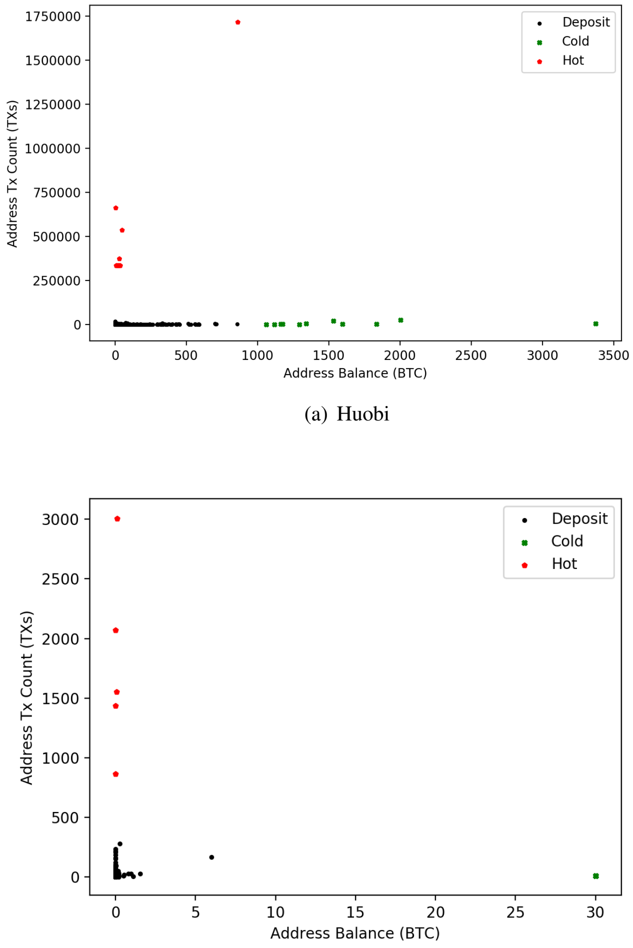
A. Experimental Setup

The search depth m in the address mining process is set to 3. The classifier for the hot wallet uses the feature of total transaction count, and the threshold is set to 0.14. Two features, balance and sent to the hot wallet, are used for the classifier of the cold wallet, and their thresholds are set to 0.3 and 0.2.

For the classifier of the deposit wallet, we have randomly selected 20000 possible addresses from Binance and manually tagged them. 12.32% of them are negative samples. 10000 of them are used for training and 10000 for testing. We use the $LogisticRegression$ module from the Python library “scikit-learn” as the classifier, with all parameters remaining default. The classification accuracy proved to be 96.12%.

B. Tagging Results

The tagging results of Huobi and Biss are presented in Fig. 3. We choose two dimensions to show their addresses: balance and transaction count. Addresses with different types are plotted in different colors. It can be seen that different types of addresses can be clearly distinguished from the figure. The hot wallet has far more transactions, and the cold wallet has higher balances than others. We can also see from the figure that Huobi is more popular than Biss because of more addresses and balances.



(a) Huobi

(b) Biss

Fig. 3. Tagging results of Huobi and Biss.

C. Exchange Information

Balances and transfers of these exchanges are further obtained through their tagged addresses, as shown in Table II. These results are calculated based on the block height of 600783, on October 24, 2019, Beijing time. The term ”24 Net Income” stands for the change of balance compared to 24 hours ago, which is useful for analyzing fund flows.

TABLE II
BALANCES AND TRANSFERS OF EXCHANGES

Exchange	Balance (BTC)	24h Net Income (BTC)	Addresses
Huobi	304,827.38	+9,115.3	597,298
Binance	249,050.34	+4,715.8	1,856,352
Bitfinex	138,629.06	+1,325.86	646,259
Bittrex	118,881.93	−295.48	1,246,496
Bitstamp	112,903.15	−259.62	376,684
OKEX	11,414.01	−4,908.74	227,706
Poloniex	5,658.45	+304.29	491,554
Bitflyer	3,359.68	−123.08	157,082
BTC.top	157.22	−57.97	883
Biss	41.61	+0.07	5,508

Exchanges are sorted by their balances. You can see that their rankings are consistent with their business scale, as we know. The difference in balances is not apparent for big exchanges, but the difference between big and small exchanges is enormous. The total balance of these exchanges is about 1 million, accounting for $1/18$ of the Bitcoin circulation.

The total net income of these exchanges is positive, which means Bitcoin is flowing into exchanges. For this observation period, the price of Bitcoin experienced a 10% drop over the past 24 hours. These two events are supposed to have a close relationship: more Bitcoin is deposited into the exchange and sold out, causing a drop in price. It hints that Bitcoin price movement can potentially be predicted through the net income of exchanges.

The number of exchange addresses can reflect the number of users in the exchange. We can see that, for some exchanges, the address count is above 1 million. These exchanges may have more users than others.

V. CONCLUSION

This paper proposes a method for tagging Bitcoin addresses of exchanges. It can be used to help users better understand the real information of the exchange, like the balance, net income, etc. Only one address is needed for each exchange to find out all the other addresses. So this method has very little dependence on the off-chain information and works for most exchanges.

We have successfully applied our method to 10 famous exchanges. But there are some exceptions, for example, Coinbase. We have observed that the behavior pattern of Coinbase is different from other exchanges. It does not follow the model of deposit, hot, and cold wallets, so our method cannot work. Maybe Coinbase does it deliberately to avoid third parties tracking their funds. We plan to study the pattern of Coinbase in our future work.

References

[1] Blockchain Transparency Institute, "Market surveillance report – April 2019," https://www.bti.live/reports-april2019/.

[2] blockchain.com, "Address tags," https://www.blockchain.com/btc/tags.

[3] M. Fleder, M. S. Kester, and S. Pillai, "Bitcoin transaction graph analysis," *arXiv preprint arXiv:1502.01657*, 2015.

[4] D. Ermilov, M. Panov, and Y. Yanovich, "Automatic bitcoin address clustering," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 461–466.

[5] blockchain.com, "Popular addresses," https://www.blockchain.com/btc/popular-addresses.

[6] M. Harrigan and C. Fretter, "The unreasonable effectiveness of address clustering," in *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. IEEE, 2016, pp. 368–373.

[7] B. Huang, Z. Liu, J. Chen, A. Liu, Q. Liu, and Q. He, "Behavior pattern clustering in blockchain networks," *Multimedia Tools and Applications*, vol. 76, no. 19, pp. 20 099–20 110, 2017.

[8] S. S. Chawathe, "Clustering blockchain data," in *Clustering Methods for Big Data Analytics*. Springer, 2019, pp. 43–72.

[9] G. Maxwell, "CoinJoin: Bitcoin privacy for the real world," https://bitcointalk.org/?topic=279249.