

Address Usage Estimation Based on Bitcoin Traffic Behavior

Hiroki Matsumoto, Shusei Igaki, Hiroaki Kikuchi

Abstract This paper studies bitcoin address usage, which is assumed to be hidden via address pseudonyms. Transaction anonymity is ensured by means of bitcoin addresses, leading to abuse for illegitimate purposes, e.g., payments of illegal drugs, ransom, fraud, and money laundering. Although all the transactions are available in the bitcoin system, it is not trivial to determine the usage of addresses. This work aims to estimate typical usages of bitcoin transactions based on transaction features. With the decision tree learning algorithm, the proposed algorithm classifies a set of unknown addresses into seven classes; provider addresses of three services for mining pool, Bitcoin ATM, and dark websites; and user addresses of four services for mining Bitcoin ATM, dark websites, exchange, and a bulletin board system. The experimental results reveal some useful characteristics of bitcoin traffic, including statistics of frequency, amount of value, and significant transaction features.

1 Introduction

Bitcoin is one of the best-known cryptocurrencies and was proposed by Satoshi Nakamoto in 2008[1]. Bitcoin is not issued by a central bank approved by a government or any single organization. Instead, it is issued by a global collaboration of distributed payment nodes linked in a peer-to-peer network architecture. One of

Hiroki Matsumoto,
Graduate School of Advanced Mathematical Sciences, Meiji University, 4-21-1 Nakano Tokyo
Japan, e-mail: cs192026@meiji.ac.jp

Shusei Igaki,
School of Interdisciplinary Mathematical Sciences, Meiji University, 4-21-1 Nakano Tokyo Japan,
e-mail: ev50516@meiji.ac.jp

Hiroaki Kikuchi,
School of Interdisciplinary Mathematical Sciences, Meiji University, 4-21-1 Nakano Tokyo Japan,
e-mail: kikn@meiji.ac.jp

the important features of bitcoin is anonymity. A bitcoin user has pseudonyms for addresses for sending and receiving bitcoins so that it is difficult to track owners with their address, which explains why bitcoin is widely assumed to achieve a high degree of anonymity.

However, many researchers claim that the anonymity of bitcoin is limited and that some heuristic approaches allow some addresses owned by a common owner to be linked. For example, Meiklejohn et al. described a heuristic approach showing that multiple addresses belonging to the same transaction are likely to be controlled by the owner who knows both corresponding private keys[2]. Ron and Shamir studied a bitcoin transaction graph and proposed a specific transaction behavior that allows unique users to be identified[3].

In addition to the linkage threat, some researchers claim that a bitcoin pseudonym is not strong enough to preserve user privacy. For example, the location where an individual moves is a privacy information but is not personal identifiable information. Dupont and Squicciarini presented a statistical method based on a distribution of transaction time to predict a time zone where a user lives[4]. Nagata et al. showed that an owner of a given address can be estimated based on the statistical property of a set of output addresses that the target user sent out previously[5].

In this work, we study a new type of private information disclosure from bitcoin transactions. We focus on the usage of bitcoin because the behavior of the bitcoin address depends much on its usage. For example, the number of transactions per day varies widely with business entities and consumers. Hence, we classify bitcoin addresses into two classes, namely service providers and users. The type of service, e.g., bitcoin exchanges and websites, is also important for distinguishing addresses.

To date, the difference between business providers and end users has not been considered in previous research, despite it being significant information to identify the usage of addresses. Therefore, in this study, we aim to explore the hypothesis that this difference can be used to estimate the usage of addresses.

To conduct an experiment to estimate the usage of bitcoin addresses, we collected 4,049 bitcoin addresses from seven major usages that are well-known services. For each service, we divide the addresses into providers and users. We propose an algorithm that classifies a set of unknown addresses into seven classes by using a decision tree learning algorithm. Based on the results of our analysis, we discuss some potential risks of revealing the usage of addresses from the characteristics of transactions including the target address.

The main contributions of this work are as follows.

- We propose *new features* to distinguish between service providers and bitcoin users based on the statistics of transaction patterns.
- We present *an algorithm to estimate the usage* of unknown addresses using a decision tree learning algorithm.
- We show *the experimental results* using 4,000 bitcoin addresses labeled for seven usages and the accuracy of the proposed method.

The structure of the paper is as follows. Section 2 describes the data used in this work. The methodology is presented in Section 3 and Section 4 presents an

Table 1 Statistics of bitcoin address dataset

| usage | # addresses | | transactions | duration |
|-----------------|-------------|-------|--------------|-----------------|
| | provider | user | | |
| Bitcointalk BBS | \ | 2,391 | 29,638 | 2019/4/1 - 9/30 |
| Bitcoin ATM | 3 | 452 | 26,849 | |
| Dark web | 26 | 67 | 35,076 | |
| Exchange | \ | 1,012 | 33,351 | |
| Mining pool | 98 | \ | 24,876 | |
| total | 4,049 | \ | 149,790 | |

overview of our approach. We conclude our work and briefly discuss future research in Section 5.

2 Data

2.1 Definition of seven Bitcoin services

In this section, we first present seven *usages* of bitcoin addresses, e.g., ATM, exchange, and mining pool. We collected all transactions that were published by *Blockchain Explorer*[6] from April 1 through September 30, 2019.

Table 1 shows the statistics of our dataset used in this research. In Table 1, we classified bitcoin addresses into two classes, namely service providers and users for each service. However, we do not distinguish the kinds of transactions between the service providers and users. For example, for Bitcoin ATM, we have three addresses for the providers and 452 addresses for the users and 26,849 transactions made by both of them. In addition, we exclude duplicated addresses that were used for more than one usage. For example, some addresses were used as Bitcointalk transactions and as transactions with exchanges.

For each of the seven usages, we classified bitcoin addresses into two subsets: those owned by a commercial service provider and those of users. The category of *provider* uses bitcoin for pamarchanymment of commercial services and goods. The category of *user* uses bitcoin for purchasing goods and services and making investments.

2.2 BBS Bitcointalk

Bitcointalk[7] is a bulletin board system (BBS) service for discussion on cryptocurrencies, including bitcoin.

Fig. 1 shows an example of a Bitcointalk profile page of a registered user. One possible reason why Bitcointalk users publish their bitcoin addresses is to receive

| Summary - FlightyPouch | Picture/Text |
|---|--------------|
| Name: FlightyPouch Posts: 3378 Activity: 1232 Merit: 287 Position: Sr. Member Date Registered: October 11, 2016, 02:15:03 PM Last Active: Today at 12:37:36 AM <hr/> ICQ: AIM: MSN: YIM: Email: hidden Website: Current Status: <input type="checkbox"/> Offline Bitcoin address: 3PyrwHe7oDdk739x78n1sUvDnadJl4fmSB <hr/> Gender: Age: N/A Location: 0x6B3A0003A273A8Bcf061cD3a611277Bec8810EDb Local Time: February 14, 2020, 07:34:55 AM <hr/> Signature: <small>Fast: % Dice Rakeback YOLOdice.com Competitions Exchange</small> <small>— BTC LTC ETH DOGE</small> <hr/> Additional Information: <small>Show the last posts of this person.</small> <small>Show the last topics started by this person.</small> <small>Show general statistics for this member.</small> | |

Fig. 1 Sample Bitcointalk user profile



Fig. 2 Bitcoin ATM machine in Toronto, Canada

donations in return for answering questions posted in the BBS. In this work, every bitcoin address that has been published in the profile pages is assumed to be the user address.

2.3 Bitcoin ATM

Bitcoin ATM[8] is a bitcoin deposit service.

Fig. 2 shows an example of a Bitcoin ATM machine. In this system, customers input their bitcoin address (public key information) via a QR code to an ATM and specify the amount of money they want to deposit in their wallet. Then, the Bitcoin ATM sends the equivalent bitcoin to the customer's address. In this work, we collected three addresses of Bitcoin ATMs in Toronto, Canada. Both users and providers are involved in the usage of a Bitcoin ATM. In a Bitcoin ATM transaction, the user's address is the recipient, while the Bitcoin ATM provider is the sender.

2.4 Dark web

The *dark web* is a website in the TOR network and has a high degree of anonymity.

Fig. 3 shows an example of the dark web. We collected bitcoin addresses published on sites that are accessible through the TOR browser. We found the address for a service provider for hacking Facebook accounts in the example. Similarly, we collected addresses from illicit services for a shop with credit card numbers. For a



Hack Facebook Account

We sell the cheapest and most reliable Facebook hacking service on the deep web.

Price per account: **0.01 BTC**

Bitcoin address for making deposit: 1MfUge8xL7hpRFDshMstUPQKvhcGcSLvJ

How does it work?

Deposit 0.01 BTC to the address above and send us an e-mail to fbstaller@torbox2.usb6wchz.onion with the victim's facebook profile url (<https://www.facebook.com/USERNAME>) and exact time of when you sent the Bitcoin so we can verify it with the blockchain. We will send you the account login info within 48 hours. 100% Money back guarantee.

FAQ

How do I send you an email?

Fig. 3 Dark website and bitcoin provider's address ¹

Table 2 List of addresses of exchanges

| exchange | # addresses |
|-----------------------|-------------|
| AnxPro.com | 4 |
| BitBay.net | 13 |
| Bitstamp.net | 40 |
| Bittrex.com | 116 |
| CoinHako.com | 2 |
| HappyCoins.com | 1 |
| Hashnest.com | 199 |
| HitBtc.com | 89 |
| Kraken.com | 26 |
| MercadoBitcoin.com.br | 130 |
| OKCoin.com | 1 |
| Poloniex.com | 110 |
| YoBit.net | 281 |

dark web user's address, we collected customer addresses published from the dark website for their promotion.

2.5 Exchange

Exchange allows their customers to trade bitcoins for fiat currencies. We collected exchange addresses from *WalletExplorer*[9] in which bitcoin addresses are classified into various categories, e.g., exchanges, pools, and gambling.

Table 2 shows the list of exchanges. We collected exchange user's addresses that have been specified in any transactions with known exchange addresses labeled by *WalletExplorer*.

¹ <http://r3cnefrmwctd6gb2.onion>

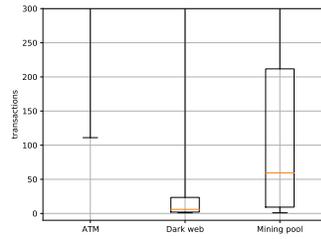


Fig. 4 Number of transactions used for bitcoin providers

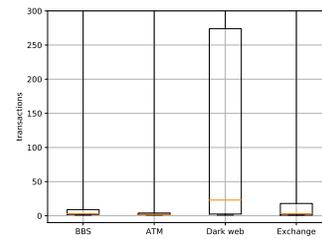


Fig. 5 Number of transactions used for bitcoin user addresses

Table 3 Number of transactions (TXs) used by bitcoin providers

| usage | Avg. TXs | Min. TXs | Median TXs | Max. TXs | SD. TXs |
|-------------|----------|----------|------------|----------|---------|
| Bitcoin ATM | 7,551 | 111 | 549 | 21,993 | 12,509 |
| Dark web | 74 | 1 | 6 | 1,272 | 250 |
| Mining pool | 271 | 1 | 60 | 4,190 | 668 |

2.6 Mining pool

Mining pool is composed of distributed miners who share their processing power over a mining network. Creating a new block is called mining, and requires a large amount of computational resources. In this work, we collected mining pool providers' addresses from bitcoin blocks in which a reward was provided.

3 Proposed Method

3.1 Transactions

3.1.1 Characteristics of service providers

Table 3 describes the statistics of providers' addresses that were specified in transactions from April 1 through September 30, 2019. Fig. 4 shows the bar plots of the number of transactions for the providers, ATM, dark web, and mining pool. Note that the Bitcoin ATM transactions are not well distributed because there were only three addresses observed in our study (see Table 3).

Table 4 Number of transactions (TXs) used by bitcoin addresses classified as users

| usage | Avg. TXs | Min. TXs | Median TXs | Max. TXs | SD. TXs |
|-----------------|----------|----------|------------|----------|---------|
| Bitcointalk BBS | 13 | 1 | 3 | 722 | 42 |
| Bitcoin ATM | 12 | 1 | 2 | 383 | 34 |
| Dark web | 503 | 1 | 23 | 7,482 | 1,228 |
| Exchange | 45 | 1 | 3 | 4,582 | 239 |

3.1.2 Characteristics of users

Table 4 describes the statistics of providers' addresses that were used in transactions from April 1 through September 30, 2019. We found that a few addresses were specified many times in the dark web, and exchanges have a small number of transactions on average. Fig. 5 shows the number of transactions that were made by bitcoin providers. Three-quarters of bitcoin users made fewer than 25 transactions in the usage of Bitcointalk BBS, Bitcoin ATM, and exchange, as shown in Fig. 5.

3.2 Proposed estimation method

3.2.1 Decision tree learning

In this work, we chose a decision tree learning algorithm to classify a set of unknown addresses into seven classes of usage because it is simple and sufficiently accurate for our purpose. To classify a set of unknown addresses, we used the CART algorithm implemented in Python with a Scikit-learn library. We performed threefold cross-validation to evaluate the accuracy of classification for avoiding distortion because of the lack of known addresses. After we had randomly sampled the dataset for 100 iterations, we estimated the usage of given addresses and evaluated the accuracy of the model in precision and recall.

3.2.2 Features of transaction patterns

For the analysis of bitcoin transaction data, we explore the features of address usage. Noting that a transaction depends on the wallet application, we try to define features of some usages. For example, the wallet BitPay creates a new address to receive change when it sends bitcoin. We count the frequencies of the four patterns to create the feature of four element vectors and apply the decision tree learning algorithm. See Fig. 6 and Table 5. The transaction pattern $S1$ is basic. The sender with address A_1 pays some amount of money to B_1 . He/she does not send the whole bitcoin charged to A_1 but specifies a part of the amount of it for B_1 and sends the change back to A_1 . The second pattern $S2$ is the same as $S1$ except the change is sent to an alternative address (say C_1) rather than back to A_1 again. The third and fourth

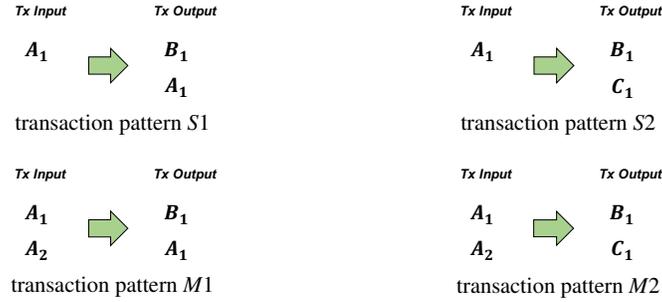


Fig. 6 Sample transaction pattern

Table 5 Definition of transaction pattern

| | # input addresses | change | description |
|----|-------------------|----------------------------|---|
| S1 | 1 | sent back to input address | basic transaction, deposit bitcoin with Bitcoin ATM |
| S2 | | new address | specific wallet applications |
| M1 | more than 1 | sent back to input address | withdraw bitcoin in exchange |
| M2 | | new address | mining pool provider pays a mining reward to miners |

patterns $M1$ and $M2$ have multiple input addresses. For example, addresses A_1 and A_2 are specified in both transactions, meaning a transfer of the sum of values of bitcoin charged to A_1 and A_2 to output addresses. Any input address is specified again at output addresses for pattern $M1$; no input addresses are used again at the outputs for pattern $M2$.

We classify all transactions into four patterns based on the number of input addresses and whether the same input address is reused to receive change. More specifically, we define patterns $S1$ and $S2$ as transactions that have a single input address, while more than or equal to two addresses are specified in patterns $M1$ and $M2$. The difference between patterns $S1$ and $M1$ is whether any of the input addresses are specified at the output to receive change. The same difference is defined for patterns $S2$ and $M2$.

A sender receives change when specifying his/her own address at the output address in a transaction. In pattern $S1$ ($M1$), a sender receives change by reusing address A_1 at the output of the transaction, as shown in Fig. 6. In contrast, patterns $S2$ and $M2$ do not use the same input address again. Note that, in this work, we assume that a sender does not receive any change even when he/she owns a new address to receive the change.

In addition to this feature, we quantify additional features shown in Table 6. Note that some features are described by a number of statistics in Table 6, such as average, minimum, maximum, median, and standard deviation.

Table 6 List of original variables in the dataset

| feature | # statistics | description |
|-----------------------------|--------------|---|
| TXs count | 5 | Total number of transactions for usages |
| TXs sending count | 5 | Total number of sending transactions for usages |
| TXs receiving count | 5 | Total number of receiving transactions for usages |
| TXs input address count | 5 | Total number of input addresses specified in transaction |
| TXs output address count | 5 | Total number of output addresses specified in transaction |
| TXs address count | 1 | Total number of addresses in transaction |
| Reused input address count | 1 | Total number of reused input addresses |
| Reused output address count | 1 | Total number of reused output addresses |

Table 7 Total number of transactions of seven usages

| usage | | transaction pattern | | | | pattern[%] | | | |
|-----------------|----------|---------------------|--------|-------|--------|------------|------|-----|------|
| | | S1 | S2 | M1 | M2 | S1 | S2 | M1 | M2 |
| Bitcoin ATM | provider | 22,319 | 135 | 174 | 25 | 98.5 | 0.6 | 0.8 | 0.1 |
| Dark web | | 1,242 | 557 | 3 | 127 | 64.4 | 28.9 | 0.2 | 6.6 |
| Mining pool | | 19,569 | 2,845 | 410 | 2,052 | 78.7 | 11.4 | 0.2 | 6.6 |
| Bitcointalk BBS | user | 6,978 | 10,704 | 1,478 | 10,478 | 23.5 | 36.1 | 5.0 | 35.4 |
| Bitcoin ATM | | 1,700 | 2,033 | 44 | 1,323 | 33.3 | 39.9 | 0.9 | 25.9 |
| Dark web | | 7,627 | 12,546 | 1,264 | 11,711 | 23.0 | 37.8 | 3.8 | 35.3 |
| Exchange | | 8,730 | 11,269 | 2,908 | 10,444 | 26.2 | 33.8 | 8.7 | 31.3 |

4 Experiment

4.1 Transactions examined

Table 7 shows the transactions summarized for the four patterns. Note that we distinguished the different kinds of transactions between service providers and users (see Table 1). This means that we classified bitcoin addresses into seven classes so that the number of transactions in Table 7 is larger than the number of transactions in Table 1.

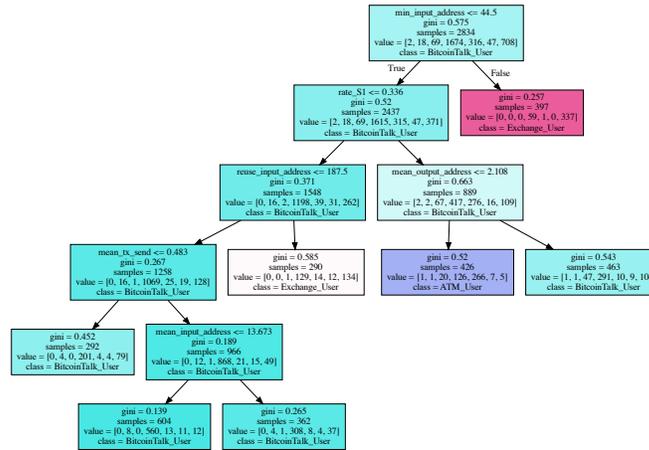
4.2 Results of classification

Table 8 shows the experimental results. Fig. 7 shows a sample decision tree that was generated by the learning algorithm. We performed pruning of this model so that the highest depth is 5 and no minor node consists of 10% of all instances.

Table 9 shows the estimated usages with the decision tree learning algorithm. Our model cannot estimate usages with ATM providers and dark web providers (see Table 9). These results indicate that Bitcointalk was detected as a false-positive

Table 8 Experimental results of estimation

| usage | accuracy[%] | | precision[%] | | recall[%] | |
|-----------------|-------------|------|--------------|------|-----------|------|
| | provider | user | provider | user | provider | user |
| Bitcointalk BBS | \ | 77 | \ | 65 | \ | 63 |
| Bitcoin ATM | 99 | 91 | 16 | 45 | 22 | 40 |
| Dark web | 98 | 93 | 6 | 49 | 9 | 36 |
| Exchange | \ | 85 | \ | 80 | \ | 79 |
| Mining pool | 92 | \ | 70 | \ | 65 | \ |
| total | 81 | | 49 | | 39 | |

**Fig. 7** Prediction model with the decision tree learning algorithm

in 112 addresses, and it is the most frequent in the set of usages. However, the estimated results of Bitcointalk were 88%, which is the highest score in the seven classes.

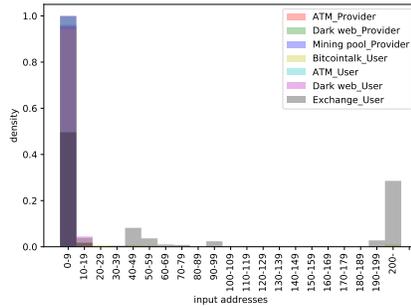
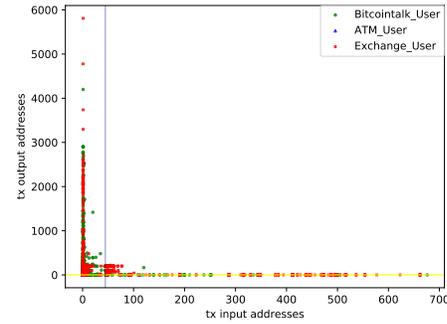
Fig. 8 shows a histogram of the features of the number of minimum input addresses. Fig. 9 illustrates the distribution of the top three usages, i.e., Bitcointalk BBS users, ATM users, and exchange users, in the scatterplot of numbers of input (x-axis) and output (y-axis) addresses in transactions. Table 10 shows the statistics of the features of the number of minimum input addresses.

4.3 Discussion

Addresses used as exchange have the highest recall and precision in the seven classes, which might be explained by the number of minimum input addresses being much larger than any other usage.

Table 9 Estimated usages with the decision tree learning algorithm

| usage | | ATM | Dark web | Mining | BBS | ATM | Dark web | Exchange | total |
|-----------------|----------|----------|----------|--------|------|-----|----------|----------|-------|
| | | provider | | | user | | | | |
| Bitcoin ATM | provider | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Dark web | | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 8 |
| Mining pool | | 0 | 0 | 2 | 19 | 8 | 0 | 0 | 29 |
| Bitcointalk BBS | user | 0 | 0 | 0 | 633 | 31 | 0 | 53 | 717 |
| Bitcoin ATM | | 0 | 0 | 0 | 16 | 119 | 0 | 1 | 136 |
| Dark web | | 0 | 0 | 0 | 12 | 3 | 2 | 3 | 20 |
| Exchange | | 0 | 0 | 0 | 56 | 9 | 0 | 239 | 304 |

**Fig. 8** Histogram of features of the number of minimum input addresses in the seven usages**Fig. 9** Distribution of the top three usages: BBS users, ATM users, and exchange users**Table 10** Number of addresses indicating the number of minimum input addresses in the seven usages

| usage | | Avg. | Min. | Median | Max. | SD. |
|-----------------|----------|-------|------|--------|------|------|
| Bitcoin ATM | provider | 1 | 1 | 1 | 1 | 0 |
| Dark web | | 1.9 | 1 | 1 | 17 | 3.2 |
| Mining pool | | 1 | 1 | 1 | 1 | 0 |
| Bitcointalk BBS | user | 7 | 1 | 1 | 676 | 40.1 |
| Bitcoin ATM | | 1.3 | 1 | 1 | 112 | 5.2 |
| Dark web | | 1.7 | 1 | 1 | 12 | 2.3 |
| Exchange | | 137.9 | 1 | 10.5 | 662 | 190 |

Table 7 shows that the usage of a Bitcointalk user has almost the same number of transactions for each pattern. We consider that the transactions of Bitcointalk have various patterns because they use bitcoin for various purposes. Therefore, Bitcointalk has need to be defined more specific classifications, such as merchant services, donations, and the others.

Transaction patterns $S1$ and $S2$ with providers (ATM, dark web, and mining pool) occupy more than 90% of total addresses, one reason for which is likely that they reused their fixed addresses many times.

5 Conclusions

In this work, we have proposed a new method to classify a set of unknown bitcoin addresses into seven usages of three services (mining pool, Bitcoin ATM, and dark websites) and four user addresses (Bitcoin ATM, dark websites, exchange, and BBS). Our results show that the clearest usage is the exchange service with 80% precision and 79% recall using the decision tree learning algorithm.

Based on the results of our analysis, we have confirmed that our proposed features distinguishes between service providers and bitcoin users based on the statistics of transaction patterns. Our proposed the algorithm estimates precisely the usage of unknown addresses using a decision tree learning algorithm. Finally, we showed the experimental results using 4,000 bitcoin addresses labeled for seven usages and the accuracy of the proposed method.

In future work, we plan to resolve the problem of skew in our dataset, which contains unbalanced usages.

References

1. Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System", <https://bitcoin.org/bitcoin.pdf>
2. Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M. Voelker, Stefan Savage, "A Fistful of Bitcoins: Characterizing Payments Among Men with No Names", [OLE7] In Proceedings of the 2013 Conference on Internet Measurement Conference (IMC'13), pp. 127–140, 2013.
3. Dorit Ron, Adi Shamir, "Quantitative Analysis of the Full Bitcoin Transaction Graph", Financial Cryptography and Data Security (FC 2013), pp. 6–24, 2013.
4. Jules Dupont, Anna C. Squicciarini, "Toward De-Anonymizing Bitcoin by Mapping Users Location", In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy (CODASPY '15), pp. 139–141, 2015.
5. Kodai Nagata, Hiroaki Kikuchi, Chun-I Fan, "Risk of Bitcoin Addresses to Be Identified from Features of Output Addresses", The 2018 IEEE Conference on Dependable and Secure Computing (DSC 2018) Workshop #4, pp. 349–354, 2018.
6. Blockchain Explorer (<https://www.blockchain.com/ja/explorer>)
7. Bitcointalk (<https://bitcointalk.org/>)
8. Coin ATM Radar Bitcoin ATM Map (<https://coinatmradar.com/>), 2020.
9. WalletExplorer.com (<https://www.walletexplorer.com/>)