Risk of Bitcoin Addresses to be Identified from Features of Output Addresses

Kodai Nagata
Graduate School of Advanced Mathematical Sciences
Meiji University, Japan.

Hiroaki Kikuchi
School of Interdisciplinary Mathematical Sciences
Meiji University, Japan.

***Abstract*—Bitcoin is a digital currency that aims to offer anonymity. However, anonymity is less secure than it is believed to be because it is based on using pseudonyms for addresses. To demonstrate this weakness, we used statistics related to transaction histories and the frequency of transactions to deanonymize a set of bitcoin addresses. In this paper, we explore the fundamental properties of Bitcoin addresses based on actual Bitcoin transaction data. We propose a new method for deanonymizing Bitcoin addresses from a set of output addresses and we demonstrate that 80.5% of addresses could be identified.**

***Index Terms*—Bitcoin, deanonymize, output address**

I. Introduction

Launched in 2009, Bitcoin is one of the oldest digital currencies in use. Because transactions occur without the help of any third party, Bitcoin appears to preserve a high degree of anonymity. However, several studies have described the deanonymization of bitcoin addresses. Meiklejohn et al. [1] combined addresses managed by the same user to demonstrate the extraction of specific transaction patterns for individual users. Dupont et al. [2] succeeded in revealing a target user's time zone by analyzing the time distribution of transactions. The accuracy of their proposed scheme was only moderate, with an average identification ratio of 72%. The main reason for errors was that different users may have similar lifestyle patterns. Moreover, it is not clear how much anonymity depends on the frequency of transactions and the output address sets used.

To address this issue, we focus on the addresses specified in transactions because the statistical properties of an address's usage, such as its mean frequency, duration, and set of output addresses, can help to identify the owner of the address. We assume that the set of output addresses used by a particular user will be very stable over time and therefore using this set of addresses can help trace the user. The purpose of this study is to investigate the proportion of addresses at risk of deanonymization by an analysis of the frequency of addresses used in transactions and the output set of addresses.
We conducted an experiment on deanonymizing Bitcoin addresses based on a large-scale dataset of transactions collected from the Bitcoin blockchain over 1.5 years since 2012.

In this paper, our contributions are as follows.

1) We propose a new method for deanonymizing Bitcoin users from the set of output addresses used.
2) We evaluate the anonymity of the addresses. One of the main results of our experiment is that 559 Bitcoin users were uniquely identified from the addresses sent from the same source address. Compared to other work using time-zone signatures [2], our proposed scheme showed improved accuracy, with 80.5% of addresses being identified.

II. Bitcoin

A. Background

Bitcoin is a digital currency that operates without a central manager such as a government. Instead of a central trusted party, a majority of the Bitcoin users approve the transactions issued within a predetermined interval. Data about Bitcoin transactions are stored in a blockchain. A new block is generated about every ten minutes and is contained in a blockchain. The mechanism for ensuring the integrity of transactions is called decentralized management.

B. Addresses

A Bitcoin address is a hash value of the form "1A1zP1eP5QGefi2DMPTfTL5SLmv7DivfNa." A Bitcoin transaction specifies the addresses of the sender and the recipient. An address is specified by applying two secure hash functions SHA256 and RIPEMD160 to a target public key and appending a checksum encoded using Base58. It is known that it is impossible to invert the address generation process to obtain the public key from the address. Despite this, it should be noted that the process is deterministic. Therefore, the owner can be identified easily even if the corresponding public key is unknown. To prevent the owner being identified via the address, a Bitcoin wallet allows a user to have multiple addresses.

C. Transactions

Table I shows the data format for the Bitcoin-transaction data structure. A Bitcoin transaction contains input and output fields specifying the sender's address(es) and the receiver's address(es), respectively.Note that multiple addresses may be specified for either or both fields.
For example, Tx_4 in Table II involves three inputs and two outputs.

TABLE I: Transaction data structure

Field	Explanation
Version	Version of transaction
Input Counter	The number of inputs
Inputs	Input data
Output Counter	The number of outputs
Outputs	Output data
LockTime	Block height or Unix timestamp

TABLE II: Transactional information in a block

ID	Input	Output	Remittance [10^{-8}]
Tx_1	N/A	a_2	2500000000
Tx_2	a_2	a_4	900000
Tx_3	a_3	a_2,a_3	60000000
Tx_4	a_2,a_2,a_5	a_1,a_2	110000000
Tx_5	a_3	a_1,a_2,a_3,a_5	40000000

The specification of output addresses depends on the input addresses. Consider the example block of five transactions Tx_1, Tx_2, Tx_3, Tx_4, and Tx_5 in Table II. Addresses a_2 and a_3 are specified as inputs in the two transactions Tx_2 and Tx_4, whereas a_3 is specified as an input in the two transactions Tx_3 and Tx_5. Address a_5 is specified as an input in the one transaction Tx_4.

III. Data Collection

In this section, we describe the method used to collect the transactional data and the address data used in our experiments.

A. Transactional Data

To collect the data, we used a *bitcoind* client, which downloads all blocks before they are ready to be used online. We collected transaction from the client datastore, implemented as an SQLite3 database. The database contains an Input table (see Table III) and an Output table (see IV). Both Input and Output tables have five attributes, including Time, Height, and Addresses. Note that a block is divided into several records classified in terms of the two tables. For example, the records in Table III and IV were retrieved from the same block, as specified by the TxHash attribute.

Table VI gives a summary of the data used in our experiments. The dataset contained 100,000 blocks describing transactions for 1.5 years and involving a total of 559 addresses.

B. Owner of an Address

Given that a Bitcoin address is a pseudonym, nobody knows who has which address. Under the assumption that the secure hash function is not invertible, it is impossible to identify the owner from an address. However, we have two ideas about identifying the genuine owner of an address.

The first method involves collecting addresses that have been specified by the coinbase for the block. The destination addresses of the coinbase are often managed by a mining pool service whose location and country are known publicly.

The second method is to collect addresses that have been published via the website Bitcointalk [5], a well-known online forum about Bitcoin. Bitcointalk provides a profile page for each user (see Fig. 1), where "Name" and "Bitcoin address" attributes are available and we can learn the genuine owner of the addresses. Not all Bitcoitalk users reveal their address, but some specify an address to which they ask donations to be sent as a gratuity. Even when a nickname is specified in the profile page instead of a genuine name, this is sufficient for our experimental purposes to identify the owner from an address. In this way, we collected more than 500 addresses and the corresponding owner's name published on these profile pages.

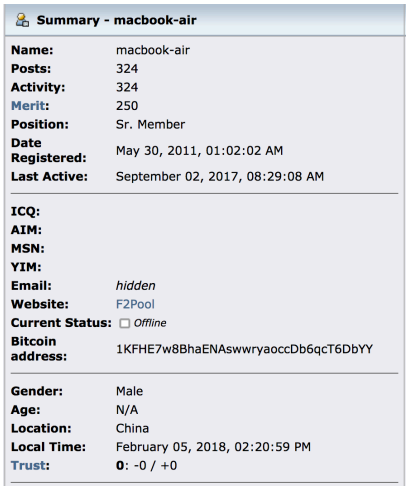

Fig. 1: Profile page in bitcointalk

Table V shows parts of addresses, names, and location data collected from the Bitcointalk website. In principle, an address is assigned indirectly from user information. However, as shown in the third record of Table V, the first six letters "Anduck" of the address after the leading "1" matches his name.

IV. Re-identification Experiment

A. Overview

In this section, we evaluated destination-address anonymity as described in Section III. We were interested in how many addresses could be identified during our re-identification experiments.

In the experiment, the dataset was partitioned into training and a test data. We evaluated the similarity of subsets of Bitcoin addresses, defined in terms of the Jaccard distance.

Let A be a set of addresses$\{a_1, \ldots, a_n\}$ for which the owner is to be identified. Let $O_i(a_j)$ be a subset of addresses specified as the outputs from input address a_j in the i-th duration. Let $T_i(a_j)$ be subset of discriminated times $\{t_1, \ldots, t_{n_{i,j}}\}$ when a transaction related to address a_i is stored in the i-th duration. For example, in Table VIIa, $O_1(a_1) = \{a_1, a_2, a_4\}$.

TABLE III: Example of input table

Attribute	Explanation	Value for example
Time	Time mined block stored in the transaction	2012/09/22 10:47:23
Height	Block number	200001
TxHash	Transaction ID	d635410b5408592d54f59a010ae77974726b2a7ccd26bc76f9a68e02babe3ee5
PreTxHsh	Transaction ID used this Tx	2d6dc2475b5ca40a081b857cc2b7e9fa29376bc299bed62c2d72244ec5a05a6a
InputAddr	Sender's address	1EEYSdwDg9Rvu7bj3AjjJ662yyDbUG1fNi

TABLE IV: Example of output table

Attribute	Explanation	Value for example
Time	Time mined block stored in the transaction	2012/09/22 10:47:23
Height	Block number	200001
TxHash	Transaction ID	d635410b5408592d54f59a010ae77974726b2a7ccd26bc76f9a68e02babe3ee5
OutputAddr	Receiver's address	1ArR7vf17C9ThWi5yt3c74TamCnPUaGb6e
Value	Amount of Bitcoin[10^{-8}BTC]	560000000

TABLE V: Sample data collected from Bitcointalk

Address	Name	Location
1KFHE7w8BhaENAswwryaoccDb6qcT6DbYY	macbook-air	China
1DNNERMT5MMusfYnCBfcKCBjBKZWBC5Lg2	BitHits	None
1Anduck6bsXBXH7fPHzePJSXdC9AEsRmt4	Anduck	None

In our experiment, we observed the change in a set of output addresses in k durations. For example, consider the subsets of output addresses $O_1(a_j)$ and $O_2(a_j)$ in Table VIIa, where we have divided the dataset into two durations. In Table VIIb, we have divided the data into three durations. The size of each divided dataset is almost the same size, based on the block numbers. Note that the size of each subset increases as the number of divided durations k decreases. In the experiment, if there are no transactions in any duration, we remove the address from the whole dataset.

We assume that the output addresses sent from an address are stable and can therefore be used as a signature to identify the owner of the address based on the similarities between sets of output addresses. The degree of similarity between sets A and B is defined by the Jaccard distance:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

where B is the template subset of a target input address and A is a subset to be tested for association with the same owner.

Algorithm 1 shows how our re-identification method works.

In their study, Dupont et al. [2] identified an address based on a set of times of transactions. Whereas our scheme exploits the signature of output addresses, their scheme involves a modification to Algorithm 1, whereby the set of output addresses is replaced by the set of times of blocks for which a relevant transaction occurred. In the example of Table III, the element 10 of 10:47:23 is included in the set of times.

B. Experimental Results

1) Change of accuracy with respect to number of partitions k: We show the change of average recall R and average precision P defined with respect to the number of partitions k in Fig 2.

TABLE VI: Summary of dataset

Period	2012.09.22 - 2014.05.10	About 1.5 years
Address number	559	
Block number	200,001 - 300,000	100,000 blocks

TABLE VII: Example of a dataset

(a) k = 2

Term i	1	2
	10 months	10 months
a_1	$\{a_1, a_2, a_3\}$	$\{a_2, a_3, a_4\}$
a_2	$\{a_2, a_5\}$	$\{a_4, a_4\}$
a_3	$\{a_3, a_4, a_6\}$	$\{a_4, a_5, a_6, a_7\}$
	Training data	Test data

(b) k = 3

Term$i$	1	2	3
	7 months	7 months	7 months
a_1	$\{a_1$,$a_2\}$	$\{a_3$,$a_4\}$	N/A
a_2	$\{a_2$,$a_5\}$	$\{a_4$,$a_5\}$	$\{a_5\}$
a_3	$\{a_3$,$a_6\}$	$\{a_4$,$a_6\}$	$\{a_5$,$a_7\}$
	Training data	Test data	

$$R = \frac{1}{n}\sum_{i=1}^{n} R_i,$$

$$P = \frac{1}{n}\sum_{i=1}^{n} P_i,$$

where R_i is the recall of input address a_i, defined as the fraction of output-address sets identified correctly out of all a_i's address sets ($= k-1$), and P_i is a_i's precision, defined by the fraction of address sets identified correctly over all outputs

Algorithm 1: Re-identification by Jaccard distance

Input: address set A, output address set O

Step 1. o is sets divided into k dataset
$O_1(a_1), \ldots, O_1(a_n), \ldots, O_k(a_1), \ldots, O_k(a_n)$
Exclude address sets $O_1(a_j), \ldots, O_k(a_j)$
such that any subset $O_i(a_j)$ is empty

Step 2. Let data $O_1(a_j)$ be a training data and $O_2(a_j), \ldots, O_k(a_j)$ be test data. For $a_j \in A$, i-th duration identify the input address a_l^* such that Jaccard distance $J(O_i(a_l^*), O_1(a_j))$ is maximized.

Output: Predicted a_1^*,... address

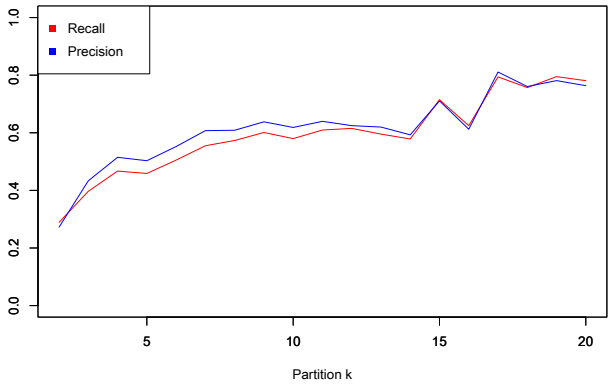

Fig. 2: Change of average recall R and precision P with respect to the number of partitions

identified as a_i. As the number of partitions k increases, both accuracies increase accordingly. As k increases, the target test addresses also decrease, leading to no transactions in the partition. We examined the potential test addresses in terms of k, showing the numbers of test addresses in Table IX. This table clearly supports our hypothesis of why the accuracy decreases with k.

2) Change in accuracy with respect to number of transactions for an address: Which dataset statistics most influence the accuracy of re-identification? To answer this question, we examine the accuracy distribution for various statistical quantities.

Figs. 3 and 4 show the distributions of recall and precision, respectively, with respect to the number of transactions. In this plot, we use the dataset with $k = 10$. We found no significant correlation between these two values in either plot. Note that the address of "BTC Guild" in Figs. 3 and 4 involves many transactions, indicated at the right edge of the figures. Because the address is cited in many coinbase transactions, the recall should be close to 1.0, but its recall is less than this because many other addresses are wrongly predicted as also being the address.

Fig. 5 gives a scatterplot of recall and precision in terms of the number of transactions when the set of addresses is divided with $k = 10$. We found a slightly positive correlation between the two quantities.

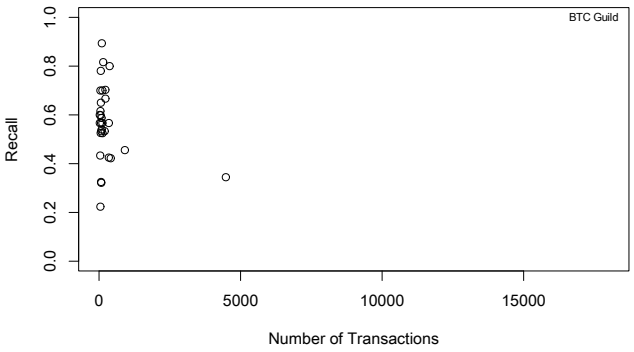

Fig. 3: Scatterplot of recalls R_i by number of transactions in address

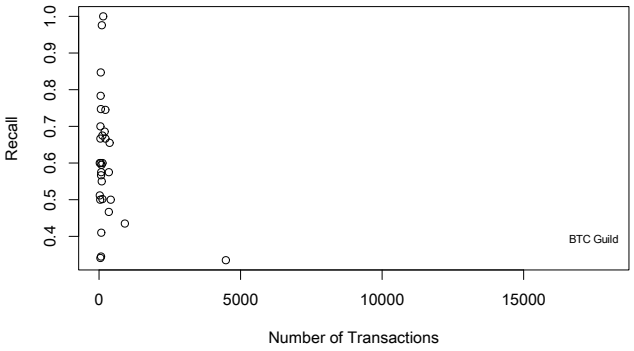

Fig. 4: Scatterplot of precisions P_i by number of transactions in address

3) Comparison of features: What are the most significant features of re-identification? We examine two features, i.e., the set of output addresses and the set of times, in terms of accuracy and discrimination distance between the genuine owner and an imposter.

To investigate the effect of features on re-identification accuracy, we measure the Jaccard distance between the genuine owner's addresses and an imposter's addresses and derive the distribution of the Jaccard distance between the two sets of addresses for the dataset of output addresses with $k = 10$ (see Fig. 6), where the most frequent Jaccard distance is at 0. The maximum Jaccard distance between imposter's addresses is 0.012, meaning that an output address is rarely encountered again among the imposter's addresses. Conversely, the Jaccard distance between genuine addresses is distributed widely. Based on this observation, we can confirm that it is possible to identify an address based on this feature.

For the second feature, Fig. 7 shows the distribution of the Jaccard distance between the genuine and imposter sets of

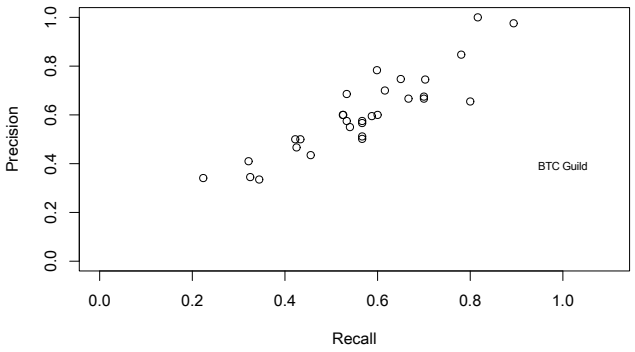

Fig. 5: Relation between recalls and precisions by number of transactions in address

TABLE VIII: Statistics for features

		Minimum	Maximum	Mean	% of 0
Output sets	Same	0	1.0	0.038	55
	Other	0	0.110	0.0001	99
Time sets	Same	0	1.0	0.264	25
	Other	0	1.0	0.155	29

transaction times for the dataset with $k = 10$. This implies that the Jaccard distance between the two sets of transaction times does not sufficiently distinguish them.

In summary, Table VIII gives the statistics for Jaccard distances for the set of addresses and the set of transaction times. The average Jaccard distance for sets of output addresses is greater than that for transaction times. Therefore, we conclude that our proposed feature of the set of output addresses is more useful in distinguishing addresses.

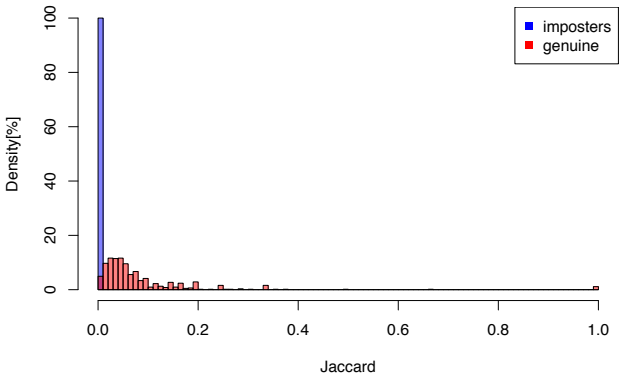

Fig. 6: The distributions of the Jaccard distance between genuine addresses and an imposter's address for a set of output addresses

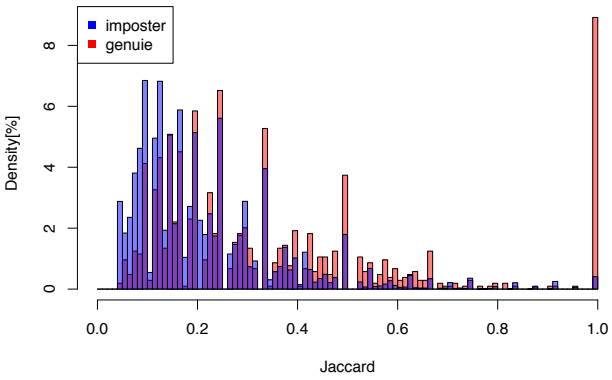

Fig. 7: The distributions of the Jaccard distance between genuine addresses and an imposter's address for a set of transaction times

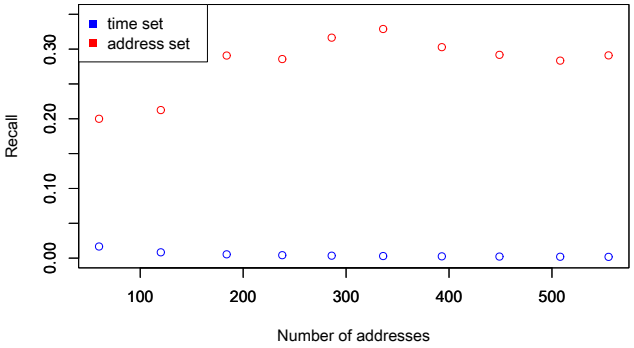

Fig. 8: The average recall R with respect to the number of addresses

4) Recall with respect to the number of addresses per owner: How does the number of addresses per owner affect recall? Heavy Bitcoin users may be at risk of identification because of their output addresses. Fig. 8 shows the average recall with respect to the number of addresses specified by the owner. We plot recalls for two features, i.e., the distances within the output set and within the set of transaction times. The average recall of the output set is always higher than that for the transaction-time set. Surprisingly, average recalls are stable, i.e., do not depend on the number of addresses. The recall for transaction times decreases by $\frac{1}{n}$ against n.

C. Anonymity

In this study, we evaluate the anonymity of the i-th input address a_i via the F value for an identification attack.

$$F_i = \frac{2R_i \cdot P_i}{R_i + P_i}$$

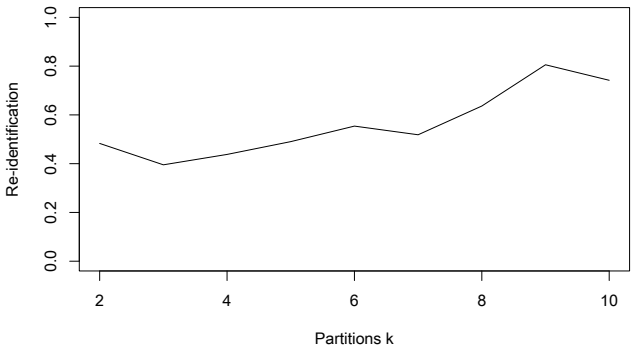

Fig. 9: Distribution of re-identification ratio with respect to the number of partitions k

TABLE IX: Target addresses with respect to partitions

k	n
2	559
3	296
4	153
5	104
6	74
7	54
8	44
9	36
10	31

A larger F value implies a lower anonymity. We suggest that an address can be re-identified if the F value is more than 0.5. Fig. 9 shows the distribution of the re-identification ratio with respect to the number of partitions k. The ratio is maximized at 80.5% for $k = 9$, whereas the minimum is 39.5% at $k = 10$. As the number of partitions increases, the re-identification ratio increases.

From our experimental results, the accuracy of the recall and precision ratio is not affected by quantities such as the number of transactions in an address. Note that the risk of an address being identified increases as the Jaccard distance increases. Therefore, to preserve anonymity of addresses, we encourage periodic address updating and avoiding use of the same output address in too many trades.

V. Conclusions

We have investigated the anonymity of Bitcoin addresses by experiments that aim to re-identify the owner from the set of output addresses that the owner uses to send or receive payments. The results indicate that 80.5% of addresses can be identified from the Jaccard distance between subsets of output addresses and neither the average recall nor the precision is affected by the number of transactions per address.

References

[1] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voeker and S. Savage, "A fistful of bitcoins: Characterizing Payments Among Men with No Names", *In Proceedings of Conference on Internet Measurement Conference (IMC'13), 2013.*

[2] J. Dupont and A. C. Squicciarini. "Toward De-Anonymizing Bitcoin by Mapping Users Location",*In Proceedings of Conference on Data and Application Security and Privacy (CODASPY'15), ACM, 2015.*

[3] S. Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System, 2008,[online] Available: https://bitcoin.org/bitcoin.pdf.

[4] Blockchain. https://blockchain.info

[5] Bitcointalk. https://bitcointalk.org/

[6] A. Biryukov, D. Khovratovich and I.Pustogarov, Deanonymisation of Clients in Bitcoin P2P Network, *In Proceedings of Conference on Computer and Communications Security(CCS14),2014*

[7] S. Delagdo-Segura, C. Perez-Sola, G. Navaro-Arribasv and J. Herrea-Joancomarti, "Analysis of Bitcoin UTXO set", *In Proceedings of Conference on Financial Cryptography and Data Security (FC'18), 2018.*

[8] P. Koshy, D.Koshy and P. McDaniel, " An analysis of anonymity in bitcoin using P2P network traffic"*In Proceedings of Conference on Financial Cryptography and Data Security (FC'14), 2014.*