Shanghai Jiao Tong University

Mobile Networks Final Project

# De-anonymization of Bitcoin Network

*Author*:
陈宏杰 515030910597
夏崇垚 5140309546

*Instructor*:
王新兵
傅洛伊

2018 年 5 月 26 日

# Contents

# 1 Abtract

Bitcoin is a cryptocurrency and worldwide payment system invented by an unknown person or group of people under the name Satoshi Nakamoto and released as open-source software in 2009.It is the first decentralized digital currency, as the system works without a central bank or single administrator. The network is peer-to-peer and transactions take place between users directly, without an intermediary. These transactions are verified by network nodes through the use of cryptography and recorded in a public distributed ledger called a blockchain. Consequently, Bitcoin has the unintuitive property that while the ownership of money is implicitly anonymous, its flow is globally visible. In this paper we explore this unique characteristic further, using heuristic clustering to group Bitcoin wallets. Some machine learning clustering methods will be implemented for the supplement of the clustering work, and then a community detection process will be applied to better understand how the transaction is running in this bitcoin network.

# 2 Introduction

We will need to represent the board states and realize some basic operations to make playing a game feasible. And to make the board state consistent with the neural network, we will have to do some more modifications with the sgf files.

## 2.1 Bitcoin Account

### 2.1.1 Elliptic Curve Digital Signature Algorithm

Bitcoin uses a particular digital signature scheme that' s called the Elliptic Curve Digital Signature Algorithm (ECDSA). ECDSA is a U.S. government standard, an update of the earlier DSA algorithm adapted to use elliptic curves. These algorithms have received considerable cryptographic analysis over the years and are generally believed to be secure.

### 2.1.2 Private Key

Private Key is a 256-bit random number generated by the bitcoin system. With this private key, you will be able to possess an account that's in the bitcoin system. You will see this a little bit casual since it's so easy to leak your bitcoins in such an account to others. However, assume you try to access a bitcoin wallet by randomly testing different combinations, you will have $2^{256}$ possible ones to check its balance. Assume each person in this world have 100 bitcoin addresses, that will add up to $6 \times 10^9 \times 100$ accounts. You will have a probability of $\frac{6\times10^{11}}{1.58\times10^{77}} = 3.8 \times 10^{-66}$ to reach an account with values if try once. And even with $10^{20}$ times of trial, the probability is still too small to make a difference. The probability makes the decentralized system safe, but it still requires absolute randomness of generating the private key to ensure the security of the system. Thanks to the bitcoin software, this randomness issue is perfectly handled.

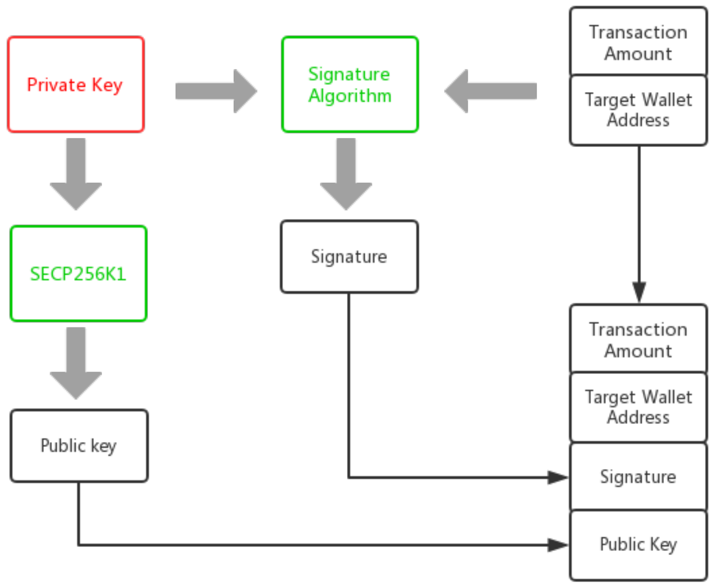


Figure 1: Blockchain Organization

#### 2.1.3 Public Key

Public Key is a value generated by the private key with a hash function. The property of hash functions make it impossible to deduce the private key from the public key, which means we don't have to worry about the security of our private key when we just use the public key as our identity in the bitcoin system. The public key will appear in the ledger of the bitcoin system named blockchain.

### 2.2 How Anonymization is Achieved

The private key is like your right hand to sign and validate a transaction, and the public key is like your name appearing on the ledger in the bitcoin system.

#### 2.2.1 Transaction

ECDSA provides the digital signature technique for the users to make valid contractions. With a contract, the user who is going to provide the amount of money has to sign the original contract or hash of the contract content to make it valid to miners. And only with private key can the user sign on the contract content. With the contract content and the private key, bitcoin system can generate a "signature" content to validate the contraction. See it in Figure1.

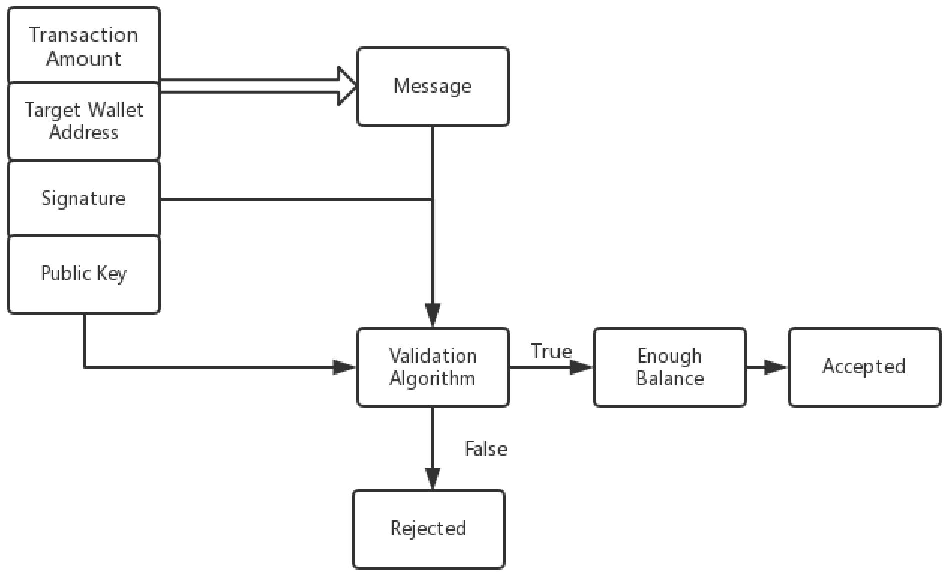


Figure 2: Blockchain Organization

#### 2.2.2 P2P Network

Bitcoin operates on a list of blocks, the block chain. Each block contains lots of transaction data. The bitcoin miners first collect all transactions not yet included in a block. Then, the miners check the signature whether it's valid according to the public key who is going to give the bitcoins and the transaction content. If valid, the miners still need to check that the transaction amount is smaller or equal to the account balance. Then, if the process is passed in over 50% percent of the whole networks, it will be written a new block and the transaction is completed. See it in Figure 2.

# 3 Related Work

In this chapter we dig more into the possible techniques of deanonymizing a bitcoin network.

## 3.1 Leakage of Personal Information

The problem of User Identity Linkage (UIL), which aims to identify the accounts of the same user across different social platforms, has been attracting an increasing amount of attention and effort due to both the significant research challenges and the immense practical value of the problem. The bitcoin blockchain system is a transaction network with no user profile or personal information, but there is still a chance that a user might leak his information by posting his public key on the social network profile or the retailer which the user bought his commodities from. The former will link your personal profile on the social network with your public address which will endanger the anonymity of all the people having transactions with this bitcoin address, the latter will link your public key to the shipping address or mobile phone number, which will still cause a lot of anonymity damage to the you and people transacting with you.

## 3.2 P2P Network Attack

This is a process linking one's public address to his temporary IP address which he logs into the bitcoin system. The bitcoin network implements P2P network, which allows attacker to connect any peer it tries to deanonymize.

## 3.3 Deanonymization of Transaction Network

A user may possess multiple public addresses in order to ensure anonymity, which is the reason why we want to find the similar behaviours between different public addresses and cluster them together.

### 3.3.1 Transaction Network

The transaction network T represents the flow of Bitcoins between transactions over time. Each vertex represents a transaction and each directed edge between a source and a target represents an output of the transaction corresponding to the source that is an input to the transaction corresponding to the target. Each directed edge also includes a value in Bitcoins and atimestamp. It is a straight-forward task to construct T from our dataset.

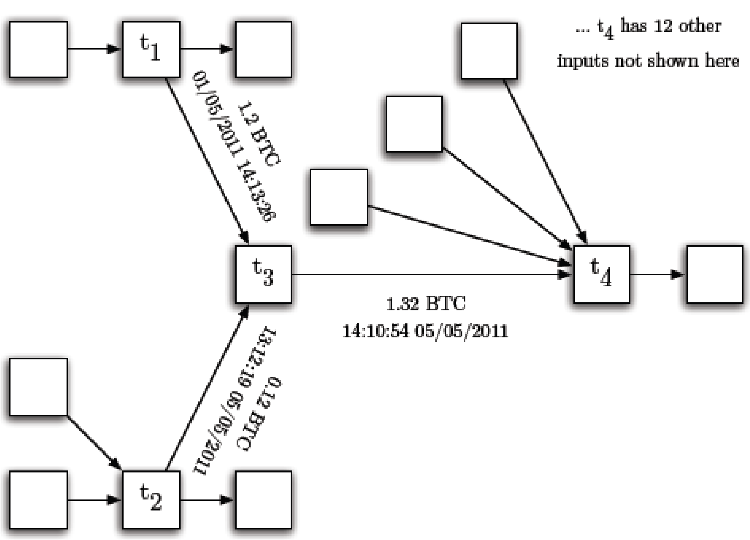


Figure 3: Transaction Network

#### 3.3.2 User Network

The user network U represents the flow of Bitcoins between users over time. Each vertex represents a user and each directed edge between a source and a target represents an input-output pair of a single transaction where the input's public-key belongs to the user corresponding to the source and the output's public-key belongs to the user corresponding to the target. Each directed edge also includes a value in Bitcoins and a timestamp.

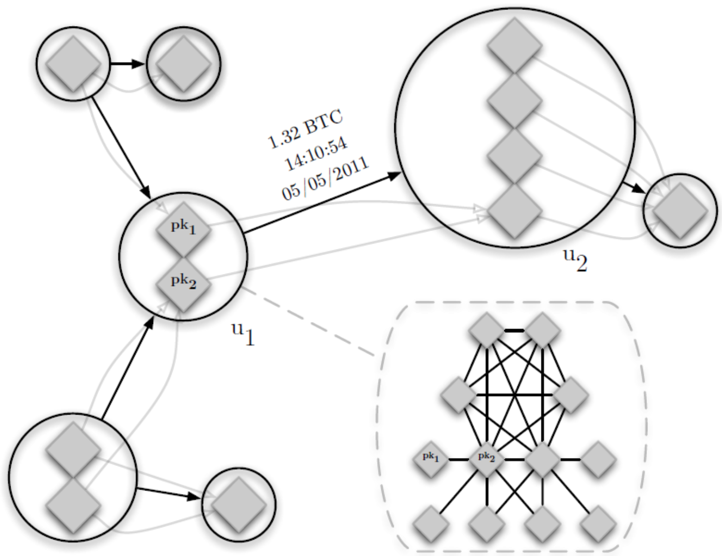


Figure 4: User Network

#### 3.3.3 Our Work

Our work is to transform the transaction network to the user network. One approach is to abstract the features of this transaction network, anther is to select features for the public address and apply machine learning method to cluster them together. Finally, we will do a community detection in this bitcoin user network, and we will see the result of this project.

## 4 Method

### 4.1 Basic Assumption

Each transaction may have multiple income and outcome addresses. One possible clustering method is to see the all the addresses on the one side of a contraction belonging to the same user. This is a very intuitive and easy but still very power assumption to cluster the addresses.

## 4.2 Machine Learning Clustering - Feature Selection

We extract the following features of a contraction, for each address or clustered addresses:

1. In-degree. The times of the address appearing on the sending side of a contraction.

2. Out-degree. The times of the address appearing on the receiving side of a contraction.

3. Average in-transaction amount and its difference.

4. Average out-transaction amount and its difference.

5. Average time interval between in-transactions.

6. Average time interval between out-transactions.

7. Active-duration. The most common interval when the client is active.(ranging from 1 to 4 representing midnight, morning, afternoon, night before 8 respectively)

# 5 Experienment

## 5.1 Data Collection

As we know, every bitcoin client can access all bitcoin transaction history data easily. From the very first bitcoin transaction to now (2018 May), there are already more than 500,000 blocks on the earth. Considering that a new blockchain is produced every several minutes, the number of blockchains is still increasing with a very large speed. As each block contains transactions in thousand, the number of transactions is even much larger. Processing such large amount of data demands mighty computation power, thus, in practice, we only use a part of all data. The original data are composed of block data. Therefore, we have to preprocess the original data to make them organized as desired. The data can be downloaded from peer-to-peer network

| Height | Age | Transactions | Total Sent | Relayed By | Size (kB) | Weight (kWU) |
|---|---|---|---|---|---|---|
| 524318 | 6 minutes | 577 | 3,977.44 BTC | F2Pool | 376.73 | 1,309.85 |
| 524317 | 9 minutes | 1375 | 5,716.06 BTC | Unknown | 855.88 | 3,043.88 |
| 524313 | 22 minutes | 430 | 428.63 BTC | SlushPool | 1,262.89 | 3,992.99 |
| 524311 | 29 minutes | 848 | 1,941.06 BTC | BitClub Network | 1,200.73 | 3,992.73 |

using open-source bitcoin client[1]. For convenience, we directly downloaded data online which are already collected by other researchers. Among about 270,000 transactions, there are around 24,000,000 different bitcoin wallet addresses, approximately 30,000,000 transactions.

[1]https://github.com/bitcoin/bitcoin

| Height | Time | Relayed By | Hash | Size (kB) |
|---|---|---|---|---|
| 524318 (Main Chain) | 2018-05-25 11:49:17 | F2Pool | 0000000000000000002abe57e410d5bb654980080ee63fb1c075852b0d529e9d | 376.73 |
| 524317 (Main Chain) | 2018-05-25 11:46:02 | BTC.TOP | 00000000000000000010e7bc57d4cc9999cea5c88391ceb3b5b4cdebbc58d4c2 | 855.88 |
| 524316 (Main Chain) | 2018-05-25 11:38:18 | SlushPool | 000000000000000000172298ab32f10ddc0029b8f60b0a234b1bf15bbdf58dc9 | 983.54 |
| 524315 (Main Chain) | 2018-05-25 11:37:22 | BTCC Pool | 000000000000000000b13e1b7992633bfe35525d1ab5c5a4dabc1b6968a0432 | 1,013.33 |
| 524314 (Main Chain) | 2018-05-25 11:34:19 | Unknown | 0000000000000000003fbaf8b31af3b7d0d7d3665ff70c8d5949357f11c45f52 | 1,129.95 |
| 524313 (Main Chain) | 2018-05-25 11:32:57 | SlushPool | 00000000000000000027a3c07fa674be944c784d3c87de9524cea45725161ecd | 1,262.89 |
| 524312 (Main Chain) | 2018-05-25 11:31:25 | BTC.com | 000000000000000000a6750a4678d7d5f0317b71852166676773121e777c4b0 | 1,182.37 |
| 524311 (Main Chain) | 2018-05-25 11:26:17 | BitClub Network | 00000000000000000017f1501a85c0e633210159abb643110e325769f881378b | 1,200.73 |

Figure 5: A glimpse of recently produced blocks

## Block #524319

| Summary | |
|---|---|
| Number Of Transactions | 1916 |
| Output Total | 15,233.7070156 BTC |
| Estimated Transaction Volume | 1,268.6519563 BTC |
| Transaction Fees | 0.71794901 BTC |
| Height | 524319 (Main Chain) |
| Timestamp | 2018-05-25 12:41:14 |
| Received Time | 2018-05-25 12:41:14 |
| Relayed By | BTC.TOP |
| Difficulty | 4,306,949,573,981.51 |
| Bits | 390158921 |
| Size | 1153.061 kB |
| Weight | 3992.834 kWU |
| Version | 0x20000000 |
| Nonce | 3073167138 |
| Block Reward | 12.5 BTC |

| Hashes | |
|---|---|
| Hash | 00000000000000000029414bd8e91de920a7c1a90ff3a934108aca5be2b0bcf6 |
| Previous Block | 0000000000000000002abe57e410d5bb654980080ee63fb1c075852b0d529e9d |
| Next Block(s) | |
| Merkle Root | 1b7c1fa4129b289405ccfc70721c3667ddcbe23f07d3d386f8b8c0f125775589 |

Transactions

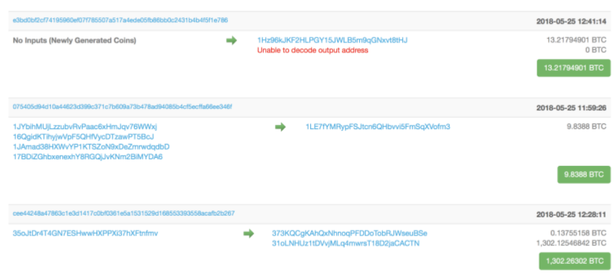

Figure 6: A glimpse of original block information

## 5.2 Deanonymization

Although the alleged anonymization of bitcoin is a very tempting, it's just pseudo-anonymization rather than absolute anonymization, which gives us opportunity to deanonymize it. In essence, denonymization is to link wallet addresses to the owner hiding behind. However, linking wallet addresses with the real-world users requires additionally external information apart from history transaction data. Considering the fact that we are only provided with transaction data, as a consequence, in this project, we just take into account how to cluster wallet addresses that potentially belong to the same user. In order to achieve this goal, we

| | addr_ID | addr |
|---|---|---|
| **100** | 100 | 111ccCf3YzzcXH6G15mukMBQ8rcmo1qCU |
| **101** | 101 | 111CH4CEu1PkTMpouRKDTK3yZPHAxz8Vv |
| **102** | 102 | 111cphrV8LtixDfWq7HbELtehH5UgRqiA |
| **103** | 103 | 111cZqKGQzMyEaPNajPmgGaHS7U9vRSWd |
| **104** | 104 | 111D7xaZHpAnJdwKknTXmZzWc6Sw8uwzH |
| **105** | 105 | 111Da3uc98pipSvS3mEjNbU12WKs578th |
| **106** | 106 | 111DADVu85myVnJ53CzZgB9kapsHwDxW2 |
| **107** | 107 | 111dDDCBvC598bkNC9qiGPDPnJ1tdBGDe |
| **108** | 108 | 111dentifieron1yLettersXXXXasmS9N |
| **109** | 109 | 111dentifiersA1waysHaveToXXZPZVdN |

| | trans_ID | addr_ID | val |
|---|---|---|---|
| **8800** | 8773 | 818085 | 5000000000 |
| **8801** | 8774 | 2960363 | 5000000000 |
| **8802** | 8775 | 4440252 | 5000000000 |
| **8803** | 8776 | 588968 | 5000000000 |
| **8804** | 8777 | 6470175 | 5000000000 |
| **8805** | 8778 | 1108046 | 5000000000 |
| **8806** | 8779 | 2927540 | 5000000000 |
| **8807** | 8780 | 2077512 | 5000000000 |
| **8808** | 8781 | 1101897 | 5000000000 |
| **8809** | 8782 | 878896 | 5000000000 |

| | trans_ID | addr_ID | val |
|---|---|---|---|
| **9100** | 50563 | 3418377 | 4900000000 |
| **9101** | 50564 | 1008527 | 80000000000 |
| **9102** | 50564 | 2518235 | 236155000000 |
| **9103** | 50564 | 4990542 | 72000000 |
| **9104** | 50564 | 6008494 | 16000000000 |
| **9105** | 50565 | 636952 | 1000000 |
| **9106** | 50568 | 5211481 | 1000000 |
| **9107** | 50569 | 5554320 | 332160000000 |
| **9108** | 50570 | 2377352 | 331660000000 |
| **9109** | 50571 | 2335172 | 331160000000 |

Figure 7: **Left**: The first column is column ID. The second column is address ID. The third column is address hash, i.e. the real address appearing in a block. **Middle**: The first column is column ID. The second column is address which receives bitcoins. The third column is the amount of $10^{-8}$ bitcoins. **Right**:The first column is column ID. The second column is address which sends bitcoins. The third column is the amount of $10^{-8}$ bitcoins.

applied two different approaches, of which the first is using heuristic to conduct clustering while the second is exploiting machine learning techniques.

#### 5.2.1 Heuristic Clustering

A bitcoin user usually owns many wallets concurrently, and each wallet contains certain amount of bitcoins. If the user wants to buy something inexpensive, for example, which can be paid by money in just one wallet, then the transaction will include just one input wallet address. However, if the user wants to buy something very expensive, which needs to be paid by money in several wallets simultaneously, then such transaction will involve several input addresses. From the two examples above, we can derive one useful heuristic, which is that if there are two or more than two input addresses in one transaction, then they may belong to the same user. Therefore, we can always cluster addresses which appear in the same transaction's. What's more, for two clusters of addresses, if they have at least one address in common, then we can merge the two clusters into a bigger one. We will keep doing this until all clusters are disjoint.

## 6 Results

#### 6.0.2 Heuristic Clustering

Among the totally about 24,000,000 bitcoin addresses, we identified about different 12,000,000 users. Part of the results are shown in Figure 4. Then we analyze such results. More than 70% users possess only one address. The mean of number of owned addresses by single user is 2.0. The maximum number of owned addresses by single user is 544754. The minimum number of owned addresses by single user is 1. Since 99% users have less than 10 addresses, it's a bad idea to draw a histogram. Instead, we draw figure 5 to visualize the data distribution. Moreover, we also count how many times an address participate in transactions, including sending and receiving bitcoins. The mean of number of transactions a single address participate

| | addr_ID | user_ID |
|---|---|---|
| **100** | 101 | 101 |
| **101** | 102 | 102 |
| **102** | 103 | 103 |
| **103** | 104 | 104 |
| **104** | 105 | 105 |
| **105** | 106 | 106 |
| **106** | 107 | 107 |
| **107** | 108 | 108 |
| **108** | 109 | 109 |
| **109** | 110 | 110 |

| | addr_ID | user_ID |
|---|---|---|
| **900** | 901 | 98 |
| **901** | 902 | 902 |
| **902** | 903 | 903 |
| **903** | 904 | 904 |
| **904** | 905 | 905 |
| **905** | 906 | 906 |
| **906** | 907 | 64 |
| **907** | 908 | 908 |
| **908** | 909 | 783 |
| **909** | 910 | 910 |

| | addr_ID | user_ID |
|---|---|---|
| **12900** | 12901 | 8704 |
| **12901** | 12902 | 12902 |
| **12902** | 12903 | 12903 |
| **12903** | 12904 | 12904 |
| **12904** | 12905 | 12905 |
| **12905** | 12906 | 979 |
| **12906** | 12907 | 64 |
| **12907** | 12908 | 4378 |
| **12908** | 12909 | 12909 |
| **12909** | 12910 | 1998 |

Figure 8: Three snapshots of results of heuristic clustering. The first column is address ID. The second column is the user ID.

Figure 9: **Left**: In this graph, each circle represents a user. And the area of a circle positively proportionally reflects the number of addresses a user owns. From this graph, we can clearly see that most users own just a small number of address, while only few users own a large number of addresses. **Right**: In this graph, each circle represents an address. And the area of a circle positively proportionally reflects the number of transactions an address participate. From this graph, we can clearly see that most addresses participate just a small number of address, while only few addresses take part in a large number of transactions.

| percentile | 50% | 60% | 70% | 80% | 90% | 95% | 99% | 99.9% |
|---|---|---|---|---|---|---|---|---|
| # of addresses | 1 | 1 | 1 | 2 | 2 | 3 | 8 | 51 |

Table 1: Percentiles of number of addresses owned by one identified user. We note that more than 70% users only own one address, and less than 0.1% users own more than 50 addresses.

in is 7.26. The maximum number of transactions a single address participate in is 1752211. The minimum number of transactions a single address participate in is 1. Similarly, we visualize the data distribution in figure 5.

| percentile | 50% | 60% | 70% | 80% | 90% | 95% | 99% | 99.9% |
|---|---|---|---|---|---|---|---|---|
| # of transactions | 3 | 4 | 5 | 6 | 9 | 16 | 59 | 277 |

Table 2: Percentiles of number of transactions a single address involved with. We note that more than 50% addresses are only involved in own 3 transactions, and less than 0.1% addresses are involved with more than 277 transactions.

# 7 Appreciation

# References


[1] A Fistful of Bitcoins: Characterizing Payments Among Men with No Names. Sarah Meiklejohn Marjori Pomarole University of California, San Diego. Barcelona, Spain. Copyright 2013 ACM.

[2] An Analysis of Anonymity in the Bitcoin System. Fergal Reid and Martin Harrigan.7 May 2012.

[3] Bitcoin and Cryptocurrency Technologies. Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, Steven Goldfeder. 2016

[4] Deanonymisation of clients in Bitcoin P2P network. Alex Biryukov, Dmitry Khovratovich, Ivan Pustogarov. 2014.

[5] Deanonymizazing Tor Hidden Service Users Through Bitcoin Transaction Analysis. HUSAM BASIL AL JAWAHERI. June 2017 HUSAM BASIL AL JAWAHERI.

[6] De-Anonymizing the Bitcoin Blockchain. Bharath Srivatsan. 2016.