Doctoral Thesis

Crypto-currency Flow among Bitcoin Users: De-anonymization and Dynamics of Persistently-Active Big Players by Network Analysis

Rubaiyat Islam

March 2021

Graduate School of Simulation Studies
University of Hyogo

To My Parents, My Wife, My Two Sons Samin and Arshan
"Together We Win"

Abstract

Crypto-currency or crypto-assets can provide a unique opportunity to perform a detailed study on financial transactions and interactions among users. Publicly available and big data accessible by the recent technology of distributed ledger, blockchain, help us to understand statistical properties and dynamics of economic network, in which users are interconnected with each other through money flow of transactions among each other and also in the exchange markets of crypto-currencies as well as fiat currencies. Users, who play dominant roles with respect to their frequencies and amounts of transactions, must have vital roles in the entire system of crypto-currencies. While anonymity of users is a core technology of blockchain, de-anonymization, if possible and even partially, helps to reveal various aspects in the ledger system of blockchain.

The purpose of this thesis is the de-anonymization of users, in particular, what we call big players and persistently-active ones, and the understanding significant properties in the dynamics of crypto-currency flow. I employ the blockchain of Bitcoin, in which all the transactions are recorded with a list of addresses, which are anonymous wallets, but can be partially identified as individual users. I constructed graphs or networks comprising of users or addresses as nodes and transactions or money flow as edges. Then I performed exploratory data analysis and network analysis in order to find significant patterns and interesting dynamics of the activities in the money flow. The thesis has the following three parts.

First I studied the daily time-series of transactions in their daily numbers and volumes during 2013 to 2018, when the generation of Bitcoin mining blocks was relatively stable. I focus on significant spikes in the transactions in the total number of transactions and total sum of volumes. By using smoothed periodogram or power-spectrum analysis for the time series, I found weekly pattern of these two variables, which implies that the financial organizations' trading systems are dominant roles giving higher activities during weekdays compared to weekends, which is similar to the exchange market of fiat currencies.

Second, following the above observation, I constructed daily networks and analyzed the network properties of the users as nodes and money flow attributed as edge flow circulated among users to focus on weekdays and weekends activities. I then performed an analysis using threshold for the flow of Bitcoin to define "big players" by proposing a method to identify financial institutions as those users satisfying certain criteria. The criteria concern about high frequency of appearance, in other words, appearing persistently on daily big transactions and showing a distinct weekly pattern of total average network flow. We were actually able to find known financial institutions as well as others.

Third I applied the method of non-negative matrix factorization (NMF) which can decompose the matrix of numbers and volumes of transactions into a certain number of components with relative weights. The purpose of such an analysis is to reveal hidden components in which users play different patterns of sending and/or receiving money. I proved that the NMF can be interpreted by a stochastic model. Then I performed simulations for a toy problem and estimated the parameters involved in the stochastic model in a framework of Bayesian estimation. From this result of simulation, one can understand that the results of NMF can be interpreted as the probabilities of relative weights and the vectors corresponding to main senders and receivers. In the real data of Bitcoin, I found that there are actually big players that were already identified as financial institutions and also as others in the second part above. Moreover, I applied the method to temporal change of the network and found that the dynamics has a stable structure corresponding to the same components as well as a slowly changing dynamics.

Contents

1 Introduction **1**
- 1.1 Background of Bitcoin blockchain: . . . 1
- 1.2 Purpose of our research . . . 8
- 1.3 Outline of the thesis . . . 9

2 Literature review **10**
- 2.1 The de-anonymization of Bitcoin blockchain . . . 10
- 2.2 The research work on Bitcoin blockchain network analysis . . . 12
- 2.3 The research work on Bitcoin exchange market analysis . . . 13

3 Mathematical notations and methods **15**
- 3.1 Definitions and notations . . . 15
- 3.2 Connected components . . . 17
- 3.3 Filtered daily graphs . . . 18

4 Dataset **21**
- 4.1 The Hungarian researchers' data set . . . 21
- 4.2 Some statistical stylized facts about Hungarian data attributes . . . 21

5 Transaction graph analysis and weekly pattern of BTC volume and transactions **24**
- 5.1 Data set of daily total transactions and BTC volume sum . . . 24
- 5.2 Auto-correlation function of BTC volume, price and number of transactions . . . 25
- 5.3 The evolution of bitcoin transactions (a bird-eye view) . . . 26
- 5.4 The weekly pattern of bitcoin volume sum and number of transactions . . . 26
- 5.5 The outliers' transaction patterns . . . 27
- 5.6 Directed transaction graph and degree correlation to visualize outliers' activities . . . 28

6 Identifying Big players in the bitcoin blockchain: (A new approach) **39**
- 6.1 Dataset for daily users' network flow graph . . . 39
- 6.2 Significant difference between weekdays and weekend . . . 40
- 6.3 Threshold and network Size . . . 41
- 6.4 Sum of edge-flow and average edge-flow . . . 41
- 6.5 Connected components of sub-graph and the BTC flow inside . . . 42
- 6.6 Examining some exchange's activities : The "Big players" market scenario . . . 43

7 Simulation, results and interpretations **58**
- 7.1 Methodology of stochastic model using NMF . . . 58
- 7.2 Stochastic model of NMF . . . 61
- 7.3 Data pre-processing on daily regular user graph for applying NMF . . . 69
- 7.4 Results and interpretation of regular user graph with NMF . . . 71

8 Conclusion **78**

Acknowledgment **80**

Appendix A Statistical analysis of monthly user graph variables **86**

Appendix B Threshold and network Size **88**

Appendix C Sum of edge-flow and average edge-flow **91**

Appendix D Total flow of subgraph **94**

Appendix E NMF use case:Topic extraction from text-document analysis **95**

1 Introduction

Money circulates in the economy as income turns into savings and investment and back again. A bookkeeping ledger is used to record the history of transactions that occur between entities [1].

Money, in the real sense, has no actual value, it is just a piece of paper or a metal coin. Its importance is conveyed when financial governing bodies acknowledge its acceptance value among users. The value of money is derived by its functions: a way of transactions, a unit of quantification, and an asset value.

Money allows people to trade goods and services indirectly; it helps communicate the price of goods. Prices are given in Yen/Dollar/Euro or other units of currency that correspond to a numerical amount in one's possession, that is, in one's pocket, purse, or wallet, and it provides individuals with a way to store their wealth in the long-term.

The increase in economic and financial network analysis has enabled the emergence and evolution of rich theories and methodologies [2], which have a distinct effect on human decision-making and are pervasive. Thus, they are becoming the link between societies and economies. Methodologically, graph theory has refined into a potent and compelling tool for complex problems. Cryptocurrency, also known as cryptoassets, which is based on the blockchain technology of non-centralized ledgers (e.g., see [3]), provides an exhaustive record of transactions in the ledger. It offers a quite unique opportunity to study how money flows among users.

Although it was difficult to construct financial networks at the start of this decade, the rise of cryptoassets has enabled significant available fortuity, for network the rise of crypto-assets has enabled significant opportunity, for network related research and analysis. Previously, information about transaction details was usually considered sensitive and not available for research [4]. The cryptoasset system comprised of a repeatedly expanding list of information reserved in a chain is publicly accessible, and enables scope to analyze transaction networks in detail.

It would be of interest to examine the structure and temporal growth of user graph, where users are nodes and the flow of Bitcoin between nodes are links, and the transaction graph, where the transactions are nodes and connection to the next transaction from the previous ones are called links. These kinds of dynamics are linked to the users' behaviors during quiet and active periods of the market value of the cryptoasset.

This study determines the identity, that is, de-anonymizing of large "wallets" and their network of peers in the Bitcoin blockchain by focusing on the structural change and flow dynamics of Bitcoin volume transacted. This study conducted rigorous exploratory data analyses on the different variables of the blockchain data. Among these, the distribution of the users' input and output, transactions, and volume transacted were significant variables.

This research utilized time-series analyses and the lowest time change used was daily timestamped data. Real-world networks have a common feature of temporal change, which means their nature deviates over time [5]. This study revealed that users' behavior or activity changes from weekdays to weekends, which compromised the security of their identity inside the blockchain. A sophisticated machine learning algorithm was implemented to classify their peer networks. Before we present the analysis, we provide a brief discussion of the background and features of Bitcoin as a cryptocurrency.

1.1 Background of Bitcoin blockchain:

Bitcoin is the pioneer of the most popular cryptocurrencies to date and also a unique example of a large-scale sustainable payment system, in which all the financial transactions are publicly available (see [6]). It is not issued by any government, bank, or organization, but rather by mathematical cryptographic protocols in a distributed network system, where users pseudo-anonymously exchange and sometimes mint bitcoins. To date, economic literature on the Bitcoin issue is quite limited. Researchers such as [7] and [8] have successfully drawn attention to the analytical aspects

related to the information contained in the blockchain. Due to its still relatively low acceptance in the foreign exchange market and its poor performance as a medium of the store of value, there has been some discussion in academia on whether Bitcoin can be considered a currency. However, the trust in this currency totally is not based on the belief in central monetary authority but rather computer technology and cryptography [9].

The blockchain is one of the revolutionary database that has evolved over the last decade. It stores any information in a decentralized computing system and once stored, data can never be altered or manipulated. It is transparently accessible to all the users logged in the database and they can view all the information published in the blockchain. Bitcoin cryptocurrency, along with the sender and receivers addresses in form of a ledger, is the financial monetary information that is stored inside a blockchain. Bitcoin is world's first successfully implemented fully-digital cryptocurrency [3]. It solved two real-world problems, double-spending [10] and "duplication problem" and created an alternative way to establish a fully functional virtual currency based financial system.

Who invented Bitcoin?

The global financial crisis in 2008 exposed financial inequalities throughout the world's economies. In January 2009, a mysterious figure named "Satoshi Nakamoto" introduced a virtual currency system called "Bitcoin", which functioned over a cryptography framework called "Blockchain" with an incentive scheme known as "Proof of work"[6]. Bitcoin is a digital currency that archives transactions and autonomously administers the generation of new units of currency inside the blockchain frame of reference. No centralized authority dominates the operation and logging in a distributed system with a private key proves the user's ownership of Bitcoins. A consensus algorithm and a public history of transactions has strengthened security to prevent duplication and double-spending [11].

In 2009, Nakamoto circulated his finished code for Bitcoin within the cryptographic community and made it open source for more development. He mined the first block, referred to as the "genesis block" on the 3_{rd} January 2009. His identity has always remained anonymous and after some collaboration with peers, he was mysteriously vanished after the genesis block had been published. His Bitcoin account worth around $19 Billion US dollars and he is considered the 44_{th} richest man in the world. There is lot of speculation in the cryptoasset community about the true identity of Satoshi Nakamoto. One thing for sure is that he is the first to successfully fully digitize the concept of fiat money and solve key problems experienced by traditional centralized monetary systems.

The shortcomings of fiat currency

Let us investigate the problems of traditional currency formally known as fiat currency and how Bitcoin was able to solve some of the key issues.

To buy a box of chocolate, we normally withdraw cash from a bank or ATM. We need to determine how and why these transactions work in our day-to-day life. Everyone accepts the metal disks known as coins and strips of paper called money, but how do these gain such value? Looking back to 1900 as an example, money or currency was traditionally considered to be equivalent to certain valuable items, for example, gold. During this period, 1 ounce of gold was valued as equivalent to 20.67$. This means that US government banks backed the currency with equivalent ounces of gold in their central reserve. Over the course of time, this system became inefficient and was ended in the US in 1971. President Nixon cancelled the direct exchange of the United States dollar to gold. This shift of policy was known as the Nixon Shock. Standardizing the way of use and the possibility of producing infinite quantity has improved the system from the previous one. The newly introduced American dollar had

no intrinsic value, but any exchange of this paper or coins between two entities is legalized by the third party, in this case the central body of the government bank. Gradually, all the world's other major economies have converted their monetary system to the fiat currency system.

One of the main problems of traditional currency is hyperinflation. This occurs when the government prints increasing amount of money, which results in devaluation of the currency. This situation occurs during times of economic crisis and when commodity prices surge. Thus, the value of the currency is determined by its supply demand. This promotes the government to ensure economic stability by monitoring, evaluating, and taking action on the credit supply, interest rates, and liquidity.

Another crucial problem is that, as the system is centralized, extensive regulation and policies are needed, which are very common in a top-down organizational structure. Alternatively, it can be said that extensive documentation is required to audit every financial transaction that involves a monetary transaction. Accordingly, when a person uses an ATM that does not belong to their bank, or when they transfer money from their account to a friend's, they often pay a fee. This transaction fee pays a key role in the financial activities of our daily lives that involve exchanging money for goods and services.

Bitcoin solves the problems faced by centralized currencies

Bitcoin was the pioneer of fully digitized currency that was envisioned to solve some of the key problems linked with the fiat currencies system. By using Bitcoin, transactions are fully digital and the fees are minimal. This is possible because the system is decentralized.

Bitcoin is a decentralized and distributed system where all the historical transactions that took place among financial agents are stored after validation. Once stored in the ledger framework technology known as blockchain, the information is totally immutable. The blocks inside the blockchain are building blocks comprised of a number of transactions validated and advertised by some super users called "miners" in the decentralized network. All the network activities and ledger system are publicly available and secured, which offers some key features.

Finite supply:
The total number of Bitcoin mined from the Bitcoin system is restricted to 21 million. In a centralized fiat currency system, government banks can print as much as currency as possible. As previously mentioned, this causes inflation, whereas the Bitcoin currency is theoretically an efficient deflationary currency.

Secured immutable information:
Immutability of information is one of the key feature of blockchain. This underlying technology ensures the information validated once in the ledger system published in the block is unalterable Even though both senders and receivers of Bitcoin may be cautious about the transaction, the information is protected in the decentralized system and thus it is theoretically impossible tasks for hackers to alter the data stored in the blockchain.

Anonymity:
In the traditional system, the sender and receiver's information is stored in the banks they are affiliated to. Bitcoin works with a completely different philosophy. It requires only the Bitcoin addresses assigned by the system to transfer money from the sender to the receiver. Although this reduces information maintenance tasks compared to the traditional system, it also has some shortcomings.

Decentralization:
This has no single information storage center, rather the full blockchain is stored on each client's device, making it a decentralized and distributively connected system of computers. Thus, it is neither controlled by a single authority, nor can any of the records be destroyed at one or a few centralized points in the blockchain. This makes the network very trustworthy and transparent for secured flow of sensitive financial information.

The protocol of Bitcoin

In this part, we explain the technology behind the bitcoin blockchain and how it solved the real world issues successfully to become the pioneer of first plausible fully digitized currency.

Concept of digital signature When we perform a bank transaction, it needs us to authenticate ourselves to the system. We may identify ourselves by our national ID card, passport or handwritten signature. In any cases, this personalized authentication system is in place so that we can verify ourselves to transfer and withdraw by ourselves only. Not by an imposter to steal our valuable financial asset from us. We have discussed above that, Bitcoin is publicly available record system and all the transactions inside are stored in chain of blocks. How Bitcoin make sure the users are authenticated by the system is achieved by the encryption technique popularly known by the term digital Signature. In the cryptographic world the technique is the implication of "Asymmetric encryption". Let us clearly understand what digital signature offers us to secure our identity.

A digital signature normally ensures the message comes to the receiver from authenticated designated sender and also confirms while reaching the receiver it has not been tampered by hackers. To understand digital signatures, two concepts of cryptography are important. One is a hashing algorithm, and the other is asymmetric encryption. A hashing algorithm ensures that an input of string or text is irrevocably converted into an output that is unique and also has fixed length. Bitcoin uses standard SHA256 as a hash algorithm [12]. The main concept of that hash algorithm can be metaphorically explained by the example of baking a pie. All the ingredients such as flour, milk, water, salt, baking powder and so on can be considered the input message. The SHA256 algorithm can be considered the pie dish, and the pie is then the output hash encrypted text. The only difference is that if a single input ingredient is changed even a fraction, then the output a completely different output product than a pie. By processing the pie, we cannot convert it into the input ingredients; the output binary hash message also has the same irreversible property. SHA256 is a mathematical tool that converts any plain text into a 256-bit fixed-length irrevocable hash text consisting of binary digits of zeros and ones.

To create a digital signature, the message broadcast to the Bitcoin blockchain network first needs to be hashed. Asymmetric encryption is then applied to the hash. It can be conceptualized with two terms called *public key* and *private key*. In this encryption method, each user has a public key and private key that complement each other. Therefore, if person 1 encrypts a message by person 2's public key and sends the encrypted message to person 2, then person 2 can decrypt the message using their own private key and vice versa. This is a similar scenario as sending email. We send a message to someone's email address, which is their public key that is known to us, and they can log in to their mailbox by using their private key to read the message.

In the Bitcoin scenario, the ledger is public. Asymmetric encryption plus hashing still

contributes to ensuring the proof of the sender's authenticity to the receiver and also the proof of the message not being tampered with.

Let us explain the Bitcoin blockchain scenario, where Alice wants to send some Bitcoin to Bob as an exchange of goods or services. In the network, Alice sends two pieces of information, one is the hashed version of transaction information, which is digitally signed with her private key, and the other is the transaction details, which remains unencrypted and contains the transaction information of the current and the previous transaction information that ensures there are sufficient funds available. Both pieces of information are sent to Bob, who receives both the message and normally does two things. First, he decrypts the digitally signed message with Alice's public key. If the action is successful, it proves the sender's identity is authentic and generates a Hash A output message. Bob also simultaneously imposes a hashing algorithm for the unencrypted transaction message sent by Alice and generates the output Hash B. If Hash A and Hash B are found to be identical, then it is verified that the message is not altered between the transmission because both hashes are originated from the same message. Furthermore, only Bob can access the messages sent to him using his private key and confirms the transmission of an authentic message from Alice through his actions.

How information are stored in blockchain The scaling problem is one of the shortcomings of the decentralized network. As there are no centralized repositories, it is difficult for each and every user to store the entire blockchain in their devices, which is a system requirement once the network starts to sync. On the contrary, by compromising with the scaling issue, it enhances the security by creating millions of copies of the blockchain and also making the network much faster. In conventional centralized systems, financial institutions, such as banks, own dedicated servers that can comprehensibly store all users' personal and ledger data. Although this central administration of data is highly secured and confidential, several examples of hacking as well as conflict of interest from the central governing body have compromised the security of the network. The Bitcoin network is distributed and decentralized. The ledger or the information inside the network is peer-to-peer based and publicly available. The network achieves this by connecting and creating a distributed network making the data available by deploying peer-to-peer sharing technology where all the participant computers considered as nodes download the entire blockchain once they enter the network; hence, there is the need to deal with the scaling issue of its information repositories.

How transactions are formed inside the blocks In our Alice-Bob case, verification is needed as to whether Alice has sufficient funds to send to Bob. In the Bitcoin blockchain, there is no alternative way to calculate the current balance, rather it is done by total Bitcoin flow coming from all previous transactions. In the system when a user creates their wallet for the first time, they receive a complete historical copy beginning from the current block of transactions to the first or genesis block. After calculating the balance, if it is clear that Alice can proceed with the current transaction, a transaction message containing the amount that going to be transferred to Bob broadcasts hash or public addresses of Alice and Bob and lastly the digital signature created by Alice. As the message is broadcast, any user can observe and collect information for mining. Along with other transactions to be verified, this transaction is temporarily stored in a pool, which is known as a mempool or memory pool of non-validated transactions. This is the place where miners pick up transactions to confirm and participate in the *"Proof of work"(PoW)* process. Miners are normally special users with larger computation powered computers and they participate in the PoW process to confirm the unverified

transactions in the mempool. In the PoW process, all the miners participate to solve a puzzle, whoever wins the competition gets the privilege of putting the current block of transactions into the live blockchain. They also receive all the transaction fee and the PoW fee summed up in the first transaction of this published block. In a normal scenario, miners are free to pick any transactions. Even though most miners attempt to pick transactions with higher transaction fees, there is no guarantee that they will solve the crypto puzzle competition, which is still proven to be completely bias free.

The amount of PoW fee updates every four years. It becomes exactly half of the previous quarter's block reward. For example, from 2009-2013 the block reward was 50 BTC, which decreased to 25 BTC from 2013-2017. In current block reward is 6.25 BTC. This is the only standard way to generate new Bitcoins from the system. Internet sources claim that 88% of 21 million Bitcoin has already been minted [13].

How blocks are verified in the blockchain In the previous section, we presented a brief overview of PoW as a verification process of blocks involving miners. Here, we discuss it as a step-by-step process.

The PoW method ensures the verification process is done before the inclusion of every new block tagged to the blockchain. The process has consensus, and it is bias free. Furthermore, as it is somewhat time consuming, there is almost no chance of double spending.

An important term in the blockchain network is *"hash rate"*, which means the rate at which the puzzle for PoW needs to be solved. In every four-year cycle, the hash rate also increases as the network grows to restrict the "51% attack." The increase of the hash rate forces miners to upgrade to more competitive computational power to participate in the PoW process. During the competition, miners need to use their computation power to solve a puzzle in which they guess a pseudo-random number, known as a "Nonce" value. Whoever guesses it first wins the reward from the PoW process.

For instance, a miner is processing a block. There are three pieces information they to put in their block. The first the hash of the previous block. Then, they need to add all the transactions they just picked to add to this block. They also need to verify the "Nonce" value that their computer needs to guess for the current block's hash value, which begins with specific number of zeros. Here, we have to recall the hashing algorithm that we discussed earlier. To achieve the goal, recall that a slight change in the input will swap in an entirely changed output. Hence, to determine a specific number, the miner needs to correctly guess the nonce value. Their computer attempts to achieve this while competing with other miners' computers. A very high computational power is required to guess this magic number, which is why specific investments are needed to engage in mining activities. The Fig. 1 shows the block verification process in general.

According to the Bitcoin protocol, this process of inserting a new block or as popularly termed, "mining a new block," normally takes no more than 10 minutes. As the hash rates changes in a periodic manner and so does the number of zeros in the new block's hash, This hash rate increases in course of time resulting continuous upgradation of computational hardware and increasing utility cost in terms of electricity. It discourages old miners to keep in control for longer period of time. From an economic perspective, this ensures the impartiality of the entire system.

We can see that each block possesses the hash of the previous block as a referral chain. Hence, if an imposter attempts to change a transaction from a published block, they have to recalculate the not only the current hash but also all the previous hashes starting from genesis block hash. This is because the slight change in the hashing algorithm results

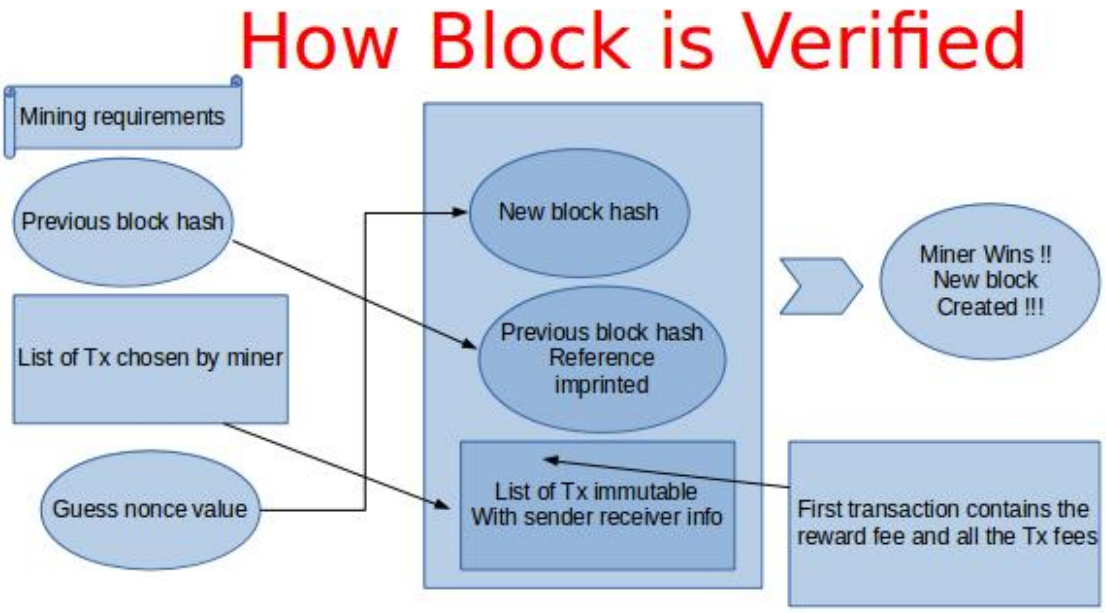


Figure 1: Block validation process

in a completely different output. Furthermore, it has to be done in 10 minutes, which makes it an almost impossible task to achieve. It also suggests that each time a new block is added, and as the hash rate is updated at regular intervals, network security becomes tighter. With the rise of quantum computing, there is a buzz in the computer science world that such computing may achieve the impossible task of gaining control over the blockchain network, but this is yet to be proven.

How bitcoin solved double spending problem We now focus on the real-world conflict scenario of buyers and sellers. In this section. . . we discuss the consequences of double-spending in the Bitcoin blockchain. . For example, Bob wants to buy an iPhone from Alice's store, which supports Bitcoin payment. Bob placed the order online and then sends the Bitcoin to Alice's wallet. The deal is Alice would ship the iPhone once she sees the payment from Bob and it is reflected in the blockchain. However, Bob knows it takes 10 minutes for every transaction to be approved and he attempts to create one transaction to send the Bitcoin to Alice and another transaction in which he sends the money to himself. We examine what happens if the illegitimate transaction is created just before the legitimate transaction is confirmed so that Bob in fact gets the iPhone for free. The best and safest practice for Alice is to wait until only the legitimate transaction survives to be added to the blockchain. In reality, sometimes more than one block is added in the chain, creating a fork. In these instances, the following blocks are added into the forked chains chosen by the miner. At some point, one branch becomes longer than the other and at that time, all the illegitimate transactions break from the shorter chain and swerve back to the memory pool or mempool. It is therefore recommended to wait for at least six more blocks to be published to obtain confirmation of the validity of the transactions in the current block. Recent transactions added in the current blocks are occasionally called *"hot transactions"*. Imposing higher computational power by miners and higher probability of selecting only legitimate transactions thus ensures the transparency of the PoW process. Let us discuss another possibility, in which Bob be-

comes the miner, and he adds the legitimate transaction in one branch the fraudulent one in another. He attempts to continue the fraudulent chain at a same rate along with the legitimate one. However, at some point his legitimate transaction from the branch will shunted back to the mempool and thus would become invalid as it gets conflicted with the fraudulent branch transaction. Furthermore, as it has the same signature as the fraudulent transaction, even if it is picked up one more time, it will be considered invalid. Theoretically, it is a possible scenario, but in reality, Bob needs to obtain control of 51% of the computational power of the full network just to attempt this. This is called the *"51% attack"* and is very difficult to achieve. In addition, it is even harder as the time duration for the consecutive block creation process is kept to 10 minutes. The bottom line of this scenario is that the fraudulent transaction will eventually be dropped down to the mempool and the authentic chain with valid transaction would keep getting longer. Ultimately, is not worth Bob investing such a significant amount of resources and effort to attempt this fraudulent transaction.

1.2 Purpose of our research

In this section, we discuss the purpose of our research focused on the Bitcoin blockchain network. Publicly attainable transaction data is the main motivation of analyzing financial networks. Several studies have examined the descriptive statistics, network expansion, network topology, and the dynamics of the Bitcoin blockchain network. The current study is motivated by the already extensive research in this area. The "User Graph" creation and analysis was based on the famous heuristic rule that states that every input in a multi-input transaction must be linked to a single user as it knows all the private addresses of those input public addresses. This was elaborated upon in [7] and also in a later section of this paper. These researchers also discussed unusual big data flows and temporal analyses for the Bitcoin blockchain. [14] performed an elaborate quantitative analysis on the major wallet exchange markets , which have a large number of public addresses, and by pinpointing the chain with a high range of threshold incoming Bitcoin value. Finally, in a detailed investigation using extensive analysis of transaction networks, a group of Hungarian researchers [15, 16] applied linear preferential attachment. In their extended work, they proposed a model that shows how structural changes in the network accompany significant changes in crypto-exchange prices. This Hungarian research group has uploaded blockchain data from 2009-2018 [17], which is also the main data source of our research. Our research mainly focuses on particular aspects of a closed economic system such as blockchain.. These are mainly concentrated on the total amount of generated cryptocurrency and the transaction patterns demonstrated by the money flow inside the blockchain. For the last 4/5 years, the Bitcoin market has gradually received increasing and significant attention from investors, technology entrepreneurs, and currency enthusiasts, which lead temporal growth of this special financial network.

In the beginning part of research, we have plans to come up with some stylized findings by our investigation on the bitcoin time-series transaction patterns. We wanted to investigate on who are the outliers' in this financial system? We also had interest on the transactions pattern. The main implication of these findings in the time-series situation provided us insights about the cryptocurrency based fully digital financial system which contributes in the field of socio-economic study. Besides in our analysis, we also keep our eyes on the cryptocurrency real world major historical events in order to link those with their impact on the network.

We have mentioned that main focus about our research is unveiling the anonymity of important users inside blockchain network. The goal is not to attack on the cryptography rather to investigate on the transaction patterns and users' activity behavior. We have found out that there are weekly patterns in a bitcoin volume to the price per day graph and there is a clear sign of economic financial trading of bitcoin flow among the transactions. The distinctive nature of transactions impulses in

the blockchain especially focused in different time slot leads to special network effect. We would like to investigate on these scenario to understand the dynamics of behavioral change of users.

Price analysis was out of the scope of our research as it is an exogenous factor of the blockchain network. In spite of that we investigated the effect of this factor on the endogenous attributes of the bitcoin blockchain. Furthermore, there are also real world users who for the nature of their businesses compromises their user identity by publishing it bitcoin addresses publicly. For example, most of the exchange markets expose their pubic bitcoin addresses in order to promote their services to the customers. In our research we also were interested about their activities inside the blockchain network and how it is reflected with their daily activities. This revealed new ways to understand the topological structure and flow dynamics among the exchange markets and the linkage with their portfolio evolutionary growth.

Another goal of our research was to understand the money flow inside the blockchain. Money flow and price mechanism are the fundamentals of the economic activities in the financial arena. How flow of money inside a enclosed financial system where the supply is limited is an interesting question to understand . This could be interpreted by understanding the bitcoin flow dynamics on the complex network of bitcoin blockchain transactions. We analyzed bitcoin blockchain data for the period between the year 2013 to 2018. Bitcoin flow attributed as edge flow circulated among users in daily/weekly/monthly and how users are located in the entire flow would be very interesting. Finally, we targeted to propose a methodology to identify the exchange market or financial institutions and distinctively classify their activities in the bitcoin blockchain on the basis of fulfilling some key criteria.

1.3 Outline of the thesis

This thesis is structured as follows: In this section 1, The basic concepts of bitcoin and blockchain in terms of technological point of view are explained . In section 2, the related previous works have been discussed. In section 3, the definitions of the important terminologies has been depicted using mathematical expressions. In section 4, the main source of data and the important variables and their statistical distribution has been analyzed. In section 5, the transaction graph analysis and the weekly pattern of number of daily transactions and daily volume sum have been investigated. In section 6, the standard user graph analysis done. Novel criteria of ”Big players” has been introduced to identify the financial institutions and non-financial institutions inside blockchain among the top frequent and daily persistently active users showing the weekly pattern for normalized BTC flow on average days of week. Finally, in , section 7, the clustering of the big players' peers has been tracked with the implementation of non-negative matrix factorization(NMF) technique and simulation result have been analyzed.

2 Literature review

In this section we will discuss about most of the research works that have been conducted till date focusing specifically on the de-anonymization of users' identity in the blockchain, complex network analysis and price analysis of the exchange market.

2.1 The de-anonymization of Bitcoin blockchain

In 1983, David Chaum, an American cryptographer envisioned *ecash*, a cryptographic anonymous electronic crypto-currency [18]. In 1995, he led the implementation of Digicash [19], an early structure of electronic payment system which required user software that ensured cryptographic withdraw of bank notes. This was the first step to withdraw a digital currency without the authorization from third parties like governments and banks.

In 1996, the Cryptography of Anonymous Electronic Cash, describing a Cryptocurrency system, first published by National Security Agency in an MIT mailing list[20]. An anonymous, distributed electronic cash system named "b-money" was introduced in 1998 [21].

The first successful, decentralized, anonymous cryptocurrency, Bitcoin, was introduced in 2009 that we discussed earlier that by the cryptography expert Satoshi Nakamoto. We also came to know that, the hash algorithm it used was SHA-256 (Secure Hash Algorithm 256), which is the standard cryptographic hash functions created in 2001 by the United States National Security Agency (NSA).

Bitcoin is pseudonymous, that is, not completely anonymous. We already have discussed that, the currency inside the wallet is tied to some particular keys (or "cryptography generated hash addresses"). Anonymity in Bitcoin is a complex issue. In the Bitcoin blockchain system the users are hidden behind their public hash keys. One of the very popular approach of unveiling these users under the public keys are mapping the multiple inputs or sender keys when used in a single transactions linked to one user or entity [7]. The heuristic got very popular in the research world and in our work we have used this famous algorithm to contract addresses to users.

Meikle et el. had explored Bitcoin network and by network analysis he found that the identity of user inside blockchain is not totally anonymous and their activities can be observed by their currency flow [22]. The researcher group also collected evidence by purchasing goods and services from real world crypto-markets and categorized those into groups. This helped to cluster the crypto-market based supplier-consumer heuristic users. This technique they implemented was termed in their work as *"re-identification attack"* in order to do classification. Apart from the crypto-currency exchanges as a third party user's personal information are not shared but the transaction details are available globally. The de-anonymization is an important research topic which as a consequence would understand the money flow among users with their peers.

The popular heuristics of Bitcoin anonymity:

There are mainly two popular heuristics of clustering public addresses or hash keys to identify a user or entity: change address detection heuristic and common input ownership heuristic. In our research, we will discuss about the later one as it is one of the technique we used in our data processing to create the "user graph" also adopted by the handful of other researchers (see [7, 23, 8, 16, 15, 4, 24, 25, 26, 27, 28, 29, 30, 31, 32] for example, and references therein). The heuristics goes like this:

"In a single transaction, the input address is all from one entity. If one of these addresses along with others participate in a different transaction, then all of the input addresses which belong to both of the transactions can be clustered to the same entity or user."

This is also known as common input ownership heuristic. There could be possibility that input addresses link to the multiple entities in real world. But as they share the same ownership they are treated as one, see Fig. 2.

Many blockchain analysis has been done with the address cluster heuristic. This is written in the white paper of Bitcoin as multi-input heuristic[6]. Reid and Harrigan [7] utilize this heuristic in order to attack on the anonymity of the Bitcoin users. With the inclusion of change heuristics, the de-anonymization technique has been expanded [33, 22, 34], to study the temporal nature of the network growth [35, 36] and transaction graph Bitcoin flow [24]. The analyses that has been done in our research are based on the multi-input heuristic only, we have extended that heuristic with some added assumptions and have created our own version of address-to-user database.

Ober et al. [23] researched about the lifetime and size of user community and important finding is that these community follows the scale-free distribution. Lischke and Fabian [26] showed that in the first four years the major hubs and authorities of the user graphs are exchanges, gambling sites, dark net users, and mining pools.

Maxwell introduced CoinJoin [37], a protocol that works as a counter measure of attacking with address-user heuristics for trust-less but mixing of Bitcoin transactions with a central authority. Multi-input heuristic becomes false positive after applying this protocol. CoinJoin needs a third party centralized server that makes the mixing of transactions. Similar protocols with same functionality are Blindcoin [38] and Mixcoin [39]. There is option available with decentralized mixing that does not require trust of third party. In this category CoinShuffle [40], CoinSwap [41], CoinParty [42] are such Bitcoin protocols. Shentu and Yu et el. [43] also have investigated on similar Bitcoin protocols.

Imwinkelreid [44] described connection of users doing crimes on the digital world by network analysis of the user graph. Similar research was done by Moser et al. [45, 46], who explored the anti-money laundering issue of the Bitcoin users by analyzing the user graph.

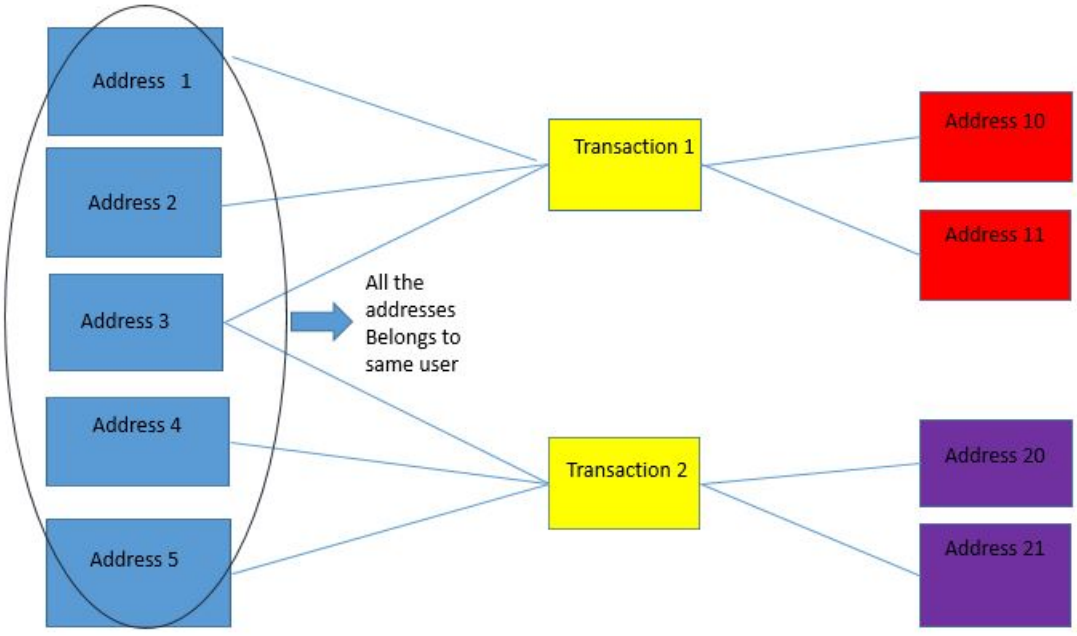


Figure 2: Bitcoin common-input-ownership heuristic

”Silkroad”: revealing of an anonymous illegal marketplace:

Nicolas [47] in his research, gathered and analyzed the blockchain data constituted for the users who are the enjoyers of *”Silkroad”*, an online market place of illegal substance and

utilized the anonymity features of Bitcoin to hide their buying and selling activities. He obtained a detailed six months' data of the year 2012 and merged those with the blockchain address-to-user data in order to track the users behind the anonymous hash addresses. The website administrator was the main architect of using both Bitcoin blockchain and *Tor* software [48] in order to hide the identity of the shoppers. The researcher showed that silk road is the market place of 24,400 separate substances related to drugs and narcotics that are sold in the site within a very busy 3 weeks of time interval. Inside the time interval of six months the researcher had tracked 120 regular users who were actively participating as buyers and sellers. A large number of commissions in Bitcoin, which is equivalent to around 92,000 USD/per month, were transacted to the site operators' wallets. A comprehensive analysis on the daily sales and revenues generated were also shown by analyzing and cross checking of the website and blockchain data.

During the researched time period of six months around 1.3 million BTC has been exchanged among the buyers and sellers in the silk road. Over this entire period the total number of Bitcoin transacted by crypto-exchange market is around 29,553,384 BTC. Comparing both, it was found that the number of BTC flown in the silk road market is $4.5\% - 9\%$ of total crypto-exchange flow inside the blockchain within the same period of time. As same coin transacted several times in different transactions the transaction flow inside silk road was higher than the entire BTC supply volume in same period. The estimation in the research was not very robust as it is hard to detect how much of the actual transactions done by exchanging fiat currency, but still the blockchain quantification of total value was almost accurate calculated from the digital footprints as evidences. The result of this research was interesting. The scope of this research actually opens up our research goal to understand the economic activities of important users' network structure and their currency flow among their community.

Silk road in the Bitcoin economy operated like an anonymous service, but later when the webmaster was arrested, it has been an established fact that Bitcoin blockchain is not entirely anonymous rather it's anonymity can be hacked partially. But on the other hand, new ways of using Bitcoin with extra anonymity strategies have also been introduced in the course of time as the network grows which made the security stronger. Putting all these into considerations in our research, in order to understand the economic activities behind the secured wall, we took the challenge of computing the other important big users like silk road in the overall Bitcoin blockchain economy.

2.2 The research work on Bitcoin blockchain network analysis

Complex network analysis [49, 50, 51], has gained an increasing recognition in financial economics as it provides further insights to understanding hidden factors. Even though a large amount of financial data, e.g., exchange market price is available, but to understand the user activities and analyses on structure of the financial network, information about transaction details is usually considered sensitive. Blockchain, where a consistently growing list of financial records stored in secured ledgers are accessible publicly in a cryptographic framework called blockchain, provides researchers a unique opportunity in this regard.

Many other crypto-currencies along with different features and algorithm have surfaced, some died and some still prevailing in the crypto world. With their rises in the fin-tech world we are gradually moving to the cashless arena. At the time of writing this thesis, there are over 6,955 cryptocurrencies in existence as of September 2020 currencies in the market. Thus, to gather proper knowledge and insights this is a high time to create research foundation on fintech data.

This subsection is all about the research work done in the Bitcoin blockchain network analysis. Transaction graph and user graph created from publicly available Bitcoin data in light of network

science are the main tools to understand Bitcoin economy. The complex network science discusses the structures and dynamics of networks or graphs mainly focusing on the exploratory statistical analysis, statistical mechanics, evolution of different attributes and parameters [52]. There are also studies on the robustness against failures and attacks, spreading processes and synchronization [5]. Econonophysics is a multidisciplinary field where complex network is studied with the help of probability theory and mathematical models developed. The complex networks consist of firms, banks, families and households governed by humans. Statistical physics and mathematical statistics are two main focus to extract socio-economic findings which helps to make macro-level policies and decisions [53, 54, 55].

Ober et el. [23] had done an empirical research on network structure of transaction graph and network dynamics. Most important finding was, several parameters of the Bitcoin transaction graph seemed to have become inactive over half years of time.

Statistical approaches to find behavioral patterns of users have been investigated in one of the most renowned research of Ron et el. and Bauman et el. [8, 4]. After the genesis block of Bitcoin network had launched the currency were used only for experimental purpose and did not have the attention for commercial use case. That is why researchers found its use case densified compared to the later in first four years [26]. Some of the key network properties of user graph such as clustering, degree distribution and power law were found by Baumann et el. [4] in their extended research. The researcher group successfully proved that like many other real world networks Bitcoin blockchain also follows a scale-free distribution and has a "small world" effect.

One of the leading network analysis of Bitcoin blockchain Hungarian research group Kondor et al. [16, 15, 56] empirically demonstrated that Bitcoin system shows the preferential attachment or "Rich get richer" phenomenon in their complete analysis of transaction and user graph. Using their own reconstructed blockchain data set, they discovered the relationship between structure of the active users' network changes with the price change in exchange. In our research work we had used their data set to construct time stamped user graph and transaction graph.

Christian et el. [29] has explored the properties of transaction graph modeling for the directed acyclic graphs of bitcoin blockchain network. They proposed the TDAG model which describes the nature of transactions focused on assets and their transfer among different entities.

Maesa et el. [31, 25, 57] in their data driven analysis for user graph considered in a network up to till December 2015, after bitcoin getting popularized and attained financial acceptance value among renowned global organizations. The set of analyses they defined includes, the dynamics and the validation of the "preferential attachment" conjecture and the detection of the key nodes which are important for network expansion.

2.3 The research work on Bitcoin exchange market analysis

Among all other currency Bitcoin is the most valuable and has the largest crypto-currency business. After launching in 2009 it experienced a jump in price value from 1 dollar to 19k dollar within just 7 years' time. The underlying technology also attracted huge interest and prospects. With the help of econophysics the reason behind price variations and predicting future price and the market effects can be focused [58, 59, 60].

There had been some work done in correlation with the blockchain network data to market price. The network usage with Bitcoin exchange price rate is the main theme of Baumann's research [4]. There was a work done also related to predicting the price of Bitcoin with latest novel machine learning technology [61]. But, in our research the exchange market analysis part was not focused completely. Price of Bitcoin is an exogenous factor among the attributes of blockchain. There is no information of price inside blockchain. In our work we had used third party websites [62] to collect daily price data for our time series analysis of the endogenous factors like daily number of transactions and daily volume of BTC transacted. As the price formation mechanism influenced

largely by outside factors of the network [63], we avoided to take deep dive in the market data rather keeping our research goals more with the blockchain data. By the way, exchange market data gave us directions in different stages of our research to infer the reasons behind the structural changes of networks.

3 Mathematical notations and methods

This section constitutes the definitions of basic terminologies we will be encountering to explaining our data analysis for Bitcoin blockchain in network science perspective. There are other mathematical definitions and derivations are mentioned in the later part of the thesis. But, this chapter has the basic ones clearly mathematically defined.

3.1 Definitions and notations

The set of all the transactions recorded in the data of blockchain, can be regarded as a giant graph or network, in which vertices or nodes are "users" mapped from addresses (see the preceding section), and links or edges are transactions among users. Let us denote by Tx : $i \to j$ a transaction from user i to j. Note that there are multiple transactions for a same pair $i \to j$. In addition, self-loops $i \to i$ can be present, corresponding to various cases, including the change in a transaction. We call this giant graph a *transaction graph*, and construct it for the period 2013 to 2018 after deleting all self-loops. For the transaction graph, we define *frequency* of individual users as follows:

$$f(i) := \text{number of Tx's such that Tx: } i \to j \text{ or Tx: } j \to i\ . \tag{1}$$

The frequency $f(i)$ can measure how frequent the user i appeared in the transactions that took place during the whole period.

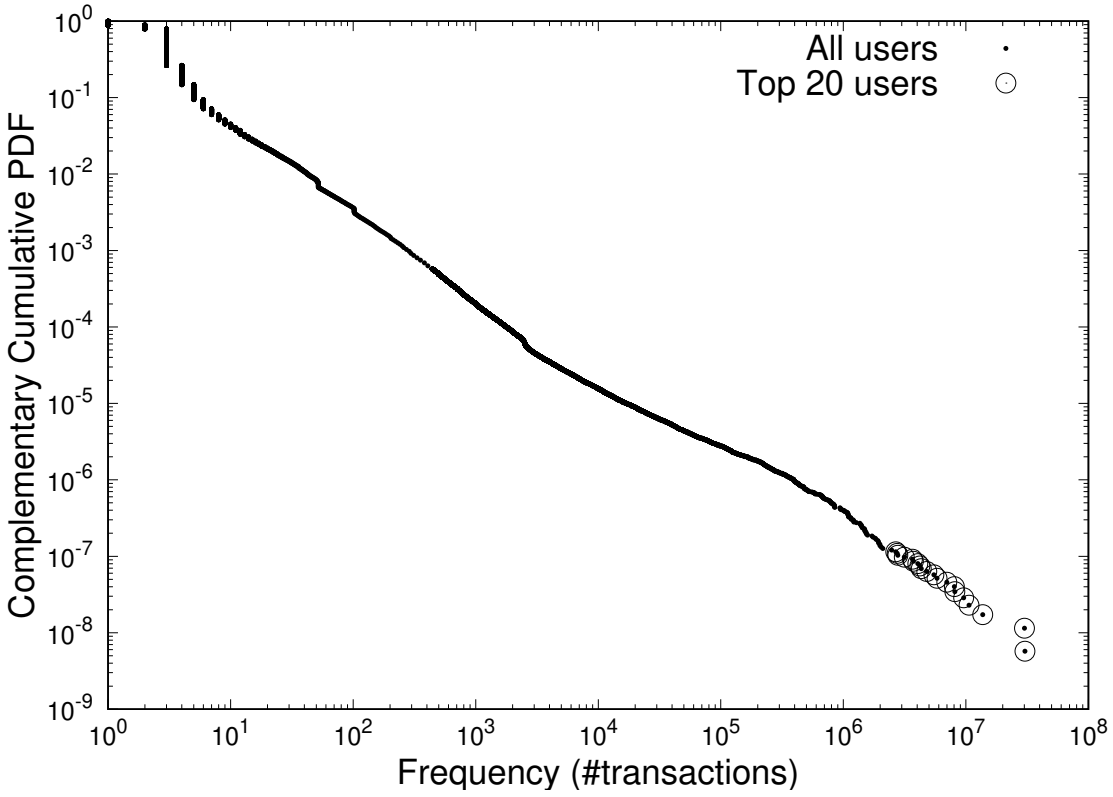


Figure 3: Complementary CDF for frequency of users(acted as input or output) from 2013 to 2018

Figure 3 shows the complementary cumulative distribution function (complementary CDF) for the frequency. One can see that the distribution has a heavy tail with approximately a power-law. We listed the top 20 users in the Table 1. We shall use the frequency to define big players.

To investigate a much shorter time-scale than years, let us construct what we call *daily graphs* from the transaction graph. A daily graph is an aggregation of all the transactions that took place in

Table 1: Top 20 users' total frequency count as input/output during 2013-2018

UserID	frequency
3366757	30329366
109540	30078473
14382265	13710199
27888617	10596775
25703559	9583373
3491614	8118837
6188061	8046046
76589853	6986710
135111428	5815920
18307826	5479502
45976983	4799924
19459450	4321546
62504973	4207824
190362585	4048302
45452467	3732724
65608404	3664173
41633902	3142230
59826853	2786511
11031719	2749122
3959839	2675207

one day. One could define weekly or monthly graphs; however, we would like to use daily graphs for our study, because we shall focus on the weekly pattern of daily activities of users. Let us denote by $G_t = (V_t, E_t)$ the daily graph at time t, where t is assumed to be a day (unless otherwise stated), V_t is the set of vertices of nodes, and E_t is the set of links or edges. An edge e_{ij} is an ordered pair (i, j), which represents all the transactions Tx : $i \to j$ during the day t. The set of all the users appearing at either end of e_{ij} is V_t such that $i, j \in V_t$. Each edge e_{ij} has the information about the amount of money transferred from i to j in the units of *satoshi* (= 1/100,000,000 BTC = 10^{-8} BTC). Note that on day t, there can be more than one transaction Tx : $i \to j$. We aggregated those multiple transactions, if present, into a single edge, and associated the sum of money flow to the edge. Let us denote the amount of money flow for the edge e_{ij} by g_{ij}. This completes the construction of daily graphs from the transaction graph. We remark that G_t does not include multiple edges nor self-loops.

Denoting the number of elements of a set A by $|A|$, in general, we can define $|V_t|$ for the number of nodes, and $|E_t|$ for the number of edges. Regarding time t, we consider two periods, as explained in the preceding section:

$$t \in T_{\text{quiet}} := [\text{1 January 2015, 30 June 2015}] , \tag{2}$$

$$t \in T_{\text{active}} := [\text{1 July 2017, 31 December 2017}] . \tag{3}$$

Subscript t for variables to be defined in what follows may be omitted when the dependence on t is obvious. G_t is a directed network in the sense that each edge has a specific direction. It is sometimes useful to ignore the direction; in such a case, we shall use the same notation $G_t = (V_t, E_t)$ for the

undirected version of G_t. In-degree $d^{\text{in}}(i)$ and out-degree $d^{\text{out}}(i)$ for a node $i \in V_t$ are defined by

$$d_t^{\text{in}}(i) := \text{\#nodes } j\text{'s such that } e_{ji} \in E_t \ , \tag{4}$$

$$d_t^{\text{out}}(i) := \text{\#nodes } j\text{'s such that } e_{ij} \in E_t \ , \tag{5}$$

respectively. For the undirected version, one can define degree $d_t(i)$ by

$$d_t(i) := d_t^{\text{in}}(i) + d_t^{\text{out}}(i) \ . \tag{6}$$

Average degree is then defined by

$$\bar{d}_t := \frac{1}{|V_t|} \sum_{i \in V_t} d_t(i) = \frac{2|E_t|}{|V_t|} \ , \tag{7}$$

where the last equality follows from the fact that each undirected edge appears twice for the two nodes at the ends of the edge. The numbers of nodes and edges, and the average degree in the two periods are shown in Figure 4 (active period) and Figure 5 (quiet period). One can observe that the average degree is relatively stable, around 3.0, much smaller than the number of nodes, which means that the network is sparse and has a small number of nodes with large degrees, namely hubs.

The network G_t is changing in time. For different times t_1 and t_2, even if they are successive in time, V_{t_1} is different from V_{t_2} as a set. However, examining the data, we found that there exist users i that $i \in V_t$ frequently at many temporal points $t \in T$ for a given period of time T. In other words, there are *persistent* users.

3.2 Connected components

A daily graph G_t is not necessarily connected as an undirected graph. In general, a connected component C_a of an undirected graph G_t is defined by

$$\begin{aligned} C_a := \{ & i \in V_t \text{ such that for any } i, j \in C_a \\ & \text{there exists at least one path from } i \text{ to } j \} \ , \end{aligned} \tag{8}$$

where a path is a set of edges, $e_{ik_1}, e_{k_1k_2}, \ldots e_{k_nj}$ connecting between i and j. One can introduce an equivalence relationship between any pair of nodes, namely, i is defined to be equivalent to j if and only if there exists a path between i and j. It is a mathematical consequence that the set of nodes V_t can be decomposed into mutually disjoint equivalence class as follows:

$$V_t = C_1 \sqcup C_2 \sqcup \cdots C_p \ , \tag{9}$$

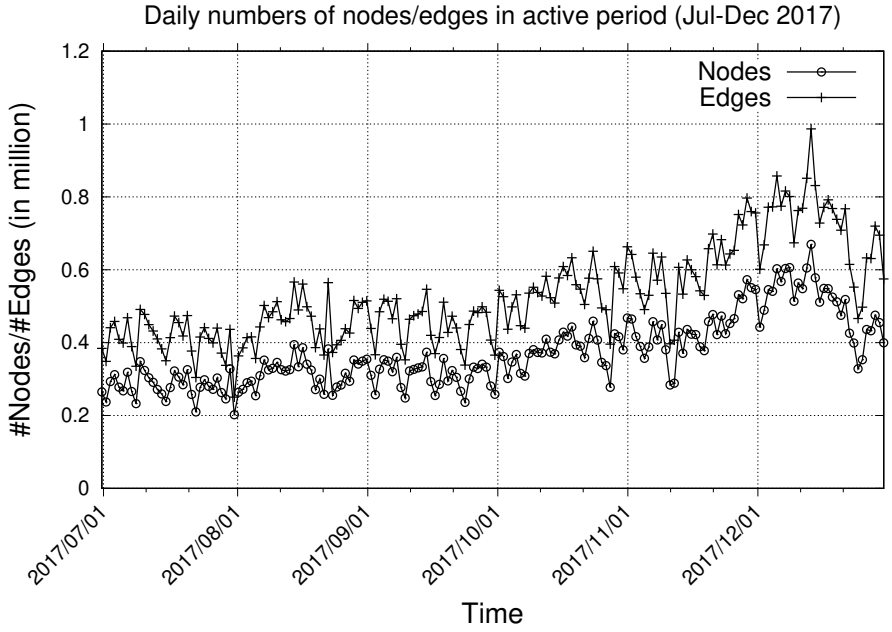
such that $C_a \cap C_b = \emptyset$ for any a, b. C_a is called a *connected component*, and is denoted by $C_1(G_t)$ when we express the dependence on G_t explicitly. p is the number of connected components. It follows from the decomposition that

$$|V_t| = \sum_{a=1}^{p} |C_a| \ . \tag{10}$$

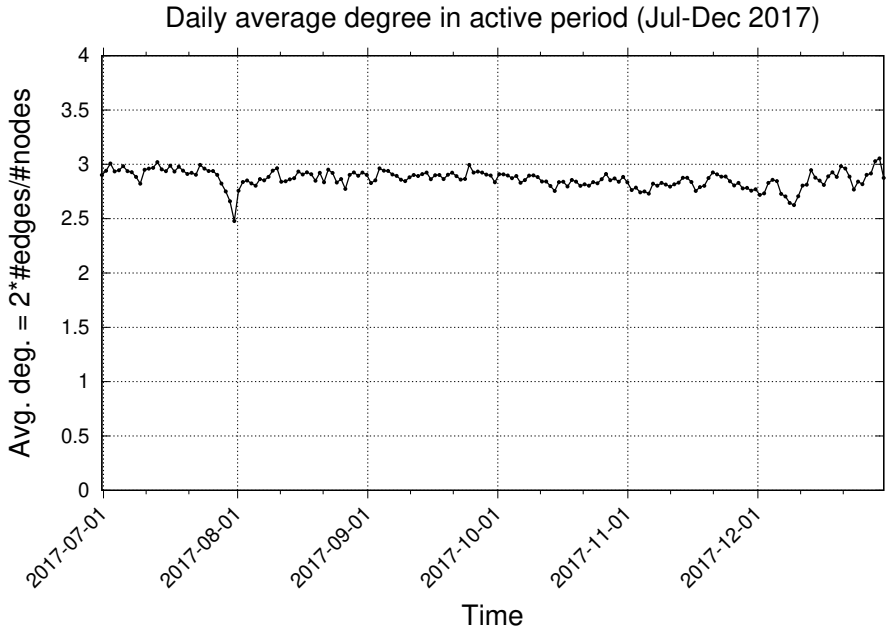
Suppose that C_a's are ordered according to size, that is,

$$|C_1| \geq |C_2| \geq \cdots |C_p| \tag{11}$$

C_1 is called *largest (max) connected component*. We denote $|C_1|/|V_t|$ as the relative size of the largest connected component. We find that often $|C_1|/|V_t|$ is relatively large, typically 0.5, or even larger.



(a) Node-edge count

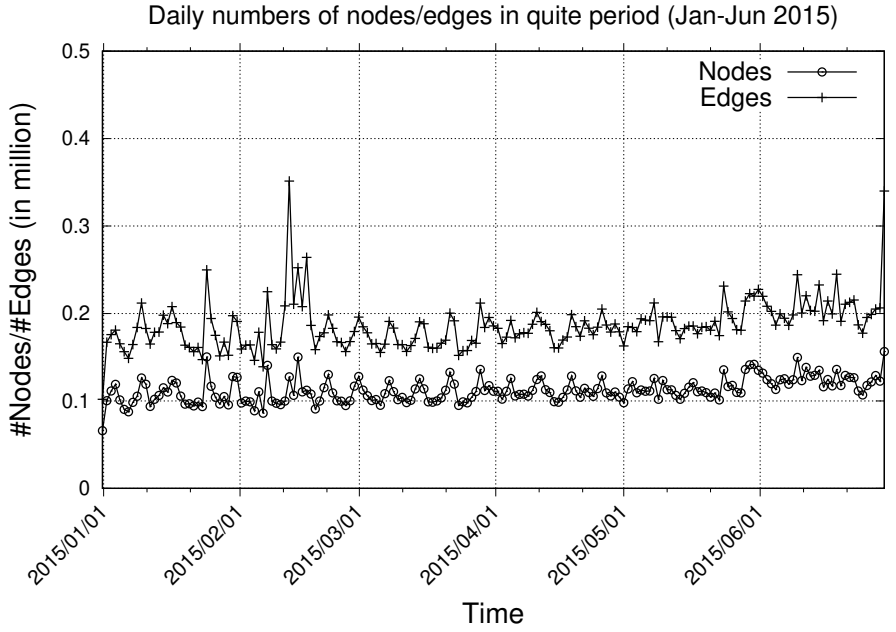


(b) Average degree

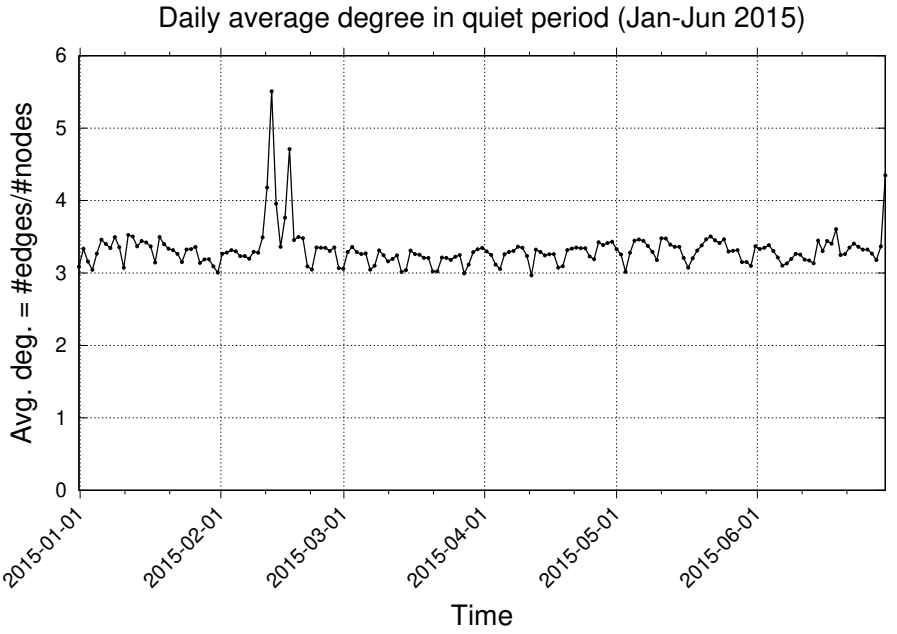
Figure 4: Node-edge statistics in active period; (a) Daily node-edge count (b) Average degree

3.3 Filtered daily graphs

To focus on large amounts of flows in the daily graphs, we shall filter G_t to obtain a subgraph $H_t \subset G_t = (V_t, E_t)$ as follows: Each edge has a certain amount of flow g_{ij} as stated above. We define a certain threshold g_*, which will be determined in the next section, and filter the edges by the following condition: $g_{ij} \geq g_*$, that is, by deleting all the edges that do not satisfy the condition. Let the set of remaining edges be $F_t \subset E_t$. Collecting all the nodes that appear at either ends of each edge, one has the set of remaining nodes $U_t \subset V_t$. This completes the construction of the



(a) Node-edge count



(b) Average degree

Figure 5: Node-edge statistics in quiet period

filtered daily graph $H_t = (U_t, F_t)$.

A filtered graph H_t can be decomposed into connected components, as described above. Let the largest (max) connected component be $C_1(H_t)$. Let the set of nodes in $C_1(H_t)$ be U'_t, and that of edge be F'_t, which is

$$C_1(H_t) = (U'_t, F'_t) \tag{12}$$

In the next section, we will compare the total amount of flows on H_t with that on $C_1(H_t)$. The

former is denoted by

$$\phi_0(t) = \sum_{e_{ij} \in F_t} g_{ij} \ , \tag{13}$$

while the latter is

$$\phi_1(t) = \sum_{e_{ij} \in F'_t} g_{ij} \ . \tag{14}$$

We will also compare the size $|H_t|$ with $|C_1(H_t)|$.

4 Dataset

4.1 The Hungarian researchers' data set

The data set used in this research work has been downloaded from publicly available Bitcoin blockchain by a group of Hungarian researchers and was restructured by them for their own research analyses and later was uploaded in their website[64]. The timeline of the historical Bitcoin transaction data set was compiled from 1^{st} of January of 2009 to 9^{th} February 2018. The number last blocks were updated was 508,241. In the initial phase of our work we have analyzed the daily transaction graph which will be discussed in Sec. 5 and in the second phase we have further reconstructed the data to construct user graph discussed in Sec. 6. But, in this section we have shared the information of the data set of Hungarian research group and their own techniques of representing the blockchain data only. It is mentioned in the website that the data set comprises of several number of text files that contains the key parameter of each transaction of blocks along with senders and receivers' information and amount of Bitcoin flow among them. The researchers had marked all these mapped as the long hash strings of Addresses ID, block ID, transaction ID and user ID with randomly generated unique numbers for convenience for their analysis. This made the computation much efficient and less memory space required to load the data.

Even though the data files description is mentioned we would like to share it here as well in order to explain the parameters more clearly to the readers.

This data set contains the following files:

1. bh.dat.gz (20 MiB): This file contains the long hash of the block ID and the mapped numerical unique identifier of each block. This also contains the POSIX timestamp of publishing the blocks.

2. txh.dat.gz (12 GiB): This is the list of all the transaction hash ID mapped to the unique numerical identifier called as TxID(transaction ID).

3. addresses.dat.gz (9.9 GiB): All the number of addresses hash ID used as inputs(sender) and outputs(receiver) in the transactions are linked to the numerical unique identifier called AddrID(address ID). The address ID are the main security that ensures transactions done anonymously by users shadowed by addresses.

4. txin.dat.xz (7.1 GiB): The transaction inputs meaning the addresses that sends Bitcoins. It also contains the previous transaction ID that helps to create the transaction graph. The volume of satoshi represents the amount the sender transacted.

5. txout.dat.xz (4.8 GiB): It contains the transactions that lists the addresses that receive the satoshi as outputs.

6. addr_sccs.dat.gz (1.6 GiB): This file contains all the identified addresses to users utilizing the heuristics that are mentioned in the previous research work [7].

4.2 Some statistical stylized facts about Hungarian data attributes

Our research data has been reconstructed from original Hungarian data[64] to each phases of our research. All the phases of dataset originate from the Hungarian research group data. We have included data processing subsections for each of the phases mentioned in Sec. 5, Sec. 6 and Sec. 7. In this section we would like to share some insight of the data attributes and familiarize the readers in broader sense with some statistical observations of the original data.

The genesis block of Bitcoin blockchain was first launched on 1^{st} January, 2009. The Hungarian research data has downloaded 508,240 blocks on a cutoff date of 8^{th} February, 2018. The number of unique output transactions, including mining transactions, is 298,325,122. The number

of unique input transactions is 297,816,881. Therefore, the mining transactions from 2009-2018 are (298,325,122-297,816,881=) 508,241, which equals the number of blocks published from 2009 to 2018. This is natural because the mining transactions have only outputs, which are rewarded by the blockchain system without having any inputs. These special transactions appeared at the first transaction of each block, summing up all mining rewards and all the transaction fees directed as an output to the miner.

In this data set, the total number of unique input address (sender) is 347,791,724. The total number of unique output address(receiver) is 369,980,378. The unique input addresses are a subset of the output address, as the structure of blockchain supports that each output of a transaction has the chance to become the input of another transaction if its spent for some purpose. Therefore, 6% of output addresses never appear as input
(369,980,378-347,791,724)/369,980,378 =0.06). These are plausibly the change addresses mentioned in the heuristics [8]. The total number of addresses contracted to users by strongly connected component method or the most popular heuristic [8] was 226,302,814, which is 65% of the total addresses (=226,302,814/347,791,724). The total number of users contracted from addresses was 35,660,272. This is understandable, as the backbone of the publicly available blockchain system is anonymity, which restricts us to map 100% addresses to users. The number of uncontracted addresses by the researcher group were (347,791,724- 226,302,814=) 121,488,910. Some of these uncontracted addresses have very large frequency of appearing in the network.

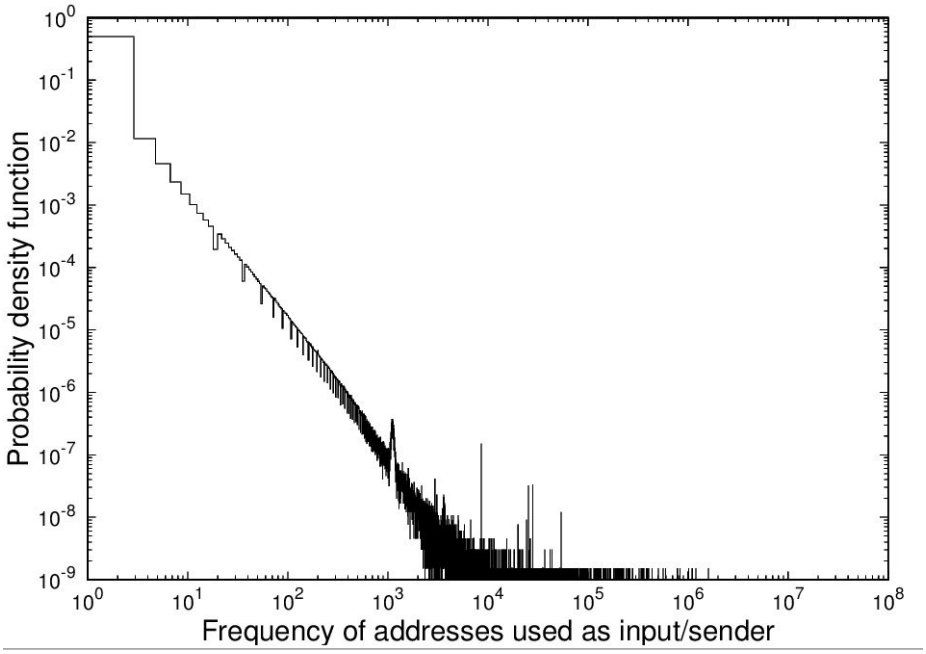
The address.dat file contains the list of hash (using SHA-256 cryptography algorithm) that represents the addresses that hides the identity making the users anonymous. The Hungarian research group represented these long size address hashes, transaction hashes and block hashes with numeric numbers for achieving faster computational processes and less memory spaces. The txout.dat file to merge with this address to user files. total number of 369,980,378 unique output addresses have been merged. similar way, then we choose the txin.dat file to merge address to user file which completed 347,791,724 of unique input address. The unique input addresses are subset of output address as the structure of blockchain supports that each output of a transaction has a chance to become as input of another transaction if its spent for some purpose.

After merging the input address to user we found out that, there are only 1 input user per transactions. The famous heuristics of address to user contraction [8] states all the multiple addresses involved in the inputs of a transactions are linked to one input. So after merging the address to users to inputs we got 297,816,881 unique input users. So these unique users list act as a key-value pair to merge with the output of 811,201,513 number of records of output transactions that has 369,980,378 numbers of unique output users.

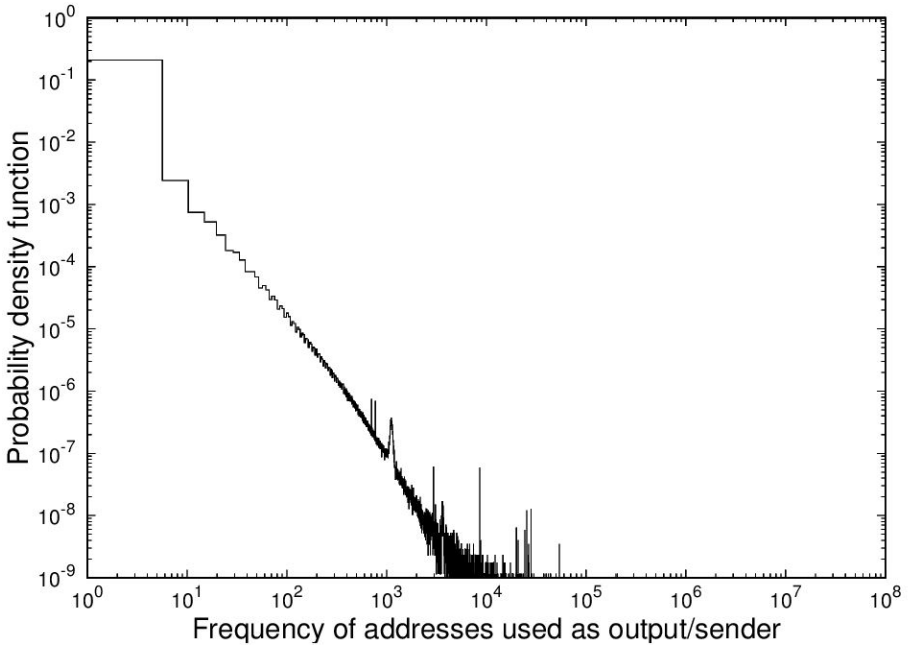
The Total number of all the input addresses are output addresses at the same time, while there are output addresses that never appear as input (6% of output addresses; 0.06 (= (369,980,378-347,791,724)/369,980,378). So these were discarded from their analysis. The total number of address to user by Strongly connected component method total number of addresses contracted are equal to 226,302,814 which is 65% of the total addresses (=226,302,814/347,791,724). The total number of users contracted was 35,660,272.

in Fig. 6 we have shown the PDF of the frequency of input and output addresses. In both cases the heavy tailed distribution suggests that there are addresses have a very large usage records in the network. In the later sections there are more analysis done for the networks reconstructed from the original data sets.

In our research, there are pre-processing of daily data for individual phases in order to measure the temporal change. We presented the detailed individual network analysis in those sections respectively. We also added some monthly network analysis in the Appendix Sec A.



(a) Input addresses or Sender



(b) Output address or receiver address

Figure 6: The PDF distribution measured in histogram (a) Input addresses (b) Output addresses

5 Transaction graph analysis and weekly pattern of BTC volume and transactions

This section constitutes the phase one of our research work. In this section we presented some findings of analysis of outliers' activities. We also found the behavioral pattern of the BTC volume and number of transactions attributes in weekdays and weekends.

5.1 Data set of daily total transactions and BTC volume sum

The data used for this part of the research have been downloaded from the website of the Hungarian bitcoin research group [64]. Their reconstructed database comprises transaction data (sending and receiving bitcoins) with sending and receiving addresses extracted from the blockchain network constituting the time duration from January 2009 to February 2018. The available data have been uploaded on the website in text files and some of the blockchain's extracted parameters have been mapped with randomly generated numbers in order to allocate those efficiently by the research group. Long characters of hashes have been mapped to random indicators, for example, BlockID which starts from numerical 0 value, representing the genesis block (first block of the bitcoin blockchain), and ends up to the value of 501418, which is the last block to download on the cutoff date of the month of February 2018. For our research purpose, we have further restructured the data. The structure of the data has been shown in Figure 7.

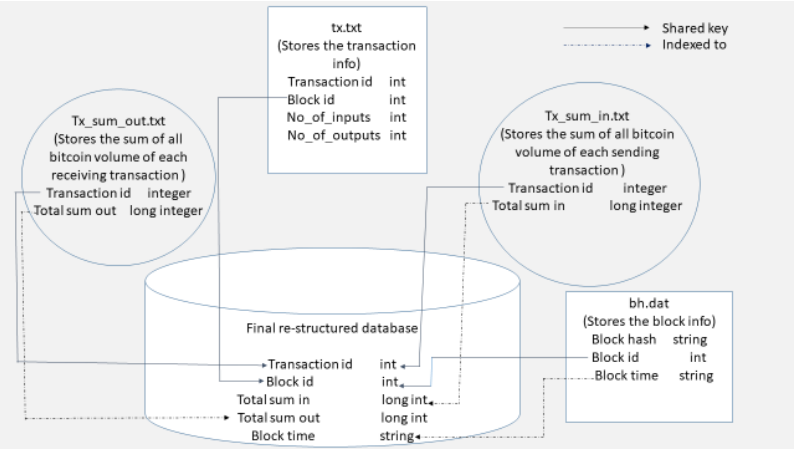


Figure 7: The final reconstructed database that generates the summation of bitcoin volume for inputs and outputs of the transactions recorded during Jan 2009 to Feb 2018 with the block time UTC timestamp.

After reconstructing the database, we had fixed our timestamp units into each day from time duration of 1st of January 2009, when the bitcoin blockchain first initiated, to the cut of date of 8th of February 2018. We summed up the transaction count for each of those days and also summed up the input volume of bitcoin for each transaction. In each block, each bitcoin transaction follows either of the two rules as an input–output relationship in terms of bitcoin volume, i.e., input = output or input = output + transaction fees. This is the reason we have summed up the bitcoin volume of each input of transactions which represents the actual volume of bitcoin exchanged through one transaction to another. Our database excludes the transactions of the miner's bitcoin generation which are called the coinbase transactions, having no inputs, which were filtered out separately to form a separate database to merge into our analysis.

A glance at final data for our analysis has been shown in Table 2.

In Table Table 3, we had the sampled price data with time duration of 9 years since bitcoin genesis block published from the beginning of January 2009. We have downloaded the market

Table 2: The sample of final data for transaction count and Bitcoin volume involved in each day transaction from a data compiled from Jan 2009 to Feb 2018

Data	Exchanged bitcoin volume/day (1 BTC=10^8 satoshi)	# transactions/day
2009-01-12	17900000000	7
2009-01-14	6100000000	1
2009-01-15	50000000000	8
2009-01-16	20000000000	2
2009-01-18	15000000000	1
2009-01-19	77500000000	2
2009-01-20	40000000000	1

Table 3: The sample of Market price in USD per BTC had been compiled from Jan 2009 to Feb 2018 Source:[65]

Data	Price/BTC (in USD)
2011-01-21	0.44
2011-01-23	0.4443
2011-01-25	0.425
2011-01-27	0.4174
2011-01-29	0.446
2011-01-31	0.5
2011-02-02	0.840099
2011-02-04	0.88
2011-02-06	0.92
2011-02-08	0.9
2011-02-10	1.1

price data from blockchaininfo website [65] and used in our analysis.

5.2 Auto-correlation function of BTC volume, price and number of transactions

In this part, using auto-correlation function to see if we could predict the direction of daily log returns. The log return can be defined as:

$$x(t) = \log \frac{z(t)}{z(t-1)} \tag{15}$$

where x(t) denoted log return of a variable z(t) on day t. We measured the log return in order to make the series stationary for the empirical analysis. Now in our case, we calculated the log return of the BTC volume v(t), and the number of transaction TX(t) and the daily price data P(t), downloaded from the blockchaininfo website.

We see in the Fig. 8 the stationary time-series log return plots of volume, the number of daily transaction and price data. The time-series data have been selected from 2013-01-01 to 2018-02-08 in order to maintain the consistency. We plotted the auto-correlation function of the three daily returns with the previous lags. The dotted line is the 95% confidence interval.

For the BTC volume and number of transactions, the time-scale for relaxation was found approximately a week as shown in Fig. 9a and Fig. 9b For price data, in Fig. 9c the ACF vanished at

the lag of 1 day. This is reasonable as otherwise, one can do a linear prediction for up or down of tomorrow's price based on that of today.

5.3 The evolution of bitcoin transactions (a bird-eye view)

The distribution of transaction count to volume with the evolution of time has been plotted. An interesting set of observable to better understand the underlying

evolution of a unique financial system has been demonstrated. We found that there is some impulse of the volume of bitcoin transaction in the different time slots. Our research focused on this evolution of the financial system is after 2013 when bitcoin is a full-fledged matured currency used by people by trading goods and services. Fig. 10 shows that the daily exchange of bitcoin volume substantially increased after 2013 and on wards. In Fig. 11, we have plotted the same graph in log scale and showed the average volume quantity flowing through the number of daily transactions. The weekly pattern of volume flow observed in the graph proves bitcoin having a solid real economic financial system that we have statistically derived in the next section. In Fig. 12, we have portrayed the volume of BTC compared with price and observed the high price volatility.We have plotted another graph in Fig. 13 where we showed the time evolution of price, the number of mined transactions and BTC volume. We observed that the number of supply mining transaction has quite stable throughout the time series.

5.4 The weekly pattern of bitcoin volume sum and number of transactions

We had observed that the volume per transaction became relatively stable after 2013, while it was so volatile before the year. Also, it is known that bitcoin mining to generate blocks has been quite stable since the year 2013. So let us use data from January 1, 2013, in the following analysis of power spectrum. Consider a time-series x_n with $n = 0, 1, \ldots, N-1$, where N is the length of the time-series. Discrete Fourier transform of x_n is given by

$$X_k = \sum_{n=0}^{N-1} x_n \, e^{-2\pi i \, nk/N} \tag{16}$$

where $k = 0, 1, \ldots, N-1$, and $i = \sqrt{-1}$, i.e. the unit of imaginary number. Obviously Eq.(16) is periodic in k with period N, so one can adopt the convention that $X_{-k} = X_{N-k}$. Because x_n is real, it follows that $X_{-k} = X_k^*$, where $*$ denotes the complex conjugate. Frequency f_k corresponding to k is defined by

$$f_k = k/N \tag{17}$$

The range of frequency can be regarded as $-0.5 \leq f_k \leq 0.5$.

Power-spectrum or periodogram is defined by

$$P(f_k) = \frac{1}{N} \, |X_k|^2 \tag{18}$$

where $|X_k|$ denotes modulus or magnitude of X_k. Because $X_{-k} = X_k^*$, one can focus on the range $0.0 \leq f_k \leq 0.5$. $P(f)$ represents how much oscillating or harmonic movement with the frequency f and, equivalently, the periodicity $T = 1/f$ is contained in the original time-series x_n. Therefore, $f = 0.5$ corresponds to $T = 2$, namely the most highly oscillating movement; $f \to 0$ is $T \to \infty$. One often uses smoothed periodogram by applying a filter to the raw periodogram. See standard textbook such as [66] and [67].

We apply the method of smoothed periodogram for the time-series of daily volume V_n and daily number of transactions T_n (where n denotes time in day) in order to find periodicity in them. From Fig.4 and Fig.5, it is obvious that the time-series of volume and transactions have a trend of

exponentially growth, it would be natural to take logarithms of them and to consider the time-series, $x_n = \log V_n$ and $x_n = \log T_n$.

First, we segment the data into different days of week, namely $n = \text{Sun}, \text{Mon}, \ldots \text{Sat}$, and calculated averages and standard error (defined by standard deviation divided by the square root of number of data in each collection). The result is given in Figure 14. One can see that the level of volume or transaction is higher during weekday than weekend; in other words, there exits a weekly pattern that is not obvious in Fig. 10 and Fig. 11.

Additionally, we performed the above method of periodogram for each of the time-series. We employed a detrending by removing the mean of the series and subtracting a linear trend, a tapering with 10% at the beginning and end of the series, and a modified Daniell smoothing with successive simple moving averages of lengths 6 and 12 (see [66] for the details of estimate spectral density of a time series by a smoothed periodogram).

The result is given in Figure 15. For both of volume and number of transactions, one can observe an obvious periodicity at $f = 1/7$ or equivalently $T = 7$ days as denoted by the dotted vertical line. Also present are higher order harmonics at $f = 2/7$ and so on. On the other hand, there is an overall increase of power spectrum towards $f \to 0$, corresponding the trend of exponential growth which one already observed inFig. 8c and Fig. 9c.

5.5 The outliers' transaction patterns

In our analysis, we had concentrated on two of the timeslots to find the outliers transaction pattern. Both the patterns have an unusually high spike of bitcoin volume within 1 month recorded from January 2016 to February 2016 shown in Fig. 16 and even though the there was not much variation of price there was a very big volume bitcoin circulated during the last week of January 2016. There was an interesting finding in Fig. 17. The number of transaction appeared in January stayed in a range of 0.18–0.24 million and the BTC volume spike

cropped up in the last week of January. This indicates that there must be some big volume of BTC fow happened on each of some specifc numbers of transactions as the total number of the daily transactions are within the regular range. This resulted in the existence of some outliers' transactions, which are responsible for the big volume of BTC transacted during that last week. By selecting this transaction pattern timeslot of 1 week (from 21st to 28th January 2016), we recalled our main reconstructed database to find out the list of individual transactions involved in that time frame.

We have created a volume to rank distribution log–log plot in Fig. 19 to understand more about the outliers activities. From the plot, it has been clearly observed that a considerable number of transactions possess the low ranks at the tail of the distribution. Besides, the steep shape of the tail suggested that there is the large rank of transactions that contained a large volume of bitcoin flow. But which are those outliers' transactions that were distinctively traceable among the rest? The most direct method we applied is to use quantiles. The quantiles are values which divide the distribution such that there is a given proportion of observations below the quantile. Mathematically, we estimate the quantile, the value such that a proportion q will be below it, as follows. We have n ordered observations which divide the scale into n +1 parts: below the lowest observation, above the highest and between each adjacent pair. We set this equal to q and get.

$$i = q(n + 1) \tag{19}$$

If i is an integer, the ith observation is the required quantile estimate. If not, let j be the integer part of i, the part before the decimal point. The quantile will lie between the jth and $(j + 1)th$ observations. The proportion of the distribution which lies below the ith observation is estimated by $i/((n + 1)$. We estimated it by x_j+(x_{j+1}- x_j) times $(i - j)$.

Now focusing on the individual transactions of the whole month of January–February 2016, we had calculated the quantile of BTC volume for each transaction which gave us the statistical insight of what percentage of daily transactions have a certain limit of BTC volume involved. As shown in Fig. 18 99% and 100% quantiles on 21^{st}–24^{th} January have very interesting and have a statistical outlier pattern. For example, there is a transaction on 22^{nd} January that has 40000 bitcoin involved in it.

On 24^{th} January, 1% of total daily transactions has 6000–10000 BTC volume at each transaction.

5.6 Directed transaction graph and degree correlation to visualize outliers' activities

The directed transaction graph represents the f;ow of BTC between transactions [7]. Each node represents transactions and each directed edge between the source that is an input (previous transaction) and a target represents an output of transactions (current transactions) as shown in the Fig. 20 Each directed edge also includes a value of BTC flow [24]. Thus, a transaction graph table can be constructed for each transaction as a node having a number of incoming connections called in-degree and outgoing connections called out-degree. Mathematically, the in-degree of node i is the total number of connections onto node i and is the sum of the ith row of the adjacency matrix:

$$k_{\mathrm{i}}^{in} = \sum_{j} a_{ij} \tag{20}$$

On the contrary, i, the out-degree of node, is the sum of connections coming out from node i and is the total number of the i_{th} column of the adjacency matrix

$$k_{\mathrm{i}}^{out} = \sum_{j} a_{ji} \tag{21}$$

The degree correlation is the relation between k_{i}^{in} and k_{i}^{out} and sometimes can make a large difference to the effective properties of the complex network. In our analysis, after considering the outcomes of the quantile plot, we have constructed a transaction graph. It includes all the transactions occurred in between 21^{st} to 24^{th} January 2016.

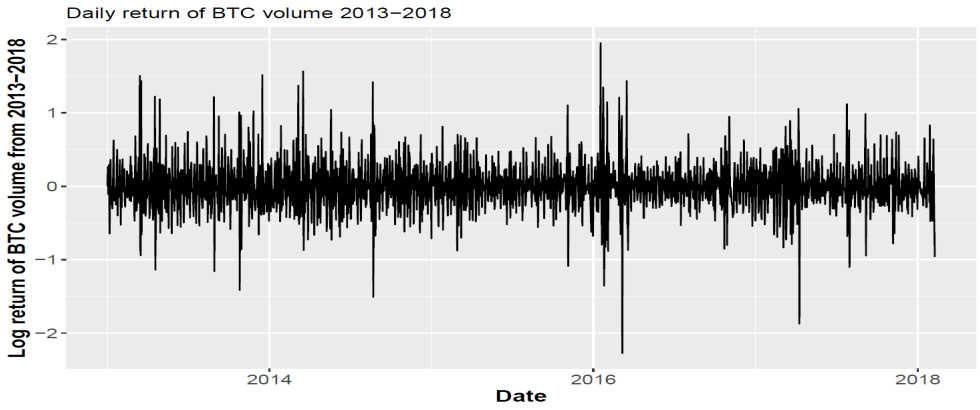
All the transactions are represented as the nodes and links are represented as the input connections from the previous transactions to the current one. We measured the degree correlation of the graph by plotting the in-degree vs out-degree of each transaction node as shown in Fig. 21 . The degree correlation graph visually described important network properties such as how many addresses involved as in-degree from the previous transactions and out-degree to the next ones. We came across an important finding that there was a transaction which has 3033 out-degrees with only 1 in-degree and also there was transaction with 633 in-degrees with 1 out-degree. This leads to the possible explanation of historical events soft fork of Bitcoin improvement protocol up-gradation called BIP-144 that was taking place during that period of time. These outlier transactions could be the result of the bitcoin developers' experiment.

We investigated on the unique transaction volume patterns and based on that we developed a methodology to extract interesting findings from a reconstructed database that has been extracted from blockchain system. We have found out that there are weekly patterns in a bitcoin volume to the price per day graph and there is a clear sign of economic financial trading of bitcoin flow among the transactions. The pattern of weekly trading shown in our analysis helped to investigate more on the specific impulses of transactions in a more focused timeslot. We have analyzed each transaction and bitcoin volume involved in that timeslot. The volume rank distribution helped us to identify outliers transactions with the largest volume of bitcoin involved in it. The SegWit (Segregated Witness) and its effect in terms of the soft fork and hard fork debate were heating up during the

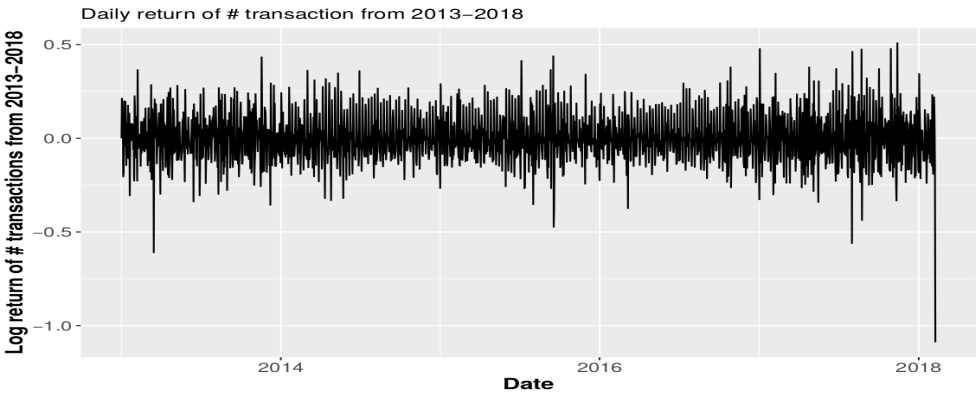
beginning of January 2016 that might be one of the causes of this large amount of bitcoin flow in some outlier transactions.

The bitcoin system has historically gone through a lot of up-gradation which was termed as BIP (Bitcoin improvement proposal). In our focused tenure of January 2016, there was a big buzz in the bitcoin community that the size of block in blockchain needed to be increased. The result would make the transactions per second become faster. So BIP-144 Proposal to increase maximum possible block size starting at 8 MB was proposed in January 2016. Increasing the block size would increase the scalability but it will reduce the transaction fee which in turn would not profitable for miners since a block can hold more data and transactions that can use that new space and thus may be resulted in cheaper fees. So, there was a big debate going on between the miners and developers' community. There was a debate even for adopting either a hard fork or soft fork. Soft forks allow compatible changes. With soft forks, the old and new software can co-exist on the network. Hard forks break compatibility of all previous Bitcoin software and require every participant to upgrade to the same rules by a deadline or risk losing money. Such events can also harm network effects. After long debate and discussion, it was never merged. So, one conjecture we can make by observing these outlier transaction patterns of January 2016 that, there might be a lot of experimentation going on by the system developer community to observe network performances which might have caused this spike of the huge daily transacted bitcoin volumes.

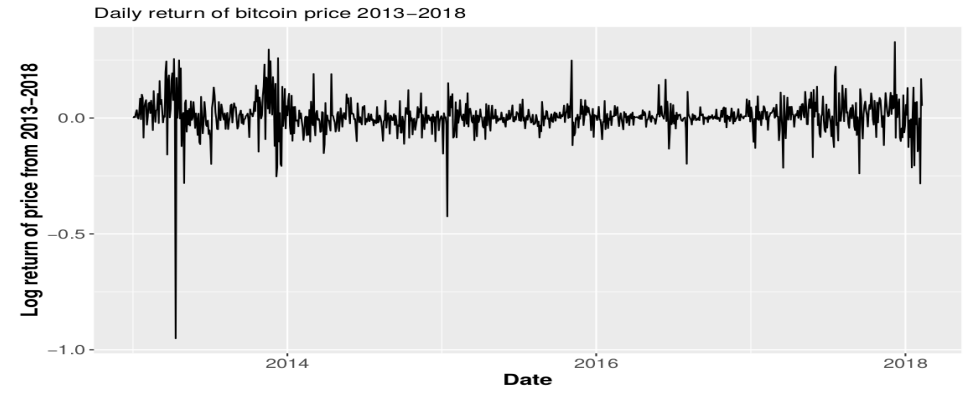
The next research goal was to interconnect the flow of bitcoin with users and transactions and find out more results that reveal new ways to understand the topological structure linkage with cryptocurrency evolutionary growth.



(a)

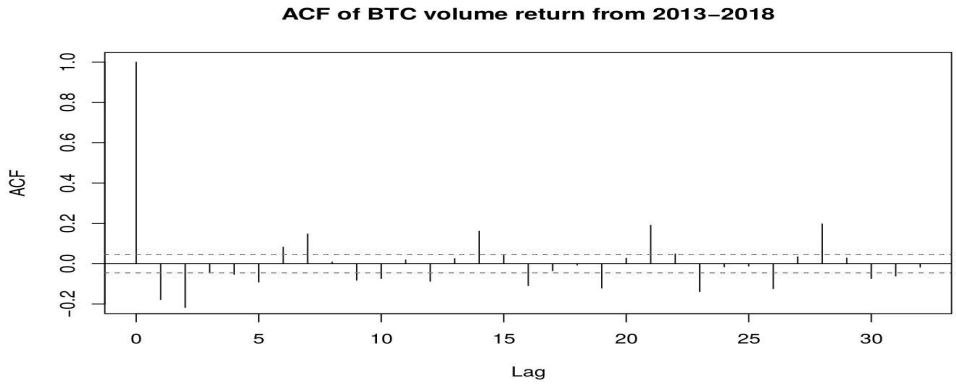


(b)

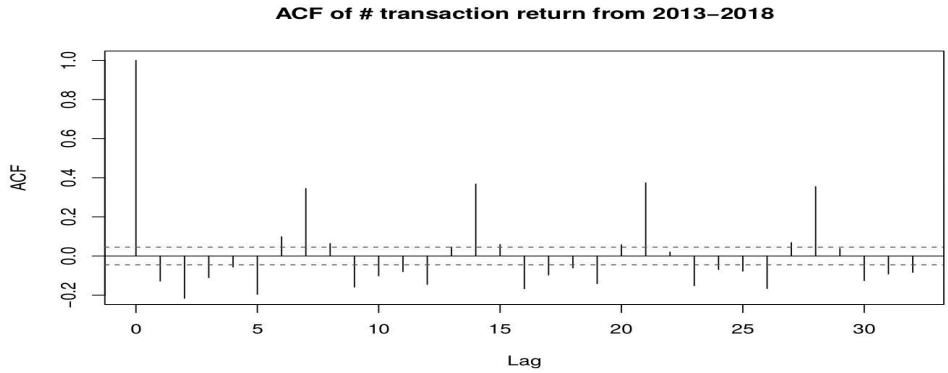


(c)

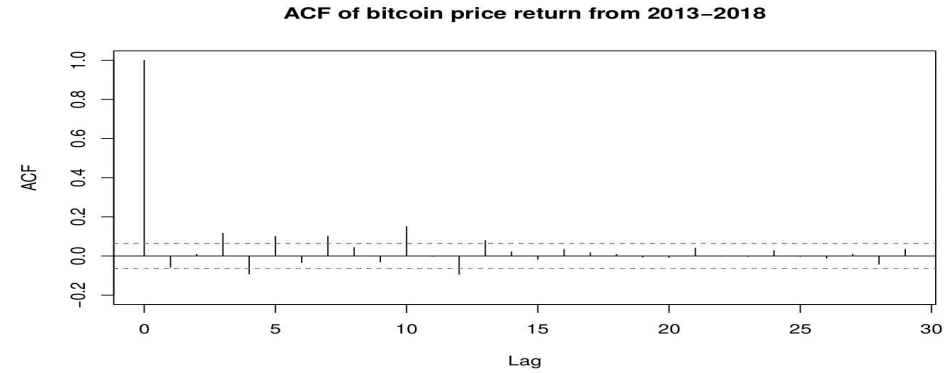
Figure 8: The log return of (a) daily BTC volume (b) daily number of transactions (c) daily price



(a)



(b)



(c)

Figure 9: Auto-correlation function of (a) daily BTC volume (b) daily number of transactions (c) daily price

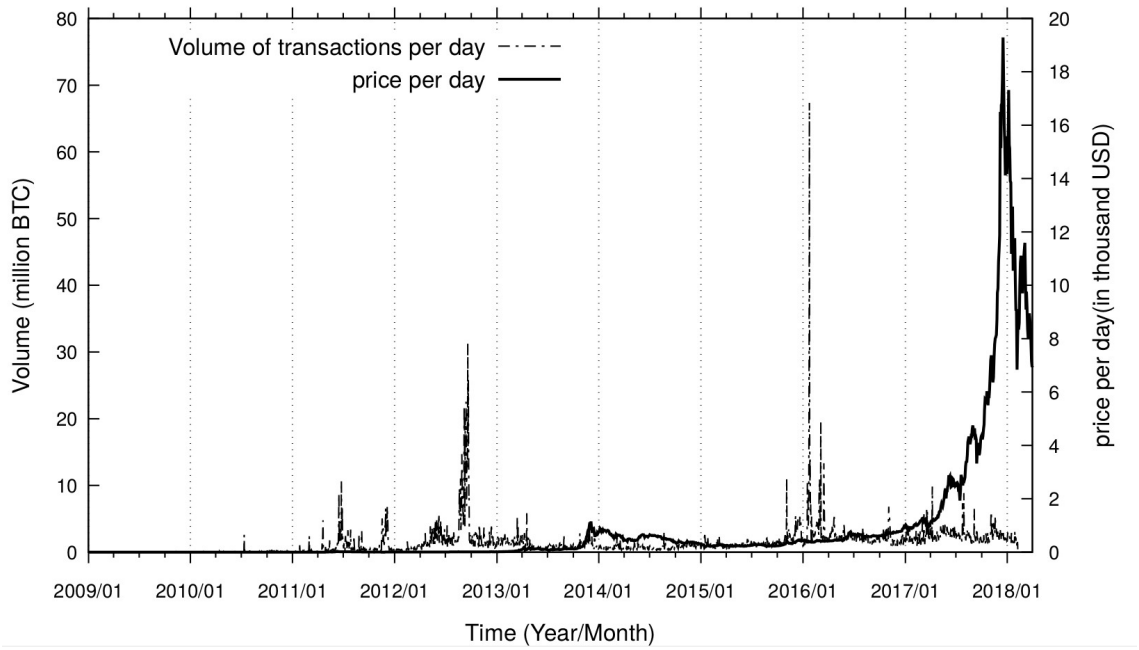


Figure 10: Price evolution of bitcoin volume transacted per day

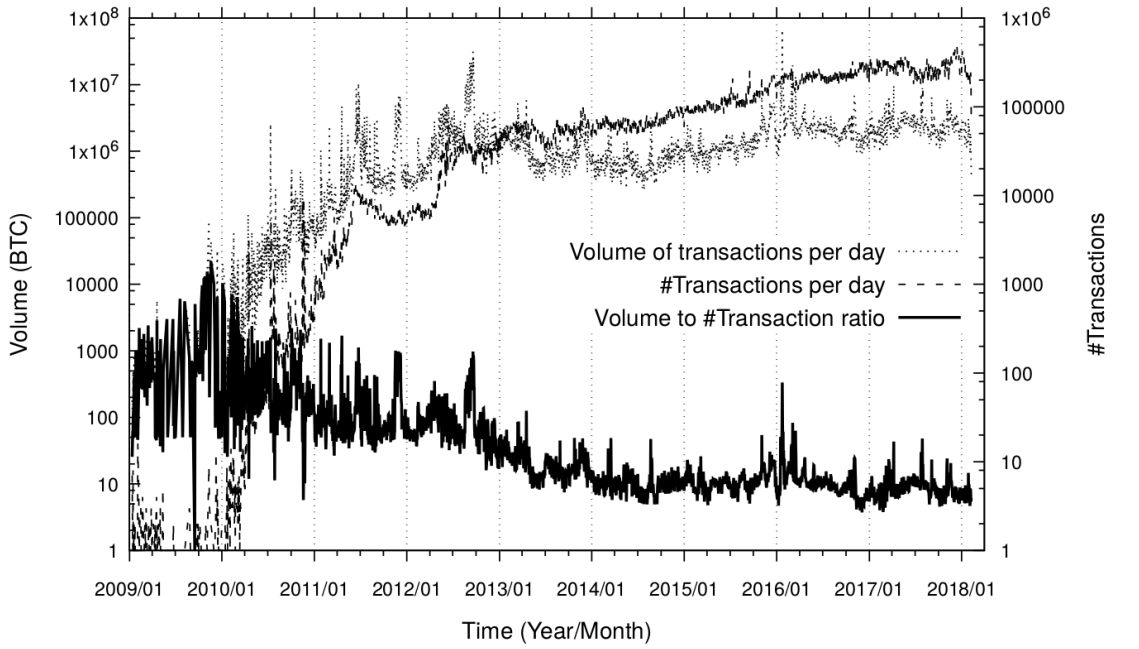


Figure 11: Log scale plot to understand the weekly pattern of exchange of bitcoin transaction.

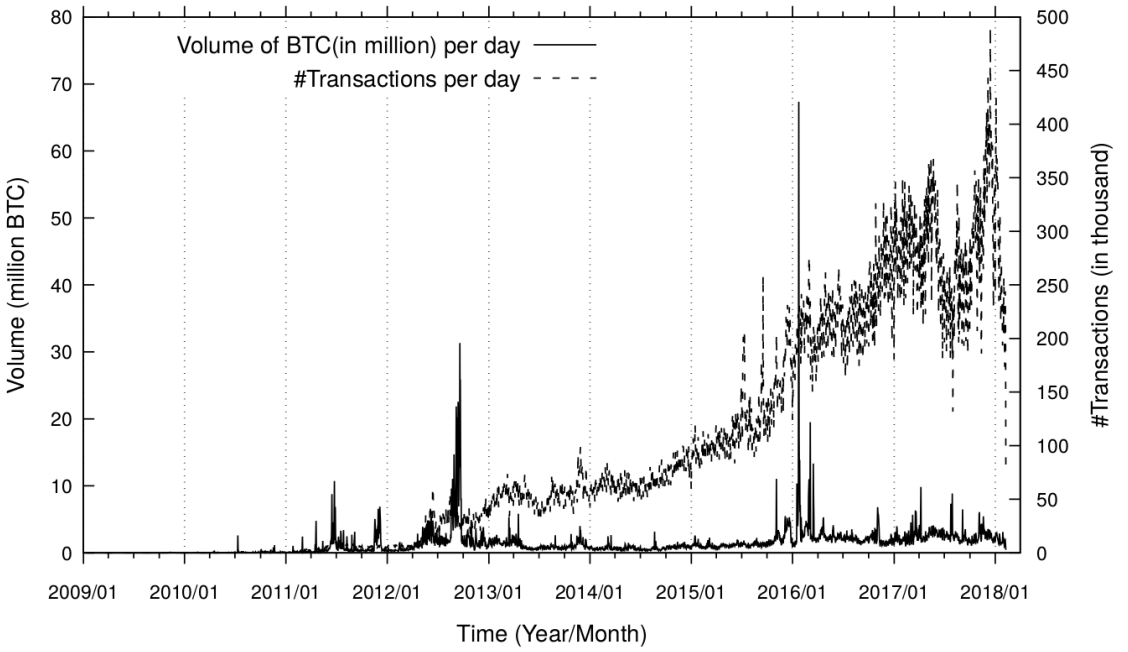


Figure 12: The time-series data of BTC volume and number of transactions per day

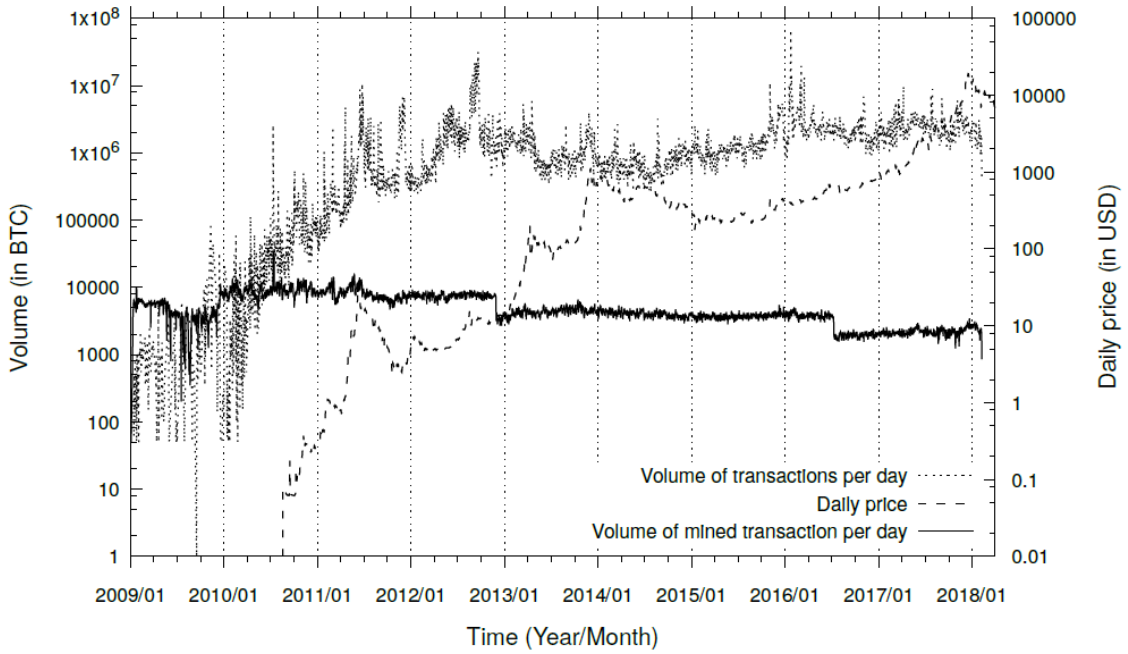
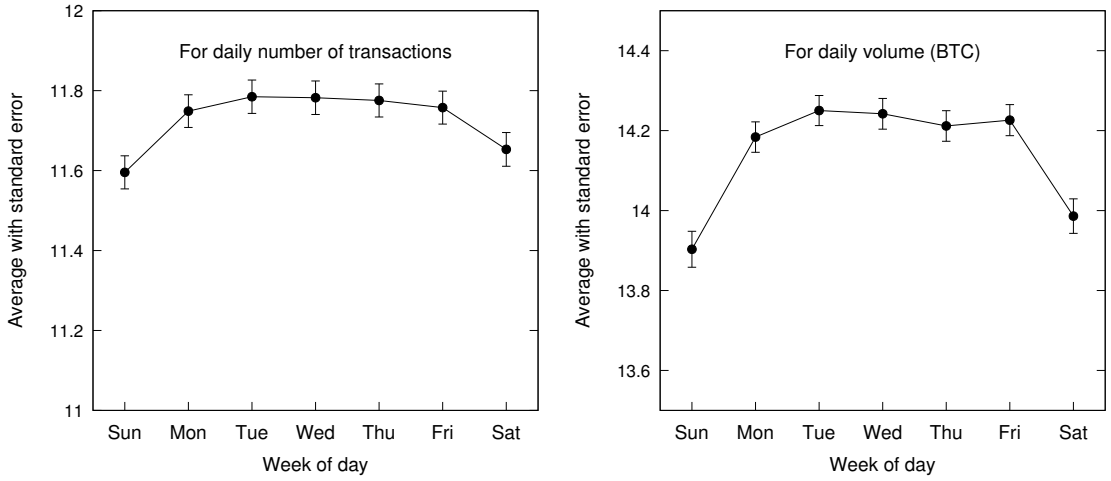
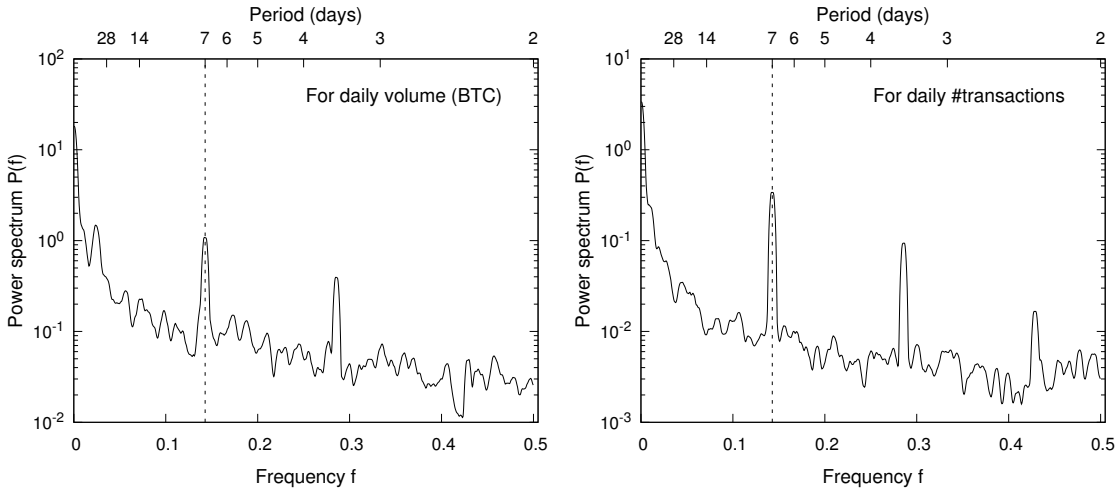


Figure 13: Price evolution of bitcoin volume transacted and mined per day in log scale



(a) (b)

Figure 14: Average of (a) logarithm of daily volume, (b) logarithm of daily numer of transaction for each day of week. Error bar is the standard error.



(a) (b)

Figure 15: Power spectrum as smoothed periodogram for the time-series of (a) logarithm of daily volume, (b) logarithm of daily numer of transaction. See the main text for the details of the method of smoothed periodogram.

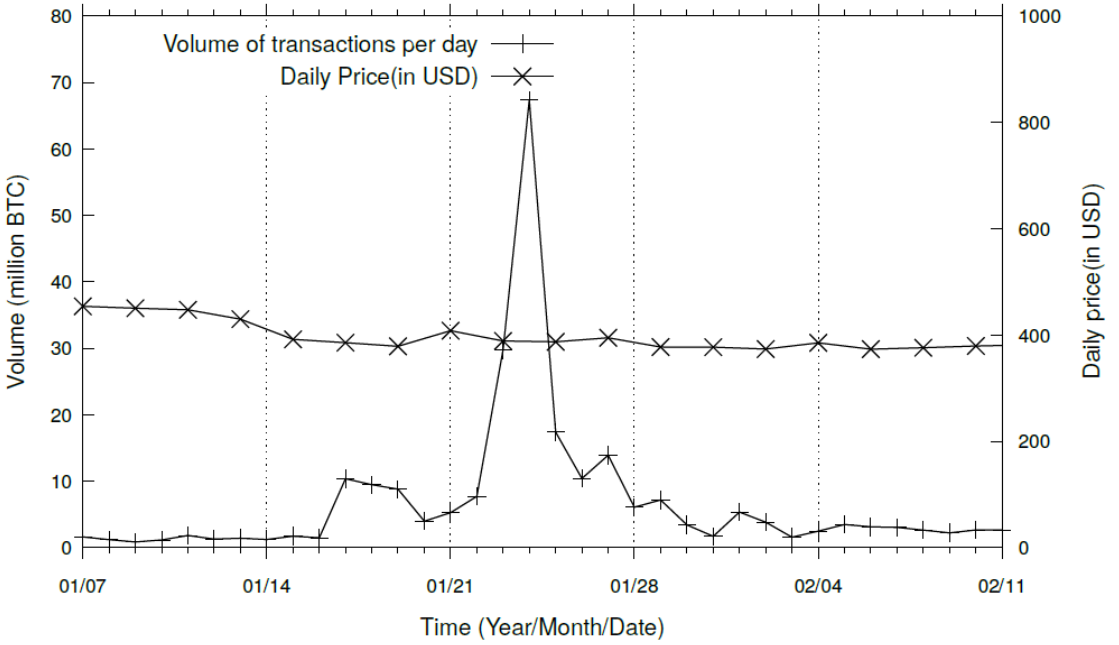


Figure 16: A closer look at the unusually high volume of transaction happened on last week of January 2016

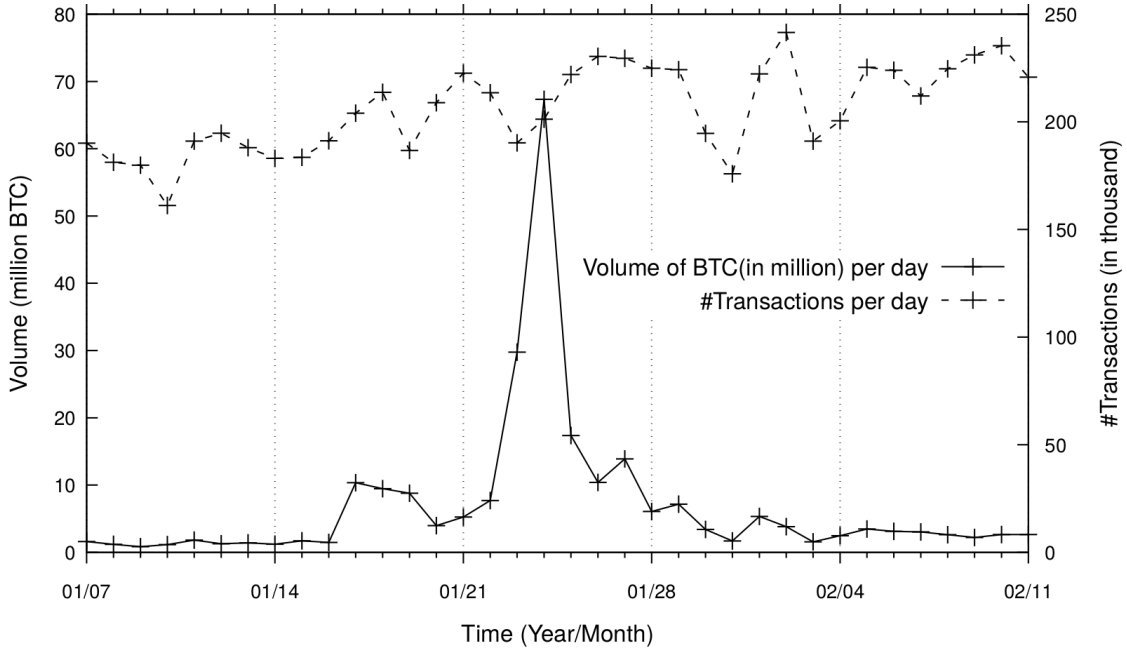


Figure 17: A closer look of BTC volume to the number of transactions during the month of January 2016

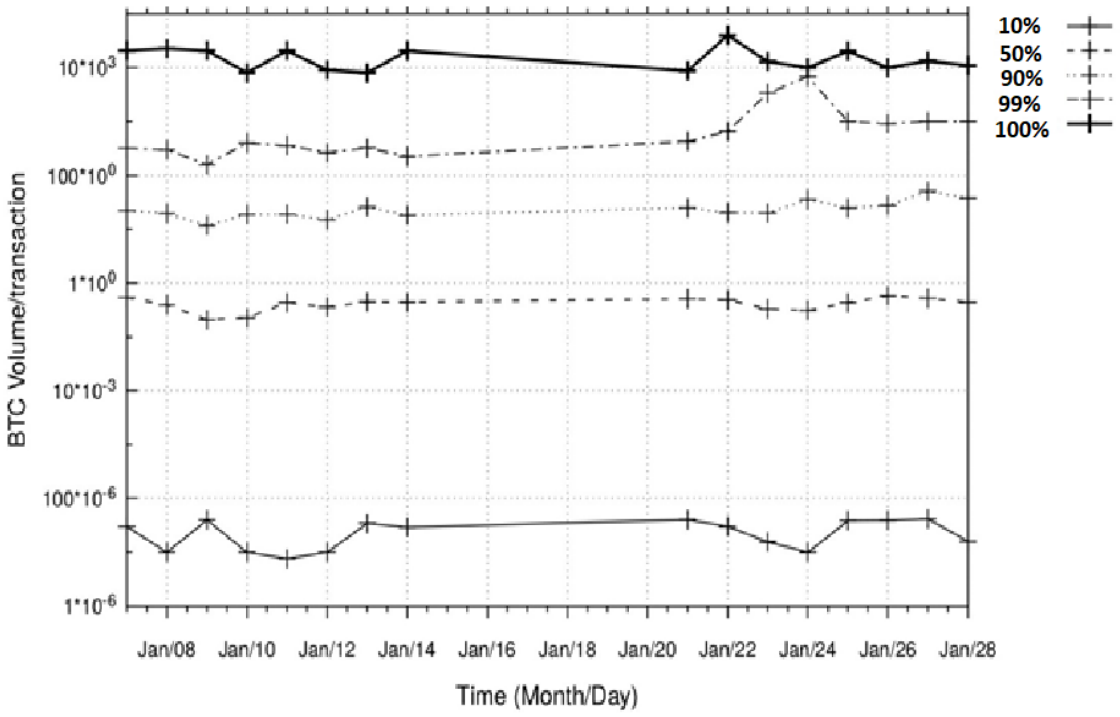


Figure 18: Quantile calculation for BTC volume/transaction for the January–February 2016

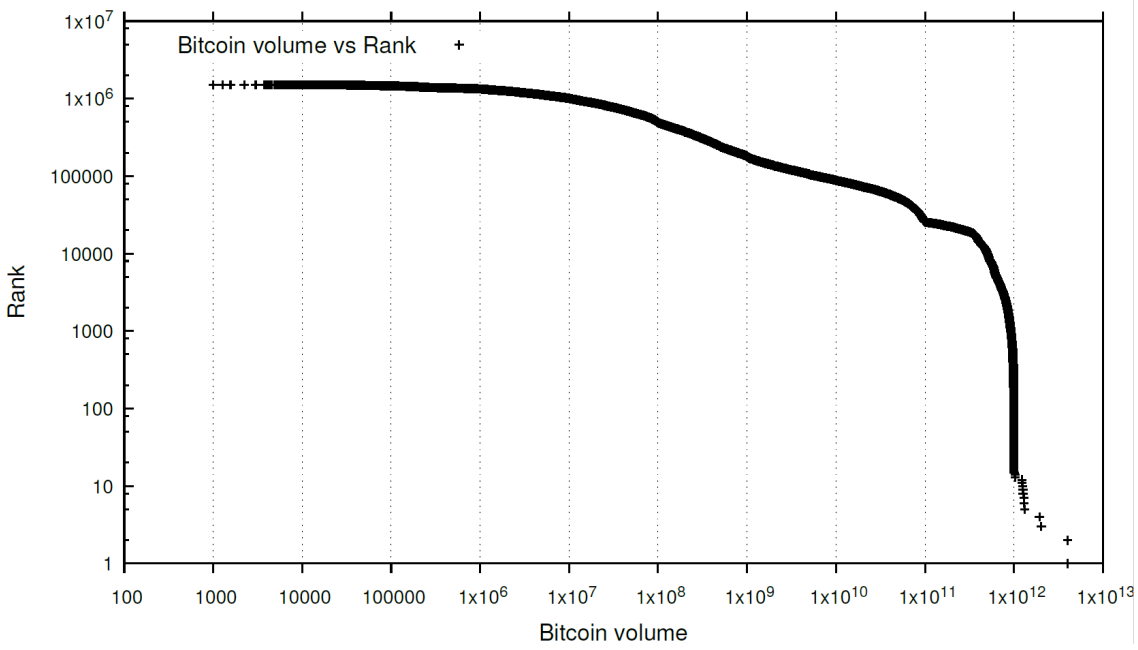


Figure 19: Bitcoin volume rank distribution (log–log scale) from 21^{st} Jan 2016 to 28^{th} Jan 2016

Transaction graph table

Previous transaction	Current transaction
Node 2	Node 1
Node 3	Node 1
Node 4	Node 1
Node 5	Node 1
Node 1	Node 6
Node 1	Node 7

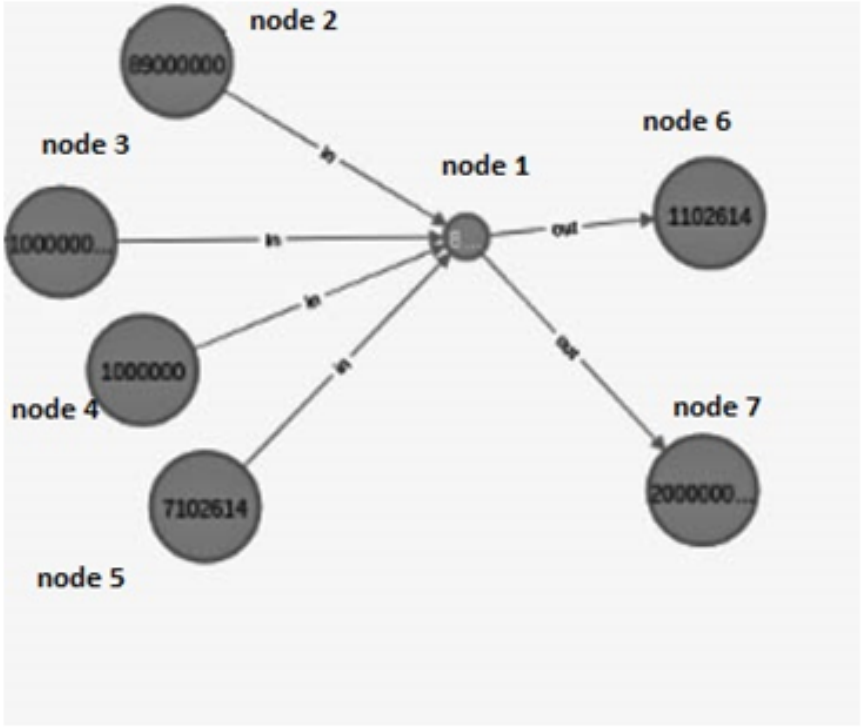
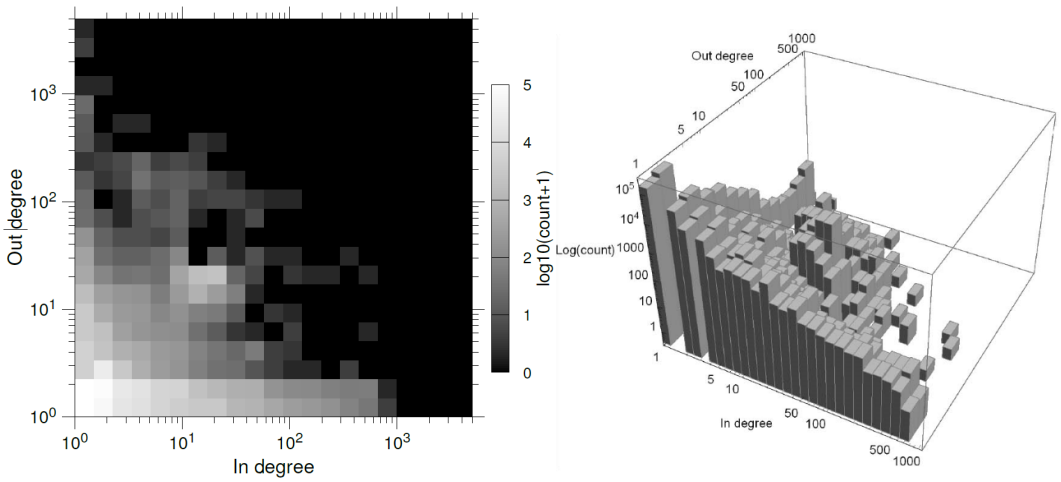


Figure 20: Transaction graph



(a) (b)

Figure 21: The degree correlation of in-degree and out-degree (a) heat map(b) 3D plot for transaction graph constructed in time series of 21^{st} to 24^{th}

6 Identifying Big players in the bitcoin blockchain: (A new approach)

This is the second phase of the research where we introduce a new approach of finding the activities of Big players inside Bitcoin blockchain.

6.1 Dataset for daily users' network flow graph

The research data for this paper has been collected from the web repository contributed by the Hungarian bitcoin research group [17]. The research group created a reconstructed database after downloading the publicly available bitcoin blockchain, where compilation of transaction information (sending and receiving of bitcoins among traders) are in secured and anonymous format holding the timestamp information from January 2009 to February 2018. They have used existing techniques [7, 22] to map addresses to user. They have marked all these mapped users with randomly generated unique numbers for analysis conveniences. Other long string ID in the blockchain has been marked with randomly generated numerical IDs for the same reason. In the blockchain there were 501,418 number of blocks which comprised of verified numbers of transactions compassed by the miners. Some statistical findings on the Hungarian research group's existing database of January 2009 to February 2018 is shared in the Appendix Sec A.

In our analysis, we have created our own address to user hash dictionary, which includes the uncontracted addresses that were filtered by the Hungarian research group. For that, we used each uncontracted address's hash as their user ID, so that we could distinguish between the two. The blockchain data contains some of the records that has non-standard transactions (that has addresses which can not be decoded by system). We had filtered out those transactions as those might lead to ambiguous results. Then we included the timestamp information to analyze the temporal change. We only used the output satoshi values as edge attributes which represents the flow of Bitcoin from input nodes to output nodes.

Between the year of 2011-2012 Bitcoin had gained commercial values and had been used more globally and acknowledged economically. Before that time, it was limited to be used as a financial innovation and were circulated only experimentally among its pioneer users.Furthermore, the first 4-5 years data after the launch were more frequently analyzed by the researchers. Considering all these, we had only focused on the data within the period from 1^{st} January 2013 to 8^{th} February 2018.We filtered out the nonstandard transactions that had contained incomplete information in order to avoid ambiguous statistical results. After separating 113,492,656 self-loops records and summing up all the multi-edges the total number of edges of our final graph during the year from 2013 to 2018 had been reduced to 432,853,828. The number of nodes of this large data set had reduced to 174,250,450.

The monthly node-edge statistics of the 2013-2018 has been demonstrated in the Fig.22.

As discussed in the previous sections, one of the main focus of this research was to investigate on the weekly pattern of the network flow, the daily timestamped data for specific duration would provide us sufficient exploratory results and would also be computationally beneficial. From the Fig.22, we divided our analysis into two periods. First one is, 'Active period', that is represented by the days between 1^{st} July, 2017 to 31^{st} December 2017. Bitcoin gained its maximum price hikes in the history during this period. The second one termed as 'Quiet Period' is comprised of the days between 1^{st} January, 2015 to 30^{th} June 2015. So, the main network has been segregated into daily snapshot of graphs where at any daily time t, the graph can be denote by the Eq.(22),

$$G_t = (V_t, E_t) \tag{22}$$

Where V_t and E_t are the vertices and edges of the graph at time t respectively. The amount of satoshi (= 1/100,000,000 BTC = $1/10^8$ BTC) flow as weight is denoted by w. Now, in order to

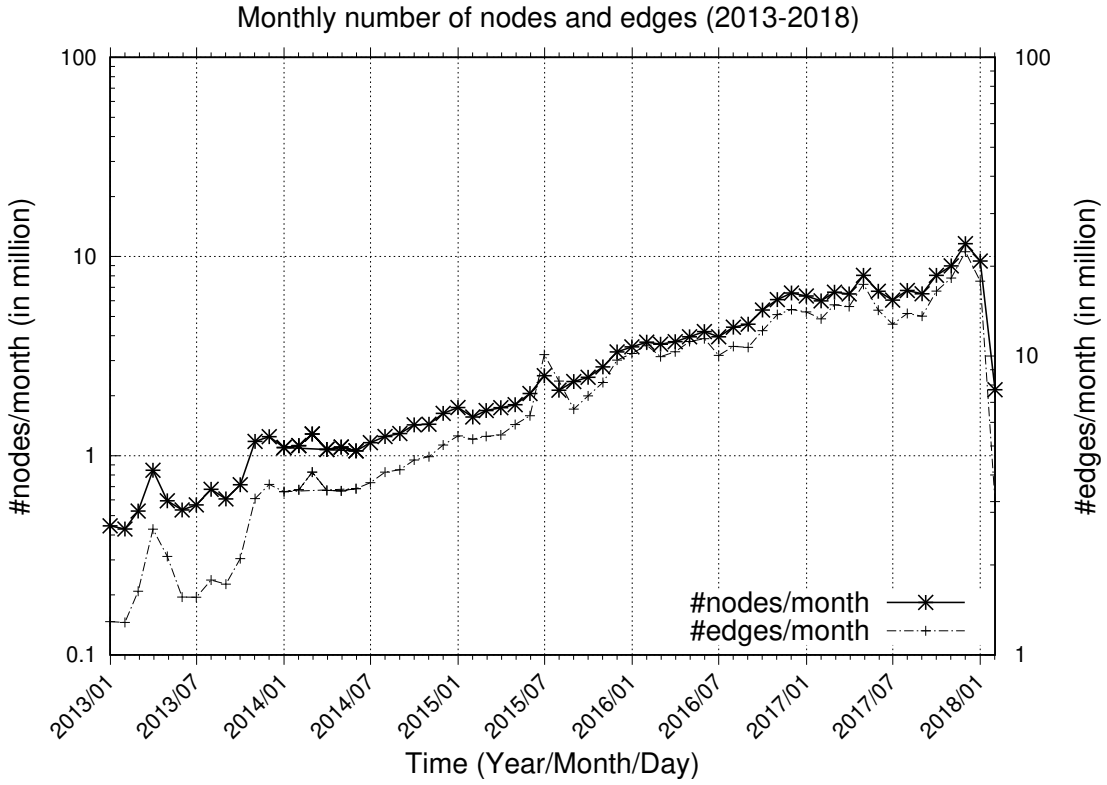


Figure 22: Monthly node-edge count from 2013 to 2018

represent the edge flow, if we consider v_i and v_j are two vertices and then weight or satoshi flow between them can be represented by w_{ij} for an edge e_{ij} = (v_i, v_j).

The node-edge count and average degree on active period are shown in Fig.23 . The same calculation for quiet period is demonstrated in the Fig.24 .

The average degree for the directed graph in both period is around 3. There are no drastic changes in the average degree in both cases. The total number of existing links existed in both cases are much smaller than the maximum possible links . Sparse linkages can be recognized in networks in which the nodes are strenuous to be linked, as shown in Eq.(23) :

$$L < L_{\max} \qquad \text{where } L_{\max} = N(N-1)/2 \tag{23}$$

Like most of the real world networks our reconstructed graphs for both time periods also demonstrated sparsity as topological property. This means the existence of few extremely linked nodes and many sparsely linked nodes within the same network.

Node frequency : The frequency distribution of the users had a key role to play in our analysis. It is one of criteria of defining 'Big players' activity which we would like to discuss in the later section of this paper. Now, we share the results of calculating the total frequency of each node being active either as input or as output. The advantage of excluding all the self-loops gave us total frequency count for each individual node appearing as input or output uniquely for each transaction. The top 20 users' frequency list shown in Table 4. We will use the information in the later section.

6.2 Significant difference between weekdays and weekend

We sub-divided the data into active and quiet period. We cleft these two period networks into daily images and dissertated the basic statistical properties of nodes and edges. Now, in this section,

Table 4: Top 20 users' total frequency count as input/output during 2013-2018

UserID	frequency
3366757	30329366
109540	30078473
14382265	13710199
27888617	10596775
25703559	9583373
3491614	8118837
6188061	8046046
76589853	6986710
135111428	5815920
18307826	5479502
45976983	4799924
19459450	4321546
62504973	4207824
190362585	4048302
45452467	3732724
65608404	3664173
41633902	3142230
59826853	2786511
11031719	2749122
3959839	2675207

in the two time domain we discuss the network properties change with the transition from weekdays to weekends for specifically for large edge flows. This converged our analysis of defining the "Big players". In order to scrutinize the Big players' activities and their network we target to find a threshold point to identify the comparatively larger edge flows.

6.3 Threshold and network Size

The "Big players" in a financial system in general case are the users or wallets which are involved in transacting very large daily volume of bitcoin and also have a tendency of frequently appearing in the network compared to normal users. The node-edge statistics gave us some insight of the difference between

size of the daily network in active and quiet period and the normalized number of nodes and edges with respect to different threshold edge flow had been shown in Figure 25 and Figure 26 respectively. Readers can also take a look at the Appendix Sec D to know the daily total flow of the subgraph for active and quiet period.

There were perceptible difference in reduction of network size which happened due to the drastic change of emulated threshold point range. In order to cynosure the large edge flows we were interested about the 20 BTC threshold point. Above this point, not only the size of the network are confined to the large flows only, but it would also give us widened opportunity to fixate the topological differences of the Big players' between weekdays and weekends. Please see Appendix Sec B which justifies our assumptions.

6.4 Sum of edge-flow and average edge-flow

In our previous research work[68], we had shown by calculating the average and standard error(standard deviation divided by the square root of recorded data) of number of daily transaction

and volume sum of transacted BTC between the period from 2011 to 2018. The both parameters had distinct proneness to follow a weekly pattern. That means the quantified volume or number of transactions is higher during weekdays than weekend. We had calculated the daily average edge flow of satoshi(lowest unit of bitcoin/edge, where 1 satoshi = 0.00000001 BTC (= 1/100,000,000 BTC = $1/10^8$ BTC). By the term "average edge flow" we meant to say the number of satoshi outflow for each unique pair of users or nodes of our graph.

From Figure 27, we can clearly observe that there is clear indication of divergence from weekends rate of flow than that of the weekdays in both active and quiet period. But even though, the weekly pattern is clear for the main graph the consistency of the pattern actually diminishes for higher edge flows as shown in the Figure 28. At different smaller threshold point the dissemblance among Sundays and Mondays were indistinguishable .

So, in our final analysis the sum of daily edge flow were taken in to consideration for big players weekly patterned activities rather than the average edge flow. In the Appendix Sec C we showed some valid reasoning for not considering the average edge flow attributes.

6.5 Connected components of sub-graph and the BTC flow inside

The "Big players" in Bitcoin exchange market are connected every day of the week with other exchanges based on transferring big volume of BTC in order to balance the demand and supply from the customers' end. For this, we analyze connected components size and the normalized BTC circulation inside the maximum connected components. We expect a deterministic differences among the weekdays and weekdays and we quantify it.

The connected components of an undirected graph G=(V,E) are the maximum subsets C_1, C_2, ..., C_k such that $V = C_1 \cup C_2 \cup C_3 \cup . \ldots . \cup C_k$, and u,$v \epsilon C_i$ if and only if u is reachable from v and vice versa . The size of connected components represents the number of nodes, who are connected with the rest at least a path. In our analysis, we concentrated on two of the aspects to explore this structural properties of our graphs. Firstly, we used the 20 BTC threshold to create daily snapshot of the undirected sub-graphs. Secondly we calculated the normalized maximum connected components and the normalized BTC flow inside the maximum connected components.

The Figure 29 shows that the daily average maximum connected components normalized size for both active and quiet periods having weekly patterns. We can explain this in terms of big edge flows and connectivity of the largest components. On weekends there were less number of nodes which remained active in the network. In real world, business certainly takes time off particularly on weekends. All the largest stock exchanges in the world maintains trading hour that follows bank's operating hours.Because of this stock markets are closed on weekdays. The crypto-asset exchanges have technical upper hand as investors are able to make trades on Saturdays and Sundays. But there are additional challenges and risks are there. The outside operational hours trading activities leads to lot of problems. As there are small number of users with relatively small volume of BTC traded, prices are more influenced by single trades and moreover, the volatility factor is always there. The mismatching between sellers' asking price and buyers' bids lead to uncertainty to complete negotiations. On the contrary on weekdays, scenario improves quiet in an extent. The main reason is, the size of maximum connected components grow on weekdays when there are more nodes to participate. The buyers and sellers have much more information and options.

In terms of flow inside the maximum connected components we quantified the circulation by taking proportion of the flow inside the maximum connected component of the daily 20 BTC sub-graph to the total flow inside the total connected components . The daily total flow of the sub-graph was calculated in order to understand the flow inside the maximum connected components and is shown in the Figure 30. We can approximate the average daily total flow in spite of having some larger fluctuation in quiet period than active period. The active period has daily total BTC flow twice as larger than that of the quiet period.

Prices quoted during after-hours sessions are not "official" and don't reflect credibility among the traders' mind. As the Figure 31 shows on every weekdays the approximately 65% or more of total flow of all connected components circulates inside the maximum connected components, where as on weekends its less than 55% approximately in active period. In the quiet period, due to some spikes of daily weekend flow push the average Saturdays' flow a bit higher but the weekly pattern still holds. So, both the size and flow inside the maximum connected components disclosed the difference among weekdays' and weekends' activities .

The sub-graph of threshold greater than 20 BTC filtered out all the small flows. The maximum connected components of this sub-graph include all the users who are persistently involved in the exchange market. Even though there might be other relatively less persistent users' influences involved in this sub-network, the quantification of currency stream inside the maximum connected components gave us good insight of flow pattern. In that context, we planned to measure some individual renowned crypto-exchanges' daily total average flow and research their seasonality weekly behavior. In that way, we can emphasize similar behavior be the identifying criteria of other active anonymous financial institutions inside blockchain.

6.6 Examining some exchange's activities : The "Big players" market scenario

In this final section, we finalize the definition of "Big players". In order to make the definition compatible we use available on line open source data of crypto-exchange markets to correlate our blockchain restructured data. In practical world, many transactions in the exchange market that has liquidity of bitcoin supply sufficient enough that customer can disengage their crypto-asset at any moment. customer can buy and sell bitcoin with equivalent exchange of fiat currency by making a bank transfer to the market's bank accounts. The purchaser's account then credited by that exchange market on their system with that fiat money. This gives a clear indication of the market's public key's daily use with very high recurrences. Besides, there are also chance of high volume of BTC flow to and from these market's wallet.

While exploring for the open source crypto-exchange data we found a website (https://bitinfocharts.com/top-100-richest-bitcoin-addresses. html) where we collected the publick key of 1000 top rich bitcoin wallets. We had merged the hash public keys to Hungary research groups list of addresses database and then consecutively merged with our restructured address to User database . In Table. 1 we have shown the public wallets that were contracted with our user database .

We found that, till the cutoff date of 9^{th} February 2018, Xapo and bitstamp wallets have very large number of edges. These two wallets were also discovered in the Table 4 of our top 20 frequent wallets that we found earlier. These two exchanges were the prime specimens to observe the money flow .

Finally, we define "Big Player" as follows : "The users that have the criteria of (1)being highly frequent in the blockchain network (2)having persistent activity (3) showing weekly pattern of total network flow". In order to authenticate the definition we present the result of applying the criteria on the top 20 frequent users in Table 6 We took both exchanges' user ID and weighed the daily average sum of inflows and outflows. In both cases the results followed the weekly pattern as shown in Figure 32 and Figure 33 .

Along with these two exchanges we calculated the average daily total inflows, outflows and total flows (In+out) of average weekdays and weekends of the other top 18 frequent users . Out of 20 users 12 users showed daily activities, very large daily in/out/total BTC volume and weekly pattern of in/out/total BTC flow. These can be identified as similar financial institution or exchange market. The rest were not shown similar results because of insufficient daily activities, no distinction of weekdays and weekend activities. These Big players can be example of crowd funding or donation accounts, gambling, gaming sites or other non-financial services.

In our previous study, we identified the weekly pattern of the daily total sum of transaction

Table 5: The contraction of some exchange with our restructured address to users' database

AddrID	USerID	Wallet name
65994621	14382265	Xapo.com-2
65994793	14382265	Xapo.com-2
65995913	14382265	Xapo.com-2
65995953	14382265	Xapo.com-2
65995956	14382265	Xapo.com-2
65995973	14382265	Xapo.com-2
65995979	14382265	Xapo.com-2
65995996	14382265	Xapo.com-2
65995997	14382265	Xapo.com-2
65996056	14382265	Xapo.com-2
65996060	14382265	Xapo.com-2
65996063	14382265	Xapo.com-2
99730379	3Nxwenay9Z8Lc9JBiywExpnEFiLp6Afp8v	Bitstamp-coldwallet
108931886	59826853	Bitstamp.net-old2
138715869	138715869	Coincheck-coldwallet
211452559	3D2oetdNuZUqQHPJmcMDDHYoqkyNVsFk9r	Bitfinex-coldwallet
219726782	59826853	Bitstamp.net-old2
269720834	59826853	Bitstamp.net-old2
301211876	3Cbq7aT1tY8kMxWLbitaG7yT6bPbKChq64	Huobi-wallet

Table 6: Revealing crypto-exchange market among top 20 frequent users

Nodes	Wallet identity	Frequency (in million)	Weekly pattern?	Persistent activity?	Financial institution?
598268531	Bitstamp	30	Yes	Yes	Yes
1432265	Xapo.com	13	Yes	Yes	Yes
109540	Unknown	30	Yes	Yes	Yes
11031719	Unknown	2	Yes	Yes	Yes
3366757	Unknown	30	Yes	Yes	Yes
13511148	Unknown	5	Yes,Only active period	Yes,only active period	Yes, new institution
25703559	Unknown	9	Yes	Yes	Yes
41633902	Unknown	3	Yes	Yes	Yes
27888617	Unknown	10	Yes	Yes	Yes
3959839	Unknown	2.6	Yes	Yes	Yes
65608404	Unknown	3.6	Yes	Yes	Yes
62504973	Unknown	4.2	Yes	Yes	Yes
45976983	Unknown	4.7	No	Yes	No
19459450	Unknown	4.3	No	Yes	No
18307826	Unknown	5.4	No	Yes	No
45452467	Unknown	3.7	No	Yes	No
190362585	Unknown	4.04	No	Yes,only active period	No
6188061	Unknown	8	No	No	No
3491614	Unknown	8	No	No	No
76589853	Unknown	6.9	No	No	No

and bitcoin volume. In this present study, we first checked whether the weekly pattern could be explained by the dynamically changing network properties. To understand this, we performed a threshold analysis aimed to identify the big flows.

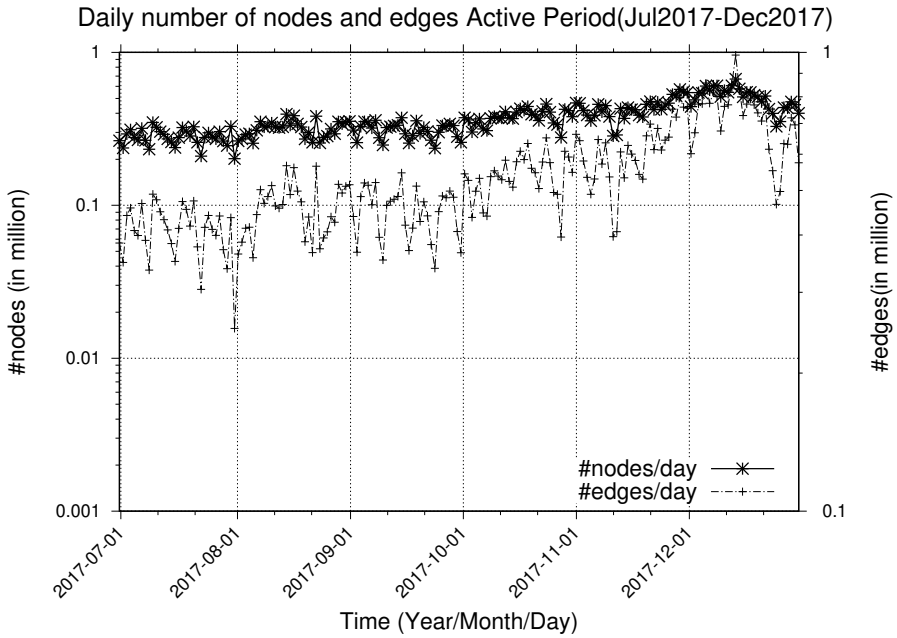
The connected component analysis of threshold sub-graph showed that the size of the maximum connected components during weekdays is larger than that of weekends. The result was per expectation for both active and quiet periods. A primary reason for the observed trends is attributed to the mismatch in the standard operating hours of banks and the crypto-asset markets. Over the weekend, not much new money comes in to support prices. In terms of flow, the normalized average edge flow inside the maximum connected components follows the weekly pattern. The two crypto-exchange institutions, Xapo.com and Bitstamp, supported the weekly pattern of daily total circulation in their own networks.

We also found that both Xapo and Bitstamp were among the top 20 frequent users in the network. Thus we defined big players in terms of high frequency, persistent daily activity, and weekly pattern of total daily average BTC flow. Among the top 20 frequent list, we tried fitting in the two criteria. The 12 out of 20 users who followed, we identified as crypto-exchange companies or financial institutions. We excluded the remaining 8 users because of insufficient persistency and not following the weekly pattern plausibly being online gambling, crowdfunding, and donation institution. The cold wallets, despite their random big flows cannot be termed as big players according to the definition.

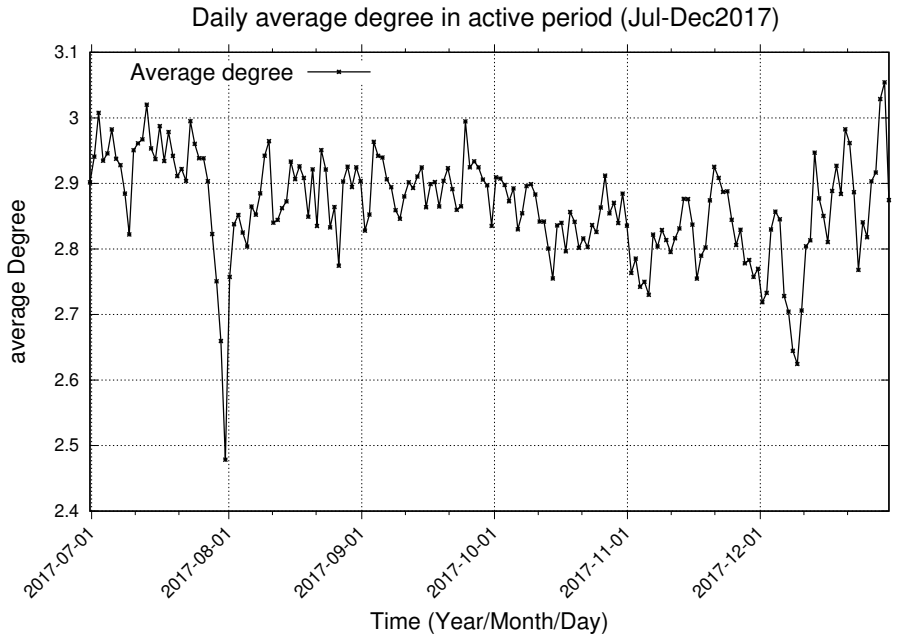
The goal of this part was to reveal the identity of some specific users who are involved in big network flow persistently in the blockchain. We proposed a methodology focusing on behavioral patterns of those users involved in the daily big circulation of money. Applying this methodology, we distinguished the big players into two hypothesized categories: financial and non-financial. The weekly patterns can help us uncover the identity of users we term financial institutions, because they have more BTC trading activities during weekdays than weekends. Most exchange markets belong to this category. A second category of big players is those with large frequency but lagging daily persistent activities and weekly patterns. We conjectured that all the crowd funding, donor organizations, gambling, or betting sites could be examples of non-financial institutions.

This part of the research has a contribution to the field of economics. The blockchain technology has been arousing a lot of interest from a variety of areas such as trade, finances, government and policy. However, because of the anonymity, it turns out to be a challenging task to quantify this engagement and the adoption by financial institutions. In this work, we aimed at understanding which are the main criteria associated with identification of the financial institutions inside a fully digitized economy. In order to do this, we applied a new technique along with the existing ones for deanonymizing the financial institution users having high frequency, appearing persistently on daily big flow of bitcoin. The financial market of crypto-asset with flow of BTC funds are the representation of saving and investing special currency through the intermediary agents like savers and investors. Like the traditional fiat currency bitcoin does not have intrinsic value. But, unlike the fiat currency, it has a store of value like nonmonetary assets for example savings accounts, stocks, bonds and real estate. In our work, the big players' connected component analysis has given us the insight of the quantification of the daily big flow of crypto-asset and acknowledges the circulation demonstrates how money moves through society. The big money flows which involves the conversion of fiat currency to crypto-asset invested by users with the help of intermediary financial institutions and flows back to them as payment for selling back for profits. In short, a digital crypto-asset economy is an endless circular flow of money. In our work we have found out that more than 50% of the total threshold flow (Above 20 BTC) involves the circulation of big flows of economic activity among the big players. In order to see the crypto-asset to be successful in future the big players need to engage and embrace reasonable and responsible regulation. The growth

of this industry depends, in part, on the establishment of safe, fair and reliable market conditions. Presently, the proper regulatory environment is still uncertain and there are a lot of provisions of research work for standardized regulatory policies .

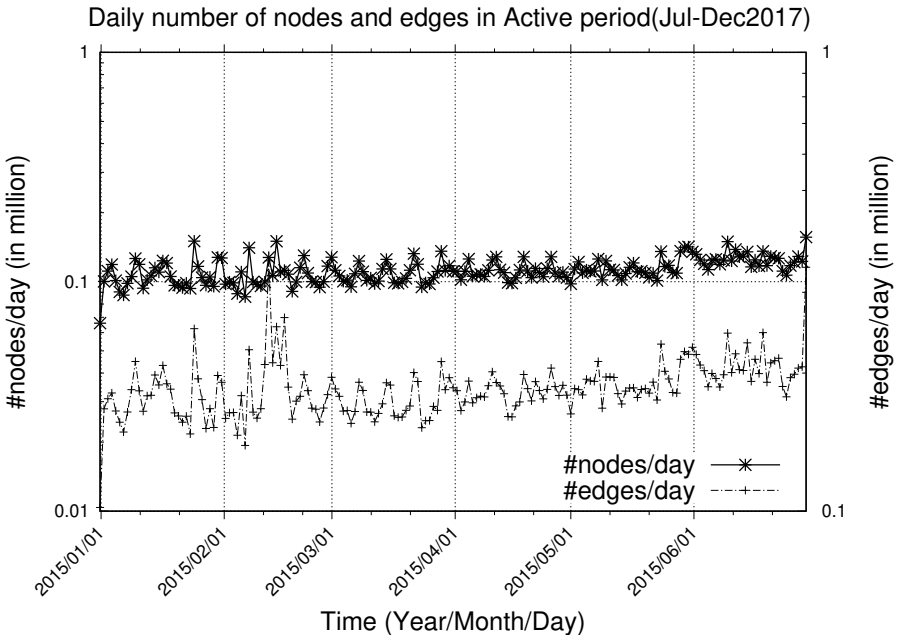


(a) Node-edge count

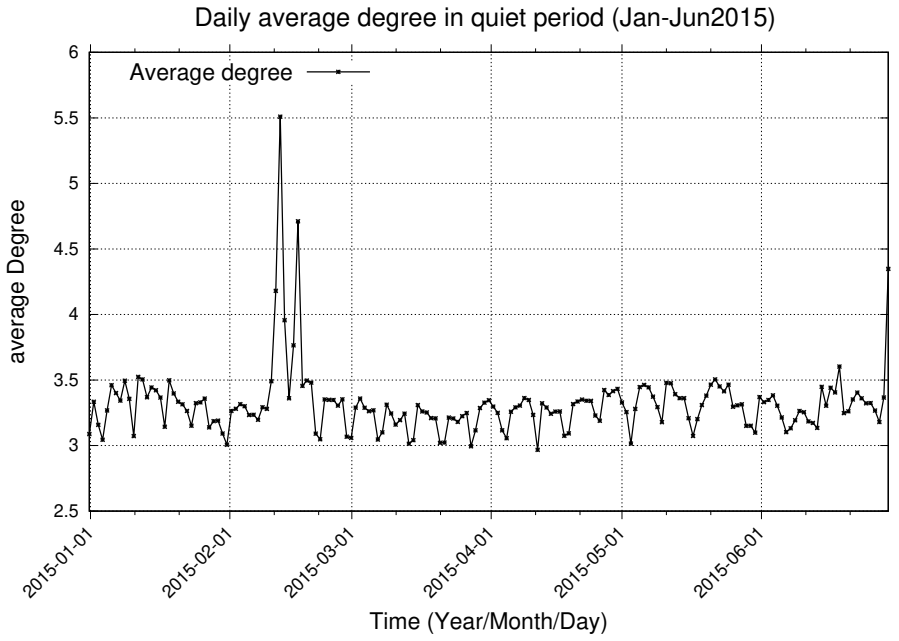


(b) Average degree

Figure 23: Node-edge statistics in active period; (a) Daily node-edge count (b) Average degree

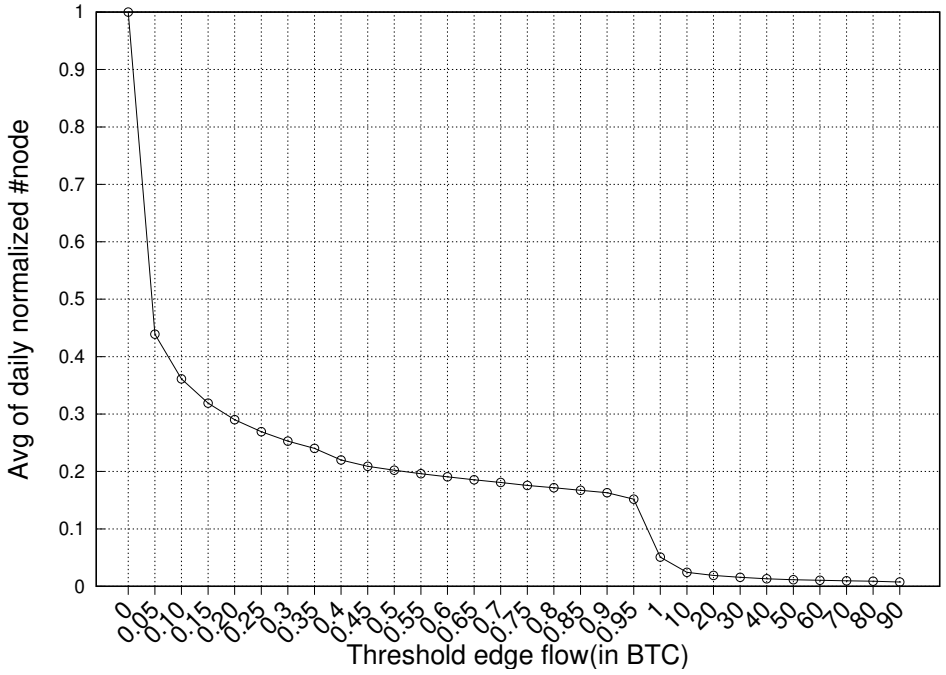


(a) Node-edge count

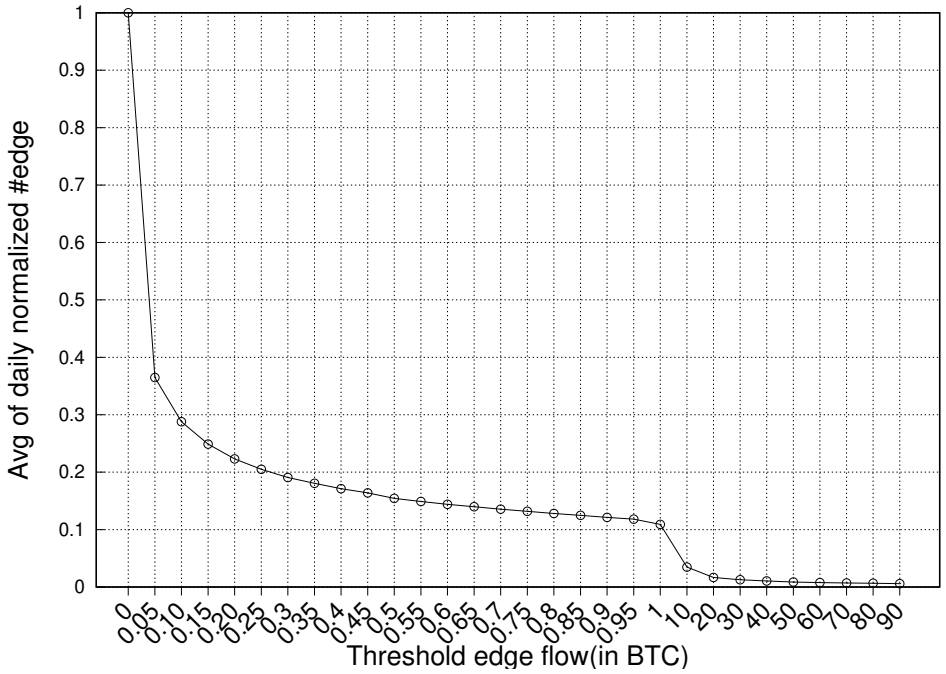


(b) Average degree

Figure 24: Node-edge statistics in quiet period

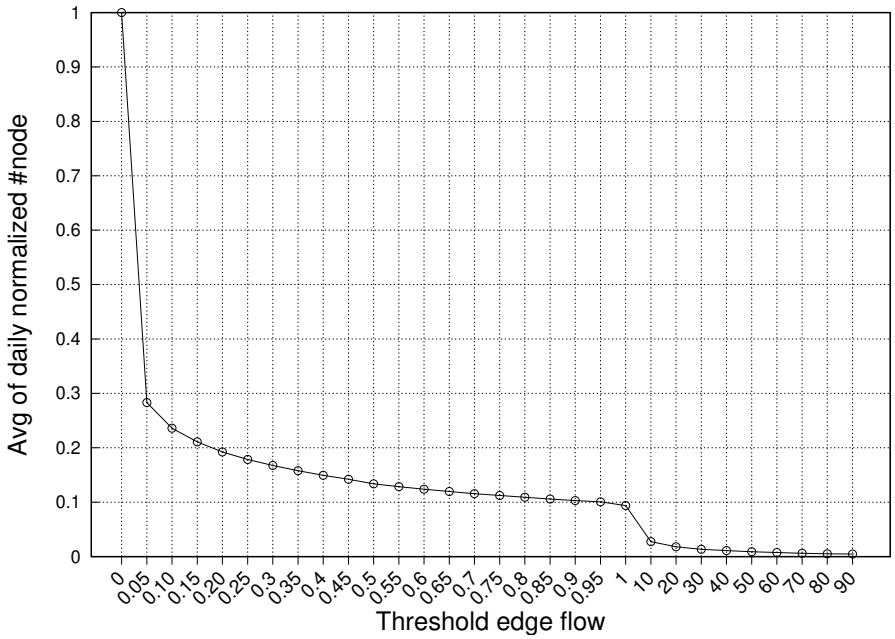


(a) Normalized Node count

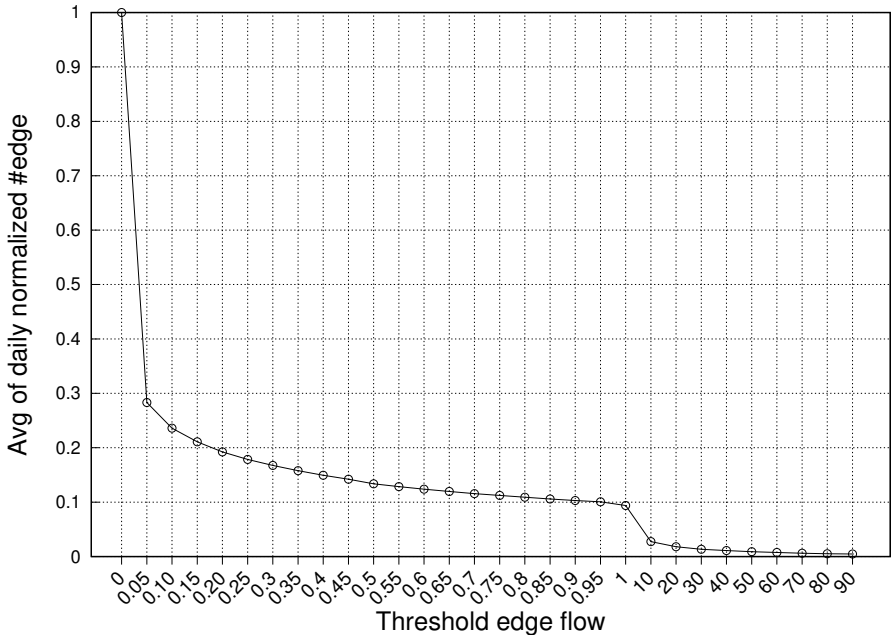


(b) Normalized Edge count

Figure 25: Normalized Node-edge count on different threshold in Active period; (a)Node count (b) Edge count

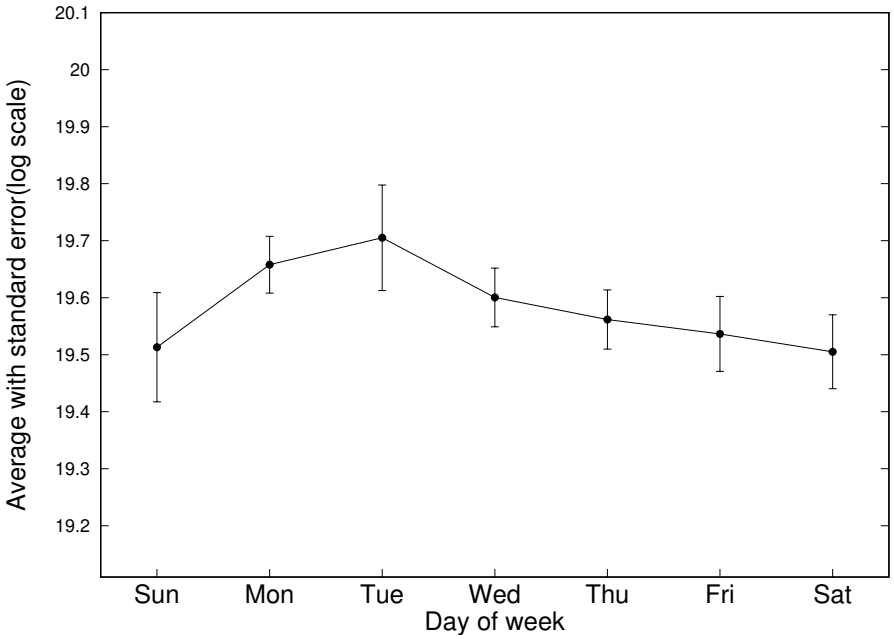


(a) Normalized Node count

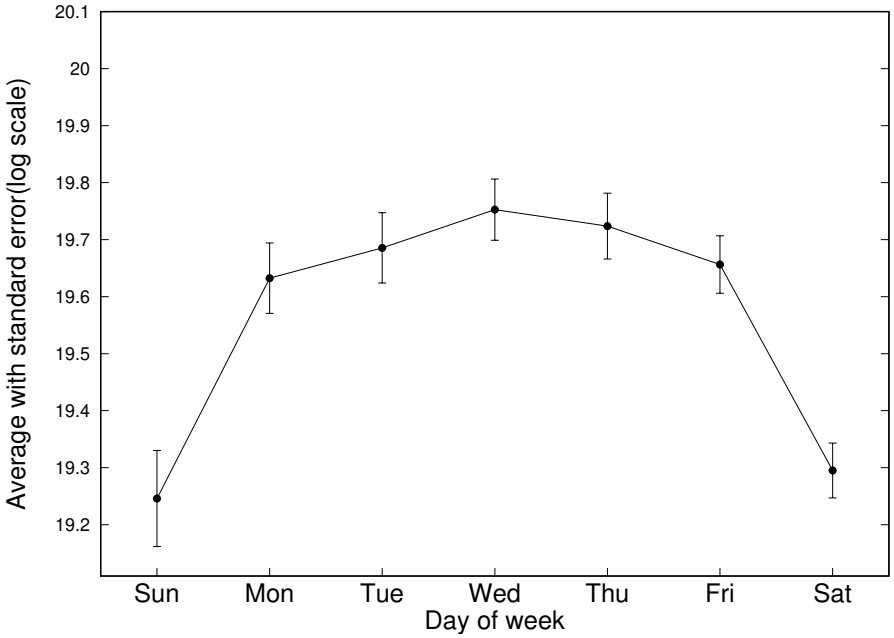


(b) Normalized Edge count

Figure 26: Normalized Node-edge count on different threshold in Quiet period

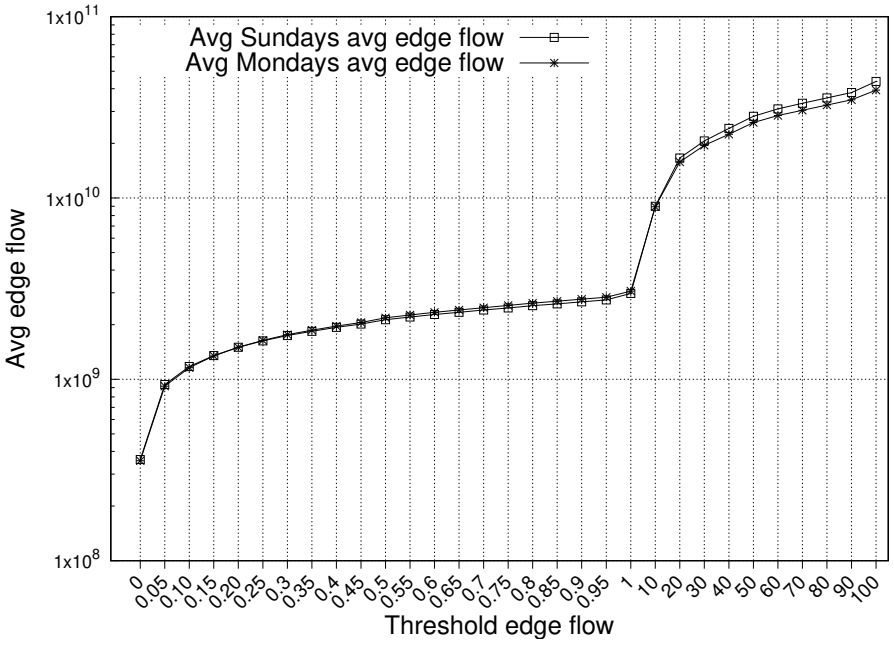


(a) During active period

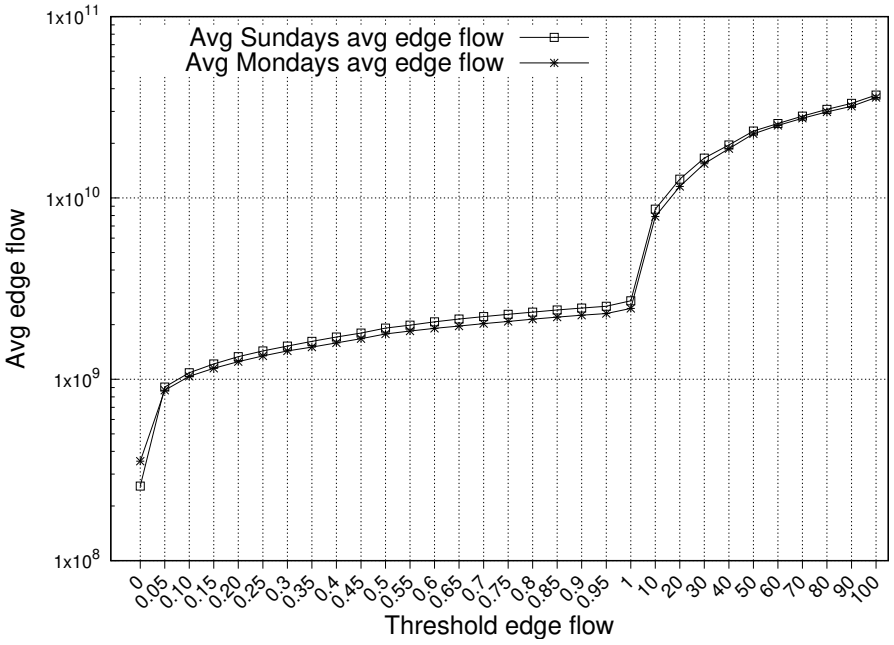


(b) During quiet period

Figure 27: average edge flow with standard error on average weekdays and weekend for;(a) Active Period (b) Quiet Period

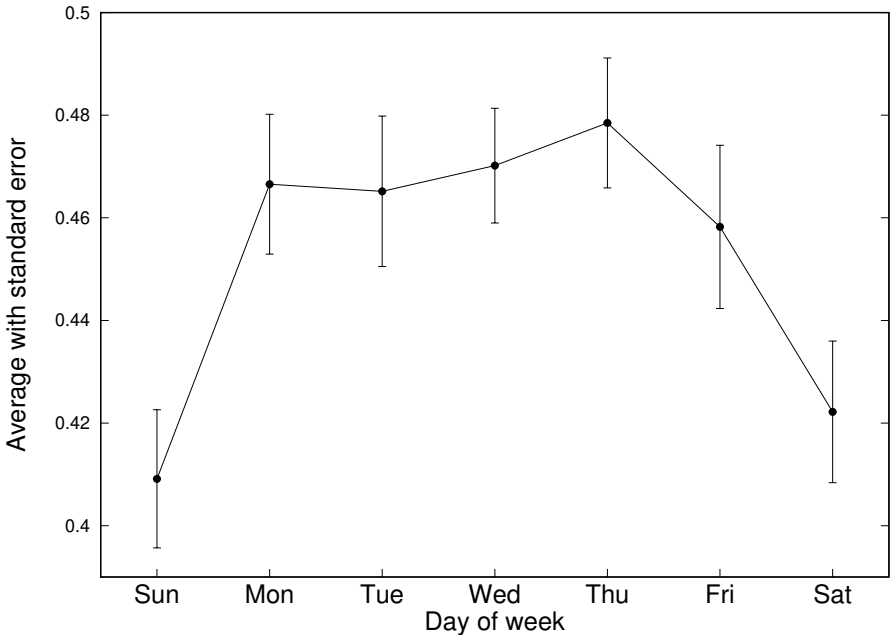


(a) During active period

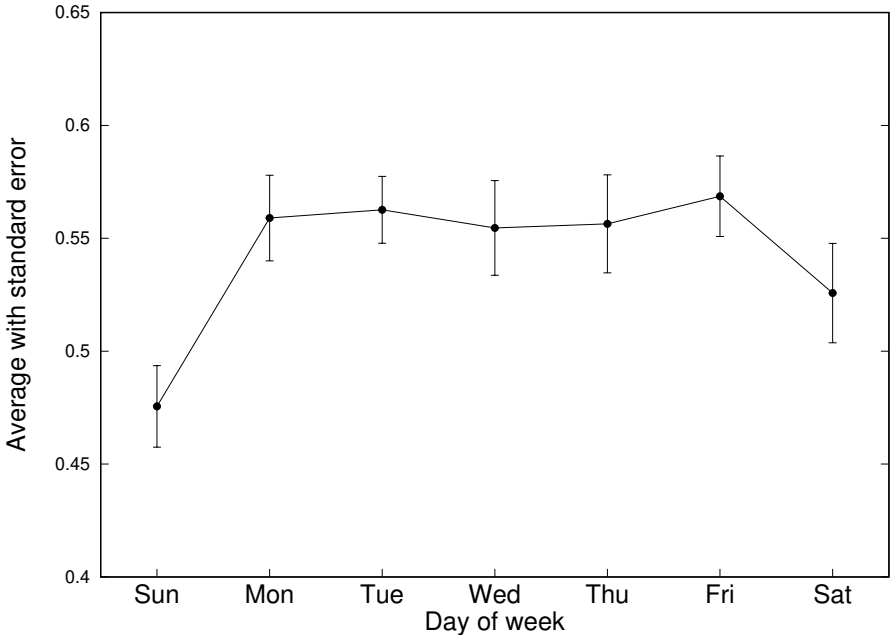


(b) During quiet period

Figure 28: average edge flow on average Sundays and Mondays at different threshold for (a) Active Period (b) Quiet Period

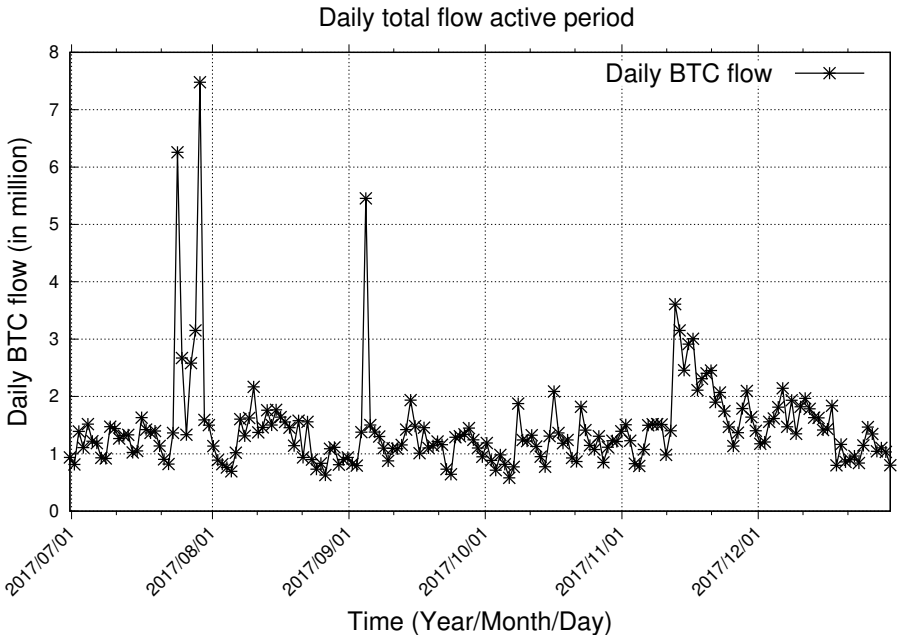


(a) Active Period

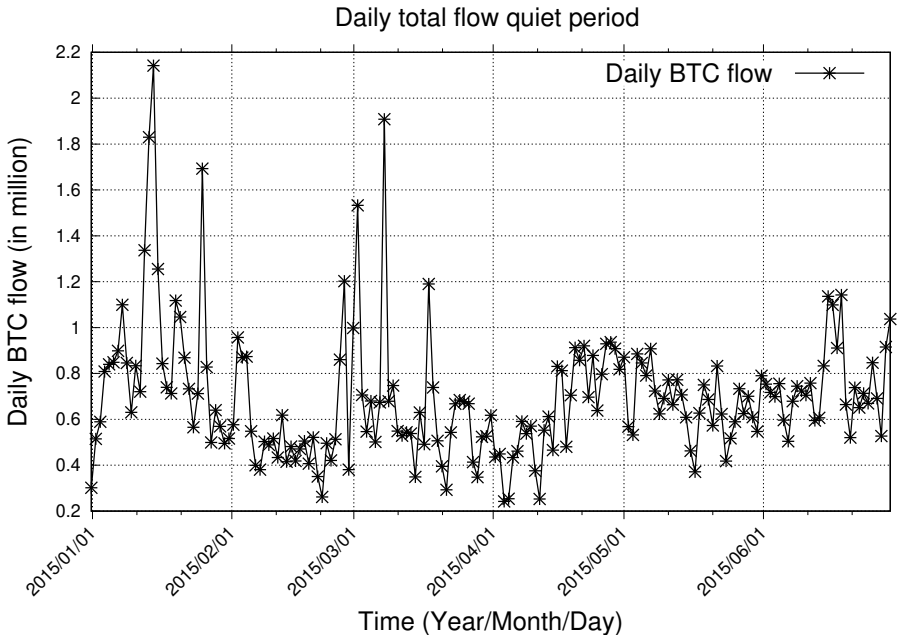


(b) Quiet Period

Figure 29: Normalized Max connected components size with threshold edge flow greater 20 BTC

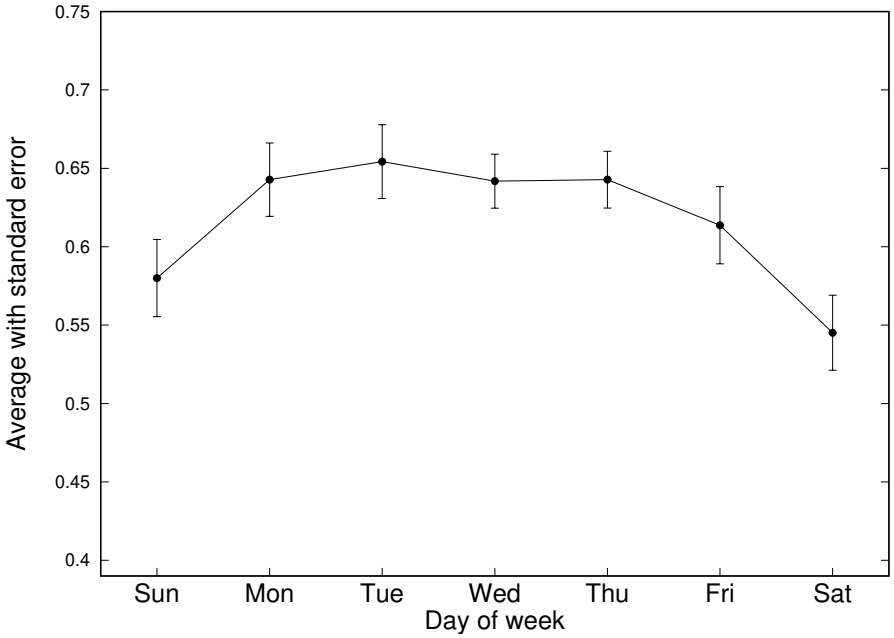


(a) Active Period

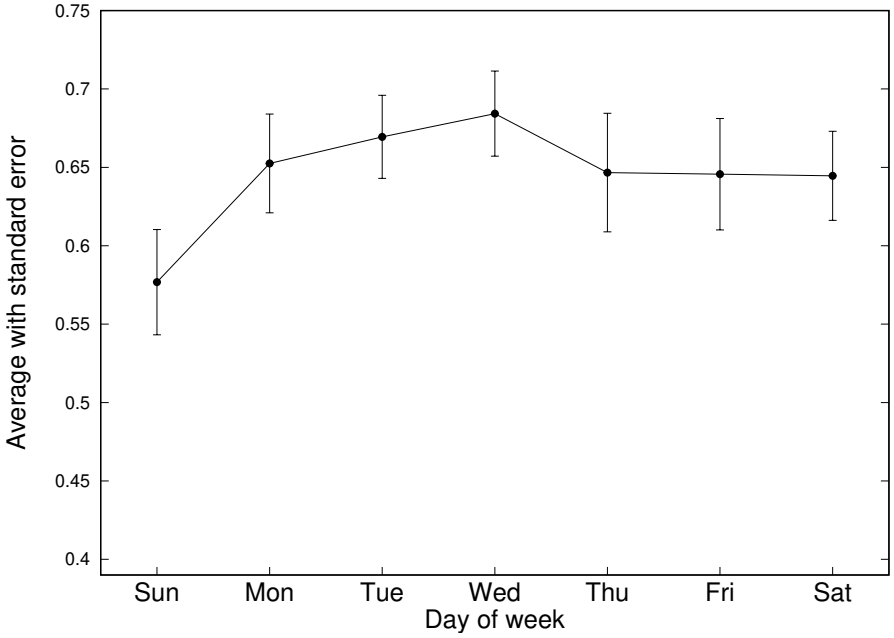


(b) Quiet Period

Figure 30: The total daily flow of sub graph in (a) Active Period (b) Quiet period

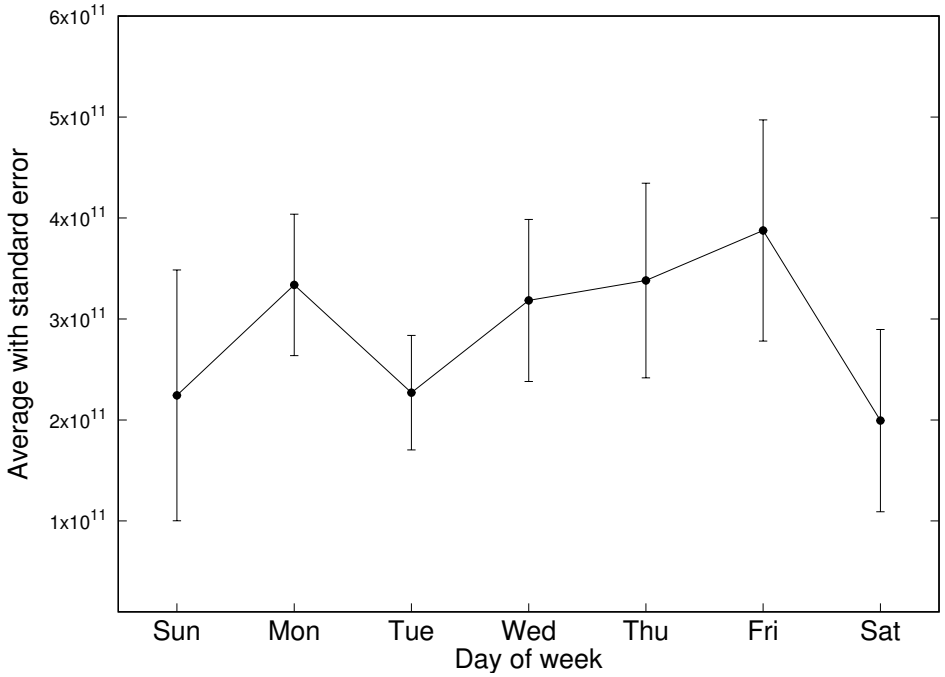


(a) Active Period

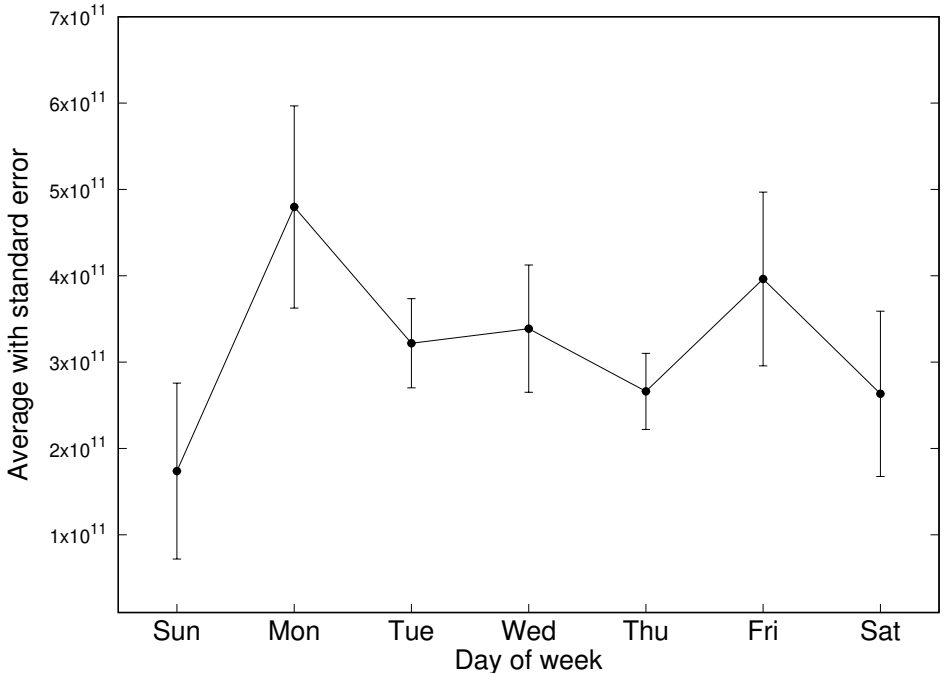


(b) Quiet Period

Figure 31: Normalized Max connected components flow with threshold edge flow greater 20 BTC

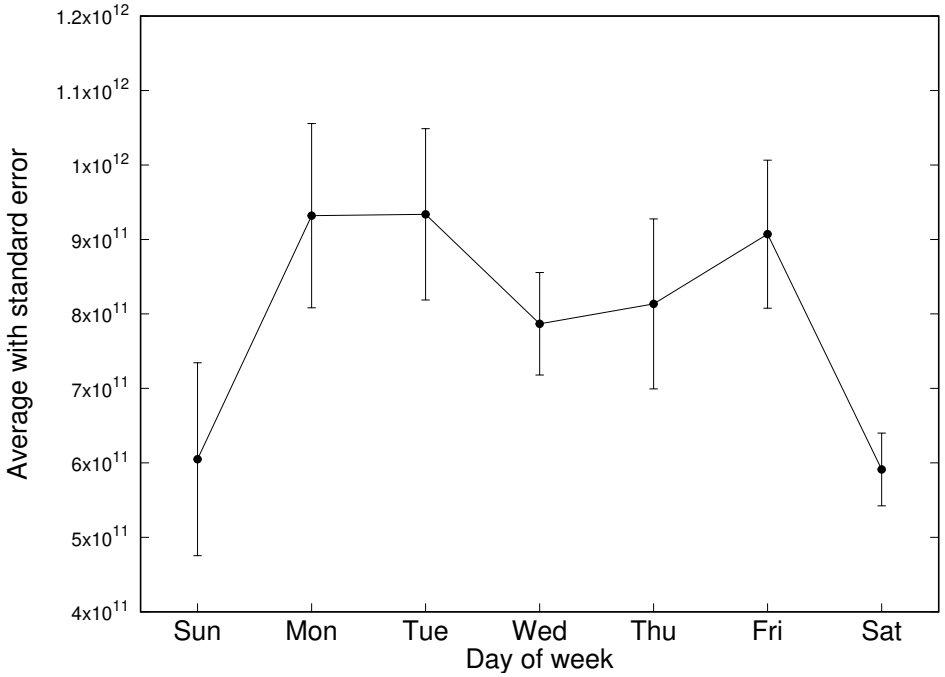


(a) In flow

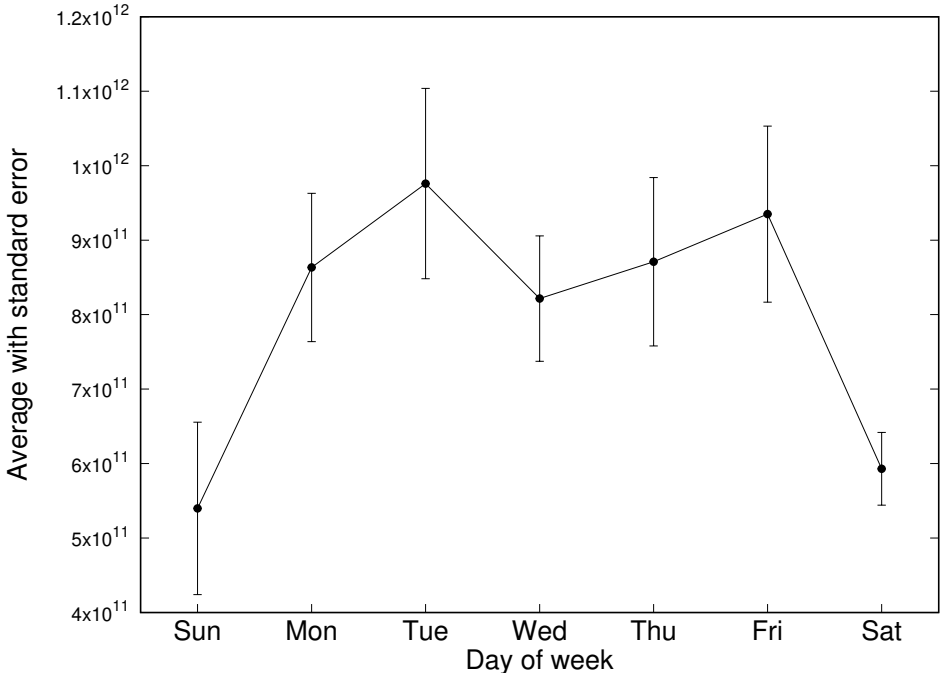


(b) Out flow

Figure 32: The average daily sum of in and out flow of xapo.com



(a) In flow



(b) Out flow

Figure 33: The average daily sum of in and out flow of Bitstamp exchange

7 Simulation, results and interpretations

In this section we would like to use the sophisticated method of non-negative matrix factorization (NMF) to extract hidden factors from the user graph to identify big players in terms of their persistent or regular activities. The hidden factor will be calculated by simple probabilistic method of Bayesian Information criterion(BIC). Finally, we will apply the NMF technique to understand the decomposed matrices of the sender/receivers' to interpret the network analysis.

7.1 Methodology of stochastic model using NMF

Background of NMF: Classification prediction is one of the practical real world issue in modern complex multidimensional datasets representing the pattern recognition and unsupervised clustering. Factorization of matrices is one of the techniques that popular decomposition methods like Principal component analysis(PCA) or independent component analysis(ICA) uses along with the comparatively new methodology of NMF which is also decompose the original higher dimensional data to a much lower dimensional basis component. But, it has a special applicability of the cases where the non-negative entries are constraints, and components are independent and thus special hidden factors with overlapping and intuitive relations are expected to be explored in the analysis. The NMF method is implemented in the recommender systems, natural language processing(NLP), audio signal processing, genetics etc. cases.

The advantages of NMF: Three properties distinguish NMF from other decomposition methods [69].

- Non-negativity
- Sparsity
- Overlapping tendency among components.

First of all, NMF is applicable on data that composed of non-negative factors. This instigates the existence of hidden interpretation among the components within the context defined by the primary data. The decomposed factors can be distributed in different proportions compared to the each observed sample. In topic-document example mentioned in Appendix Sec E the top frequent words are linked relationships of the determining the topics of the document.

Secondly, sparse results are very common in NMF, i.e the decomposed basis components have very small number of nonzero elements. This makes the classifications are very localized and compact compared to other ones [70, 71]. The sparseness of the results is such that the most dominant components is easily identifiable compared to the less supreme ones.

Lastly, NMF aims for finding underlying relations among the entries irrespective to the orthogonality or dependence. For example in case of PCA next component of the highest variance or the principal components has a chance to be localized to be orthogonal to the principal ones. So, for cases where unexplored relationships categorization is needed, NMF is comparatively more effective.

Mathematical definition of NMF: This part of the discussion provides a formulation for NMF. The problem solving feature of NMF is linear dimensionality reduction (LDR), which is a key tool in data analysis, and is widely used for example, in the use cases like compression, visualization, feature selection and noise filtering. The original data matrix X is decomposed by two low-rank matrix W and H which approximately regenerates the original data when matrix multiplication of those are implemented. An error or loss is involved in doing so.

This introduces the concept of *Error or Loss function*. It indicates the measure of the quality of approximation. Let X be a $n \times p$ non-negative matrix, meaning all the row and column item are either zero or positive values. This original data source can be in applications such as audio spectrograms or muscular activity, Netflix users' movies rating, or in our cases the users' currency flow ledgers. Then the Non-negative matrix factorization can be defined as the approximation of as follows:

$$\begin{aligned} X \approx WH \qquad & \text{where } X_{ij} \geq 0 \quad for \quad all \quad i,j, \\ & W_{ij} \geq 0 \quad for \quad all \quad i,j, \\ & H_{ij} \geq 0 \quad for \quad all \quad i,j. \end{aligned} \tag{24}$$

and can be rewritten as,

$$X_{i,j} = \sum_{k} W_{ik} H_{kj} \tag{25}$$

where W, H are the decomposed matrices with dimension $n \times k$ and $k \times p$ respectively. The term k is a hidden feature positive value, i.e. $k > 0$. In order to implement the NMF two important factors needed to be considered.

Optimization: The main target of NMF is to estimate the matrices W and H in such a way that it reaches to the local minima. It is also called the factorization rank k which is often chosen in such a manner that $k \ll min(n,p)$. It takes more computational time as the main data represented in the matrix form in order to update the rules of NMF. Instead of converging to global minima it stabilizes in the local minima as per the rules update which is also based on the randomly initialized for W and H. It is important to choose the initial condition as because of the stochastic nature of initialization might produce the same output in every run. The algorithm binds the samples into k features or components, where k is the pre-specified factorization rank.

$$\min_{W,H \geq 0} [D(X, WH) + R(W, H)] \tag{26}$$

where

- D is the loss or error function that measures the approximation. There are two types of loss functions which are based on either the *Frobenius distance*

$$D_{FD}(X, WH) = \frac{1}{2} \sum ||X_{ij} - (WH)_{ij}||^2 \tag{27}$$

or the *Kullback-Leibler divergence*.

$$D_{KL}(X, WH) = \sum_{ij} \left[X_{ij} \log \frac{X_{ij}}{(WH)_{ij}} - X_{ij} + (WH)_{ij} \right] \tag{28}$$

To be minimized, measuring the distance between X and the dot product WH. Note that, there is another option called *"Itakura-saito"* different from *"Frobenius"* and *"kullback-leibler"* lead to significantly slower fits and have use cases in the audio signal processing.

- Here, an optional regularization function is R, which can be imposed to extract desirable outcomes in the matrices W and H, for example smoothness and sparsity[72]. Regularization is very important as it involves some known factors that is observed with the experience of handling the data and that might influence the outcome. In our stochastic model development as we are dealing with the unsupervised nature of our Bitcoin user pool data we were not influenced about using the regularization parameters. We have done an analysis with the known use case of text-document NMF analysis based on topic as the hidden factor included in the Appendix Sec E, where we used the *Scikit-learn NMF* python library and its builtin NMF package (more information , please see [73]). In there we showed how the regularization parameters can be ignored. And for the loss function we had selected the *Kullback-Leibler* and multiplicative updater("mu") as solver parameter.
  The readers will get a basic understanding of NMF in terms of the influence of both the *loss function* and no influence of the *regulaization* with the experimental analysis done on the topic-document example.

Initialization: As discussed earlier initialization or seeding (i.e. a fixed value for W_0 and/or H_0) is important, from which iteration process can be started. It is very crucial to set the initialization. As the data can be highly dimensional and optimized into only local minima, the initialization parameter for NMF is in fact very important to ensure meaningful results. Even though there are common methods in our case we have selected the *nndsvd* (Nonnegative double singular value decomposition) method. No randomization is needed for this run and normally it has two SVD processes. Data matrix is approximated by one process, the other process focus on rounding up the non-negative portion producing partial SVD terms by deriving the rank matrices. For sparse matrix dataset this algorithm is well fitted [74].

NMF with stochastic approach We have seen the basic NMF equation in Eq. (24) and Eq. (25) where the main matrix X is the approximation of multiplication between the feature matrices W and H. This is a trivial case. Now we want to see if we can interpret the big players' network flow in our restructured user graph and whether the factorization approximation of NMF holds in a probabilistic case.

Now, if we consider the probabilistic context we can define our data as the following,

$$\tilde{X} := c \cdot X, \qquad \text{where} \quad c = \left(\sum_{ij} X_{ij} \right)^{-1} \tag{29}$$

here, $\sum X_{ij}$ represents the total sum of frequency or total flow of BTC from user i to j. Now, in Eq. (24) can be rewritten by multiplying the scalar quantity c in both side as :

$$cX = aW \cdot bH \qquad \text{where} \quad c > 0 \quad \text{and} \quad c = a \cdot b \tag{30}$$

Now,The normalizing the data in accordance with the probabilistic trivial aspect, we can define that :

$$\sum_{i,j} \tilde{X}_{ij} := 1 \tag{31}$$

from Eq. (25) we can rewrite as :

$$\sum_{i,j} \tilde{X}_{ij} = \sum_{i,j} \sum_{k} W_{ik} H_{kj} \tag{32}$$

$$= \sum_{k} (\sum_{i} W_{ik})(\sum_{j} H_{kj}) \tag{33}$$

Here, in accordance with Eq. (31) we define W and H in similar stochastic approach:

$$\sum_{i,j} \tilde{W}_{ij} := 1 \tag{34}$$

$$\sum_{i,j} \tilde{H}_{ij} := 1 \tag{35}$$

and thus we can write from that,

$$\tilde{W}_{ik} = \frac{\sum_{ik} W_{ij}}{a_k} \tag{36}$$

$$\tilde{H}_{kj} = \frac{\sum_{kj} H_{ij}}{b_k} \tag{37}$$

So the final equation turns out to be as follows:

$$\tilde{X}_{ij} = \sum_{k} (a_k \tilde{W_{ik}}).(b_k \tilde{H_{kj}}) \tag{38}$$

$$= \sum_{k} r_k \tilde{W_{ik}} \tilde{H_{kj}} \tag{39}$$

where $r_k = a_k.b_k$

And, r_k, $\tilde{W_{ik}}$ and $\tilde{H_{kj}}$ satisfy the following conditions,

$$\sum_{k} \tilde{W_{ik}} = 1 \qquad \text{for all } k, \tag{40}$$

$$\sum_{k} \tilde{H_{kj}} = 1 \qquad \text{for all } k, \tag{41}$$

$$\sum_{k} \tilde{r_k} = 1 \qquad \text{for all } k. \tag{42}$$

7.2 Stochastic model of NMF

At this stage, we propose a stochastic model of interpretation backed by simulation that verifies the reasoning of NMF outcomes. In order to do that, we want to understand NMF in such a way that could be validated by the way the estimation and simulation is done. The proposed simulation model is a toy-model for a small set of example which gives some insight to do the interpretation from the NMF technique. At the end of this discussion we will be able to observe a model that is identical to NMF.

Poisson's distribution Let us start with an example. Suppose we have a matrix having dimension 10×10, where the number of observations are all positive and denoted by $Y_{ij} \geq 0$. Here the Y_{ij} represents the frequency of the observation or real data from sender i to receiver

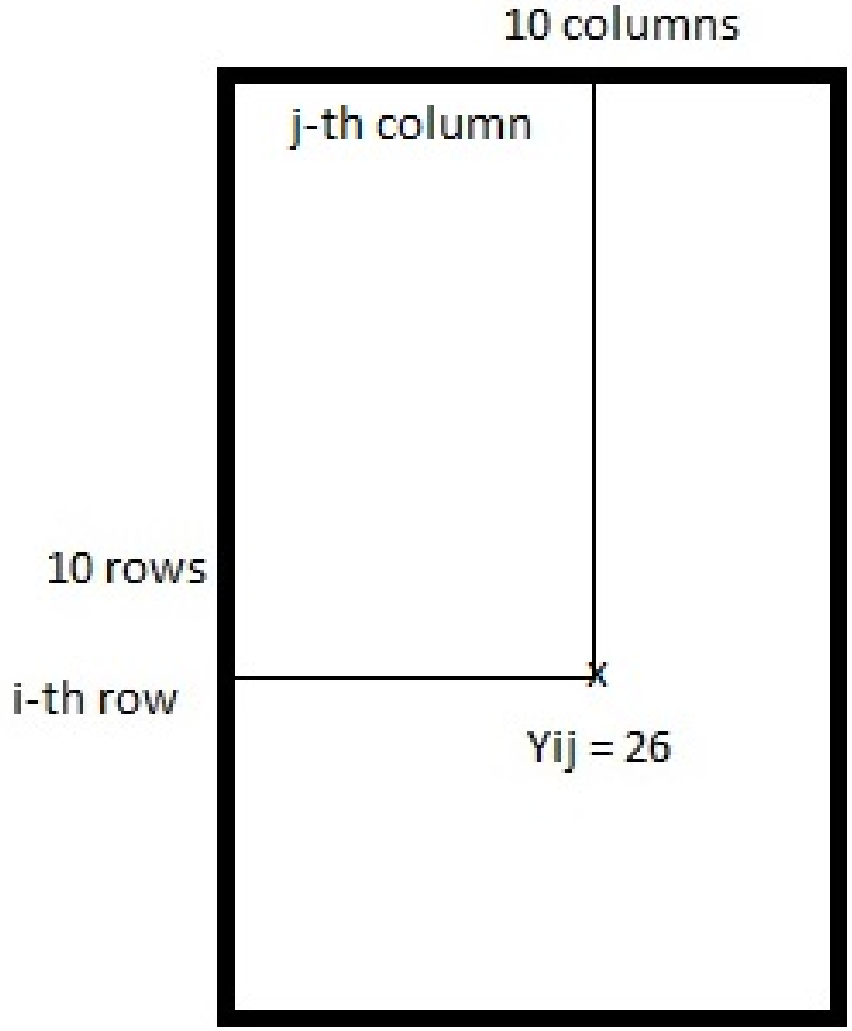


Figure 34: A user graph of 10x10 matrices of frequency observation from sender to receiver

j as shown in Fig.34. Now, our goal is to construct a probabilistic model where we want to estimate X_{ij}, which can be defined by Poisson distribution. The Poisson distribution estimates the probability of a number of independent events occurring a particular period of time. It can be defined as:

$$P(Y = y|X) = e^{-\mu}.\frac{\mu^y}{y!} \qquad \text{where} \qquad \sum_{y=0}^{\infty} P(Y = y, \mu) := 1 \tag{43}$$

Here, μ is the Poisson distribution parameter. And by the definition of the expected value of the 10x10 dimension matrix having 100 observations we know by definition, it is equal to the average of the all probabilities, which in this case is equal to be μ:

$$E[Y] := y \cdot \sum_{y=0}^{\infty} P(Y = y, \mu) = \mu \tag{44}$$

Conditional probability of model parameters: So, if our goal is to find inference from observations or data D and construct a model M, one of the way to achieve this is by doing simulation

where we can generate artificial data D from model parameter M that supports stochastic Poisson distribution. From there we can apply Bayes theorem of conditional probability to gain the reverse outcomes, which can be clarified the following definition of Bayes theorem:

$$P(M|D) := \frac{P(D|M) \cdot P(M)}{P(D)} \tag{45}$$

where the terms $P(M|D)$ is the likelihood, $P(D|M)$ is the posterior problem distribution $P(M)$ is the prior problem distribution and $P(D)$ is the observation. As, our target is to generate simulated data from the likelihood function which consist of a single model parameter $X_{i,j}$ upheld by the above Eq. (45).

Thus we get a framework, where we can look at the Likelihood function and find out the maximum likelihood estimate, which is a standard estimate to minimize the error function of our posterior problem distribution. By taking a small example, we can derive the standard Poisson distribution as mentioned in Eq.(44):

$$P(Y = y|\mu) = e^{-\mu}.\frac{\mu^y}{y!} \qquad \text{where} \quad \text{y=0,1,2,..} \tag{46}$$

$$\log(P(y|\mu)) = -\mu + y \cdot \log \mu \tag{47}$$

$$\frac{\partial[log(P(y|\mu))]}{\partial \mu} = -1 + \frac{y}{\mu} \tag{48}$$

So, based on the conditions of the maximum log-likelihood (if we take the derivatives of the left hand side of) it turns out to be the μ becomes exactly equal to the observation:

$$\mu = y \tag{49}$$

This is trivial case. For example, if we consider the average frequency observation value y is 26 then the log-maximum likelihood is also becoming 26. Now, going back to our main case mentioned in Fig.34, we would like to predict the occurrence of our observations $Y = y$ and instead of μ we replace the term with an independent stochastic terms, for example X_{ij} for all rows and column values of $i, j = 0, 1, 2, ..., n$ where X_{ij} obeys the Poisson distribution. If the observations are replaced by $Y_{i,j}$ instead of $Y = y$ from the above derivation the conditional probability for our entire entries of our matrix are assumed to be the product of all the corresponding observations and parameters because of their nature of independence. So, in our problem scenario, we have a 100 parameter times 100 observations which are independent of each other.

$$P((y_{ij})_{ij}|(x_{ij})_{ij}) = \prod_{i,j}(y_{ij}|x_{ij}) \tag{50}$$

By taking the maximum log-likelihood termed as $L(x_{ij})$, we can derive it similar to the example mentioned earlier:

$$L(x_{ij}) = \log[P((y_{ij})_{ij}|(x_{ij})_{ij})] \tag{51}$$

$$= \sum_{i,j} \log P(y_{ij}|x_{ij}) \tag{52}$$

$$= \sum_{i,j} (y_{ij} \log x_{ij} - x_{ij} \log y_{ij}) \tag{53}$$

$$\frac{\partial L(x_{ij})}{\partial\, x_{ij}} = -x_{i,j} + y_{i,j} \log x_{ij} - \log y_{ij}$$
$$\frac{\partial L(x_{ij})}{\partial\, x_{ij}} = -1 + \frac{x_{ij}}{y_{ij}}$$

Thus, The condition turns out to be for the maximum likelihood is:

$$x_{ij} = y_{ij} \tag{54}$$

Now, calculating the error function :

$$\begin{aligned}
\text{Error Function} &= \sum_{i,j} [L(x_{ij} = y_{ij}) - L(x_{ij})] \\
&= \sum_{i,j} [(y_{i,j} \log y_{ij} - y_{ij} + \log y_{ij}!) - (y_{i,j} \log x_{ij} - x_{ij} + \log y_{ij}!)] \\
&= \sum_{i,j} (x_{ij} - y_{ij} + y_{ij} \log \frac{y_{ij}}{x_{ij}})
\end{aligned}$$

So, the equation turn out to be like :

$$\text{Error function} = \sum_{i,j} (y_{ij} \log \frac{y_{ij}}{x_{ij}} + x_{ij} - y_{ij}) \tag{55}$$

which is the same as the **Kullback-leibler** error function of Eq. (28) mentioned earlier to discuss about the optimization of NMF.

So, we can finally interpret that, our stochastic model which involved Poisson distribution is validated by the KL-error function which is exactly equal to the log-likelihood function and according to the Beysian estimation framework we can use that to estimate the model parameters X_{ij}. And the final investigated findings in a nutshell is that NMF is trying to find the local minima for this error function which is exactly identical to the maximum of the posterior problem distribution or the log likelihood function.

Simulation with Bayesian Information Criterion In this part we would like to do a simulation trial to estimate some parameters with the framework that we have just discussed. For this we want to introduce a small toy model. We have taken some assumptions for this example as shown in Fig. 35.

So, we have a scenario where we have assumed number of sender and receivers of Bitcoin is $n = 10$ and for that we have $n^2 = 100$ number of observations. The main issue is to estimate the number of components for NMF according to this scenario with Bayesian framework. We found that, for data following Poisson distribution if decomposed with NMF then the error or loss function is equal to the maximum log-likelihood taken for that distribution. For simulation purpose our initial estimation for the number of components are 2 which are normalized by assuming $r_1 = 0.6$ and $r_2 = 0.4$. We have two decomposed matrices of dimensions $[10 \times 2]$(W matrix) and $[2 \times 10]$(H matrix) where all the rows and columns are normalized(i.e sum is equal 1).To understand the money flow we assumed that 1^{st} user of the 1^{st} component of W sends money that is received by 2^{nd} and 3^{rd} user of 1^{st} component of H matrix and 2^{nd} and 3^{rd} users of the 2^{nd}component of W matrix sends money to receiving

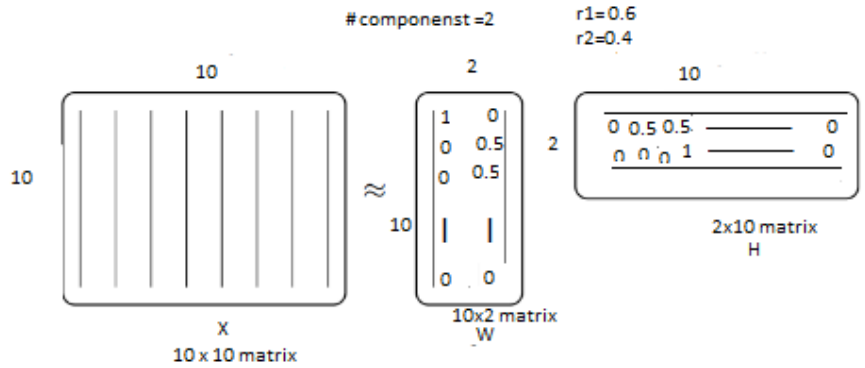


Figure 35: A 10x10 matrix decomposed into dimension of 10x2 and 2x10 matrices where no. of components taken is assumed 2.

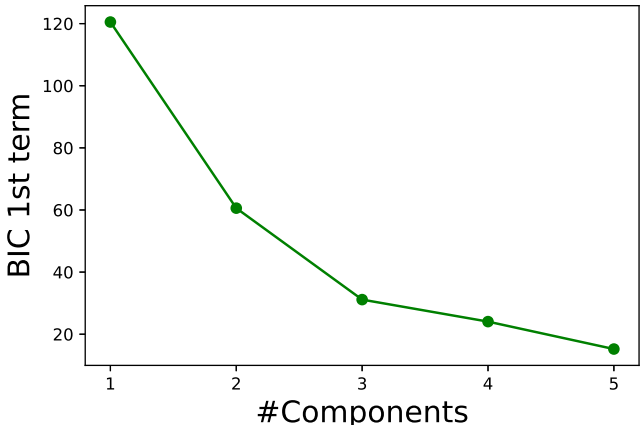
users of 4^{th} user of the 2^{nd} component of H matrix. This is arbitrary and can be tested with random configurations. Now, we transformed this to Poisson distribution observations X_{ij} to run the simulation. The complete codes and results with graphs are uploaded and available in Github (readers can take a look, [75]). Now, we explain and discuss about the NMF outcomes and the estimation of number of components. The NMF decomposition for this scenario is quite straight forward as we have set everything. But, the main problem is to check whether our number of component estimation holds or not. We implemented Bayesian Information Criterion (BIC) to validate this. From Eq. (45) we have to choose different values of $P(D)$ we need to choose and from that the larger values of $P(D)$ should be preferred [76]. This is also termed as *Laplace Approximation* and can be defined as follows:

$$\underbrace{\log P(D)}_{\text{BIC}} \approx \underbrace{\log(P(D|\hat{\theta})}_{BIC \quad 1^{st}} - \underbrace{\frac{D}{2}\log N}_{BIC \quad 2^{nd}} \tag{56}$$

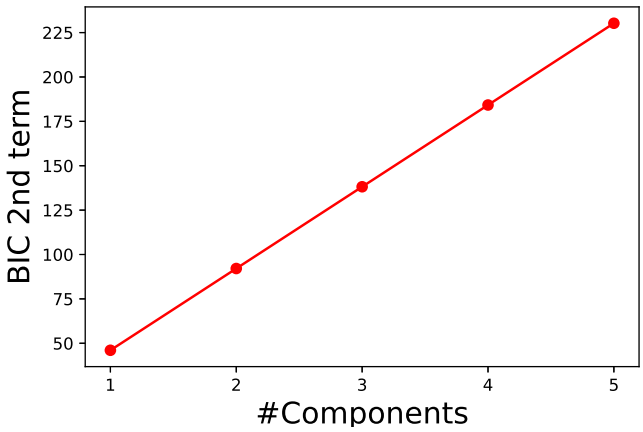
where $\hat{\theta}$ is termed as maximum likelihood estimate, D is the number of parameters of the model and N is the number of observation. Here, the number of parameters are $2*n*k = 40$ and total number of observations are $N = n^2 = 100$ for our case. The advantage of Eq. (56) is, it can do a trade-off between maximum likelihood estimate (1^{st} term) and change of model parameters, observations (2^{nd} term). If we now consider we do not know the ideal number of components and run NMF with KL error function for 5 times, we found that our result for estimating the number of components by measuring BIC has a minima when the component number is 2 as shown in Fig. 36.

Experimentation with BIC toy model We have done some experimentation with this toy model to predict the higher number of components. The main focus is to observe the performance of this model and gather important insights. We started our experiments to estimate the number of components for simple interactions among sender receiver to more complex one. In the following discussion we demonstrated some experiments on the above toy model and all the programming codes are available in Github[77].

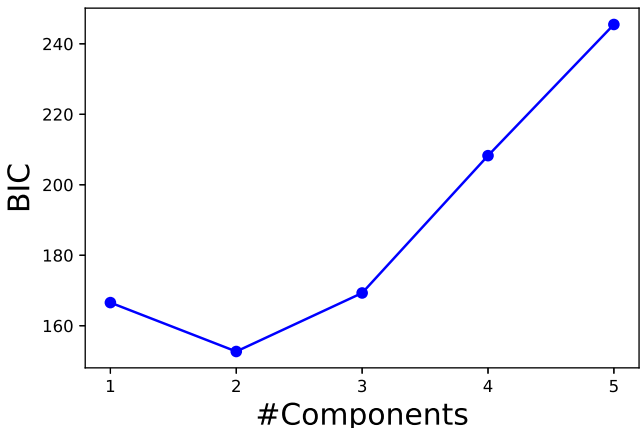
- Predicting number of components k=3: In the Fig. 37 we showed that the scenario where the sending and receiving interaction among 6 nodes. The edges defines the interaction between the nodes and the weights defines the normalized frequency of interaction or money flow. Like the standard toy model in our experiments there were some



(a)



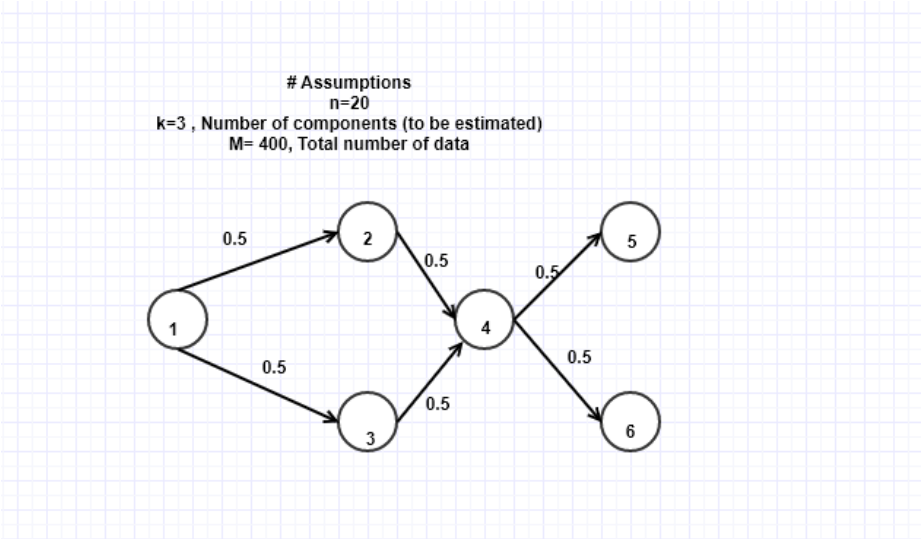
(b)



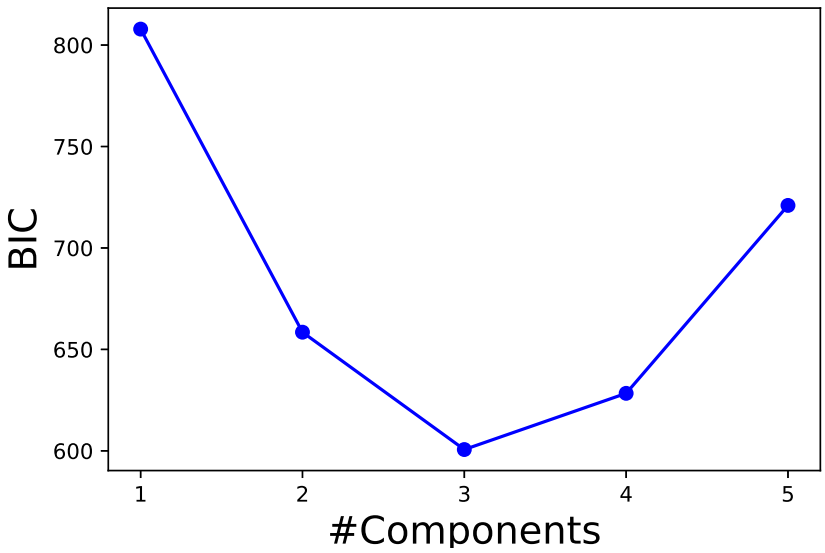
(c)

Figure 36: Estimating number of components with BIC (a) BIC 1^{st} term (b) BIC 2^{nd} term (c)BIC

assumptions we had taken. In the experiment we have found that, assuming higher dimension gives us better result while estimating higher number of components. For example, if we take n=10 and for that the n^2=100 number of observation the accuracy depends on the relative probability r_k values speculated for each components. It also depends on the normalized frequency values taken for the factorized sending and receiving matrices. But increasing the dimension by taking n=20 and thus n^2=400 it showed that, the prediction of the components number is accurate irrespective to the change of relative probabilities or sending/receiving normalized frequencies.



(a) Experimental assumptions with interaction between sender receiver

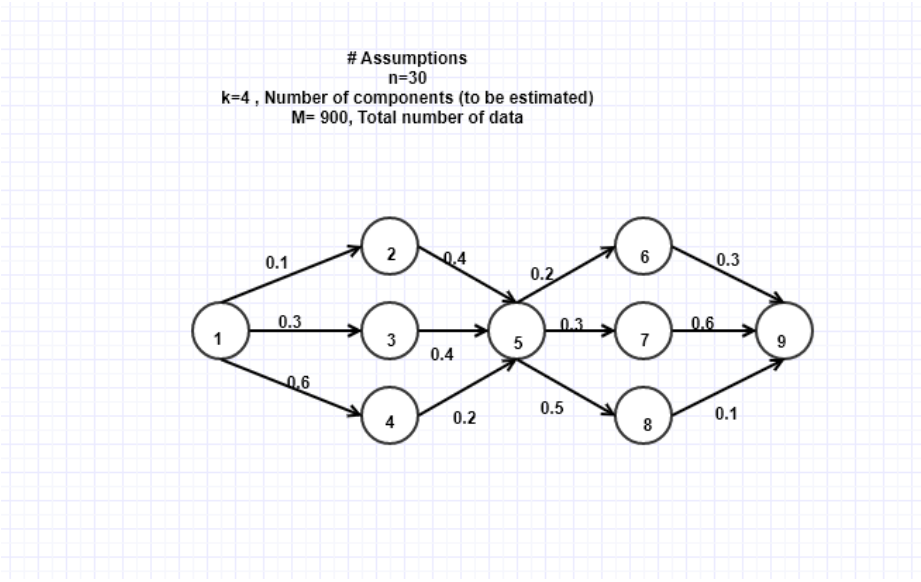


(b) Successful estimation number of components by BIC

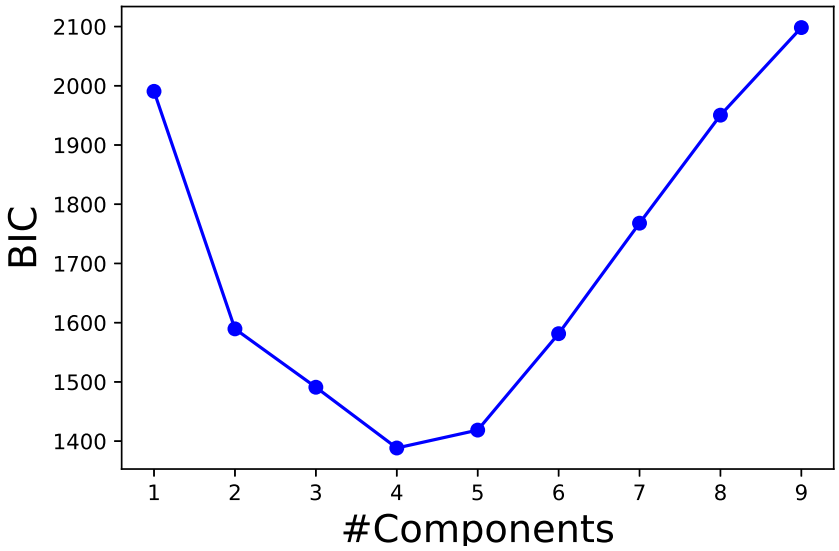
Figure 37: simulation result of estimation of number of components when k=3

- Predicting number of components k=4: In the Fig. 38a the prediction of the number of components k=4 experimental case has been showed. In this case there are more than

two inputs and outputs cases have been considered. In Bitcoin blockchain real user or entity graph data there are most cases single input from one user to multiple outputs to different users. In this example we assumed and tested this kind of scenario. Like the previous results in this case the higher dimension of n=30 and n^2=900 observation provides better estimation.



(a) Experimental assumptions with interaction between sender receiver



(b) Successful estimation number of components by BIC

Figure 38: simulation result of estimation of number of components when k=4

- Predicting number of components k=5: In Fig. 39 the number of components k=5 has been estimated with more complex interactions among sender and receiver nodes. Here the final outputs of group of nodes with their normalized frequency acted as an input of the main sources(node 1), this is not similar to the user graph of real Bitcoin blockchain

scenario, but we implied it in this experiment just to check the models performance. The BIC could successfully could estimate the number of components with higher dimension of data. So, from the experiments conducted so far we found that, dimension of the observation plays a vital role in estimating the model parameter.

Summary: In this section we proposed a stochastic model of NMF that decomposes the observations that follows the Poisson distribution and as the optimization of NMF error function is equivalent of the log likelihood function of Bayesian Information Criterion (BIC) which is helpful estimating the number of components parameter of NMF. Estimation of number of components is one of the key issue for solving classification problems in real world machine learning and data analysis.

7.3 Data pre-processing on daily regular user graph for applying NMF

In this part, from simulation we return to real world data analysis with NMF for daily networks for Bitcoin blockchain. The daily network constitutes of a special user we define as "regular users". For this we recall the definitions of nodes and edges of daily graph that we defined mathematically in the beginning of Sec 3. We only need to define the $\hat{V}$ in such as follows:

$$\hat{V} := \text{Set of nodes/users who appear in every day persistently such that } t \in T_{Month} \tag{57}$$

To accumulate all the regular nodes, we randomly selected two weekdays of the month of February of 2017. We would like to create daily graph created among only *"Regular users"*. So, let us create the daily graph which is in this case is $G_t = (\hat{V}_t, E_t)$, where t is assumed to be each day of February, 2017. For creating daily network matrix, we chose much smaller time scale, i.e a consecutive days of a week. For this reason, we selected two days 21^{st} February,2017(Tuesday) and 22^{nd} February, 2017.

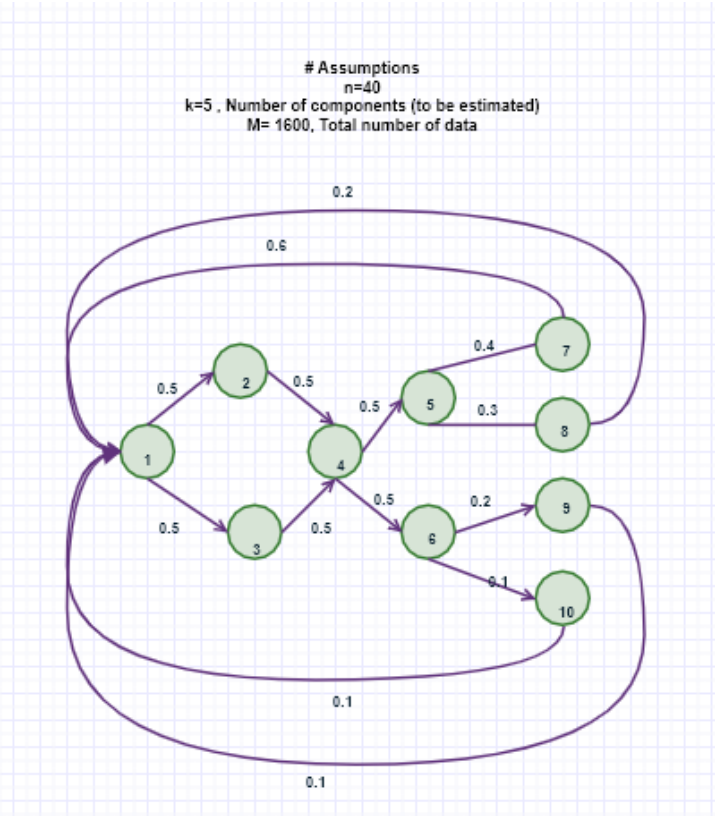
We found that, The total number of overlapping regular users between 21^{st} February,2017 (Tuesday) and 22^{nd} February, 2017 is 3502. And we would like to create the *"Among regular users only"* graph by selecting the edge lists among the regular users. The E_t is the set of links e_{ij} or ordered pair (i, j), which represents all the transactions Tx : $i \to j$ during the day t. The set of all the users appearing at either end of e_{ij} is $\hat{V}_t$ such that $i, j \in \hat{V}_t$. Each edge e_{ij} has the information about the amount of money transferred from i to j and number of frequency of transactions aggregated for unique pair among regular users or mathematically we can say, links or edges are transactions among users denoted by Tx : $i \to j$ total number of transaction from regular user i to j.

$$f_{ij} := \text{number of Tx's among unique pair } e_{ij} \tag{58}$$

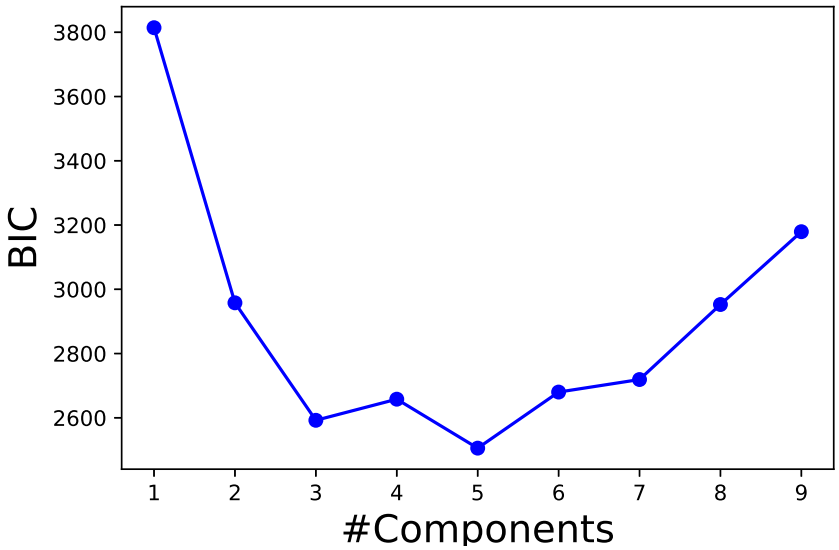
$$B_{ij} := \text{sum of all BTC in Tx's among unique pair } e_{ij}. \tag{59}$$

where we have considered only the cases like Tx: $i \to j$ or Tx: $j \to i$. There are two justifications of creating "Regular Users graph" for NMF analysis. First, regular users are presumably the big players of the network as one of the criteria of them is they are persistently active. Second, we expect that the activities of the regular users are stable during weekdays. We will use NMF to interpret the stability of the temporal change. As the regular users' edge list are associated with the frequency of sending BTC from sender to receiver as unique pair we collected this information. So we have "Frequency" and "Amounts" as edge attributes.

We have created two separate matrices for each edge attributes. The total number of edges for $X_{ij}(t_{21})$ is 9,196 and for $X_{ij}(t_{22})$ is 9,247.The total dimension of the both days are kept same , i.e (3502,3502), same users as senders and receivers to keep the structure stable.



(a) Experimental assumptions with interaction between sender receiver



(b) Successful estimation number of components by BIC

Figure 39: simulation result of estimation of number of components when k=5

7.4 Results and interpretation of regular user graph with NMF

In this section we discuss and interpret about the final analysis of daily "Regular Users'" graph with NMF. The purpose of applying NMF is to find out the hidden factors of the sender/receiver users where they show common patterns of frequency and money flow distributions. We are also interested about the stability of temporal changes of the daily network these regular users. We also like to see whether the financial institutions are actively doing their weekdays' activities or not.

Both days network has been decomposed by NMF with selecting "Kullback-Leibler" as optimization parameter and "nndsvda" as an initialization parameter. And after applying NMF decomposition the matrices also had daily components, i.e $W_{ik}(t_{21})$, $H_{ik}(t_{21})$ and $W_{kj}(t_{22})$, $H_{kj}(t_{22})$, respectively. Here the k value has been assumed to be 20.

Temporal change of network: We have decomposed the 21^{st} February and 22^{nd} February daily matrices constructed with edge attributes as frequency. We have decomposed it with NMF with the conditions mentioned earlier and we wanted to understand whether the decomposed components are drastically changing or not. For this we have taken inner product between W_{21} and W_{22} for "Sender-Component" matrices (6916×20) respectively for those consecutive days. Similarly, inner product between H_{21} and H_{22} for "Receiver-Component" matrices (20×6916) has been calculated. The result were promising, as shown in the Fig. 40.

Some interpretation can be drawn from these results. We can see that there are quite a number components which constitute the users (as senders or receivers) have inner product equal to 1. That means, the users of those components behave similarly in the consecutive days of Tuesday and Wednesday. This is quite natural during weekdays as we have seen in our weekly pattern analysis. The temporal change is slow, the users doing their regular activities and no drastic changes.

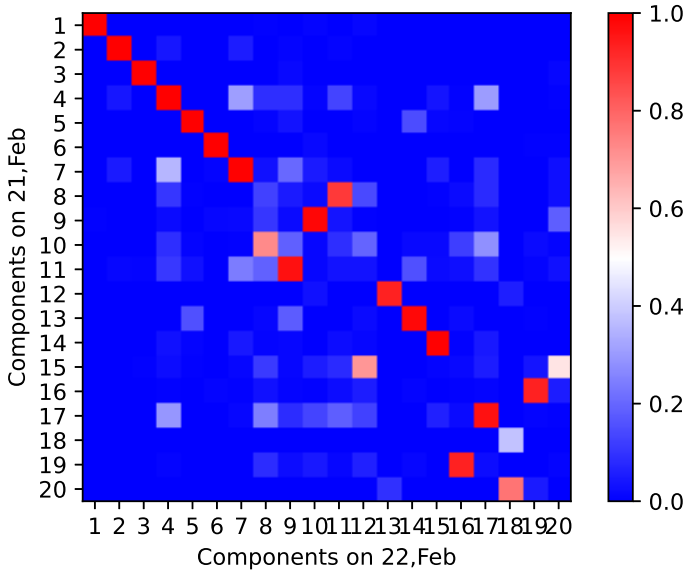
The relative probability of the components: So we have known about the components that are matching or unchanged for the consecutive days. We can now find out the relative probability of the components of the consecutive days as shown in the Fig. 41.

The interpretations we would like to make out of this is, the first component for the consecutive days has the highest probability to occur. We would like to see this component more closely to identify the normalized scores. This is shown in Fig. 42 The first component constitutes the users that has the highest relative probability (r_k, calculated in the Eq. (38)) of appearing as senders and receivers in the network. It will be interesting to look into component members' activities.

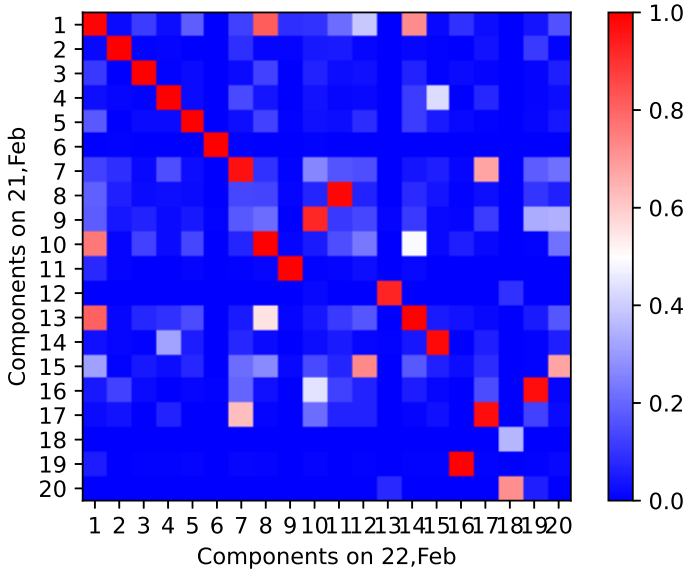
Financial institutions are top users in NMF result: We have looked into the top users in terms of their normalized frequency rank available in the relevant components and got some results as shown in the figure Fig. 43 and Fig. 44

We found big players acting as financial and non-financial institutions as the topper of all the components. observing the results we can draw some important interpretations:

- In both days, the top users of most of the NMF decomposed components and based on their relevant component's normalized frequency are the big players as they matches in list we have discussed in Table 6,Table 5 and Table 4.So, frequency criteria that is one of the criteria of the big players plays a vital role to identify financial institutions that we can prove at this point.
- As for some of the components of sender W_{21} and W_{22} and receiver H_{21} and H_{22} has cosine similarity for example component 1 and component 4 of the both days, we can

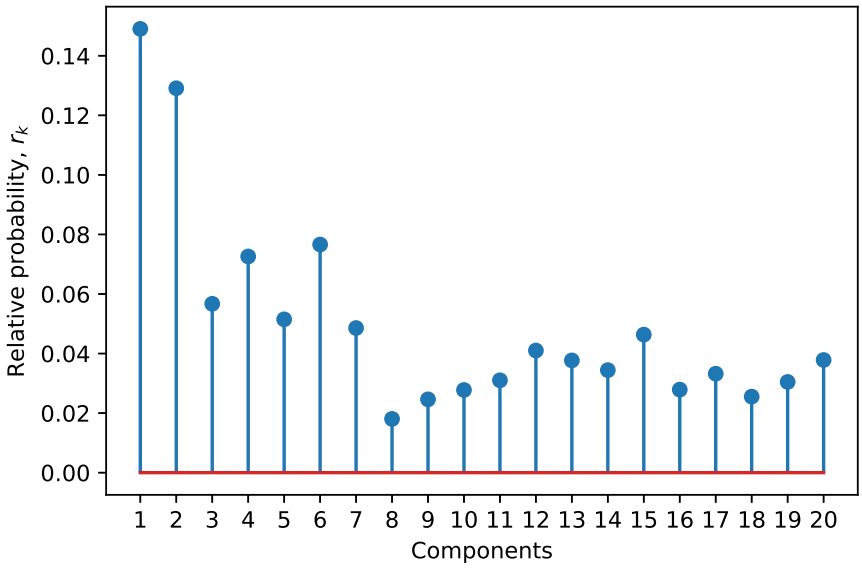


(a) Sender-component

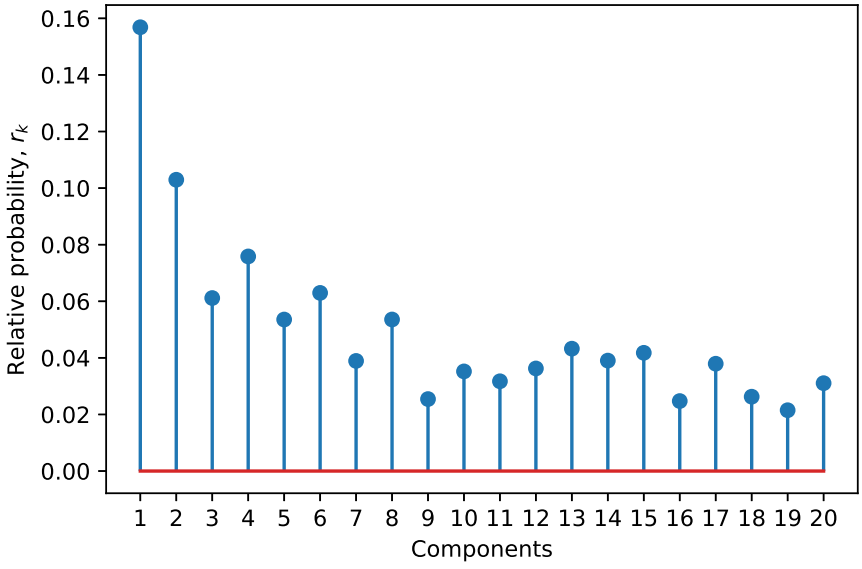


(b) Receiver-component

Figure 40: The temporal change of the components in 21^{st},Feb and 22^{nd},Feb.

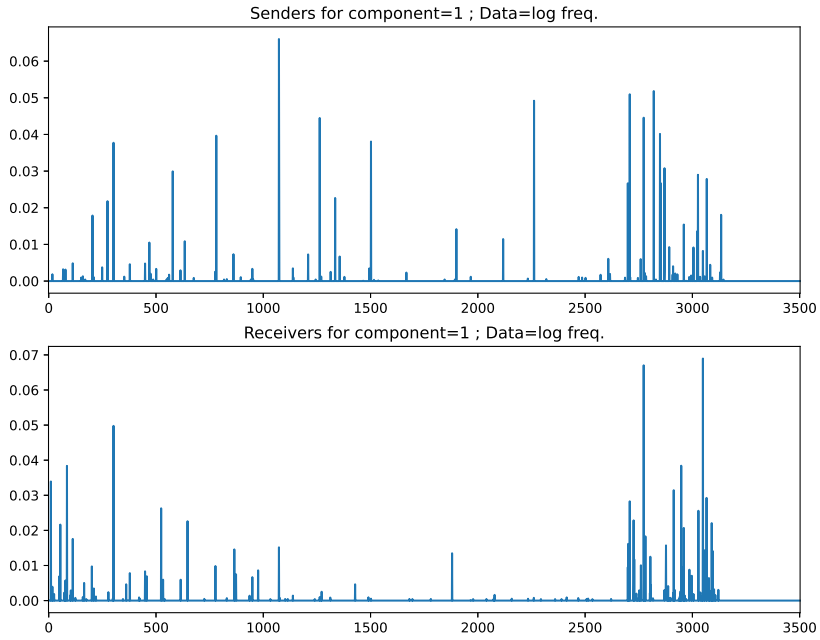


(a) 21_{st},Feb

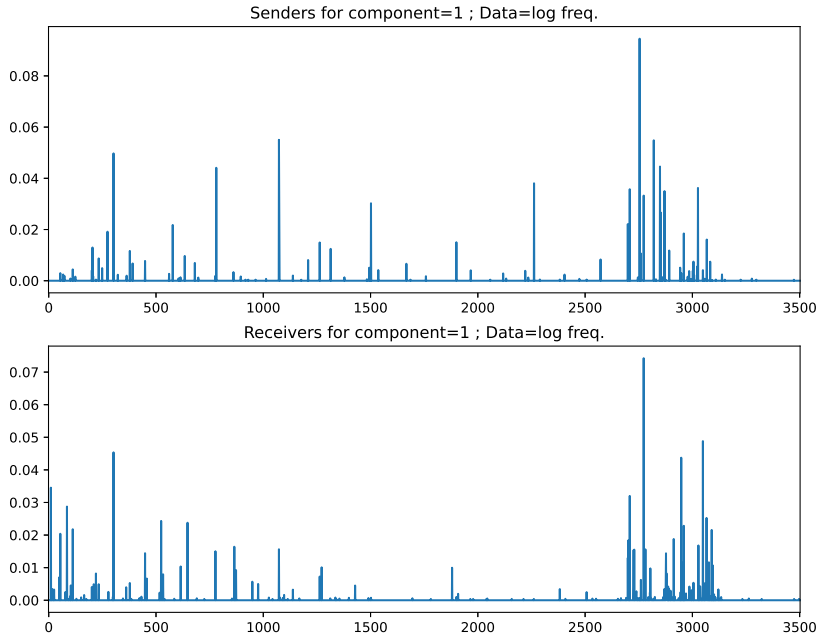


(b) 22^{nd},Feb

Figure 41: The relative probability for components of 21^{st},Feb and 22^{nd},Feb.



(a) 21^{st},Feb



(b) 22^{nd},Feb

Figure 42: The First components scores of 21^{st},Feb 22^{nd},Feb

	top1	top2	top3
comp1	18307826	38GKaU5uRBv8kSPvYqLQy2ikUv1bEyDF78	25703559
comp2	3366757	76589853	14382265
comp3	109540	141669011	172214498
comp4	190362585	134774366	106952630
comp5	183842430	180284474	220075551
comp6	25703559	13vHWR3iLsHeYwT42RnuKYNBoVPrKKZgRv	167253481
comp7	135111428	59826853	104262392
comp8	3LvS6r1mFvxjK7dc2zTfgL3Jq7hUxpjsmw	199376643	3366757
comp9	132252476	204176880	184146522
comp10	216227624	178853729	70410849
comp11	166674129	95811511	174273744
comp12	218694953	201838090	1Ghb1yTzeqd4wwDHB2a8vyQZQVy5Wp6fiA
comp13	211177224	173555579	74108048
comp14	70086283	199178973	123194937
comp15	151063657	175697027	166949925
comp16	183220491	3BUbejc5LDfwg4zCTavmeFRexT7YZ8KTwn	130215517
comp17	48254245	207712945	71169849
comp18	217997389	14382265	210756895
comp19	45452467	154433031	49078080
comp20	192618238	155539540	78989971

(a) Top 3 senders 21^{st},Feb

	top1	top2	top3	top4	top5	top6
comp1	76589853	3366757	14382265	56427193	116291458	101401877
comp2	190362585	215462913	170294300	18307826	24563868	216825400
comp3	45976983	16033666	103031823	49538463	3366757	63310938
comp4	32369669	210795311	23743588	86929094	201571301	16265909
comp5	42913255	74108048	86929094	3366757	201571301	14382265
comp6	25703559	27888617	65608404	59826853	104262392	199784969
comp7	79975574	148938232	106988610	177282733	174443029	25703559
comp8	164720225	95811511	101401877	12168693	56427193	24711277
comp9	45452467	107326172	131684425	18307826	41633902	210486448
comp10	25703559	42913255	32369669	74108048	3366757	14382265
comp11	3366757	14382265	56099204	76589853	25703559	165426040
comp12	190362585	32369669	14382265	17934254	94153183	152268885
comp13	42913255	11031719	188332135	33974220	74108048	190362585
comp14	174443029	135111428	59826853	119968883	11031719	188689386
comp15	45976983	212752069	16033666	75248066	51695864	127461440
comp16	157931929	58601568	87348010	57984505	190362585	72220935
comp17	3366757	41633902	149071712	25703559	29310636	14382265
comp18	3366757	190602236	43337625	36960179	157931929	178393336
comp19	41633902	165454643	214810920	47299944	113298852	100648210
comp20	215462913	14382265	88232286	3366757	81117251	25703559

(b) Top 6 receivers 21^{st},Feb

Figure 43: The top scorer of users having largest normalized frequencies for each components for 21^{st}, Feb

	top1	top2	top3
comp1	31792353	18307826	38GKaU5uRBv8kSPvYqLQy2ikUv1bEyDF78
comp2	3366757	76589853	14382265
comp3	109540	166949925	167652232
comp4	190362585	182118679	158717899
comp5	183842430	187104536	43002642
comp6	25703559	13vHWR3iLsHeYwT42RnuKYNBoVPrKKZgRv	58601568
comp7	135111428	3366757	92758581
comp8	36960179	70086283	17934254
comp9	132252476	204176880	184146522
comp10	216227624	3LvS6r1mFvxjK7dc2zTfgL3Jq7hUxpjsmw	3MGAP2FWrDrxRxZ5zSYQjsZLmtotron2sd
comp11	45452467	98392933	107326172
comp12	155539540	71169849	113114754
comp13	151063657	175697027	166949925
comp14	48254245	65608404	104262392
comp15	211177224	173555579	59826853
comp16	183220491	1Ghb1yTzeqd4wwDHB2a8vyQZQVy5Wp6fiA	177298596
comp17	207712945	218694953	201838090
comp18	14382265	18p9Ftp3m4435tdpZTvoBsm3yjUgkvTF2b	19697478
comp19	192618238	137456481	95811511
comp20	217997389	153874755	154433031

(a) Top 3 senders 22^{nd},Feb

	top1	top2	top3	top4	top5	top6
comp1	3366757	76589853	14382265	56427193	101401877	25703559
comp2	56099204	109540	18307826	170294300	40759549	37113487
comp3	49538463	103031823	16033666	45976983	48047011	77766238
comp4	32369669	76589853	14382265	3366757	210795311	86929094
comp5	42913255	74108048	201571301	86929094	3366757	14382265
comp6	65608404	58601568	199784969	23743588	55139807	27888617
comp7	79975574	41633902	148938232	106988610	177282733	128405868
comp8	215462913	190362585	3366757	25703559	14382265	174443029
comp9	45452467	107326172	131684425	41633902	210486448	216825400
comp10	42913255	74108048	3366757	32369669	157931929	14382265
comp11	174443029	135111428	41633902	119968883	165426040	104707138
comp12	14382265	78989971	36960179	81117251	72220935	88232286
comp13	45976983	212752069	16033666	75248066	51695864	132424640
comp14	25703559	59826853	104262392	3366757	27888617	41633902
comp15	76589853	11031719	42913255	74108048	190362585	188332135
comp16	190362585	17934254	152268885	11031719	14382265	43337625
comp17	32369669	3366757	48225116	14382265	48650557	190362585
comp18	3366757	197098271	160628247	214258796	183220491	19697478
comp19	74108048	47299944	188332135	33974220	113298852	100648210
comp20	157931929	3366757	87348010	57984505	14382265	11031719

(b) Top 6 receivers 22^{nd},Feb

Figure 44: The top scorer of users having largest normalized frequencies for each components for 22^{nd}, Feb

say that the top users' appearing frequency as financial institution during weekdays are relatively similar. Even though the users are not the same for the components having the highest cosine similarity(nearly 1), as being at the top at consecutive days but they have a stable pattern.

- As we have included the self-loops in our analysis we can see that same users are topper as receiving users as well as the senders. These are natural as for financial institutions send and receive money behind a pool of addresses, which results self-loops when they are converted into user graph.
- The money flow attributes would provide more insights about the financial institutions if the flow matrix of the consecutive days is decomposed.
- One short coming of NMF that we had experienced is that, as the decomposed matrices consists of very small number of non-zero items the highest scores are way to larger than the rest lower ones. To overcome this, we used log scale in our analysis.

8 Conclusion

The Bitcoin blockchain with the historically immutable ledger system provide us a great scope to analyze big data and extract important insights. In the field of socio-economic complex network analysis Bitcoin blockchain is one of the researched trusted frontier financial emerging technology that has evolved recently. The scope of research work of this kind of technology is still immense as the network is expanding as users' interest in adopting innovative and trust-worthy technology are increasing.

In my research tenure, I have tried gathering some understanding about the de-anonymization and dynamics of Bitcoin blockchain with network science point of view. We have conjectured some hypothesis about the exchange market who are the outliers inside the network. They have vital roles to play in terms of their frequent activities and large volume of network flow. We successfully de-anonymized identity not by breaking cryptographic codes, rather by exposing behavioral pattern they show. In light of dynamics of network science, we showed that, their behavioral pattern changes from weekdays to weekends. Our simulation techniques, providing scopes to estimate the value of key parameters of solving clustering problem, in this case, NMF, which we interpreted in terms of probability. We analyzed regular user graph and found out financial institutions existence as top users based on NMF scores and found the temporal change of their activities are quite stable during weekdays.

As the research work has been divided in three phases, a phase by phase research activities and finding can be discussed in details in this concluding section.

In the first phase, time series analysis of the number of daily transactions and total daily exchanged BTC were scrutinized. The significant findings of this phase was the weekly pattern of the mentioned parameter, which indicates that even if the blockchain based crypto-currency is theoretically fully functional on 24/7 days of the week, the activities during weekdays are higher on average compared to weekends. Besides, when focused on the significant spikes in the transactions time series data, we found activities of outliers whom I conjectured to be hidden crypto-exchange markets at that stage.

In second phase, daily users' graphs were created and network properties with threshold analysis were conducted for different endogenous attributes. I proposed criteria to identify the "big players" and categorized them as financial and non-financial institutions those who are complacent with the criteria. The criteria are high frequency, persistent activities and weekly pattern of the daily total average BTC flow. The big players following fully these criteria are speculated to be identified as financial institutions. Another important finding at this stage was, more than 50% of the total threshold flow (Above 20 BTC) involves the circulation of economic activity among the big players.

In the final stage, I had implied decomposition technique known as NMF, that factorizes a transaction frequency matrix X into certain lower rank matrices W and H, with the characteristics of all the three matrices having non-negative relative weights. The role of users in the decomposed sender and receiver matrices are distributed into components revealing hidden features of classification. I found the financial institutions that I categorized in the second phase of my research actually exists having top ranked relative weights. This successfully proved our "big player" hypothesis. I also explained that NMF can be interpreted as a probabilistic model. A simulation toy model had also been introduced that can successfully estimate model parameter, the number of components in this context, in Bayesian estimation framework. I had run several experiments with different transactions pattern scenario n that toy model. The results of the experiments instigated that NMF outcomes can be explained by probabilities of the relative weights of the main matrix comprised of interaction frequency. The dimension of the main matrix plays an important role in estimating the parameter in our experimental findings. I also analyzed the temporal change of the

consecutive weekdays graph for regular users. I found the dynamics is quite stable as expected and very slowly indication of change was observed.

There were some caveats experienced while using the NMF. The factorized matrices comprised of very small number of non-zero scores. In order to overcome this limitation, I had to measure in log-scale. Another problem I experienced that, the approximation or the reconstruction error of NMF in our analysis was not satisfactory. Techniques can be implemented in order to improve this. Other LDR method implementation and comparative performance studies can be explored as well.

The practical real world daily data classification issue was partly resolved for estimating the model parameter of number components. Our findings regarding that was, in spite of BIC being very promising in showing accurate experimental results in the simulated data, it has limitation in correctly estimating in case of the real world data. We need to further investigate this and find out more sophisticated suited method in solving the problem.

We observed that few big players appeared in different components in multiple times and influenced their existence in terms of possessing high relative weights. In order to quantify their influence in the main matrix, in future I can explore TFIDF (Term Frequency-Inverse Document Frequency).

Acknowledgment

This research of my doctoral thesis has begun in April 2017 at the Graduation School of Simulation Studies at the University of Hyogo. The main objective was to analyze a closed economic-financial network in order to achieve insights of the structural properties and flow dynamics. Many respectable teachers, peers, co-researchers have helped and guided me during the events of this journey and there are some persons I especially would like to thank.

First of all, I would like to thank my supervisor Professor Yoshi Fujiwara of the University of Hyogo for his endless guidance, support, and advice. Furthermore, his collaboration links with crypto-exchange companies and researcher groups of different Universities had opened up new directions and knowledge pursuing opportunities. I have learned a lot about academic research under his supervision and guidance.

I am grateful to acknowledge that my Ph.D. study was funded by the Japanese government (MEXT) scholarship program in the year of 2016. It was a great opportunity for me to pursue a Ph.D. research in Japan. I was also supported financially by JSPS KAKENHI Grant Numbers, 17H02041 and 19K22032, by the Nomura Foundation (Grants for Social Science). On the verge of this global pandemic there were serious impact on my academic research and the final year was hard. I would like to show my heartfelt gratitude to the authorities of the University of Hyogo for all the financial support they kindly provided and part time position arranged for me (Research Assistant,University of Hyogo), and also to my supervisor for managing extra financial support. During this difficult time, our graduate school officials had helped me immensely for which I will be indebted to them for the rest of my life.

I would also like to thank my co-researchers, Mr. Shinya Kawata, Fabeee, Inc. and Mr. Hiwon Yoon, CMD Holdings Inc., Tokyo amd Mr. Yuji Fujita Manager at Turnstone Research Co., Ltd. for their collaborative knowledge and experience sharing. Their price analysis and experiences on behavioral patterns of vital crypto-exchanges activities assisted us to interpret our research results in accordance with the real market scenario.

I am sincerely acknowledging Professor Hideyaki Ayoama, Professor Naoyuki Iwashita, and Professor Yuichi Ikeda at Kyoto University, for their advice and suggestions in the collaborative meetings with the crypto-currency research group of the University of Hyogo and for sharing valuable insights into the blockchain-based financial network.

I would like to thank Ms. Emi Yoshikawa, Ripple Labs, Inc. for her research interest through the University Blockchain Research Initiative (UBRI), a research collaboration framework funded by Ripple Lab, with the crypto-asset research group of Kyoto University and the University of Hyogo.

I would like to thank the Japan Society for Evolutionary Economics for their kind permission for using the texts and figures of my published papers to use in this thesis. I am also grateful to the reviewer team of the "Evolutionary and Institutional Economics Review" journal for assessing my papers and gave directions to make the research better.

Finally, no word can express my deepest gratitude to my wife Dr. Saadia Binte Alam, my sons Samin and Arshan and my parents. Without their immeasurable love, encouragement, sacrifice and sympathy, this thesis could never be possible. I am grateful to all the support and motivation provided by all the neighborhood families and communities during my stay in Japan.

References

[1] F. Martin. *Money: The Unauthorized Biography*. Knopf, 2014.

[2] Mark Ed Newman, Albert-László Ed Barabási, and Duncan J Watts. *The structure and dynamics of networks.* Princeton university press, 2006.

[3] A. M. Antonopoulos. *Mastering Bitcoin: Programming the Open Blockchain*. O'Reilly Media, 2 edition, 2017.

[4] Annika Baumann, Benjamin Fabian, and Matthias Lischke. Exploring the bitcoin network. In *Proceedings of the 10th International Conference on Web Information Systems and Technologies – Volume 2: WEBIST*, 2014. DOI:10.5220/0004937303690374.

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424, 2006.

[6] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Technical report, Manubot, 2019.

[7] F. Reid and M. Harrigan. An analysis of anonymity in the bitcoin system. In Y. Altshuler, Y. Elovici, A. Cremers, N. Aharony, and A. Pentland, editors, *Security and Privacy in Social Networks*, pages 197–223. Springer, New York, 2013.

[8] Dorit Ron and Adi Shamir. Quantitative analysis of the full bitcoin transaction graph. In *Financial Cryptography and Data Security*, 2013. DOI:10.1007/978-3-642-39884-1_2.

[9] Ultimate bitcoin guide. `https://www.vpnmentor.com/blog/ultimate-guide-bitcoin/`. Accessed: 2020-12-28.

[10] Diego Romano and Giovanni Schmid. Beyond bitcoin: A critical look at blockchain-based systems. *Cryptography*, 1(2):15, 2017.

[11] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder. *Bitcoin and cryptocurrency technologies: a comprehensive introduction*. Princeton University Press, 2016.

[12] Dian Rachmawati, JT Tarigan, and ABC Ginting. A comparative study of message digest 5 (md5) and sha256 algorithm. In *Journal of Physics: Conference Series*, volume 978, page 012116. IOP Publishing, 2018.

[13] Hungary research group : Elte bitcoin project. `https://www.buybitcoinworldwide.com/how-many-bitcoins-are-there/`. Accessed: 2020-12-9.

[14] Dorit Ron and Adi Shamir. Quantitative analysis of the full bitcoin transaction graph. In *International Conference on Financial Cryptography and Data Security*, pages 6–24. Springer, 2013.

[15] D. Kondor, I. Csabai, Szüle J., M. Pósfai, and G. Vattay. Inferring the interplay between network structure and market effects in bitcoin. *New Journal of Physics*, 16(12):125003, 2014.

[16] Kondor, D. and Pósfai, M. and Csabai, I. and Vattay, G. Do the rich get richer? an empirical analysis of the bitcoin transaction network. *PLoS ONE*, 9(2):e86197, 2014.

[17] Hungary research group : Elte bitcoin project. https://senseable2015-6.mit.edu/bitcoin/. Accessed: 2020-12-03.

[18] David Chaum. Blind signatures for untraceable payments. In *Advances in cryptology*, pages 199–203. Springer, 1983.

[19] Julie Pitta. Requiem for a bright idea. *Forbes*, 164(11):390–392, 1999.

[20] Laurie Law, Susan Sabett, and Jerry Solinas. How to make a mint: the cryptography of anonymous electronic cash. *Am. UL Rev.*, 46:1131, 1996.

[21] Wei Dai. B-money. *Consulted*, 1:2012, 1998.

[22] Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey M Voelker, and Stefan Savage. A fistful of bitcoins: characterizing payments among men with no names. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 127–140. ACM, 2013.

[23] Micha Ober, Stefan Katzenbeisser, and Kay Hamacher. Structure and anonymity of the bitcoin transaction graph. *Future internet*, 5(2):237–250, 2013.

[24] Michael Fleder, Michael S. Kester, and Sudeep Pillai. Bitcoin transaction graph analysis. http://arxiv.org/abs/1502.01657, 2015.

[25] Damiano Maesa, Andrea Marino, and Laura Ricci. Uncovering the bitcoin blockchain: An analysis of the full users graph. In *Conference: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016. DOI:10.1109/DSAA.2016.52.

[26] Matthias Lischke and Benjamin Fabian. Analyzing the bitcoin network: The first four years. *Future Internet*, 8(1):7, 2016.

[27] Cuneyt Gurcan Akcora, Yulia R. Gel, and Murat Kantarcioglu. Blockchain: A graph primer. http://arxiv.org/abs/1708.08749, 2017.

[28] Massimo Bartoletti, Andrea Bracciali, Stefano Lande, and Livio Pompianu. A general framework for bitcoin analytics. http://arxiv.org/abs/1707.01021, 2017.

[29] Christian Cachin, Angelo De Caro, Pedro Moreno-Sanchez, Björn Tackmann, and Marko Vukolić. The transaction graph for modeling blockchain semantics. https://eprint.iacr.org/2017/1070, 2017.

[30] Rémy Cazabet, Baccour Rym, and Matthieu Latapy. Tracking bitcoin users activity using community detection on a network of weak signals. http://arxiv.org/abs/1710.08158, 2017.

[31] Damiano Maesa, Andrea Marino, and Laura Ricci. Data-driven analysis of bitcoin properties: exploiting the users graph. *International Journal of Data Science and Analytics*, 2017. DOI:10.1007/s41060-017-0074-x.

[32] Stephen Ranshous, Cliff Joslyn, Sean Kreyling, Kathleen Nowak, Nagiza F. Samatova, Curtis L. West, and Samuel Winters. Exchange pattern mining in the bitcoin transaction directed hypergraph. In *Financial Cryptography Workshops*, 2017. DOI:10.1007/978-3-319-70278-0_16.

[33] Elli Androulaki, Ghassan O Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Capkun. Evaluating user privacy in bitcoin. In *International Conference on Financial Cryptography and Data Security*, pages 34–51. Springer, 2013.

[34] Michele Spagnuolo, Federico Maggi, and Stefano Zanero. Bitiodine: Extracting intelligence from the bitcoin network. In *International Conference on Financial Cryptography and Data Security*, pages 457–468. Springer, 2014.

[35] John V Monaco. Identifying bitcoin users by transaction behavior. In *Biometric and Surveillance Technology for Human and Activity Identification XII*, volume 9457, page 945704. International Society for Optics and Photonics, 2015.

[36] Marc Santamaria Ortega. *The bitcoin transaction graph—anonymity*. PhD thesis, Master's thesis, Universitat Oberta de Catalunya, 2013.

[37] Gregory Maxwell, Andrew Poelstra, Yannick Seurin, and Pieter Wuille. Simple schnorr multi-signatures with applications to bitcoin. *Designs, Codes and Cryptography*, 87(9):2139–2164, 2019.

[38] Luke Valenta and Brendan Rowan. Blindcoin: Blinded, accountable mixes for bitcoin. In *International Conference on Financial Cryptography and Data Security*, pages 112–126. Springer, 2015.

[39] Joseph Bonneau, Arvind Narayanan, Andrew Miller, Jeremy Clark, Joshua A Kroll, and Edward W Felten. Mixcoin: Anonymity for bitcoin with accountable mixes. In *International Conference on Financial Cryptography and Data Security*, pages 486–504. Springer, 2014.

[40] Tim Ruffing, Pedro Moreno-Sanchez, and Aniket Kate. Coinshuffle: Practical decentralized coin mixing for bitcoin. In *European Symposium on Research in Computer Security*, pages 345–364. Springer, 2014.

[41] Gregory Maxwell. Coinswap: Transaction graph disjoint trustless trading. *CoinSwap: Transactiongraphdisjointtrustlesstrading (October 2013)*, 2013.

[42] Jan Henrik Ziegeldorf, Fred Grossmann, Martin Henze, Nicolas Inden, and Klaus Wehrle. Coinparty: Secure multi-party mixing of bitcoins. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, pages 75–86, 2015.

[43] QingChun ShenTu. Jianping yu. *Research on anonymization and de-anonymization in the bitcoin system. CoRR, abs/1510.07782*, 2015.

[44] Jason Luu and Edward J Imwinkelried. The challenge of bitcoin pseudo-anonymity to computer forensics. *Criminal Law Bulletin*, 52(1), 2016.

[45] Malte Möser, Rainer Böhme, and Dominic Breuker. Towards risk scoring of bitcoin transactions. In *International conference on financial cryptography and data security*, pages 16–32. Springer, 2014.

[46] Malte Möser, Rainer Böhme, and Dominic Breuker. An inquiry into money laundering tools in the bitcoin ecosystem. In *2013 APWG eCrime researchers summit*, pages 1–14. Ieee, 2013.

[47] Nicolas Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd international conference on World Wide Web*, pages 213–224, 2013.

[48] Tor software. https://en.wikipedia.org/wiki/Tor_(anonymity_network). Accessed: 2020-12-10.

[49] Giulio Cimini, Tiziano Squartini, Diego Garlaschelli, and Andrea Gabrielli. Systemic risk analysis on reconstructed economic and financial networks. *Scientific reports*, 5:15758, 2015.

[50] Francis X Diebold and Kamil Yılmaz. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134, 2014.

[51] Frank Schweitzer, Giorgio Fagiolo, Didier Sornette, Fernando Vega-Redondo, Alessandro Vespignani, and Douglas R White. Economic networks: The new challenges. *science*, 325(5939):422–425, 2009.

[52] A. L. Barabasi and M. Posfai. *Network Science*. Cambridge University Press, 2016.

[53] Hideaki Aoyama, Yoshi Fujiwara, Yuichi Ikeda, Hiroshi Iyetomi, Wataru Souma, and Hiroshi Yoshikawa. *Macro-Econophysics*. Cambridge University Press, 2017.

[54] H. Aoyama, Y. Fujiwara, Y. Ikeda, H. Iyetomi, W. Souma, and H. Yoshikawa. *Macro-Econophysics: New Studies on Economic Networks and Synchronization*. Cambridge University Press, 2017.

[55] H. Aoyama, Y. Fujiwara, Y. Ikeda, H. Iyetomi, and W. Souma. *Econophysics and Companies: Statistical Life and Death in Complex Business Networks*. Cambridge University Press, 2010.

[56] P. Juháaz, J. Stéger, D. Kondor, and G. Vattay. A bayesian approach to identify bitcoin users. *PLoS ONE*, 13(12):e0207000, 2018.

[57] Damiano Di Francesco Maesa, Andrea Marino, and Laura Ricci. An analysis of the bitcoin users graph: inferring unusual behaviours. In *International Workshop on Complex Networks and their Applications*, pages 749–760. Springer, 2016.

[58] Pavel Ciaian, Miroslava Rajcaniova, and d'Artis Kancs. The economics of bitcoin price formation. *Applied Economics*, 48(19):1799–1815, 2016.

[59] Jamal Bouoiyour, Refk Selmi, Aviral Kumar Tiwari, Olaolu Richard Olayeni, et al. What drives bitcoin price. *Economics Bulletin*, 36(2):843–850, 2016.

[60] Ladislav Kristoufek. What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PloS one*, 10(4):e0123923, 2015.

[61] Sean McNally, Jason Roche, and Simon Caton. Predicting the price of bitcoin using machine learning. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 339–343. IEEE, 2018.

[62] Market price of cryptocurrency. https://www.blockchain.com/charts/market-price. Accessed: 2020-12-03.

[63] Alex Greaves and Benjamin Au. Using the bitcoin transaction graph to predict the price of bitcoin. *No Data*, 2015.

[64] Hungary research group: Elte bitcoin project. http://www.vo.elte.hu/bitcoin/;https://senseable2015-6.mit.edu/bitcoin/. Accessed: 2020-01-03.

[65] Blockchaininfo :market price. https://www.blockchain.com/charts/market-price. Accessed: 2020-12-03.

[66] Peter Bloomfield. *Fourier analysis of time series: an introduction*. John Wiley & Sons, 2004.

[67] William N Venables and Brian D Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[68] Rubaiyat Islam, Yoshi Fujiwara, Shinya Kawata, and Hiwon Yoon. Analyzing outliers activity from the time-series transaction pattern of bitcoin blockchain. *Evolutionary and Institutional Economics Review*, 16(1):239–257, 2019.

[69] Renaud Gaujoux and Cathal Seoighe. A flexible r package for nonnegative matrix factorization. *BMC bioinformatics*, 11(1):367, 2010.

[70] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[71] Karthik Devarajan. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol*, 4(7):e1000029, 2008.

[72] Yuanqing Li, Andrzej Cichocki, and Shun-ichi Amari. Analysis of sparse representation and blind source separation. *Neural computation*, 16(6):1193–1234, 2004.

[73] Scikit-learn python library: Nmf algorithm. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html. Accessed: 2020-12-03.

[74] Christos Boutsidis and Efstratios Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, 41(4):1350–1362, 2008.

[75] Simulations code uploaded in github. https://github.com/Rubz2018/simulation/blob/main/Simulation_with_BIC_estimating_NMF_components.ipynb. Accessed: 2020-12-03.

[76] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

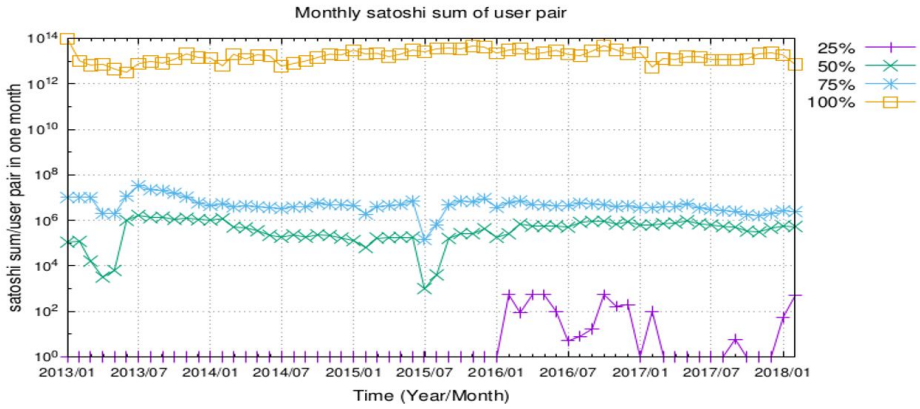
[77] Simulations experiment uploaded in github. https://github.com/Rubz2018/simulation. Accessed: 2021-01-17.

[78] An example of nmf for nlp. https://github.com/Rubz2018/Nonmatrix-Factorization-for-document-terms-analysis/blob/main/Document_terms_analysis_with_Nonmatrix_factorization.ipynb. Accessed: 2020-12-03.

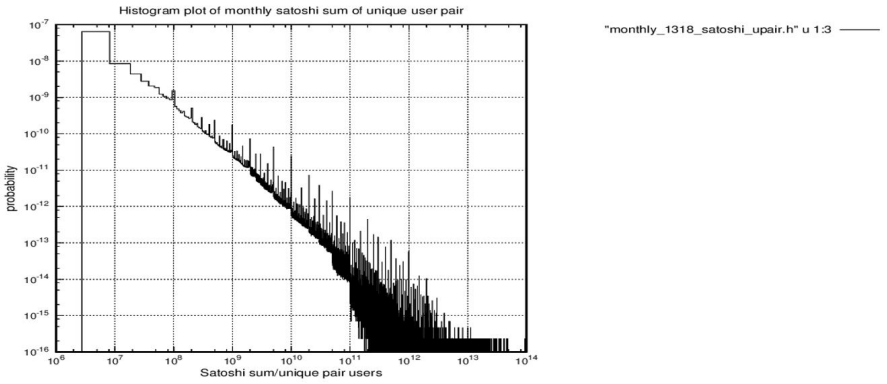
A Statistical analysis of monthly user graph variables

This section consists of the monthly user graph formed during 2013-2018. Monthly data in our research was out of scope. But we would like to share the descriptive analysis of monthly data so that we can demonstrate the comparative situation.

The unique pair volume of the user graph The unique user pair volume in the user graph shows the summation of BTC volume in the time series user graph taken from 2013-2018. In the monthly distribution of unique edge flow distribution shown in Fig. A.1 we can see the maximum 10^6 BTC/unique user. This indicates there are consistent big money flow among unique user pairs throughout the time period. The fat tail PDF distribution shows considerable existence of big BTC flow.



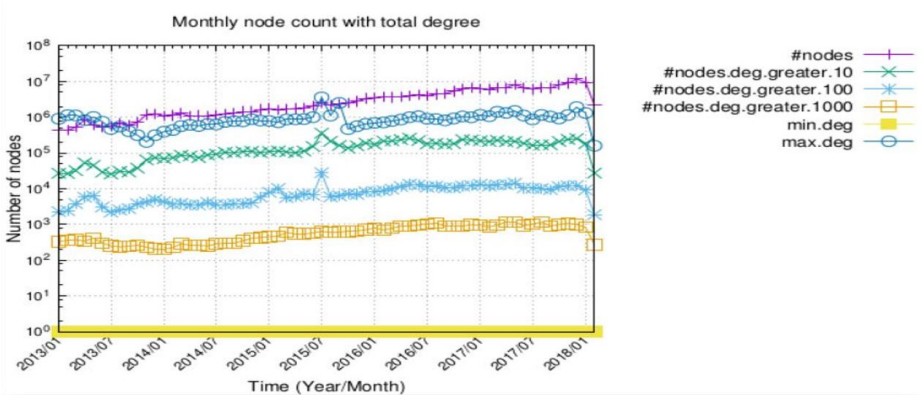
(a) Quantile



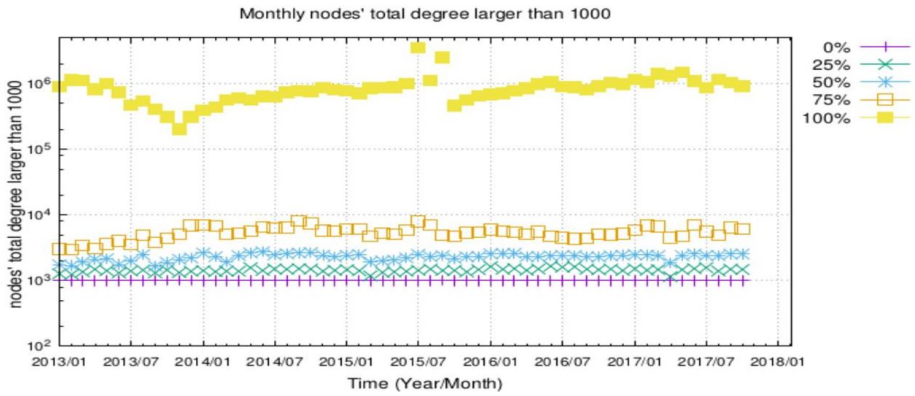
(b) PDF

Figure A.1: Satoshi flow of unique user pair(monthly data)

The degree distribution of monthly user graph The degree distribution of monthly user graph has been demonstrated in the Fig. A.2. The maximum total degree is increasing over 1 million. If we consider to look at the quantile distribution for nodes whose total degree larger than 100, we can see the outliers in the monthly user graph has very large total degree.



(a) Total Degree distribution for monthly node count



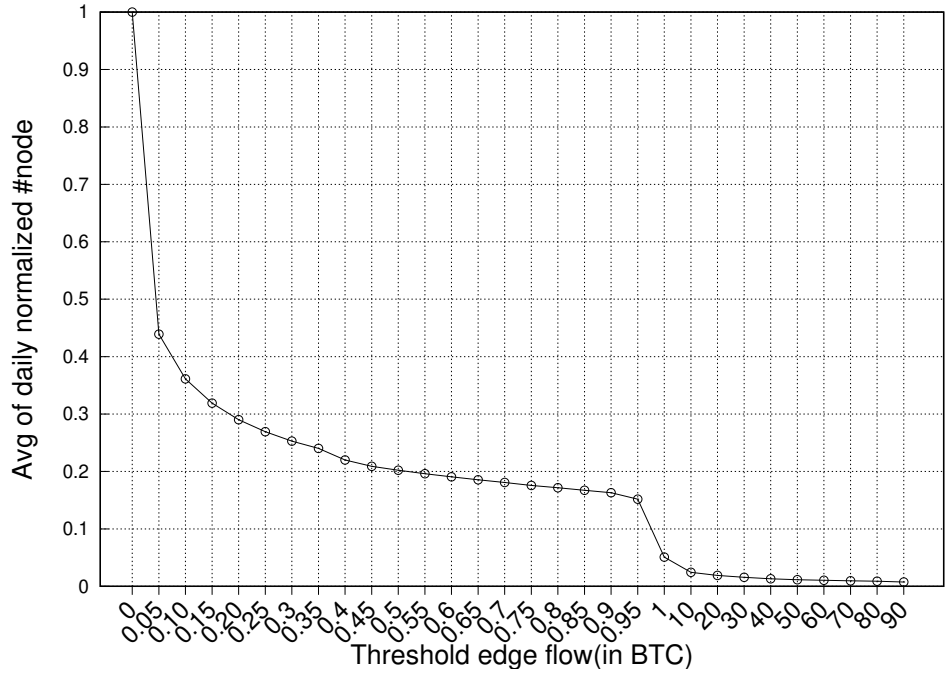
(b) Total Degree distribution with node larger than 1000 counts

Figure A.2: Quantile distribution of total degree to node count over monthly data

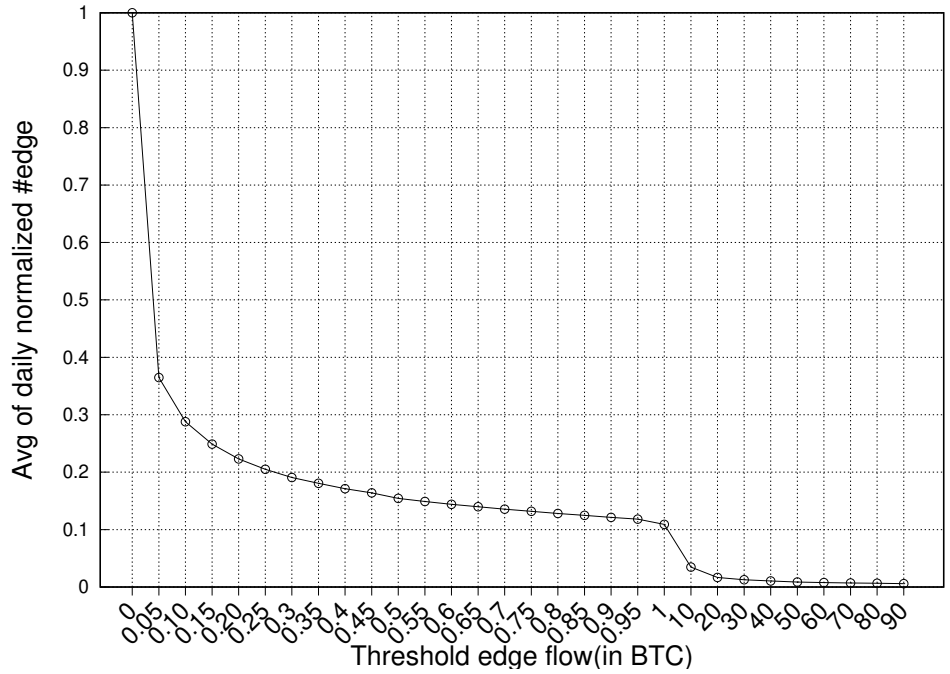
B Threshold and network Size

The node-edge statistics give us some insight on the difference between size of the daily network in active and quiet periods and the normalized number of nodes and edges with respect to different threshold edge flow were shown in Fig. B.1 and Fig. B.2, respectively.

There was perceptible difference in the reduction of network size due to the drastic change in the emulated threshold point range. To cynosure the large edge flows, we focused on the 20 BTC threshold point. Above this point, not only was the size of the network confined to large flows, but it would also give us greater opportunity to identify the topological differences among the big players between weekdays and weekends.

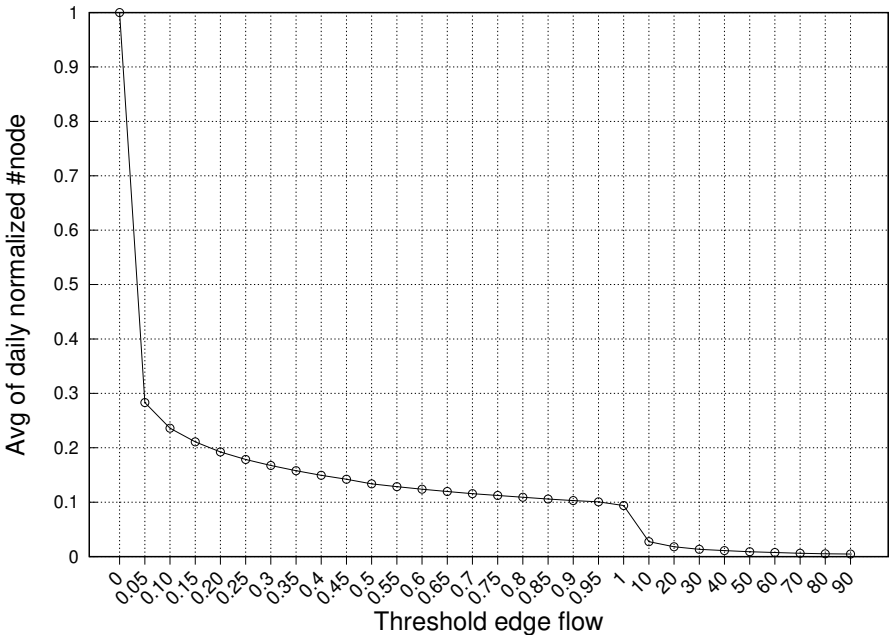


(a) Normalized Node count

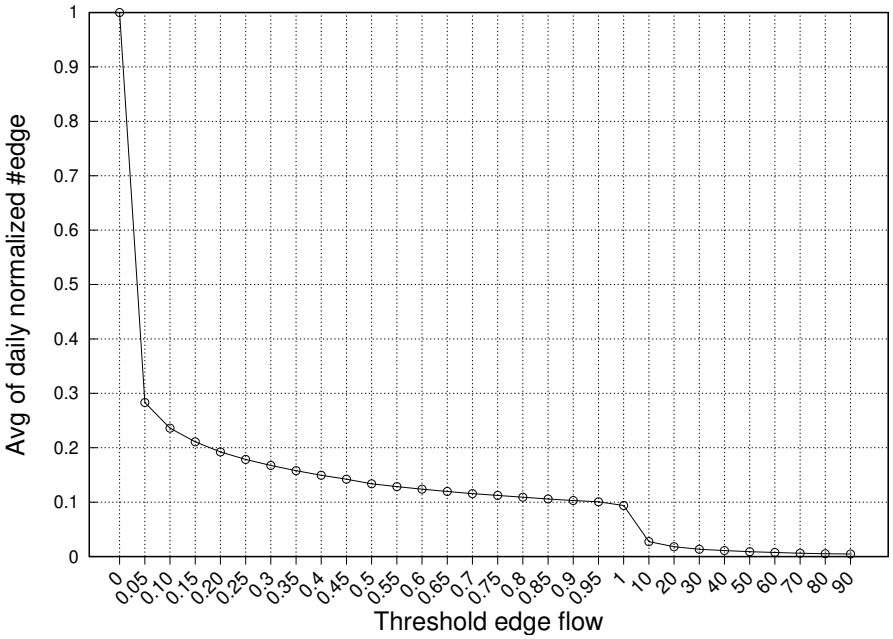


(b) Normalized Edge count

Figure B.1: Normalized node-edge count on different threshold in active period; (a)Node count (b) Edge count



(a) Normalized Node count



(b) Normalized Edge count

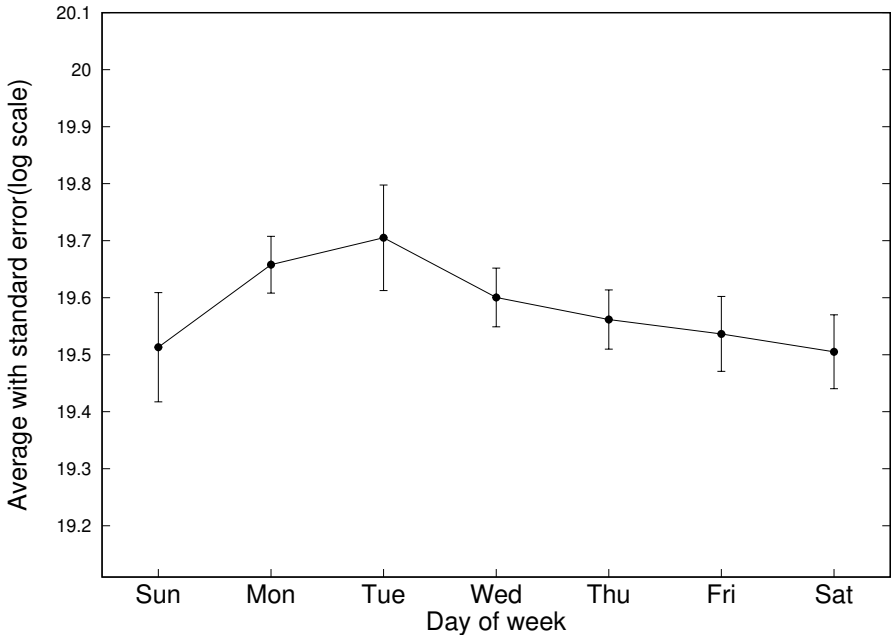
Figure B.2: Normalized node-edge count on different threshold in quiet period

C Sum of edge-flow and average edge-flow

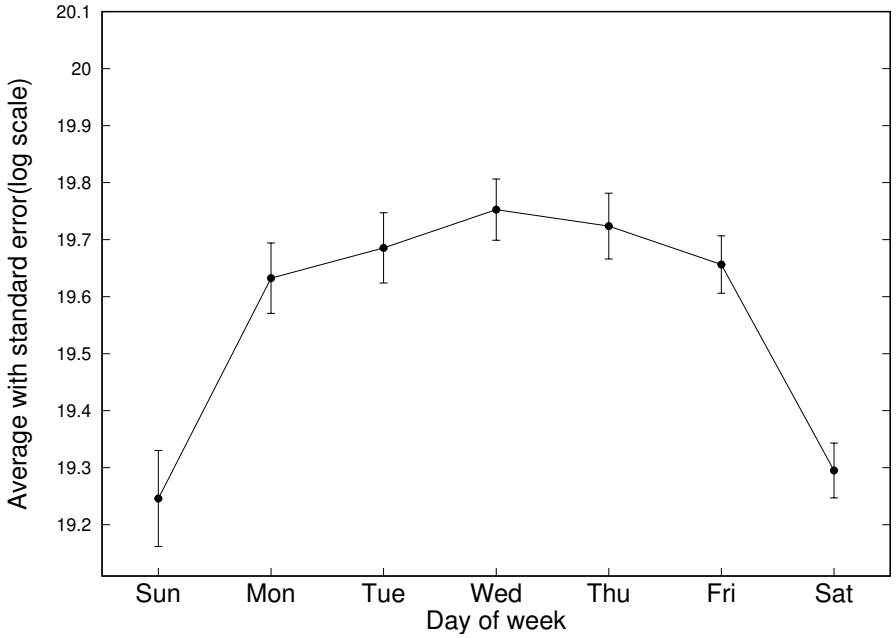
In our previous study[68], we had shown by calculating the average and standard error (standard deviation divided by the square root of recorded data) of the number of daily transactions and the volume sum of transacted BTC between the period 2011 to 2018. Both parameters had distinct proneness to follow a weekly pattern. Therefore, the quantified volume or the number of transactions is higher during weekdays than weekends. We had calculated the daily average edge flow of satoshi(lowest unit of Bitcoin/edge, where 1 satoshi = 0.00000001 BTC (= 1/100,000,000 BTC = $1/10^8$ BTC). By the term "average edge flow" we meant to say the number of satoshi outflow for each unique pair of users or nodes of our graph.

From Fig. C.1, we can clearly observe an indication of divergence from the weekend rate of flow than that of the weekdays in both active and quiet periods. Although the weekly pattern is clear for the main graph, the consistency of the pattern actually diminishes for higher edge flows, as shown in the Fig. C.2. At different smaller threshold points, the dissemblance among Sundays and Mondays were indistinguishable.

Therefore, in our final analysis, we considered the sum of daily edge flow rather than the average edge flow for the weekly patterned activities of big players.

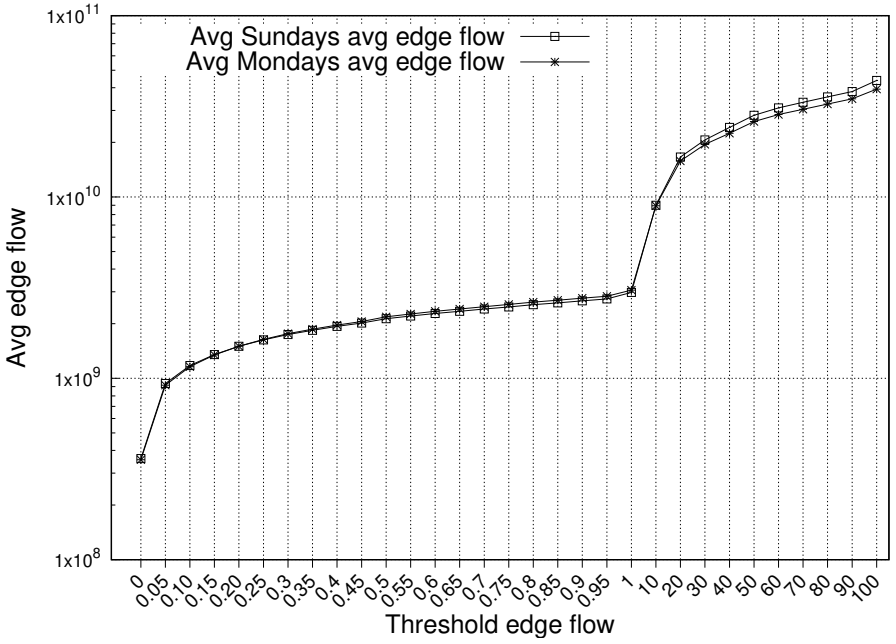


(a) During active period

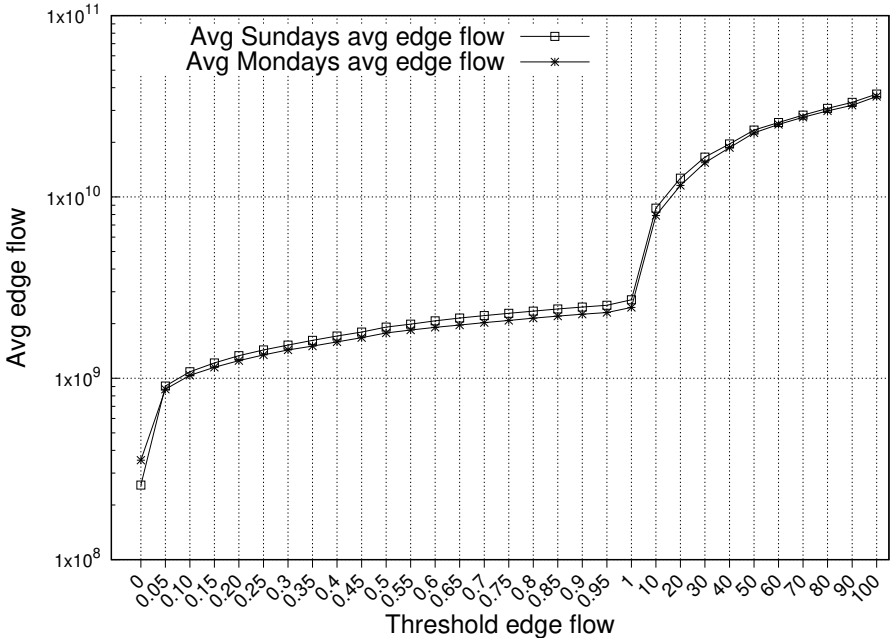


(b) During quiet period

Figure C.1: average edge flow with standard error on average weekdays and weekend for;(a) Active Period (b) Quiet Period



(a) During active period

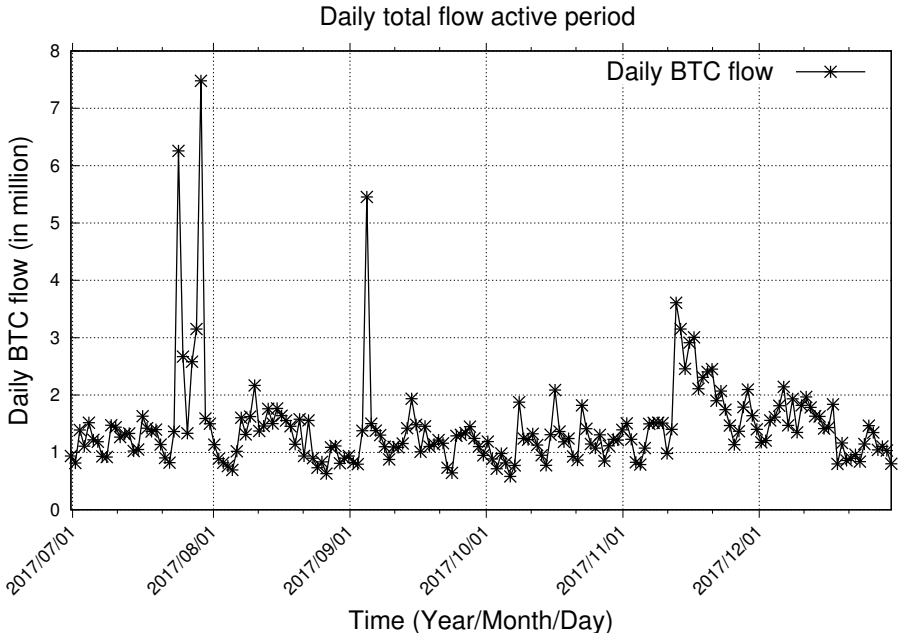


(b) During quiet period

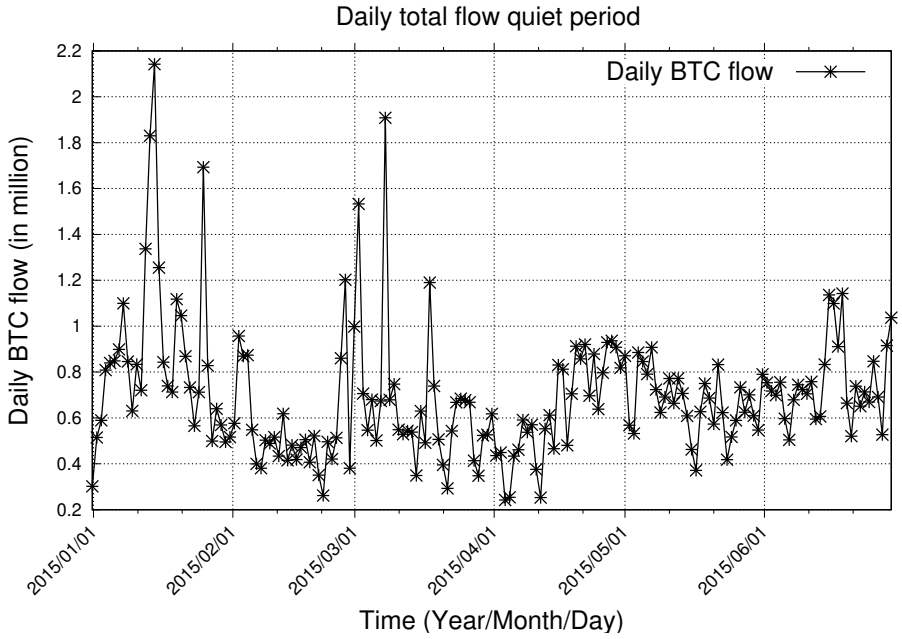
Figure C.2: average edge flow on average Sundays and Mondays at different thresholds for (a) Active Period (b) Quiet Period

D Total flow of subgraph

The sub-graph was formed by taking all the edge flows that were equal to and greater than the 20 BTC threshold. For quantifying the daily total edge flow, we have added the Fig. D.1. This helps the reader to calculate the relative daily total flow inside the largest connected component.



(a) Active Period



(b) Quiet Period

Figure D.1: The total daily flow of sub graph in (a) Active Period (b) Quiet period

E NMF use case:Topic extraction from text-document analysis

This is an example use case of NMF in the field of Natural Language processing (NLP). In this section, we discussed in order to give a basic idea of applying NMF with the scikit-learn python package (class:sklearn.decomposition.NMF) with a tiny example on small number of documents and understand the inner working of the NMF with some exploration on key factors like error function and initialization. In case of the error function, Non-negative Matrix Factorization is applied with two different objective functions, the Frobenius norm, and the generalized Kullback-Leibler divergence. The latter is equivalent to Probabilistic Latent Semantic Indexing. In case of the initialization there are several ways, but we utilized the 'nndsvda' method for our purpose. The complete codes with analysis has been uploaded in the Github [78]. We provide the readers an insight of NMF through this small example. It is useful to understand the parameters that influences the outcomes of this machine learning technique.