Tampere University

Sonja Pakarinen

NETWORKS OF BITCOIN INVESTOR WALLETS

Master of Science Thesis

Faculty of Engineering and Natural Sciences

May 2020

Examiners: Postdoctoral researcher Kęstutis Baltakys,

Professor Juho Kanniainen

ABSTRACT

Sonja Pakarinen: Network of Bitcoin Investor Wallets

Master of Science Thesis

Tampere University

Master's Degree Programme in Information and Knowledge Management

May 2020

Bitcoin is a cryptocurrency which has been on the surface lately. Bitcoin bases on a peer-to-peer decentralized network maintained by the bitcoin users. In this thesis we use statistically validated networks method to validate links between bitcoin investor wallets, a bitcoin system equivalent to bank accounts, to identify clusters and understand the investing behavior by characterizing the bitcoin investor wallets. We characterize the investor wallets based on their hourly activity status to study the degree of synchronization in the decision of when to trade and their links. The analysis is based on the bitcoin transaction data from July 2017 to May 2018. The time period was chosen, because in the middle of the analyzed period the bitcoin price reached its highest point so far and then decreased to a quarter of the highest price.

The study finds that the networks consist of multiple investor wallet clusters, where the wallets have statistically validated links to each other. There is continuity in the behavior of investors between months in terms of the links they have to the other wallets and how they react to the price changes. We also notice, that the investor wallets are likely to transact in the same quantities in different months.

Keywords: Bitcoin, Wallet Identification, Network analysis

TIIVISTELMÄ

Sonja Pakarinen: Bitcoin-sijoittajaverkostot

Diplomityö

Tampereen yliopisto

Tietojohtamisen diplomi-insinöörin tutkinto-ohjelma

Toukokuu 2020

Bitcoin on viime aikoina ilmiöksi noussut kryptovaluutta. Bitcoin perustuu käyttäjien ylläpitämään hajautettuun vertaisverkkoon. Tässä diplomityössä käytämme tilastollisesti validoitua verkostoanalyysimenetelmää bitcoin-investoijien käyttäjälompakoiden välisten linkkien validoimiseksi, tunnistaaksemme verkostosta klustereita sekä ymmärtääksemme sijoittajien käyttäytymistä. Käyttäjälompakot ovat bitcoin-järjestelmän vastineita pankkitileille. Karakterisoimme käyttäjälompakot perustuen niiden tuntikohtaiseen aktiivisuuteen tutkiaksemme lompakkojen välisen yhteyden vahvuutta ja synkronointiastetta. Tämä analyysi perustuu bitcoinien transaktiodataan heinäkuusta 2017 toukokuuhun 2018. Ajanjakson puolessa välissä bitcoinin hinta nousi sen tähänastiseen huippuunsa ja tippui lyhyen ajan sisällä tästä neljäsosaan huippuhinnastaan. Analysoitava ajanjakso valittiin tästä syystä.

Tutkimuksessa havaittiin, että verkosto koostuu useista klustereista, joiden sisällä tilastollisesti validoidut sijoittajalompakoiden välisten linkkien samankaltaisuus on suurempaa, kuin muiden klustereiden lompakkoihin yhdistyvien linkkien. Verkoston lompakoiden linkeissä toisiin lompakkoihin on havaittavissa jatkuvuutta ja näiden käyttäytymisessä hinnanmuutostilanteissa eri kuukausien välillä. Havaitsimme myös, että käyttäjälompakoiden sijoitusmäärät transaktioissa laskettuna pysyvät hyvin samoissa lukemissa eri kuukausien aikana.

Avainsanat: Bitcoin, käyttäjälompakoiden tunnistaminen, verkostoanalyysi

PREFACE

Writing this thesis besides a full-time job and with a 9-month-old baby has not been the easiest task. I would like to say thank you to this thesis' supervisor, Kęstutis Baltakys, for his invaluable help during this process. I also thank you Professor Juho Kanniainen for the supervision and help.

The biggest thanks belong to my husband Oskar who made all this possible and supported me throughout this thesis, and to Klaus, the loveliest son to have.

Helsinki, 8.5.2020

Sonja Pakarinen

CONTENTS

1. INTRODUCTION

Bitcoin and other cryptocurrencies have been on the surface in recent years. The payment digitalization and online payments have become more common, and as a result, many different systems utilizing the internet and digitalization have evolved. Bitcoin can be classified as the original cryptocurrency and it is therefore an interesting subject for analysis. As a result of the economic boom investing has become popular among ordinary - not financially orientated - people. Bitcoin has been one of the investment methods that benefited from the boom. Although the biggest bitcoin hype is over, there still are estimated 7.1 million bitcoin investors (Lielacher, 2020), and the bitcoin transaction data includes interesting information about the hype and the factors behind it. In this thesis we are interested to find major players and investor clusters in order to understand the currency and the methodology behind it. The first part of this thesis is a literature review on bitcoin. Bitcoin is a suitable investment and transaction currency, because all of its transactions are stored in ledger and that is publicly available in the Internet.

Network analysis has been on the surface lately as well. Network analysis is used broadly in social science and sociology, physics, computer science, finance, and economics. In this thesis we use so called complex network analysis to create bitcoin investor networks. We use the methodology to understand the structure of bitcoin investing wallet networks and the investing behavior. This analysis is based on open access bitcoin transaction data from July 2017 to May 2018. The data contains all the transactions made in this time window. During that time there were over 91 million transactions transacting over 73 million bitcoins. That is equivalent to 625 billion dollars (at 1.3.2020's exchange rate).

The network analysis bases on the bitcoin ownership breakdown to bitcoin user, bitcoin user wallet, and bitcoin address. We use heuristics determined in the literature to connect user addresses visible in the bitcoin transactions to the user wallets to understand the investor network structures and relations. We leave the combining to the wallet level, though it is possible to address the wallets to the user level (Meiklejohn, et al., 2013).

In this thesis we categorize investors into synchronization networks. The thesis follows methodically research about Nokia investors (Identification of clusters of investors from their real trading activity in a financial market (Tumminello, et al., 2011)). As a reference for bitcoin user identification methods is an analysis of "Anonymity in the Bitcoin System" (Reid & Harrigan, 2011).

This thesis aims to analyze the investing behavior by characterizing the bitcoin investor wallets in the created networks. The thesis search for answers for the following research questions:

- How the networks of bitcoin investor wallets is structured?
- How the price of bitcoin affects the investors' decision to trade?
- How the different synchronizations of investors' relationships correlate and how the activity changes over time?

The hypothesis for the results is that the networks are heterogeneous, and there are found continuity from the investor's behavior. We assume the price changes have an impact of the trading behavior. As the bitcoin can be classified at least partially a hype investment, the record high bitcoin prices and the significant increase effects on the decision to buy. On the contrary, when the price decreases, the investors refuse to sell because they don't want to lose their money. Also, we assume that the most of the wallets invest only few times a month, and then there are small number of wallets trading frequently. Furthermore, we expect the links between remain the same, i.e. wallets tend to trade with the same wallets.

The structure of this thesis is as follows: Chapter 2 presents bitcoin and cryptocurrency methodology. Chapter 3 examines bitcoin user anonymity and identification. Chapter 4 presents methodology used in this thesis. Chapter 5 describes the data used. Chapter 6 presents the results of the analysis. Chapter 7 discusses the results of the analysis and possible areas of further analysis.

2. BITCOIN

Bitcoin is probably the best-known cryptocurrency and the first “proof of concept” of blockchain. Satoshi Nakamoto, an anonymous “Banksy of the Internet”, published a paper about bitcoin in 2009. (Böhme, et al., 2015; Darren, 2018). Bitcoin is a monetary system that has no physical dimensions and works only online. It is a combination of cash and existing online payment system features. Bitcoin as a system does not identify the payee or the payer in the transactions. Transactions are signed cryptographically with public keys to transfer funds from one to other. (Meiklejohn, et al., 2013) Bitcoin's reliability and functionality is based on the open availability of all the transactions ever made in the history stored in the transaction ledger (Bovet, et al., 2018).

Bitcoin is a completely decentralized currency achieved by p2p (peer-to-peer) architecture and it is independent from central authorities. This shows as following:

- The Bitcoin network has no central server
- Bitcoin do not have data storage; the ledger is stored in the users’ computers
- The bitcoin ledger is public, and everyone can download it
- Bitcoin do not have a single administrator; a network of miners maintains it, and changes in ledger require majority (51%) support
- Anybody can become a bitcoin miner.

Because there is no central authority in bitcoin system, the cryptocurrency only can cease to exist by itself; it is not possible to regulate or eliminate by force. The only way bitcoin can be ceased is when the users lose their confidence in it and stop investing. The confidence can be lost by technical attacks, hacks, or loss of value. In spite of this, it is possible to decide for a form of regulation voluntarily by the users. (Lansky, 2018)

Bitcoin bases on the anonymity of users and that is called pseudonymity: transactions don’t need confirmations from a third parties to take place, and the bitcoin investors cannot be linked to a real users or identities. Cash works the same way, but transaction made with it are not documented as bitcoin's. (Bovet, et al., 2018) Transaction tracks only users’ public addresses in which one user could have an unlimited amount. This is one of the reasons bitcoin has been criticized for close connection to illegal activities (Commission, 2014). There is an estimation that around 46% of bitcoin transactions are related to illegal activities (Foley et al. 2018).

Bitcoin has been a topic lately for many reasons on top of its relation to illegal activities. It has been criticized for its high energy consumption (Vranken, 2017) price volatility, thieving, and the possibility of being an economic bubble (Chaim & Laurini, 2019; Bovet, et al., 2018). For these reasons investing in bitcoin is a risk and multiple authorities have issued warnings about it. Bitcoin and its markets have been criticized by several banks for their volatility and unregulated nature (Lam, 2017). For example, Finnish financial group Nordea has prohibited its employees from investing in bitcoin because of the fear of employees being exposed to criminal activity (Lehmusvirta, 2018).

The bitcoin exchange rate has been strongly unstable during its history. The exchange rate between the US dollar and Bitcoin was determined for the first time in October 2009 and then one dollar was equivalent for 1309 bitcoins meaning one bitcoin was $ 0.00076. The New Liberty Standard Stock Exchange was the launcher of that public sale. (Bitbay.net, 2019) In summer 2011 bitcoin experienced its first peak and one bitcoin was $ 31.91 and the price reached this point next time in November 2013.

Bitcoin experienced its highest point so far in 17. December in 2017 when its price rose to $ 19 783. The price more than doubled in one month as in 17. November it was $7 885. This is a good example of bitcoin's instability. Another major change in bitcoin's price was in August 2016 when the price dropped as a result of cyber-attack and associated theft by Hong Kong bitcoin broker Bitfinex where nearly 120 000 bitcoins worth of $ 72 million were reported lost (Wilson, 2019).

Although initially developed as a cost-effective and independent third-party payment system, bitcoin I mainly a speculative investment (Hileman & Rauchs, 2017). Nevertheless, bitcoin is widely accepted as an online payment method.

2.1 Cryptocurrencies

Cryptocurrencies are digital tokens or asset, a digital equivalent to banknotes, built on blockchain technology (Chohan, 2017; Xunhua, et al., 2018). Cryptocurrencies are objects of exchange as currencies in general, and they are secured with strong cryptography that verifies transactions and controls the creation of additional units i.e. mining. Cryptocurrency methodology bases on decentralized controlling as an opposite to traditional currencies and the central banking systems. (Chohan, 2017)

The first cryptocurrency, *ecash*, was created in 1983. It was an early form of electronic payments and used specified software on user's computer storing money in digital format. The security of

that system based on an early form of cryptographies and public-private key pairs. (Chaum, 1998)

Lansky (2018) defines cryptocurrencies as following: “Cryptocurrency is a system that meets all the following 6 conditions:

- The system does not require a central authority, distributed achieve consensus on its state.
- The system keeps an overview of cryptocurrency units and their ownership.
- The system defines whether new cryptocurrency units can be created. If new cryptocurrency units can be created, the system defines the circumstances of their origin and how to determine the ownership of these new units.
- Ownership of cryptocurrency units can be proved exclusively cryptographically.
- The system allows transactions to be performed in which ownership of the cryptocurrency units is changed. A transaction statement can only be issued by an entity proving the current ownership of these units.
- If two different institutions for changing the ownership of the same cryptographic units are simultaneously entered, the system performs at most one of them.”

Cryptocurrencies are validated through blockchain, a continuously growing list of transaction records linked and secured through cryptography. Blockchain is defined in the next chapter. Bitcoin and other public crypto-currencies base on peer-to-peer networks, public and distributed ledger, and the network has no single administrator (Lansky, 2018). The main differences between centralized and distributed ledgers are visualized in Figure 1. Briefly, the main difference is that centralized ledger uses a common payment system controlling the ledger and payments. It often sets rules and monitors and supervises payment activity.

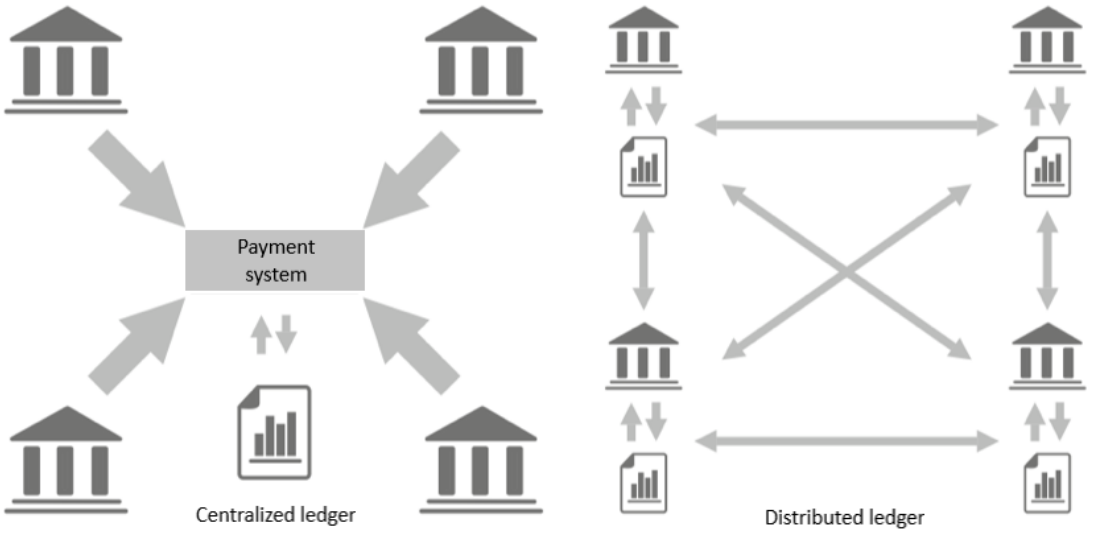


Figure 1. *Distributed ledger system versus Centralized ledger. Edited from (Belinky, et al., 2015).*

2.2 Blockchain

Blockchain is a list of uniquely identified records, blocks, containing recorded transactions in a chain (Treleaven, et al., 2017). The block consists of a parent block linking to the previous block, a timestamp and transaction information (Zheng, et al., 2018). Blockchain grows continually as amount of transactions increases and each hash is "chained" to the previous hash by referring to its hash value. Blockchain is resistant to data modification by design. Blockchain is a method to record made transactions between two or more parties in an efficient, verifiable, and permanent way. (Iansiti & Lakhani, 2017) Each transaction ever made is readable and verifiable from the chain in the right order.

The first idea of blockchain technology was described in 1991 by S. Haber and W. Stornetta (Narayanan, et al., 2016). The blockchain technology was processed into bitcoin by anonymous Satoshi Nakamoto in 2008. In Nakamoto's idea the blockchain stores the bitcoin transaction data ledger. As the distributed ledger is stored in the blockchain, the cryptocurrency does not need any central authority or data hubs to store the transaction data, but it is distributed to users' computers and managed collectively by a peer-to-peer network. Once the blockchain's block is created, it cannot be changed retroactively without a consensus of investor majority and change in the blocks after in blockchain (Lamport, et al., 1982). Figure 2 presents an example of blockchain.

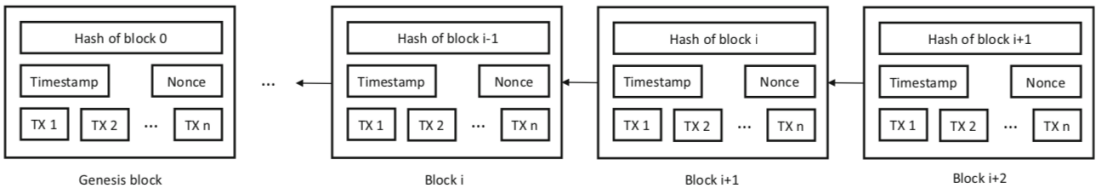


Figure 2. *An example of a blockchain. Edited from Zheng et al. (2018)*

Anonymity, auditability, decentralization, and persistency are blockchain's key characters according to (Zheng, et al., 2018). *Anonymity* occurs when users interact with the blockchain with generated addresses that a user could generate as many as they need to avoid identity exposure. Since in blockchain there is not any central authorization ensuring users' privacy, users themselves need to create the necessary level of privacy. This mechanism ensures privacy to a certain point on the transactions in the blockchain, but there is no way to achieve perfect privacy. *Auditability* is a feature that ensures that users can trace and verify the previous records by just accessing any transaction in the distributed ledger. This is possible because each transaction on the blockchain is validated and recorded with a timestamp and each transaction is connected to the previous one. The *decentralization* of the blockchain shows as the transactions in the blockchain don't need any central authority's authentication or verification. Compared to traditional centralized transaction system decentralized blockchain can significantly reduce operational costs and mitigate the performance bottlenecks because transactions do not need to authenticate by authority. *Persistency* is the character ensuring the reliability and durability of the system. Since each transaction is confirmed and recorded in blocks of the blockchain and the network is stored in every users' computer, it is nearly impossible to alter the ledger. Besides, as every transaction is validated by the following nodes, each falsification could be detected easily,

Blockchain technology is attracting massive attention and triggers for multiple applications and attributes for different industries, for example for banking and financial services markets. Other fields, such as real estate, app development, gaming, data storing, digital identity, are interested to develop applications on top of the blockchain. However, the most known application for blockchain is bitcoin and other cryptocurrencies, which blockchain was developed for. (Treleaven, et al., 2017; Nower, et al., 2017)

2.3 Transaction verification and mining

Bitcoin transactions are verified through the mining process, which is maintained by the bitcoin network community. The transaction is a digital recipe from bitcoin transfer, and it is validated by miners. Miners get rewarded with newly created bitcoins as an incentive to validate the transaction ledger, to store it and contribute to the processing power to the system.

Transactions and verification

A bitcoin transaction is bitcoin value transfer from seller(s) to buyer(s). Transaction is a receipt announced to the bitcoin network about a transfer and then collected into blocks and stored into the blockchain. Transaction (see Figure 3) contains information about

- a bitcoin value (number of bitcoin)
- the address of the owner of bitcoin
- a link to the previous transaction where the bitcoins were gotten
- and a timestamp that tells when the transaction was confirmed (Chuen, 2015)


```
{"hash": "a0caeb61b042f1b1ff90e6e1f729ab1a2edce813e843620819464d9640eeb261", "mined":
"2012-12-17 21:18:50", "locktime": 0, "block_height": 212576, "block_hash":
"00000000000000417607783e5329869f158477a6c692facc9b8ae40c8fcbc2032", "inputs":
[{"transaction_hash": "591cc388dae8979cd66a8226913077e348a8e06b1de7ed15c645867d4d32db2e",
"index": 0, "address": "17aL3YrqYo63dnQWw1RohU8x6rK5AEHV9D", "value": 4632635476}],
"outputs": [{"address": "12WwyhCjptK9HPZ7xteuwtyB4VGyoM47GK", "value": 220000, "index": 0,
"type": "pubkeyhash"}, {"address": "1JiJ7md9gKZpif8qWaAsRK4ZeZNhKgRubs", "value":
4632365476, "index": 1, "type": "pubkeyhash"}]}
```


Figure 3. *An example of bitcoin transaction stored in the blockchain.*

Bitcoin transaction consists of one or more *inputs* and one or more *outputs* (McCorry, et al., 2017). Each in- and output can be understood as bank account and balance: in one transaction it is possible to transfer money (bitcoins) from one or many bank accounts (addresses) to one or many bank accounts and in that transaction, money decreases in the source account (input) and increases in the target account (output). Though transaction can have multiple input accounts, they all belong to the same user (see chapter 4.1). However, output addresses can belong to many users, and in one transaction there is a possibility to send bitcoins to many receivers. (Meiklejohn, et al., 2013)

Transaction does not take place immediately but must be verified by the community first. Transactions are packed into block estimated in every 10 minutes, but the time can increase up to 24 hours. The verification is part of the mining process and takes place when the next blockchain block is created.

Mining

Bitcoin mining is from a bitcoin system point of view a process of adding and verifying transaction records to bitcoin blockchain i.e. bitcoin ledger. From the users' point of view, it is a process invented to give users an incentive to contribute transaction verifications through solving complex computational math problems. (Vilim, et al., 2016)Transaction verifying is crucial for keep-

ing the bitcoin blockchain consistent, complete, and unalterable. Miner gets a reward from mining is dependent on the amount of power user uses and luck, because only the first miner solving the problem gets the reward (Thum, 2018, p. 19).

The primary of mining is to set the history of transactions in a way that is complete, equivalent, and impossible to alter. Miners do the work of verifying previous bitcoin transactions. Every ten minutes all of them in that time announced transactions get grouped into one block. To get that new block accepted by the rest of the network (which is in distributed ledger system the authority), a newly created block must contain a *proof of work*, which requires miners to find a specific number that will be the block's hash. When miners have solved the mathematical problem and confirmed the new block, it will be linked to the previous block and the existing blockchain. The intention of mining from a system point of view is to "create a digital wax seal" (Hopkins, 2018) and make sure that no individual can hack the blockchain.

For a user to benefit from bitcoin mining three things have to occur: firstly, they have to verify 1MB worth of transactions, secondly, they have to solve a complex mathematical problem, *proof of work*, and thirdly, to be the first to get an answer. The first part is easy: to verify one megabyte of transactions require work with several thousand transactions.

To meet the second requirement the miner must solve a mathematical equation, *proof of work,* and get a 64-digit hexadecimal number as a result. This number is called a *hash* that will be used to identify the created bitcoin transaction block. Often miners create mining pools where miners combine their computational power to get ahead into "computing competition" and then split the possible reward (Krishnan, et al., 2015).

Bitcoin is designed with a hard limit of 21 million bitcoins. In October 2019 about 18 million (86 %) of bitcoins were mined (Blockchain.com, 2019). The bitcoin system is built to ensure miners interest over time and to maintain a growth in new bitcoins. For this the difficulty of solving hash increases and the reward decreases. The reward is halved every 210 000 blocks, which take about 4 years (Thum, 2018, pp. 43-45). Tindell (2013) compares the increasingly difficult bitcoin mining to search for prime numbers: while it is easy to find the small ones, it becomes increasingly more difficult the find the larger ones.

2.4 Bitcoin phenomenon and critique

Bitcoin has been a global phenomenon almost since it was created in 2009, but it captured significant investor interest and media attention in 2013 when the value reached $1 000 for the first time and shortly after decreased to around $300 (Marr, 2017). Bitcoin has got exposure from being an easy investment and speculation property for the masses and “ordinary people”.

Bitcoin’s value has been extremely volatile, and it has been claimed to be a bubble (Bovet, et al., 2018). In addition to this bitcoin has been subject to criticism for example for its high electricity consumption, its use in illegal transactions, and thefts from exchanges. Several regulatory agencies have issued investor alerts about bitcoin (Commission, 2014) nonetheless it has been used as an investment.

Bitcoin has been criticized for its technical challenges. For example, the up to 10-minute verification time is considered too long. Also, its close connection to illegal activity and its illegal nature than enables anonymous transactions for example in drug business have been criticized. On the other hand, the fact that bitcoin is not as anonymized as it was meant to be, has also been under critique. Lastly, bitcoin’s energy consumption is questioned for efficiency.

3. BITCOIN USER IDENTIFICATION

Bitcoin and other cryptocurrencies base on the idea of anonymity, despite this, there are ways to identify sets of the public addresses users use. This identification allows us to analyze bitcoin transactions on a higher level and find patterns from wallet use.

3.1 Ownership

Bitcoin ownership is a multi-layer system (presented in Figure 4) where a single investor can have one or more wallets, which are proportional to actual real-life wallets. Each wallet has multiple addresses that are used to transact and store bitcoins.

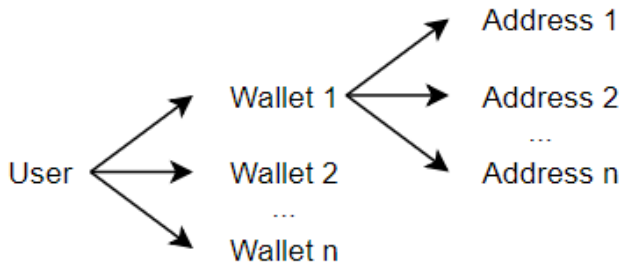


Figure 4. *Bitcoin framework from user to address.*

Addresses

A bitcoin address is an identifier that ensures the functionality of bitcoin transactions and is used to identify the bitcoin seller and the buyer or receiver. An address contains 26 to 35 numbers and letters and can be generated by the bitcoin users. One user can have an unlimited amount of addresses to trade with.

Bitcoin addresses is a private and public key pair. A public key is by name public and a hashed version of the bitcoin address. A private key is used by the owner and it allows bitcoins to be spent. A private key works like a password. The key has a mathematical relation to the bitcoin address, but impossible to obtain because of strong encryption.

Wallets

A bitcoin wallet is a storage for the needed bitcoin transaction information. It “stores the digital credentials for bitcoin holdings” and allows the user access and spend bitcoin (Villasenor, 2014). Wallets have a strong relation to bank accounts, physical, and web wallets. A person could own multiple bank accounts for different intentions: one is for daily usage, one is for saving, one is

for rent, and one is for long-term investing. Having multiple bitcoin wallets works similarly (Volety, et al., 2019). The number of bitcoin wallets is not limited, but practical matters set limitations. Bitcoin wallets come in multiple forms: the four main types are desktop, mobile, web, and hardware.

Volety et al. (2019) describes three main differences between physical and bitcoin wallet: control, accessibility, and receive only. Control: not like a physical wallet that can be used or possessed by one person at the time, a bitcoin wallet can be copied, and its ownership is easily transferred from one to another by stealing it. As physical wallet's stealing needs physical contact with the previous owner, bitcoin wallet fraud can take place by copying the wallet, which transfers the ownership to the copier. Accessibility: as the physical wallet can be only used in one place by one person at the time i.e. one wallet can exist, bitcoin wallets can have multiple copies and it can be accessed from multiple devices. Receive only: while it is always possible to spend money from a wallet that has money and what is under the user's possession a bitcoin wallet can be "receive only". That means the funds can only be received and not spent.

The necessary information consists of public and private key pairs for every bitcoin address stored in that wallet. Simply put, a wallet is a collection of these keys. Wallets facilitate sending and receiving bitcoins and give ownership of bitcoins to the user of the wallet. The public key can be thought of as an account number and private key as an online banking password.

User

A bitcoin user could own multiple wallets and further multiple addresses to trade bitcoins with. As bitcoin-system bases on anonymity, trading and owning bitcoins do not require any kind of identification. Connecting the wallets into users is not in the scope of this thesis, although that information could be interesting to analyze. It must also be considered that it is not possible to connect bitcoin transactions to real-life investors or organizations, only into "anonymous users".

3.2 Pseudonymity

Bitcoin bases on the anonymity of users defined as pseudonymity. The anonymity is achieved by absence of a third party's confirmation in a transaction, and users cannot be directly linked to real users or an identity. A transaction does not identify the user in any way. (Bovet, et al., 2018; Meiklejohn, et al., 2013). Bitcoin identities are thus pseudoanonymous: all transactions are completely transparent. All transactions in the blockchain ledger are public, but the users are not identified. Regardless it is possible to trace and cluster the public addresses owned by the same investor wallet with exploiting the protocol features, like the fact that the transaction history is

publicly available. (Bovet, et al., 2018) Two wallet identification heuristics are presented in Chapter 4.1.

3.3 Bitcoin features used in this thesis

The main bitcoin feature for this thesis is the public availability of bitcoin transactions. The availability is necessary for three reasons. Firstly, it provides an easy source data to analysis while the ledger is freely downloadable from the Internet. Secondly, publicness is necessary for validating and preventing double-spending by bitcoin users. Thirdly, while the transaction data is public and therefore validated with the users, we can be sure the results from the data are real and not just mistakes made in the data mining process. The publicness of the transactions is a feature that only a few currencies can have. It would be interesting to follow transactions made with euro currency, but it would be impossible for political, legislative, and practical reasons. Bitcoin is not a regulated currency and for that reason this analysis is possible.

Secondly, for this thesis is crucial that the recorded bitcoin transaction has multiple input and output addresses. The heuristic we use takes advantage of the bitcoin attribute that all input addresses belong to the same wallet. With this information, we can run through the transaction history and create wallet – address connections. With these connections, it is possible to create networks and to start analyzing them. The third feature related to the previous one is the re-use and co-use of public keys, which ensures that as time passes the addresses are used in multiple transactions with different address combinations, our wallet – address connections expand and get stronger.

4. METHODS

In this thesis, we use multiple methods to process and analyze the open transaction data. The wallet identification heuristic method is specialized for analyzing and combining bitcoin transactions and wallets. The methods presented in 4.2, 4.3, and 4.4 follow the cluster identification process introduced in Tumminello et al. (2011) paper investigating clusters and trading activity of Nokia stocks from 1998 to 2003. The Jaccard Index calculates the similarity and diversity of wallet activity between trading months.

4.1 Wallet identification heuristics

The base of this thesis' analysis is identification heuristics or principles, which are methods to combine fundamentally anonymous bitcoin transactions and find connections to analyze. The bitcoin transaction does not identify either the payee or the payer, as mentioned. However, it is possible to cluster and trace addresses shown in the transaction data that are controlled by the same user via the same wallet. This could be done by exploiting the open availability of all transactions and some features the bitcoin methodology uses.

Bitcoin transaction (presented in Chapter 2.3) consists of two sides: input and output and on both sides are enumerated all addresses participated in the transaction either sell or receive side. The identification approach is that all sell-side addresses belong to one wallet (heuristic I) and receive side addresses can belong to the same sell-side wallet but also the completely new one, but the same address belongs to the same wallet (and user). The approach introduced by Meiklejohn et al. (2013) describes two heuristics, but we only use in this thesis the first one.

The identification method is incomplete, and all addresses could not be connected to the wallets, because firstly, it cannot be sure if two addresses in the receive-side belong to the same or different wallet. In this case, we assumed that all addresses in receive -side belong to different wallets. This assumption might cause situations, where one wallet is split into two or more wallets and some connections cannot be found. Secondly, some wallets are only found in receive-side, not as an input, and therefore not connected to wallets.

Meiklejohn et al. (2013) introduce two heuristics to identify bitcoin users in their paper as follows:

Heuristic I (input-based): If two (or more) addresses are inputs to the same transaction, they are controlled by the same user; i.e. for any transaction t, all public keys *pk* ∈ inputs (*t*) are controlled by the same user.

Heuristic II (one-time change addresses): The one-time change address is controlled by the same user as the input addresses; i.e., for any transaction t, the controller of inputs (t) also controls the one-time change address *pk* ∈ outputs(*t*) (if such an address exists). (Meiklejohn, et al., 2013)

To put it shortly, according to the first heuristic, two (or more) public keys are controlled by the same user if they are used as an input to the same transaction. (Meiklejohn, et al., 2013) The second heuristic detects addresses in the output of a transaction that can be added to the input wallets (Bovet, et al., 2018).

4.2 Categorical variables characterizing the transaction activity

One of the problems of high heterogeneity of bitcoin investors is that it might be difficult to compare different investor wallets with each other. This occurs when an investor could be a single person trading once a year or an automated trading bot trading every day in large volumes. Since this thesis' interest is in comparing the transaction position taken by an investor on a given hour, not the absolute traded volume, we use categorical variables that describe wallet's transaction activity.

More specifically, we define for each bitcoin investor *i* and each time period *t*, a sell- and purchase volumes $V_s(i, t)$ and $V_b(i, t)$. The time period is one hour in this thesis. This information is converted into a categorical variable with three states: primarily buying *b*, primarily selling *s*, and both buying and selling *bs*. We do the conversion with the Equation 1:

$$r(i,t) = \frac{V_b(i,t) - V_s(i,t)}{V_b(i,t) + V_s(i,t)} \quad (1)$$

We say an investor is on a primarily buying state *b* when $r(i, t) > \theta$, in a primarily selling state *s* when $r(i, t) < -\theta$, and in a buying and selling state, *bs* when $-\theta < r(i, t) \theta$ with $V_b(i, t) > 0$ and $V_s(i, t) > 0$. In this thesis the threshold θ is set to 0.05.

With this categorization it is possible to map the data in the network, where the links between nodes (wallets) follow the *b, s,* and *sb* categorization. A link exists if a pair of wallets is active in the same time slot and statistically validated to exist. That process is defined in the following chapters.

4.3 Statistically validated networks of investors

In this thesis, we use a statistical validation method to validate a network of investors and to identify co-occurrences in trading actions. A co-occurrence is determined in the presence of a particular type of link between two nodes (wallets) in a network.

With this method, we can statistically validate with the p-value the co-occurrence of states defined in the previous chapter. We validate the co-occurrence of states P (investor *i* in state b, s, or bs) and Q (investor *j* in state b, s, or bs). First, we split the data into month review periods and then identified the investor activity within that period. Then for each investor pair *i* and *j* we started to focus on the intersection of the equivalent activity periods in the reviewed month. The activity period is one hour. The parameters needed in the validation are defined in Table 1:

Table 1. *Description of the variables of network validation.*

Variable	Description
T	Number of possible activity periods ie. number of hours in an examined month
N^P	The number of hours when investor *i* was observed in state P
N^Q	The number of hours when investor *j* was observed in state Q
N^{PQ}	The number of hours when investor *i* was observed in state P and *j* was observed in state Q at the same time (the co-occurrence)

We can describe with hypergeometric distribution, H(X|T, NP, NQ), the probability of observing x co-occurrences of the investigating states P and Q of the two investors in T observations. With the distribution, we can calculate a p-value for each combination of states for each pair of investors. Specifically, for each kind of co-occurrence of states P and Q we can define the p-value as follows:

$$p(N_{P,Q}) = 1 - \sum_{X=0}^{N_{P,Q}-1} H(X|\mathrm{T}, N_P, N_Q). \quad (2)$$

The nine possible combinations of the three states between investors *i* and *j* are (b, j), (b, s), (b, bs), (s, b), (s, s), (s, bs), (bs, b), (bs, s) and (bs, bs) where the first refers to an investor *i*'s state and second to *j*'s state. In this thesis, we only took (s, s), (b, b), and (s, b) under consideration.

To solve the multiple test problem in this thesis we use the Bonferroni correction. The method is the simplest way to adjust p-values generated with hypergeometric test and it multiplies the determined alpha by the number of tests performed:

$$p_B = \frac{\alpha}{N_{tests}} \quad (3)$$

In the thesis' case the number of tests performed is the estimation on different co-occurrence networks for one month.

4.4 The Jaccard Index

The Jaccard index (also known as an intersection over union and the Jaccard similarity coefficient) is a statistical similarity method for comparing and analyzing the similarity or diversity of two sets (Li, et al., 2018). More specifically, it is a measure of the relative size of the overlap of two finite sets A and B. (Kosub, 2016). It is determined as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}, \quad (4)$$

where $0 \leq J(A,B) \leq 1$.

In this case, the Jaccard index is calculated to estimate the continuity of wallet activity between months. Jaccard Index analysis is not a direct part of the network analysis but gives additional information about trading activity what is the similarity of monthly sets of active wallets.

5. DATA

This thesis analyses bitcoin transaction data found from its open access blockchain. The data is used to create connections between transactions and the user wallets behind transactions. The data consists of all transactions on the 11-month period starting from July 2017 and ending May 2018. There are 335 days in the timespan. The transaction data consists of 91 million (91 348 126) transactions, 145 million (145 447 675) addresses participating in the transactions and 80 million (80 417 205) wallets identified from addresses. During the time period estimated 73 million (73 147 196) bitcoins were transacted. Bitcoin transaction volume and price distribution are visualized in Figure 5.

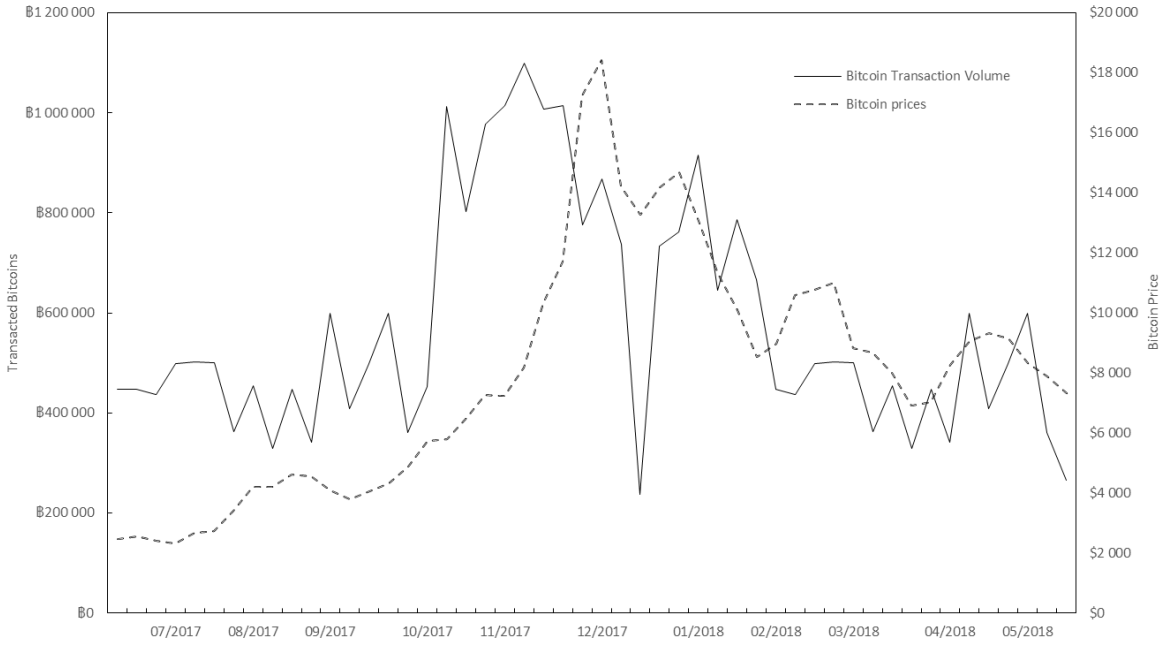


***Figure 5**. Bitcoin price and bitcoin transaction volume over time*

Bitcoin price has been very volatile in the analyzed period as well as the transaction volume has had its ups and downs from 200 000 bitcoins to more than a million bitcoins in the one-week-period. It is visible in the picture that the price's changes follow the transaction volume changes a little behind and the major changes have leveled off. The highest peak in transactions have been on October – November 2017 and the price peak has followed in late November – December 2017.

5.1 Original data

Transactions (presented in Chapter 2.3) include sell- and receive-side addresses, timestamp, and address-based transacted bitcoin amounts. One transaction could have one to as many as needed addresses as input and output addresses. In this thesis, we identify wallets from addresses. One step forward would be identifying users from wallets and that is one area for further analysis.

Table 2. *Description of the variables included in the transaction*

Variable	Description
Hash	Identifies a single transaction
Input address	Identifies a seller or one of them, all belong to one wallet
Output address	Identifies a bitcoin receiver or one of them
Bitcoin amount	Number of bitcoins transacted (sold or bought) from/to one address
Time	Time the transaction was confirmed (delay 1 min to 1 day)

About 80 million wallets were found from the dataset. One wallet participates on average in 2.4 transactions during the observed time period. The maximum amount of transaction observations is 58 244 and the median is 2.

The first part of the analysis was to create address – wallet – connections with heuristics presented in Chapter 4.1. This was done by going through all transactions in the bitcoin ledger. All addressed in input-or sell-side were connected to one wallet and addresses in output or receive side were connected either existing wallet if one was found or to the newly created one. As a result, we got 80 million wallets.

Secondly, all wallets participating in one transaction were identified and listed as wallet id – transaction time – transaction way (sell or receive) – bitcoin amount one wallet had. With this data, we calculated how many transactions one wallet was participating in.

5.2 Categorical variables

Categorical variables describe the wallets' activity on the market in the chosen hourly slot. To determine the variables for wallets, firstly, all wallets were resampled to hourly slots to which the volume was calculated with Equation (1). In the resampling, all the wallet's transactions within one hour were combined, and for that combination we calculated a volume value from the bitcoin values. The categorical variables deduced from volumes show us whether the wallet

was more selling or buying bitcoins at one-time point. The volume values are b (primary buying), s (primary selling), and bs (selling and buying almost in the same quantities). That resampling gave us material to start generating co-occurrence pairs and the co-occurrence networks.

With the hourly resampled data, we sorted out all wallets active at least in one hour in every observed 11 months. The wallets left are interesting because with these the investing is continuous, and it is possible to do time series analysis with the same data over time. After the delineation, there were 16 159 wallets each month.

5.3 P-value for co-occurrence pairs

P-value is a random probability distribution variable from statistic test used to test the null hypothesis (Hung, et al., 1997). In this thesis, we used p-value to estimate the strength of connection between co-occurrence pairs and see if the link is real or just statistical coincidence.

To obtain the information of link strength we statistically validated the co-occurrence of state P of investor *i* and state Q (b or s) of investor *j* with the methodology presented in Chapter 4.3. Shortly, we calculated how often investor I was in state P and how often investor j in state Q and what is the p-value of that co-occurrence. In this p-value estimation we took only wallets that were active in all observed 11 months.

The p-value threshold was determined with the Bonferroni correction (Equation 3). The alpha was set to 0.01. The Bonferroni p-values for different months are listed in Table 3.

Table 3. *Bonferroni test p-value thresholds.*

Month	Bonferroni threshold
07/2017	0.00000004197
08/2017	0.00000004461
09/2017	0.00000004482
10/2017	0.00000004217
11/2017	0.00000004350
12/2017	0.00000004482
01/2018	0.00000005460
02/2018	0.00000005810
03/2018	0.00000005573
04/2018	0.00000005844
05/2018	0.00000011089

We created tables for each month where all statistically validated connections were listed. We call these tables *statistically validated co-occurrence networks*. In Table 4 is listed the variables we found interesting in the network analysis.

Table 4. *Description of the variables in network analysis*

Variable	Description
Degree distribution	The probability distribution of network's degrees
Path length	The number of links a path contains
Avg. clustering coefficient	The overall level of clustering in a network
Density	Actual connections/potential connections, %
Connected components	Number of network components connected

We performed for statistically validated co-occurrence networks a sequential degree correlation analysis, where we took degree sequences and from network A and B and calculated a correlation with Pearson's sample correlation coefficient:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n s'_x s'_y}, \quad (5)$$

where $\bar{x}$ and $\bar{y}$ are sample means and s'_x and s'_y are sample standard deviations.

6. RESULTS

During the analyzed time period the number of identified investor wallets varied from 2.3 million to 6.7 million, with an average of 4 million wallets in a month. The change in wallet numbers is presented in Table 5. The number of active wallets does not give a clear vision of how many actual investors there are, but it gives an insight of the size and the relative change in time.

Table 5. *Number of wallets in observation activity group based on how many transactions a wallet is participating monthly.*

Time period	Over 1	Over 10	Over 50	Over 100	Over 500	Over 1 000
07/2017	4 016 778	74 114	8 062	3 140	646	398
08/2017	4 346 990	76 979	8 236	2 934	620	376
09/2017	4 145 879	78 224	7 890	2 839	653	378
10/2017	4 793 515	96 401	10 089	3 378	688	399
11/2017	5 130 783	101 462	9 142	2 995	658	389
12/2017	6 655 445	99 517	7 236	2 572	616	382
01/2018	5 215 876	70 206	5 494	2 089	600	385
02/2018	2 854 681	51 860	4 436	1 781	515	322
03/2018	2 355 155	57 382	4 948	2 038	537	342
04/2018	2 324 999	52 228	5 282	2 291	540	328
05/2018	2 375 674	57 680	5 356	2 282	574	336

The table shows how the most of the investor wallets only participated in one transaction monthly, and around 98 % of wallets less than 10 transactions. This confirms the hypothesis related to the network structure, where we assumed most of investor wallets are used for couple of transactions in a month.

The selection based on the monthly number of transactions is not suitable for the actual network analysis, because that does not tell whether the wallet has been transacting all the transactions in one day or one transaction per hour. The latter seems to refer to a trading bot. Regardless, it gives a useful insight for the overall structure of wallets.

For the rest of the analysis we made a selection of wallets that have been active for at least in one hour in a month. The hourly resampling is explained in Chapter 4.2. In the transaction dataset, there were 16 159 wallets identified having transaction activity at least once a month. As

we performed a hypergeometric test to the wallets, the number of wallets decreased as the links wallets have were not statistically strong enough.

6.1 Jaccard Index

We use the Jaccard Index for presenting the wallets' similarity over time and it gives a picture of the investors who invest more than 10 or 1 000 transactions monthly. In Table 6 is presented the Jaccard Index in an activity group of more than 10 observations in a month. The similarity between two sequential months is between 23 % and 32 % meaning that about 25 % of the wallets investing in month 1 is also investing in the next month. When the time span expanded into one month, the similarity decreases to 12-21 %. With the longest time span (9 months between) the Jaccard Index is 3 %. It is visible in the Figure 6 that the Jaccard Index over time follows the $\frac{1}{x}$-function.

Table 6. *Jaccard Index investor wallet similarity between two time periods with wallets having more than 10 observations in a month.*

Time period	07/17	08/17	09/17	10/17	11/17	12/17	01/18	02/18	03/18	04/18	05/18
07/2017		23 %	15 %	11 %	9 %	7 %	4 %	4 %	3 %	3 %	3 %
08/2017	23 %		26 %	17 %	13 %	9 %	6 %	4 %	4 %	4 %	3 %
09/2017	15 %	26 %		29 %	19 %	13 %	7 %	5 %	5 %	4 %	4 %
10/2017	11 %	17 %	29 %		28 %	16 %	8 %	6 %	5 %	5 %	4 %
11/2017	9 %	13 %	19 %	28 %		29 %	13 %	8 %	7 %	6 %	5 %
12/2017	7 %	9 %	13 %	16 %	29 %		24 %	12 %	10 %	7 %	7 %
01/2018	4 %	6 %	7 %	8 %	13 %	24 %		25 %	18 %	12 %	10 %
02/2018	4 %	4 %	5 %	6 %	8 %	12 %	25 %		32 %	19 %	15 %
03/2018	3 %	4 %	5 %	5 %	7 %	10 %	18 %	32 %		29 %	21 %
04/2018	3 %	4 %	4 %	5 %	6 %	7 %	12 %	19 %	29 %		32 %
05/2018	3 %	3 %	4 %	4 %	5 %	7 %	10 %	15 %	21 %	32 %	

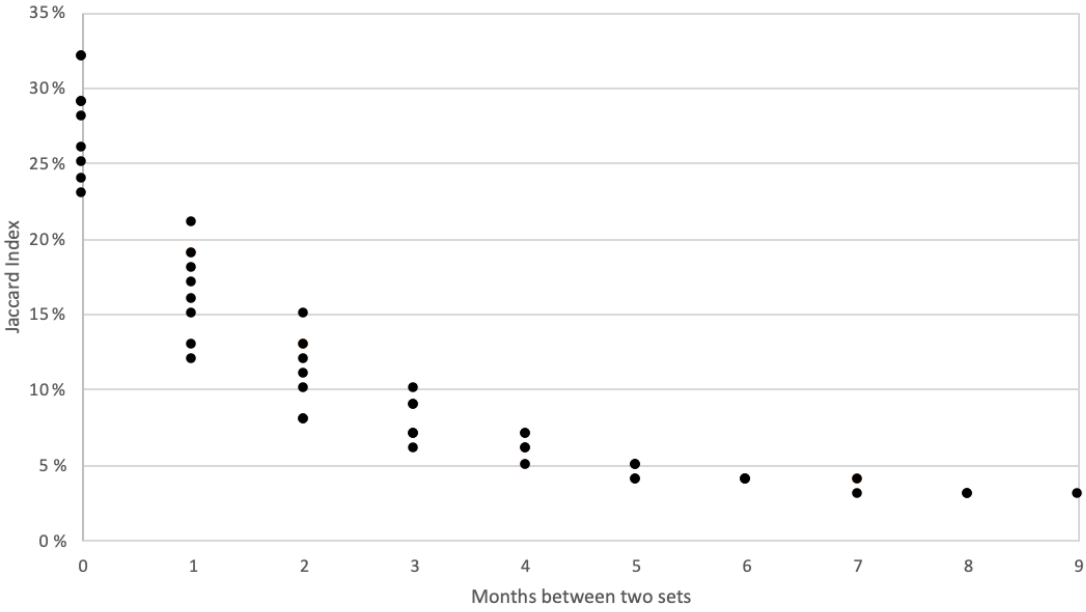


Figure 6. *Jaccard Index on a function of time distance between two monthly wallets sets where a wallet has been active in more than 10 observations in a month.*

When expanding the threshold to 1 000 observations, the similarity percent in sequential months increases average to 67 % with a scale from 59 % to 76 %. With the longest time span (9 months between) the similarity is 20 % meaning that 80 percent of investor wallets have changed from the wallets investing in 9 months ago. The Jaccard Index percent are presented in Table 7. Also, in this activity group, the Jaccard Index over time follows the $\frac{1}{x}$-function (see Figure 7).

Table 7. *Jaccard Index wallet similarity between two time periods with wallets having more than 1 000 observations in a month.*

Time period	07/17	08/17	09/17	10/17	11/17	12/17	01/18	02/18	03/18	04/18	05/18
07/2017		59 %	50 %	41 %	34 %	31 %	26 %	24 %	22 %	21 %	20 %
08/2017	59 %		64 %	50 %	42 %	36 %	30 %	28 %	26 %	25 %	23 %
09/2017	50 %	64 %		64 %	48 %	41 %	33 %	31 %	29 %	25 %	23 %
10/2017	41 %	50 %	64 %		63 %	49 %	38 %	35 %	31 %	29 %	25 %
11/2017	34 %	42 %	48 %	63 %		67 %	49 %	42 %	37 %	33 %	31 %
12/2017	31 %	36 %	41 %	49 %	67 %		63 %	49 %	43 %	37 %	35 %
01/2018	26 %	30 %	33 %	38 %	49 %	63 %		63 %	53 %	47 %	42 %
02/2018	24 %	28 %	31 %	35 %	42 %	49 %	63 %		74 %	63 %	55 %
03/2018	22 %	26 %	29 %	31 %	37 %	43 %	53 %	74 %		73 %	60 %
04/2018	21 %	25 %	25 %	29 %	33 %	37 %	47 %	63 %	73 %		76 %
05/2018	20 %	23 %	23 %	26 %	31 %	35 %	42 %	55 %	60 %	76 %	

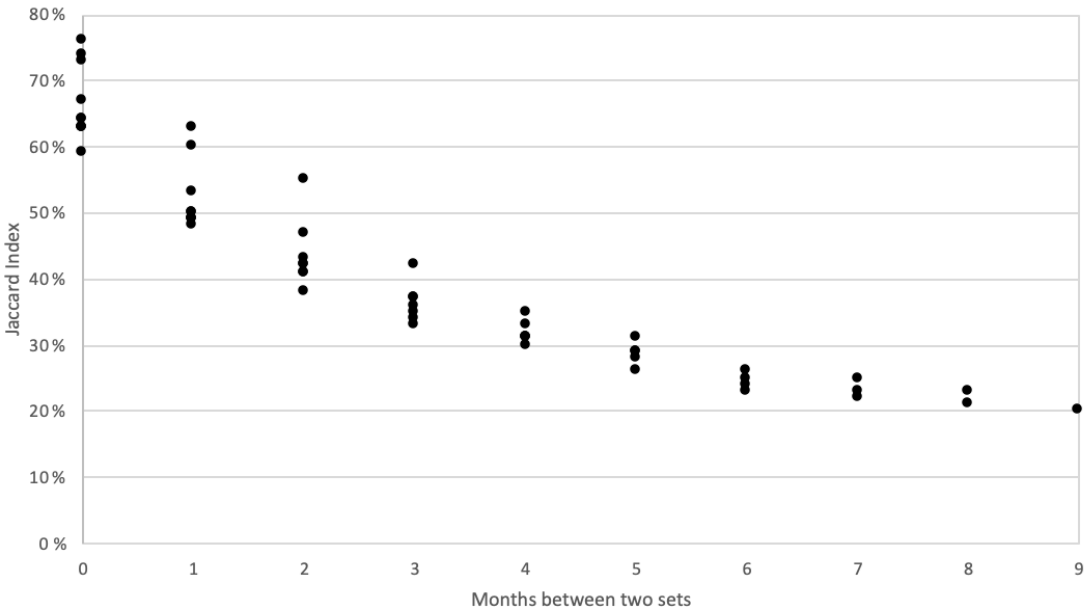


Figure 7. *Jaccard Index on a function of time distance between two monthly wallet sets where a wallet has been active in more than 1 000 observations in a month.*

The results show there is continuity with wallets investing in both 10 and 1 000 times a month. The wallet population in the over 1 000 group is smaller and the similarity is higher meaning the high frequency investor wallets are used in investing in the same way over time more than the over 10 group. It must be taken into account that the over 1 000 group is a subgroup of the over

10 transactions. The scatter plots of Jaccard Index for both activity groups show that the Jaccard Index similarity decrease the same formula over the time distance between months.

Jaccard similarity between two synchronization network links (ss and bb networks) shows us the links in sell- and receive-networks do not have a close relation. This is explained by the significant difference in sizes of the bb and ss networks. The aggregated bb network has 1 166 539 links and the ss only 251. We call aggregated network a network where all the wallets and the links between have resampled into one network. In other words, in that network we have all the wallets having transactions during the analyzed period and the links between these wallets with the activity limitation. The activity limitation refers to that in this thesis we have taken into analysis the 5 000 most active wallets based on the monthly hours they were having transactions, and the links the active wallets have. However, the Jaccard percentage (in Table 8) show the stronger similarity between February, March, April, and May 2018 links, which infers that the investors have started to sell their bitcoins after the price sunk in December - January (see Figure 5).

Table 8: *Jaccard similarity between ss and bb synchronization network links. Ss time periods are on columns, bb time periods on rows for the 5 000 most active wallets based on monthly activity. All wallets have at least one active hour (see Chapter 4.2) in every month.*

Time period	07/17	08/17	09/17	10/17	11/17	12/17	01/18	02/18	03/18	04/18	05/18
07/2017	0.007 %	0.009 %	0.008 %	0.008 %	0.013 %	0.010 %	0.007 %	0.006 %	0.005 %	0.006 %	0.008 %
08/2017	0.009 %	0.007 %	0.006 %	0.006 %	0.010 %	0.007 %	0.005 %	0.004 %	0.004 %	0.005 %	0.006 %
09/2017	0.008 %	0.006 %	0.004 %	0.005 %	0.007 %	0.005 %	0.004 %	0.003 %	0.003 %	0.003 %	0.004 %
10/2017	0008 %	0.006 %	0.005 %	0.005 %	0.007 %	0.005 %	0.004 %	0.003 %	0.003 %	0.003 %	0.004 %
11/2017	0.013 %	0.010 %	0.007 %	0.007 %	0.009 %	0.006 %	0.005 %	0.004 %	0.004 %	0.004 %	0.005 %
12/2017	0.010 %	0.007 %	0.005 %	0.005 %	0.006 %	0.010 %	0.007 %	0.006 %	0.006 %	0.007 %	0.008 %
01/2018	0.007 %	0.005 %	0.004 %	0.004 %	0.005 %	0.007 %	0.008 %	0.006 %	0.006 %	0.007 %	0.009 %
02/2018	0.006 %	0.004 %	0.003 %	0.003 %	0.004 %	0.006 %	0.006 %	0.012 %	0.012 %	0.013 %	0.017 %
03/2018	0.005 %	0.004 %	0.003 %	0.003 %	0.004 %	0.006 %	0.006 %	0.012 %	0.016 %	0.018 %	0.023 %
04/2018	0.006 %	0.005 %	0.003 %	0.003 %	0.004 %	0.007 %	0.007 %	0.013 %	0.018 %	0.021 %	0.027 %
05/2018	0.008 %	0.006 %	0.004 %	0.004 %	0.005 %	0.008 %	0.009 %	0.017 %	0.023 %	0.027 %	0.027 %

6.2 Network analysis

The networks consist of investor wallets that have been active in every observed month (nodes), and co-occurrence pairs filtered with Bonferroni threshold (edges). To the networks is selected the top 5 000 most active wallets. In Table 8 and Table 9 are collected the key attributes from the undirected *bb* and *ss* networks, and the directed *sb* network for the observed 11-months

period (July 2017 – May 2018), including network's number of nodes, node degrees, link and node Jaccard similarity, network's average degree, average path length, density, average clustering coefficient, and the number of connected components.

Table 9. *Network attributes for ss and bb synchronization and sb trade relationship tables for the 5 000 most active wallets for July 2017 – May 2018.*

Variable	bb network	ss network	sb network
Number of nodes	3 510	183	1 611
Number of edges	1 166 539	251	2 498
Network density	0.18924	0.0151	0.001926
Number of connected components	18	30	33
Average clustering coefficient	0.803	0.254	0.01698
Average degree	664.70	0.1004	0.9992
Average path length	2.231	4.406	4.339

Number of nodes and edges are variables that tell how many wallets the network consists of, and how many links there are between wallets. Network density is a measure of the proportion of possible connections that are actualized among the members of a network (Giuffre, 2015). Number of connected components tell how many wallet clusters the network has. Average clustering coefficient is a measure of a degree to which nodes in a graph tend to cluster together. Average degree is a statistic of the probability distribution of network's degrees over the whole network. In the undirected synchronization networks (ss, bb) the degree represents the number of connections to the other wallets. In directed network node has two degrees, the in- and out-degree. The average degree for directed network is an average of wallet's in- and out-degrees. Path length is the shortest path between two nodes in the network.

The size of the three networks have a high degree of variation. The bb network is the largest one with 3 510 wallets (nodes) out of 5 000 having links (edges) that are statistically strong enough to pass hypergeometric test. The high number of links in bb network shows that during the analyzed time period the bitcoin receiving have been on a high level. The high number of links in bb network shows that during the analyzed time period the bitcoin receiving activities have been on a high level. It is understandable from the bitcoin hype point of view. The bitcoin price had a peak in December 2017 and attracted a lot of media's attention causing an investment spike. The network densities are proportional for the relative size of the network and are obtained as a function of links and nodes.

The networks have connected components, clusters, a groups of wallets. Having clusters in network mean some of the wallets have a closer relation to the other clusters' wallets and weaker to the wallets outside the cluster. The number of clusters vary from 18 to 33 in this thesis' networks. As the transaction data is anonymous, the clusters cannot be linked to the real clusters there is. However, as cluster is a set of wallets that are more similar to each other than the wallets outside the cluster, we can assume one cluster is centralized around bitcoin brokers or bitcoin trading service provider that the investors use. As a result from a timeline analysis of the network, we detected that the wallets stayed in the same cluster as the months changed.

Clustering coefficient is a measure telling how complete the neighborhood of the node is. The average varies from 0 to 1. Average clustering coefficient for bb network is 0.803. This shows the fact that the level of buying synchronization has been very high during the analyzed period. For ss network the average clustering coefficient is 0.1004, showing the opposite synchronization levels. The assumption with this is that the sudden price sunk has made the investors and other bitcoin users feel the price will come up again and is not worth selling. Again, the bb network has the highest degrees because the size of it. In Figure 8 is visualized the average node degree distributions for each network.

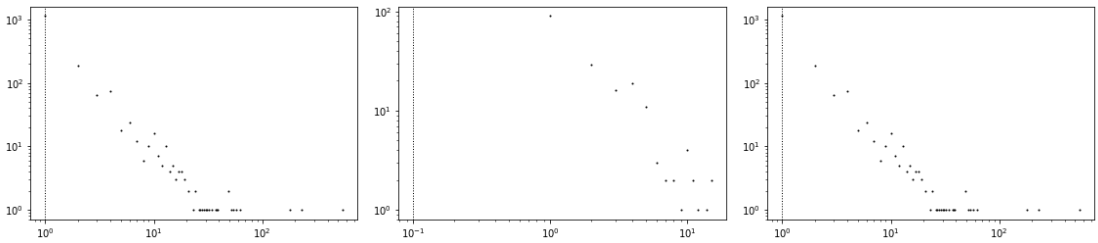
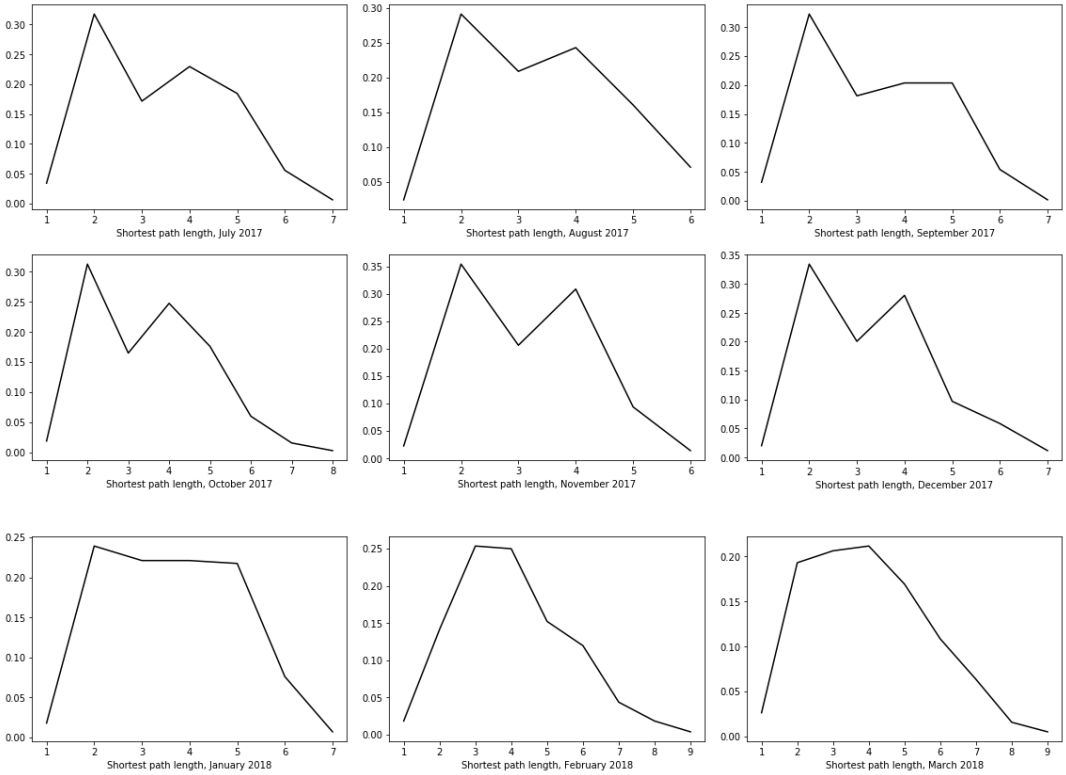


Figure 8. *From left to right bb, ss, and sb synchronization network's average node degree distribution. Dotted vertical line represents average node degree.*

In Figure 9 we have visualized we have visualized the aggregated networks for each co-occurrence pairs. As mentioned, the aggregated networks are combinations of all the monthly networks for co-occurrence pairs for the analyzed period. The pictures visualize the whole aggregated co-occurrence network to more understandable form and gives insight of the clusters. Figure 9 also gives a good insight of the network's density and size.

Figure 9. *From left to right sb, bb, and ss network visualizations for July 2017 – May 2018.*

Shortest path length distributions for each month and network are shown in the Figures 11, 11, and 12. The visualizations show how in ss around 30 % of wallets' distance from others is 2 transactions and the maximum distance is around 6 and 7 transactions. An exception is wallets, which are isolated from others i.e. having their own cluster with no links to other clusters' wallets. In ss network, the main pattern presenting the shortest path distribution stays the same between months.



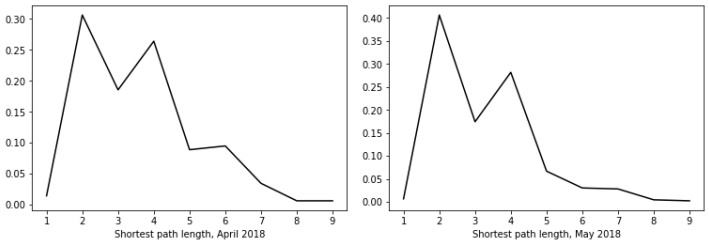


Figure 10. *Shortest path length distributions for ss network for the 5 000 most active wallets.*

In sb network visualizations there are more variation than in the ss network. The highest peak settles to length of 4, meaning the distances between wallets are longer than in bb and ss networks. This can be interpreted as the wallets selling are not so close to each other. This can be explained with that the wallets have just few connections (average degree is around one), and the network is not dense.

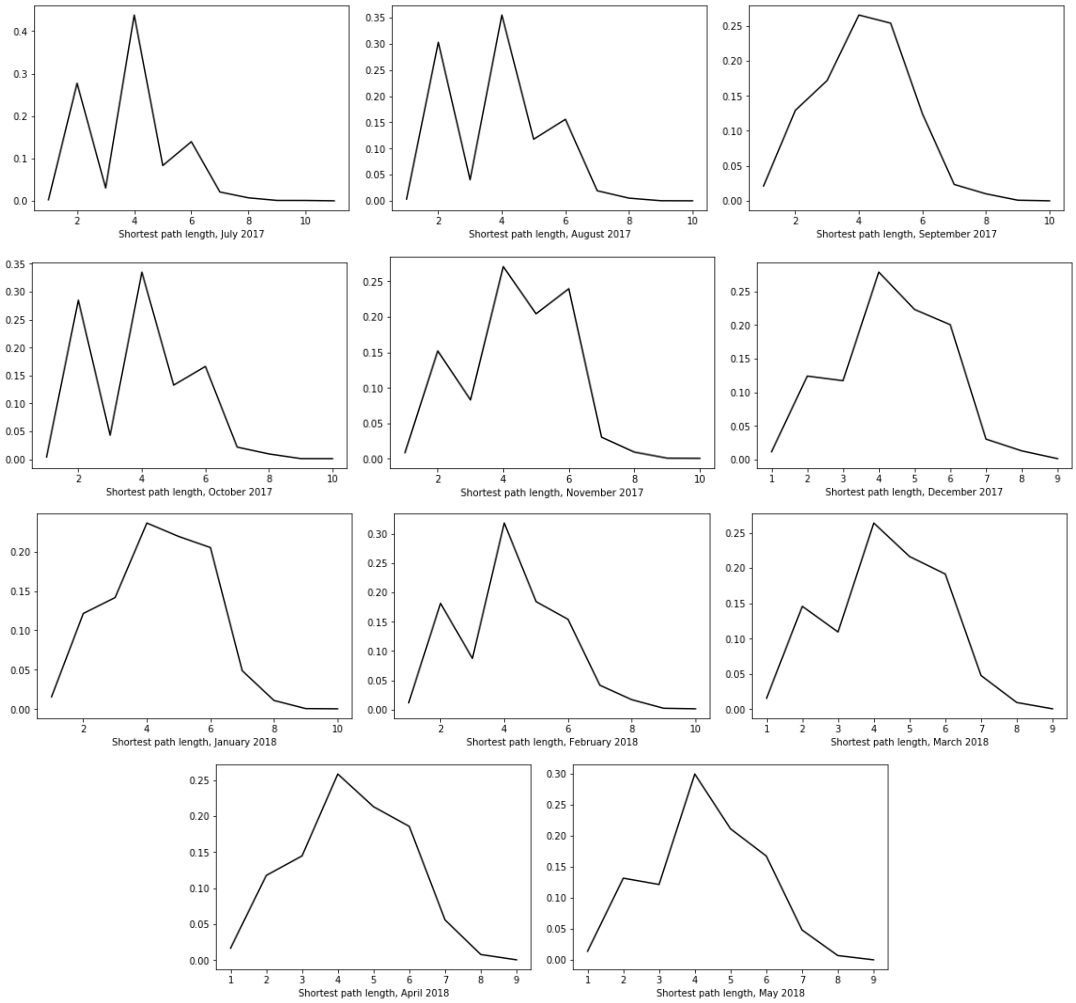


Figure 11. *Shortest path length distributions for sb network for the 5 000 most active wallets.*

The bb network remains the same over the analyzed period. The highest peak's probability in the 4 transaction length is around 50%, meaning every other of the wallets shortest path length is 4 transactions and on the other words between these wallets there are three other wallets. The fact that the pattern stays the same over time reflects the similarity of the network on the time period. It also means, the top 5 000 investor wallets kept investing in similar way, regardless the price changes or which month it was.

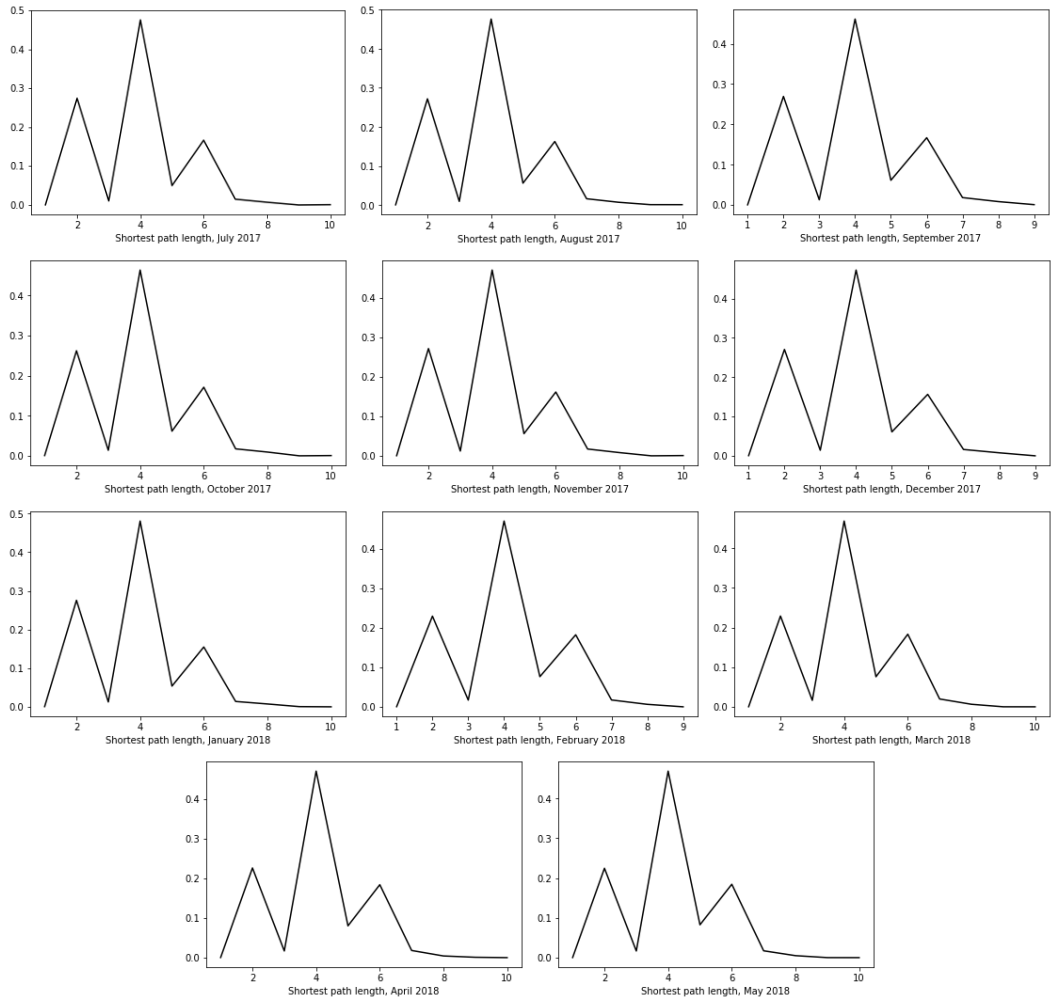


Figure 12. *Shortest path length distributions for ss network for the 5 000 most active wallets.*

6.3 Network degree correlation analysis

Network degree tells about the similarity of the network structures. The correlation is calculated with Pearson's sample correlation (Equation 5). The Pearson's correlation measures the correlation between two sets from -1 to 1. 1 refers full correlation between sets and, 0 full independence from the other set, and -1 inverse correlation. The correlations shown in Table 10 for ss network seem not to be as strong as on the other networks. The strongest correlation is between

March, April, and May 2018, where up to 53 percent of previous month's sell activities explain for the next month's similar activities. As shown in Figure 5, the bitcoin price had a major decrease in December 2017 and right after that a major increase. This has affected the selling activities and the correlations have turned to negative. This means, when the price dropped, the January's selling activities were very different from December's activities, as in December the price could be expected to rise further. Also, in February, when the price started to increase again, the selling activities compared to January were different. Again, in March, when the price started to decrease, the selling activities started to differ from the last month's activities. When we compare December's and February's selling activities, the correlation is 0.54, meaning the investors sold at the time of price increasing quite similarly. This shows how the selling is proportional to the price. It is also visible that the correlation weakens significantly the longer the time between examined months is.

***Table 10**. Ss network's correlations between sequences of wallet degrees.*

Time period	07/17	08/17	09/17	10/17	11/17	12/17	01/18	02/18	03/18	04/18	05/18
07/2017	-	0.58	0.41	0.65	0.31	0.17	0.50	0.00	1.00	-0.54	0.00
08/2017	0.58	-	0.45	0.66	0.40	-0.07	0.09	-0.32	0.00	-0.57	-0.50
09/2017	0.41	0.45	-	0.62	030	0.52	0.16	0.11	0.13	0.76	-0.03
10/2017	0.65	0.66	0.62	-	0.57	0.35	0.27	0.00	-0.58	-0.02	-0.07
11/2017	0.31	0.40	0.30	0.57	-	0.21	0.27	0.35	-0.31	0.21	-0.66
12/2017	0.17	-0.07	0.52	0.35	0.21	-	-0.12	0.54	-0.87	0.87	0.87
01/2018	0.50	0.09	0.16	0.27	0.27	-0.12	-	-0.17	-0.11	-0.33	-1.00
02/2018	0.00	-0.32	0.11	0.00	0.35	0.54	-0.17	-	-0.08	0.02	-0.63
03/2018	1.00	0.00	0.13	-0.58	-0.31	-0.87	-0.11	-0.08	-	0.46	0.27
04/2018	-0.54	-0.57	0.76	-0.02	0.21	0.87	-0.33	0.02	0.46	-	0.53
05/2018	0.00	-0.50	-0.03	-0.07	-0,.66	0.87	-1.00	-0.63	0.27	0.53	-

In the bb network the correlations between two sequential months is up to 77 %, and the correlations are higher than in ss network. In bb network the correlation weakens as well the longer the time between months increase, but not as strongly as in ss network. We cannot find similar price dependents from the bb network as we have in ss network. We can assume, the investors keep buying regardless the price, but they sell only when the price is high and assumed to be rising.

Table 11. *Bb network's correlations between sequences of wallet degrees.*

Time period	07/17	08/17	09/17	10/17	11/17	12/17	01/18	02/18	03/18	04/18	05/18
07/2017	-	0.50	0.63	0.61	0.42	0.44	0.32	0.24	0.17	0.15	0.16
08/2017	0.50	-	0.51	0.44	0.32	0.30	0.16	0.09	-0.01	0.02	0.05
09/2017	0.63	0,51	-	0.77	0.55	0.56	0.51	0.39	0.19	0.18	0.22
10/2017	0.61	0.44	0.77	-	0.56	0.57	0.52	0.37	0.28	0.18	0.22
11/2017	0.42	0.32	0.55	0.56	-	0.52	0.48	0.40	0.35	0.27	0.27
12/2017	0.44	0.30	0.56	0.57	0.52	-	0.57	0.36	0.28	0.18	0.21
01/2018	0.32	0.16	0.51	0.52	0.48	0.57	-	0.55	0.30	0.27	0.26
02/2018	0.24	0.09	0.39	0.37	0.40	0.36	0.55	-	0.54	0.55	0.49
03/2018	0.17	-0.01	0.19	0.28	0.35	0.28	0.30	0.54	-	0.57	0.60
04/2018	0.15	0.02	0.18	0.18	0.27	0.18	0.27	0.55	0.57	-	0.61
05/2018	0.16	0.05	0.22	0.22	0.27	0.21	0.26	0.49	0.60	0.61	-

In sb trade network the links present a connection between a sell and receive side wallets. As the sb network is directed i.e. the link between two nodes is directed from the sell wallet to the receive wallet. It is possible, that the link is either one-way or two-way, meaning both wallets have been at the sell and receive side during the observed month. In Table 12 we have calculated the correlation between wallet's in-degrees, link from sell wallet to receive wallet. In

Table 13 the links are the opposite, links from receive wallet to sell wallet. The correlations are stronger in out-degree than in-degree sequences. That means a wallet is more likely to have a regular sell link than buy link i.e. the investor alienates bitcoins to the same investor wallets regular, but receives bitcoins from a larger group of wallets, though the correlation is in both cases quite high.

Table 12. *Sb network's correlation between sequences of wallet in-degrees.*

Time period	07/17	08/17	09/17	10/17	11/17	12/17	01/18	02/18	03/18	04/18	05/18
07/2017	-	0.44	0.63	0.85	0.38	0.11	0.47	0.60	0.07	0.61	0.69
08/2017	0.44	-	0.75	0.56	0.81	040	0.68	0.75	0.63	0.74	0.60
09/2017	0.63	0.75	-	0.55	0.57	0.12	0.44	0.91	0.51	0.89	0.78
10/2017	0.85	0.56	0.55	-	0.68	0.53	0.77	0.46	0.19	0.61	0.57
11/2017	0.38	0.81	0.57	0.68	-	0.60	0.77	0.63	0.72	0.72	0.46
12/2017	0.11	0.40	0.12	0.53	0.60	-	0.77	0.03	0.27	0.09	-0.05
01/2018	0.47	0.68	0.44	0.77	0.77	0.77	--	0,41	0.57	0.49	0.24
02/2018	0.60	0.75	0.91	0.46	0.63	0.03	0.41	-	0.60	0.83	0.77
03/2018	0.07	0.63	0.51	0.19	0.72	0.27	0.57	0.60	-	0.78	0.65
04/2018	0.61	0.74	0.89	0.61	0.72	0.09	0.49	0.83	0.78	-	0.87
05/2018	0.69	0.60	0.78	0.57	0.46	-0.05	0.24	0.77	0.65	0.87	-

Table 13. *Sb network's correlation between sequences of wallet out-degrees.*

Time period	07/17	08/17	09/17	10/17	11/17	12/17	01/18	02/18	03/18	04/18	05/18
07/2017	-	0.53	0.88	0.92	0.69	0.84	0.60	0.99	0.57	0.43	0.62
08/2017	0.53	-	0.87	0.72	0.97	0.17	0.84	0.62	0.95	0.93	0.17
09/2017	0.88	0.87	-	0.73	0.49	0.43	0.75	0.19	0.07	-0.03	0.33
10/2017	0.92	0.72	0.73	-	0.85	0.72	0.62	0.99	0.76	0.63	0.60
11/2017	0.69	0.97	0.49	0.85	-	0.32	0.33	0.78	0.97	0.91	0.35
12/2017	0.84	0.17	0.43	0.72	0.32	-	0.76	0.67	0.17	0.02	0.54
01/2018	0.60	0.84	0.75	0.62	0.33	0.76	-	0.25	0.17	0.02	0.15
02/2018	0.99	0.62	0.19	0.99	0.78	0.67	0.25	-	0.68	0.55	0.66
03/2018	0.57	0.95	0.07	0.76	0.97	0.17	0.17	0.68	-	0.95	0.33
04/2018	0.43	0.93	-0.03	0.63	0.91	0.02	0.02	0.55	0.95	-	0.11
05/2018	0.62	0.17	0.33	0.60	0.35	0.54	0.15	0.66	0.33	0.11	-

The correlation tables show that the network trade link's correlation can be anything between 0.01 and 0.99. Significant in the results is that the out-degrees of January 2018 correlate very little and make an exception to the other observed months. When this is combined with price data, we can detect that December 2017 and January 2018 are the times when the bitcoin price sunk after the great increase in the price. The observation related to in-degrees and the price increase is that the wallet link correlation decreased exceptionally in December 2017 when the

price increased. Though as in out-degrees, the correlation decrease was more permanent in in-degree links.

7. CONCLUSIONS

The results of this thesis' analysis were consistent with the hypotheses. The network of bitcoin investor wallets is clustered and the links between wallets and the wallets have continuity between time periods. In the buy-buy synchronization we found 18 from the aggregated network and sell-sell network 30 clusters. This results show the structure of these networks is not homogenous in terms of how the wallets are located in the network. In the sell-buy network there are 33 clusters. Of all identified wallets, 14 159 were active in every month on the 11-months period. This is only 0.02 % of all wallets identified, which means, most of the wallets are used only rarely or even just once.

Clustering means that in the cluster the wallets are more similar to each other than to others in other clusters, in this case the wallets in clusters have higher trade timing similarity. As we are not able to connect the wallets to the real-life identities, we don't know who or what the center of the cluster is or what is the factor that gets the wallets trade with each other. However, the fact the bitcoin network has clusters, is presumptive conclusion to draw from the data. According to the analysis, the wallets are likely to stay in the same cluster i.e. connected to the same wallets as the months change. This strengthens the assumption that the wallets in the cluster are connected to certain broker or system provider.

The bitcoin price had a peak in December 2017, middle of the analyzed time period. As assumed in the hypothesis, this affected to the investors' behavior. It was visible, that the selling activities decreased after, which can be assumed to be at least partially due to the price. In the correlation and Jaccard analysis the buying and selling transaction continuity decreased. Behind it is probably many reasons, but the connection to the price changes is there. As well, the number of transactions, especially after the price drop, follow the price curve. The correlation between wallets decreased for two months after the price drop, but reverted back as the price rose again. This shows the price's affection to the trading activity and traders' behavior.

The levels of wallet synchronizations are quite predictable. The Jaccard analysis shows how the frequently investing wallets have a high similarity with the wallets investing in the sequential months. Up to 76 % of the wallets have at least 1 000 observations in a month have the same amount in the following month as well. This shows the high continuity of these investing wallets. The similarity between sell and buy network's synchronized links show that the wallets are likely to continue trading with the same wallets that they are used to.

In this thesis we connected bitcoin transaction addresses to bitcoin wallets. As mentioned, a single bitcoin investor could, and very likely do, have multiple wallets for different purposes. One obvious subject for further analysis would be extend the analysis to user level. That could reveal better connections and clusters between investors and reveal big actors in the bitcoin field. This analysis could also be extended to motif analysis. That would reveal interesting patterns and give more results to analyze.

8. BIBLIOGRAPHY

Bech, M. & Garratt, R., 2017. Central Bank Cryptocurrencies. *BIS Quartely Review.*

Belinky, M., Rennick, E. & Veitch, A., 2015. The Fintech 2.0 Paper: rebooting financial services. *Santander Innoventures.*

Bitbay.net, 2019. *Bitbay.* [Online]
Available at: https://bitbay.net/en/exchange-rate/bitcoin-price-usd
[Accessed 2 December 2019].

Blockchain.com, 2019. *Blockchain.com.* [Online]
Available at: https://www.blockchain.com/charts
[Accessed 2 December 2019].

Bovet, A. et al., 2018. Network-based Indicators of Bitcoin Bubbles. *Physics and Society; Social and Information Networks.*

Böhme, R., Christin, N., Edelman, B. & Moore, T., 2015. Bitcoin: Economics, Technology, and Govenrance. *Journal of Economic Perspectives,* 29(2), pp. 213-238.

Chaim, P. & Laurini, M. P., 2019. Is Bitcoin a Bubble?. *Physica A: Statistical Mechanics and its Applications,* Volume 517, pp. 222-232.

Chaum, D., 1998. *Blind signatures for untraceable payments.* University of California, Deparment of Computer Science.

Chohan, U., 2017. *Cryptocurrencies: A Brief Thematic Review. Discussion Paper Reviews: Notes on the 21th Century.* Canberra, University of New South Wales.

Chuen, D., 2015. *Handbook of Digital Currency: Bitcoin, Innovation, Financial Instruments, and Big Data.* 1st ed. s.l.:Academic Press.

Commission, U. S. a. E., 2014. *Forbes.* [Online]
Available at: https://www.forbes.com/sites/bernardmarr/2017/12/06/a-short-history-of-bitcoin-and-crypto-currency-everyone-should-read/#5d441a913f27
[Accessed 26 October 2019].

Darren, H., 2018. Introducing Bitcoin. *Governance Directions,* pp. 247-252.

Foley, S., Karlsen, J. R. & Putnins, T. J., 2018. Sex, Drugs, and Bitcoin: How Much Illegal Activity is Financed through Cryptocurrencies. *Review of Financial Studies.*

Giuffre, K., 2015. Cultural Production in Networks. *International Encyclopedia of the Social & Behavioral Sciences,* Volume II, pp. 466-470.

Hileman, G. & Rauchs, M., 2017. *Global cryptocurrency benchmarking study.* s.l., University of Cambridge.

Hopkins, D., 2018. *Understanding Blockchain in 5 Minutes.* [Online]
Available at: https://www.mcgrathnicol.com/insight/understanding-blockchain-5-minutes/
[Accessed 26 October 2019].

Hung, H. M., O'Neill, R., Bauer, P. & Kohne, K., 1997. *The Behavior of the P-Value When the Alternative Hypothesis is True,* s.l.: 10.2307/2533093.

Iansiti, M. & Lakhani, K. R., 2017. The Truth about Blockchain. *Harward Business Review,* Issue January-February.

Kosub, S., 2016. *A note on the triangle inequality for the Jaccard distance,* Konstanz, Germany: Department of Computer & Information Science, University of Konstanz.

Krishnan, H., Saketh, S. & Vaibhav, V., 2015. Cryptocurrency Mining - Transition to Cloud. *IJACSA – International Journal of Advanced Computer Science and Applications,* 6(9).

Lam, E., 2017. *Bloomberg.* [Online]
Available at: https://www.bloomberg.com/news/articles/2017- 12-15/what-the-world-s-central-banks-are-saying-about-cryptocurrencies
[Accessed 26 October 2019].

Lamport, L., Shostak, R. & Pease, M., 1982. The Byzantine Generals Problem. *ACM Transactions on Programming Languages and Systems,* 4(3), pp. 382-401.

Lansky, J., 2018. Possible State Approaches to Cryptocurrencies. *Journal of Systems Integration,* Issue 1.

Lee, T., 2013. *The Washington Post.* [Online]
Available at: https://www.washingtonpost.com/gdpr-consent/?destination=%2fblogs%2fthe-switch%2fwp%2f2013%2f08%2f21%2ffive-surprising-facts-about-bitcoin-2%2f%3f&utm_term=.9f62a4d483b5
[Accessed 26 October 2019].

Lehmusvirta, A., 2018. *Kauppalehti.* [Online]
Available at: https://www.kauppalehti.fi/uutiset/nordea-kieltaa-tyontekijoiltaan-bitcoinin-sama-kuin-nokia-kieltaisi-skypen-kayton/e286b1d6-7c75-3a75-87ed-43268c355c7d
[Accessed 2 December 2019].

Li, C., Yu, K. & Wu, X., 2018. Co-clustering Analysis of Mobile Users' Usage Behavior on Apps. *ICTCE 2018: Proceedings of the 2nd International Conference on Telecommunications and Communication Engineering,* pp. 214-219.

Lielacher, A., 2020. How Many People Use Bitcoin in 2020?. *Bitcoin Market Journal.*

Marr, B., 2017. *A Short History of Bitcoin and Crypto Currency Everyone Should Read.* [Online]
Available at: https://www.forbes.com/sites/bernardmarr/2017/12/06/a-short-history-of-bitcoin-and-crypto-currency-everyone-should-read/
[Accessed 2 December 2019].

McCorry, P., Shahandashti, S. & Hao, F., 2017. Refund Attacks on Bitcoin's Payment Protocol. *Lecture Notes in Computer Science,* Volume 9603, pp. 581-599.

Meiklejohn, S. et al., 2013. A Fistful of Bitcoins: Characterizing Payments Among Men with no Names. *IMC Proceedings of the 2013 Conference of Internet Measurement Conference,* pp. 127-140.

Narayanan, A. et al., 2016. *Bitcoin and cryptocurrenct technologies: a comprehensive introduction.* Princeton: Princeton University Press.

Nower, M., Gomber, P., Hinz, O. & Schiereck, D., 2017. Blockchain. *Business & Information Systems Engineering,* 59(33), pp. 183-187.

Raval, S., 2016. *Decentralized Applications: Harnessing Bitcoin's Blockchain Technology.* s.l.:O'Reilly Media.

Reid, F. & Harrigan, M., 2011. An Analysis of Anonymity in the Bitcoin System. *IEEE Third International Conference on Privacy, Security, Risk and Trust,* pp. 1318-1326.

Thum, M., 2018. The Economic Cost of Bitcoin Mining. *CESifo Forum, Special,* 19(I).

Tindell, K., 2013. *Business Insider.* [Online]
Available at: https://www.businessinsider.com/how-bitcoins-are-mined-and-used-2013-4?r=US&IR=T
[Accessed 26 October 2019].

Tiwari, N., 2018. The Commodification of Cryptocurrency. *Michigan Law Review,* 177(3), pp. 611-634.

Treleaven, P., Brown, R. & Yang, D., 2017. Blockchain Technology in Finance. *Computer,* 50(9), pp. 14-17.

Tumminello, M., Lillo, F., Piilo, J. & Mantegna, R., 2011. Identification of clusters of investors from their real trading activity in a financial market. *Trading and Market Microstructure.*

Vilim, M., Duwe, H. & Kumar, R., 2016. *Approximate Bitcoin Mining.* Austin, TX, s.n., pp. 1-6.
Villasenor, J., 2014. *Forbes.* [Online]
Available at: https://www.forbes.com/sites/johnvillasenor/2014/04/26/secure-bitcoin-storage-a-qa-with-three-bitcoin-company-ceos/#4adeb6b95cdd
[Accessed 2 December 2019].

Wilson, T., 2019. *Reuters.* [Online]
Available at: https://www.reuters.com/article/us-crypto-currencies-bitfinex/u-s-returns-to-bitfinex-exchange-fraction-of-bitcoin-stolen-in-2016-heist-idUSKCN1QE11S
[Accessed 26 October 2019].

Volety, T. et al., 2019. Cracking Bitcoin Wallets: I want what you have in the wallets. *Future Generation Computer Systems,* Volume 91, pp. 136-143.

Vranken, H., 2017. Sustainability of Bitcoin and Blockchains. *Current Opinion in Environmental Sustainability,* Volume 28, pp. 1-9.

Xunhua, W., Tjaden, B. & Mata-Toledo, R. A., 2018. Cryptocurrency. *Access Science,* Issue AccessScience, McGraw-Hill Education.

Zheng, Z. et al., 2018. Blockchain Challenges and Opportunities: a survey.. *International Journal of Web and Grid Services,* 14(4), pp. 352-375.