Masaryk University
Faculty of Informatics

Darknet market analysis and user de-anonymization

Master's Thesis

Tomáš Šíma

Brno, Spring 2018

Masaryk University
Faculty of Informatics

Darknet market analysis and user de-anonymization

Master's Thesis

Tomáš Šíma

Brno, Spring 2018

This is where a copy of the official signed thesis assignment and a copy of the Statement of an Author is located in the printed version of the document.

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Tomáš Šíma

Advisor: RNDr. Martin Stehlík, Ph.D, Mgr. Jaroslav Šeděnka

Acknowledgements

I would like to thank my supervisor RNDr. Martin Stehlík Ph.D for guiding me and providing technical support for my work.

I would also like to thank Mgr. Jaroslav Šeděnka for his continuous stream of helpful comments and ideas.

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

Abstract

The goal of this thesis is to create a tool to find, analyze and visualize publicly available data, which can be helpful to deanonymize users of drug markets available via TOR on dark web. The aim of this tool is to help investigators with collecting intelligence on entities related to these drug markets. Users and operators of these markets employ multiple means to prevent their deanonymization. The markets are operated ad TOR services, PGP encryption is often required to use in communication between multiple parties and bitcoin is used as a way to pay for goods or services.

We scraped multiple publicly available social sites and websites related to bitcoin(twitter,bitcointalk, reddit, blockchain.info...) and drug markets thereself using python. Westored all these data into AgensGraph database, which is a graph database based on PostgreSQL. Wecreated a tool, which uses these data and multiple heuristics to analyze and visualize data and metadata of users,drug markets, social media and blockchain. Tool can also for given adress find the nearest adresses or transactions related to drug markets and also find the nearest adresses that are mentioned in scraped websites.

To test the efficiency of this tool, we created multiple profiles on these dark markets and performed multiple transactions to deposit and withdraw bitcoins. The tool identified these and these percent of transactions.

Keywords

blockhain, bitcoin, darknet, drug market, TOR, cryptocurrency, anonymity, metadata, de-anonymization

Contents

List of Tables

List of Figures

1 Introduction

The relative anonymity of internet offer an incentive for criminal parties to use internet as a tool for their activities. Internet facilitated some forms of existing crimes(Selling drugs and guns, counterfeits selling, Ponzi schemes) and also enabled many new types of frauds like hacking, phishing and carding. Police statistics show, that crime happening online is much less likely to be discovered and criminals persecuted. Criminals value their anonymity very high and use various means to make them even more anonymous, like VPNs and TOR. The big problem for criminals were getting the money they got from criminal activity to their possession(In banknotes, or to their bank account), since that requires some form physical presence. Also, it was hard for two anonymous entities engaging in criminal activity to transfer money to each other, since none could be sure about the origin of money they are receiving.

For bitcoin, there is no central authority requiring bitcoin address (bitcoin equivalent of bank account number) to be linked to person's identity. Criminals can use their anonymous connection to internet to both recieve and send bitcoins and therefore not disclose their identity. This feature of bitcoin and other cryptocurrencies gave rise to drug markets, which can be publicly accessed via TOR. However, the history of all bitcoin transactions is publicly available and so each bitcoin can be tracked through the whole transaction history.

In this work we collect multiple public sources of data about bitcoin transactions, bitcoin adresses and drug markets. We examine these data in order to describe the behaviour of drug markets users (distribution of sellers, availability of drugs, number of users, revenues) and also to see, if these data can be used for disclosing identity of users and operators of drug markets.

1.1 Goals

The main goal of this thesis is to analyze one available drug market and try multiple approaches for deanonymization of identities related to this drug market. The outcome of this work is gathering data about the trades, which happened on the drug market and computing interesting statistics about the whole market as well as actors operating there. Another outcome of this work is a tool, that uses the data mentioned above to help investigator to disclose transactions, addresses and identities related to online drug markets.

1.2 Structure of thesis

The following text describes individual chapters forming structure of this thesis. The chapter Related works give overview of work already done on similar topics and how this work differ or extend the previous done research.

The chapter Technology and terms gives quick introduction to bitcoin and blockchain, which is used for paying on crypto markets. Then it describes how the dark markets operate within dark web.

The chapter Methods and tools to get and analyze data describe the proccess of collecting the data from bitcoin blockchain, drug markets and pubicly available sites(Mainly forums and social networks).

The Deanonymization techniques chapter descibe heuristics and methods that are later used by the proof of concept application to detect addresses used by drug markets and link the users of drug markets to publicly found identities.

The chapter Statistics of drug markets describes various statistics about drug markets, that were gathered during drug market website scraping. It contains two parts, the first is focused on statistics related to money, the second part is giving insight about non-money related statistics.

POC application chapter describes the functionality, implementation and possible future development of application for investigating bitcoin addresses, which was created as part of this thesis.

Testing and verification of the created tool descibes the proccess by which the proof of concept application was tested and the results

The last chapter Discussion is about achieving goals, problems of implementation and future work.

2 Related works

2.1 Pairing blockchain transactions with public data

Multiple papers were published regarding analysis of blockchain graph. The (Reid and Harrigan 2013) was published in 2013 and dealt with much smaller number of people using bitcoin and smaller transaction graph. Their analysis also focus on danonymization through multiple aspects of bitcoin protocol, while this thesis focus on deanonymization from transaction graph and public data. The (Ron and Shamir 2013) focus on bringing interesting statistics about bitcoin transaction graph and track only really big(>50000 BTC) transactions on the network. The authors of this paper also had to deal with much smaller blockchain graph.

Similar work to this thesis was done by (Fleder, Kester, and Pillai 2015). This paper use data from bitcointalk, the most popular bitcoin forum. They apply simple algorithm to group multiple bitcoin adresses belonging to one user together. Than they use the scraped data to show that some of the bitcointalk users were using silkroad marketplace or other popular services accepting bitcoin.

Advanced and similar work was done by (Spagnuolo, Maggi, and Zanero 2014). They downloaded the blockchain, transformed to the database and performed clustering to get graph of transaction between users. Than they developed a tool, which scraped data from multiple locations(bitcointalk and bitcoin-OTC forum) to link off-chain data and identities to bitcoin adresses. They tested the tool on few popular transactions related to seizure of silkroad marketplace.

All of the previously mentioned works had to deal with much smaller transaction graph, as the usage of bitcoin grew exponentionally over the last year. My work is unique in that way, that it utilize much more sources of data, than the works previous mentioned. Also, the aim of this tool is to be able to identify even just regular users of drug markets, not just big and important transactions.

2.2 Behaviour of drug markets users and operators

Papers describing the drug market users,vendors and the dynamic of the online drug marketplace economy mostly focused on data related to silkroad marketplace seizure. Few authors described, how is the whole drug trafficking crime changing overtime with the coming of the new technologies. There are only few articles focusing on describing the economy of fully operating drug market at the time of data collection. In this work, we bring analysis of the micro-economy of two fully operating drug markets and present interesting statistics about vendors, size and frequency of the deals, sortiment and availability.

3 Related terms

In this chapter, I explain the terms and technology related to online drug markets. The online drug markets use several technologies, that are crucial for their anonymous operation. The Bitcoin enables different parties to exchange value in an anonymous way. TOR allows users and administrators of marketplace to hide from any third party doing packet sniffing on network, that they are accessing drug marketplace. It also hides the location on drug market webserver from it's users. PGP enables sellers and vendors to communicate between them in encrypted way, so that drug market administrators can not eavesdrop on that communication. Drug markets also use bitcoin mixers, services designed to mix their funds with others, in order to obstruct analysis of their cashflow and improve anonymity of users and administrators.

3.1 Cryptomarket

3.2 Bitcoin and blockchain

Bitcoin is the first decentralized peer2peer cryptocurrency, created by anonymous author(s) known by pseudonym Satoshi Nakamoto in 2009. Bitcoin transactions are not verified by central authority, they are processed by peer2peer network of bitcoin nodes instead. The entire history of transactions is stored in distributed public ledger called blockchain.

The nodes collect transactions broadcasted by users and send them to other nodes. The source code of bitcoin nodes is open source and can be downloaded and run locally.

3.2.1 Transactions and addresses

In order to recieve bitcoins, user need to have a bitcoin address. In order to send bitcoin from bitcoin address, user needs to have private key associated with the given bitcoin address. Storing and using bitcoin addresses and associated private keys is automatically managed by software called bitcoin wallet. There exists many third party software wallets.

Bitcoin address is string of 26 to 35 alphanumerical characters. All the transactions of every address are stored in blockchain, the balance and all transactions related to address are publicly available. In order to not see the whole history of transactions of address's owner, the bitcoin wallets generate new bitcoin address for each new incoming transactions and when spending bitcoins, it use one or more of the addresses the wallet generated. Therefore, when pairing the address to identity, we can directly obtain just the history of transactions related to the address, but can not get all transactions and balance of the user, as he is likely to own multiple bitcoin addresses.

3.2.2 Mining

Users of bitcoin can be roughly separated in two groups, end users and miners. End users use bitcoin wallets to recieve and send bitcoins. Miners are verifying transactions. When the transactions is send from the address, it is signed with private key associated with that address and send to the neares bitcoin node. The transaction is immidiately broadcasted to other bitcoin nodes. Miners are running bitcoin mining software, which enables them to create a new block of transactions, add it to blockchain and broadcast new, longer version, of blockchain to other nodes. Finding new block of transactions is a hard problem from computanional perspective. The difficulty of algorithm is adjusted every X block, so that new block is generated roughly every 10 minutes.

When miner finds new block, he can claim all of the fees of transactions included in that block, also he is able to create a special transaction called coinbase transaction, that sends bitcoins from nowhere to his address. By these coinbase transactions, new bitcoins are emmitted into network.

3.2.3 Decentralisation

Since anyone can run bitcoin node or mine bitcoins, and every node

3.3 TOR - the onion routing

Tor is an free open source software, that provides access to tor network. Tor network is a network of TOR nodes. The goal of TOR project is to provide it's users encrypted access to internet in order to to prevent third parties from evesdropping and analysis of the transmitted data. The communication of the user's computer with network is encrypted and rerouted through multiple TOR nodes using onion routing technology. The usage of TOR can be detected by third party, but the third party can not decryptsa user's data, that are transmitted via tor. Some websites restrict access from TOR, due to many risks involved.

Communication between browser and webserver is usually done via encrypted HTTPS protocol. This protocol use assymetric cryptography. The webserver and browser exchange their public keys at start of communication and encrypt the data using these keys. Decrypting the data is possible only by corresponding private keys, which the browser and webserver keep locally. This protocol is suspectible to man in the middle attacks. If the attacker has control over the transmission from the start of communication, he can place himself in the middle of communication and act as webserver for user and as a user for webserver. To prevent these types of attact a certification authority is needed, which is a institution, that sign public keys, belonging to webserver. When browser recieves the public key, it automatically checks, if it is signed by any authority from it's list of authorities and if not, it displays warning or error message.

The HTTPS protocol encrypts data, but doesn't hide the identity of the user from webserver, and also the internet provider can see, where is the user connecting to. In TOR, the user's identity is hidden from webserver, and internet provider can only see, that user is connecting to TOR, but can not see where is the destination of the data that are transmitted via TOR. Tor uses Onion routing technology. When user visits website, there TOR software picks randomly few TOR nodes from the network and estabilish a circuit, as we can see on 3.1. The packet of data is encrypted with the each public key of the node in circuit, starting from last node as on 3.2. Each node of the network only knows the previous node he recieved the data from and it gets the address of next node by decrypting the packet and reading the added metadata.

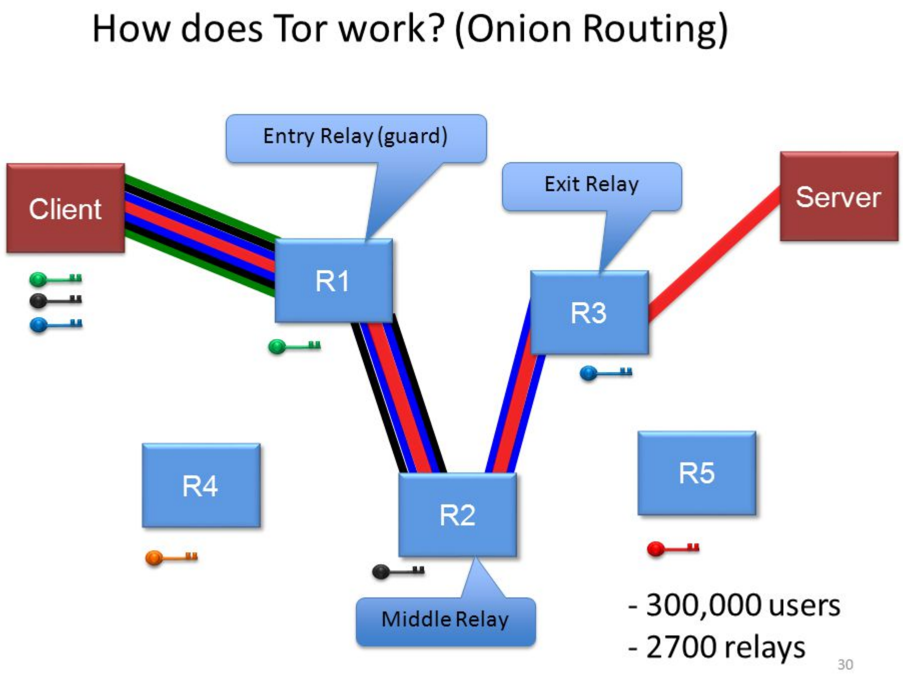

Figure 3.1: TOR routing schema

3.4 PGP

PGP is a program for encrypting data and communication between two parties using public key cryptography. PGP is used for signing, encrypting and decrypting messages, mostly e-mails. PGP was developed in 1991 as open source, with the intention to provide an open widely used standart for encrypted communication. Nowadays, PGP program is not open source any more, but the standart is used by open source GPG software.

PGP uses public key cryptography, unlike symmetric cryptography, pgp uses two different keys for encrypting and decrypting. The user generate a pair of keys, public key for encrypting mails sent to them and private key, which the user keeps for himself and use for decrypting messages encrypted with associated public key. The user also publish his public key, so other users can send him necrypted messages.

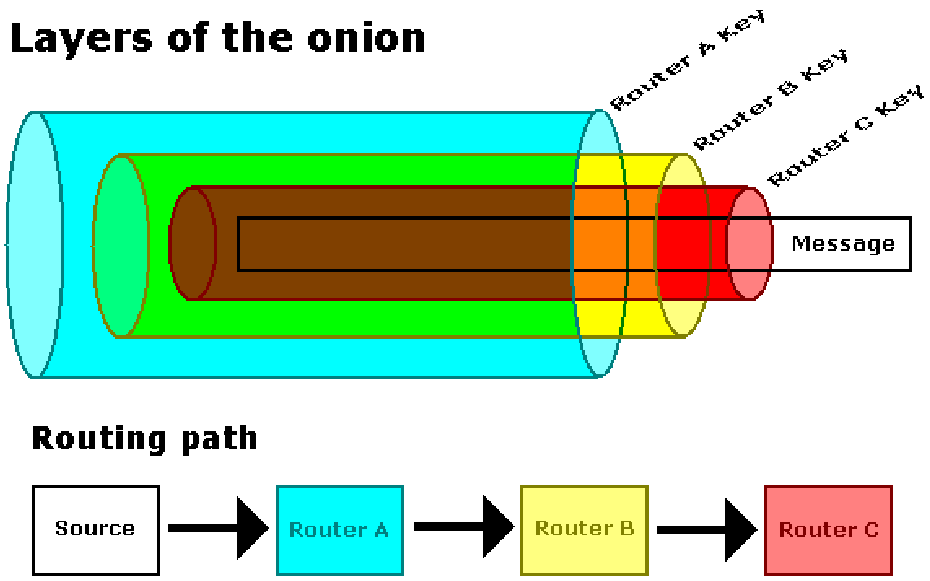

Figure 3.2: TOR packed encryption schem

PGP is used in the context of online drug markets as a means of communication between vendors and customers. Both vendor and customer has their public keys published on their profile page and use the public key of the other party to encrypt messages to them. This enables vendors and sellers to keep their communication private also from the administrators of marketplace.

3.5 Online dark markets

A dark market is

3.5.1 Escrow

3.5.2 Tumbler

3.5.3 Feedback

4 Methods and tools to get and analyze data

4.1 Obtaining,storing and analyzing blockchain data

In order to create a tool, that will find data related to bitcoin adresses, we need to store the blockchain locally in that way, that common graph algorithms can be applied. We ran the official bitcoin daemon (further referenced as bitcoind), to obtain a copy of bitcoin blockchain. Bitcoind store blockchain in multiple *.blk files. These files have structure, which is unfit for searching, processing and analysis of blockchain, so I used rusty-parser to parse these files and create csv files of transactions, outputs and adresses.

Than we imported these files into neo4j graph database, to have whole transaction graph in one place and be able to compute statistics and heuristics. All entities in the 4.1 are represented as graph nodes, the relationships between them are edges.

4.2 Drug markets web scraping and data collection

We scraped data from dream market and valhalla, 2 big popular drug markets available via TOR. Wescraped the vendor nicknames, buyer reviews and the sortiment that each vendor sells. We tested, if every transaction that is happening on drug market has its counter transaction in bitcoin blockchain. Wesent 0.05 bitcoins to both markets, bought a virtually deliverable legally service(link to secret forum) and checked, if the bitcoins that I have sent to deposit adress left. For both markets, there was no transaction happening for days after the transaction was done. This means, that markets don't transfer bitcoins, when there is filled order, all the transactions that these drug markets do are just for depositing bitcoins on drug market account, withdraw bitcoins and money laundering bitcoins. We made multiple deposits and withdraws from drug markets in order to track, where were the deposited bitcoins transfered and where the withdrawn bitcoins originated. These deposits and withdrawals are used to test the resulting application We scraped 158 vendor PGP keys from dream market and 70 PGP keys from walhalla. Wetested these keys, if they are vulnerable to ROCA attack, via python module roca-detect. None of these

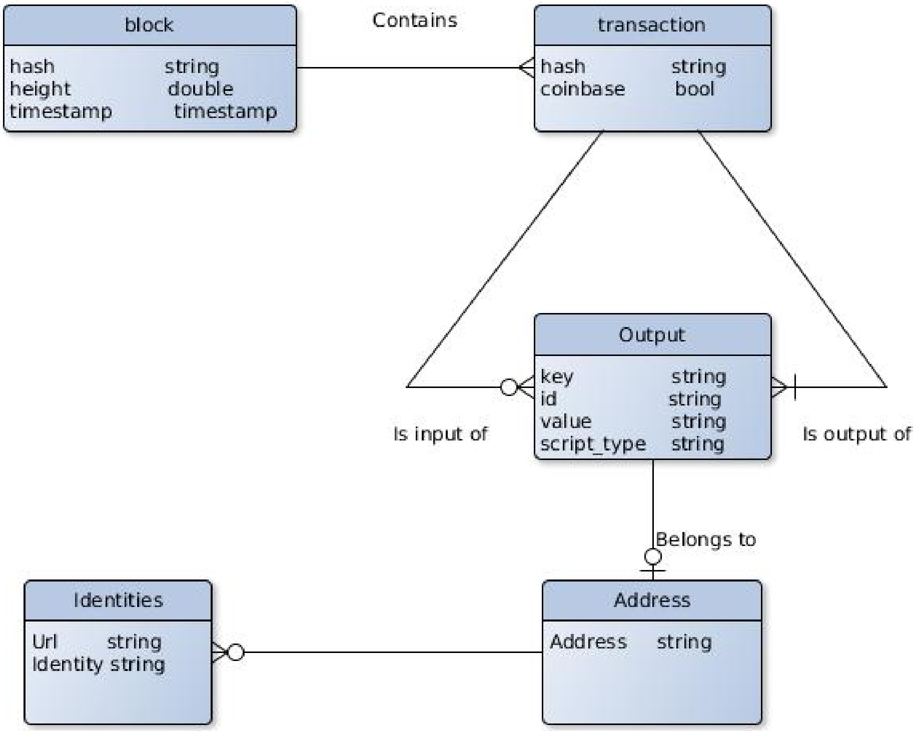

Figure 4.1: Neo4j database ER diagram

keys were vulnerable. All these PGP keys were searched for User-Id in metadata of PGP key and these user-Ids were seached by google. None of thesearches for user-Ids(both nicknames and mail addresses) returned any results.

We thought that metadata from the photos of drugs, which are available on the drug markets might be useful. We downloaded hundreds of pictures both from walhalla and dream market. Only metadata directly dependending on image content(like amount of red, green and blue colors) differ, metadata that could potentionally help dislosing user identity(date of creation nad modification, signature, software version) were the same. The software version contained line: $ImageMagick6.8.9-9Q16x86_642017-07-31http://www.imagemagick.org$ We created vendor account on both markets and uploaded an image with custom made metadata to see, if the metadata were scraped and same version of software version appears. It happened so for both markets, therefore we believe, that markets automatically scrape metadata from uploaded images in order to protect privacy of the users.

4.3 Drug market server fingerprinting

We tried to scan ports of drug markets servers and fingerprint their webserver, in order to find any vectors of further information gathering. We scanned both drug markets servers using netcat, finding, that the only opened port is number 443(HTTPS), which is used by webserver. We used httprecon to fingerprint used HTTP server. The fingerprinting consists of sending multiple malformed HTTP requests and comparing the webserver output with the database of responses by different webservers. The results of fingerprinting can be see in figure xxx, the best matches are various modern versions of apache webserver. The results of port scan and webserver printing doesn't indicate any way how to gather data about drug markets servers.

4.4 Publicly available data scraping

In order to have some bitcoind addresses and bitcoins linked to identities, We searched internet for pages, where are bitcoin adresses tied to real or virtual identities. The interesting sites that I decided to scrape were bitcointalk forum, bitcoin-OTC, reddit, twitter, bitcoin.info. The bitcointalk and bitcoin-OTC are the most popular internet forums related to cryptocurrencies. The script bitcointalk-scraper.py visits profile pages of all profiles on both forums (even those without any posts) and matched with bitcoin address regular expression.

The reddit and twitter were scraped by twitter-reddit-scraper.py. The script contain several hardcoded phrases like "Donate bitcoin" and "bitcoind address" and scrapes the results of search page. Bitcoin.info is a webpage that serves primarly as bitcoin blockchain explorer, secundary, it gathers multiple statistics about bitcoin blockchain and also offers for third parties to have their bitcoin address and identity listed on their webpage. Some of these identities are verifies by signaturing custom made message with the bitcoin address associated private key.

We scraped data with the intention to link identities to bitcoin addresses. The data scraped from public sources are rows with thre collums: bitcoin addres, URL where was the addres scraped and username of the associated identity. All data scraped from the public sources(bitcointalk, reddit,twitter, bitcoin-OTC) are imported to the same neo4j graph database as metadata belonging to the nodes representing given address.

5 Deanonymization techniques

5.1 Detecting wallets owned by drug markets

5.2 Using own transactions to get market wallets

6 Statistics of drug markets

Since we are trying to identify

6.1 Methodology

The data was collected from walhalla drug market on 20.1.2018. This url's of all market listings are in pattern http://valhallaxmn3fydu.onion/products/xxx where xxx is incrementing with each new listing. We wrote a small script in bash to iterate through all the listings and download them using wget. To be able to download via wget from .onion links, I had to use privoxy, to redirect the wget through locally ran TOR daemon. After downloading all the pages of products, we parsed the downloaded files using python and common linux command line tools(cat,grep,cut,sed) From the listing, we parsed vendor's nickname, the subcategory where the listing was placed and title of listing.

By this, we got 666 unique vendors name, so we downloaded and scraped the vendor's profiles pages from the walhalla market in similar way. From the vendor's profile pages, we scraped name of vendor, his total revenue, number of positive and negative reviews and the countries from which the vendor ships. The shortcoming of this method is, that we can download and analyze only sellers, that have at least one active listing at the time of data collection. Hovewer, we managed to download 20000 listings out of 100000.

The statistics, tables and plots in this chapter were produced by statistical and data analysis software R. The exact commands to generate these figures and plots can be found in attachments in file named 'valhalla-r.txt'.

6.2 Overall statistics of Walhalla drug market

Walhalla was originally founded as local Finnish market, that seems the reason for surprisingly many vendors shipping from Finland. The reader can see the frequency of countries the vendors are shipping from in table 6.1.

Table 6.1: Countries vendors are shipping from

Countries vendors are shipping from	
Belgium,Bulgaria,Hungary,Ireland,	1
Philippines,Romania,Russia,Serbia,Switzerland	1
Austria, Czech Republic, India,Spain,Sweden, Argentina	2
Australia	3
Poland	4
Canada	5
France	6
Norway	7
Netherlands	10
Germany	13
United States	17
United Kingdom	24
Finland	34
Unknown	511

Each circle in 6.2 represents one neighbour and axis represent the amount of positive and negative reviews that vendor recieved. We can see, that vast majority only 2 vendors out of 666 have recieved more negative feedback than positive. Only 19 vendors out of 666 managed to get more than 50 negative feedbacks, while all of the these 19 vendors had more than 400 positive reviews. Only 40 vendors got more negative feedbacks than positive feedbacks. If we look at statistics of reviews from popular e-shop amazon(http://minimaxir.com/2017/01/amazon-spark) and consider one and two star reviews as negative, we can see, that amazon sellers on average gets between 5-25% negative reviews, depending on category of the goods. On the walhalla market, vast majoririty of sellers have >95% of positive reviews, as is shown on 6.1. Also, only 40 vendors have less than 80% positive reviews and out of that 36 have less 50 reviews in total. These numbers indicate, that the customers of valhalla market are much more picky about the vendor they choose than regular e-shop cuystomers. If 6.4

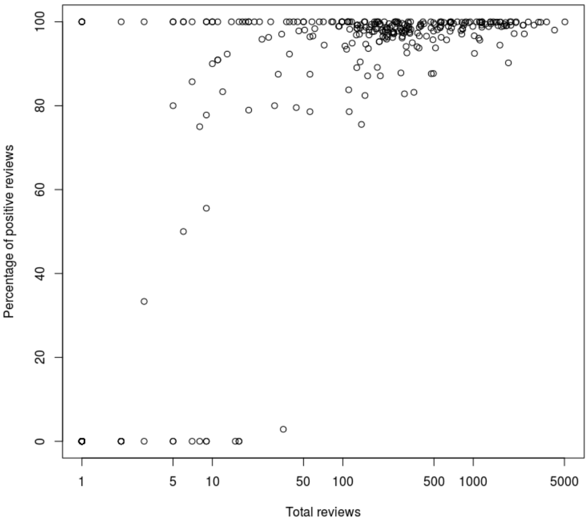

Figure 6.1: Positive reviews of vendors

asfd asdf

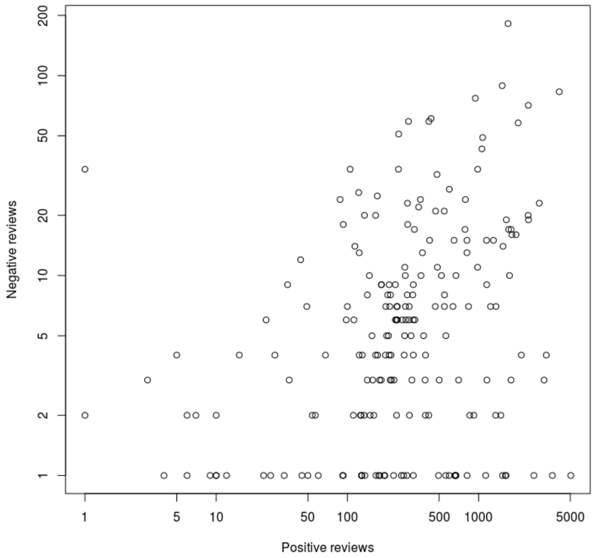

Figure 6.2: Positive/negative reviews of vendors

6.3 Statistics about vendors, drugs availability and distribution and buyers satisfaction

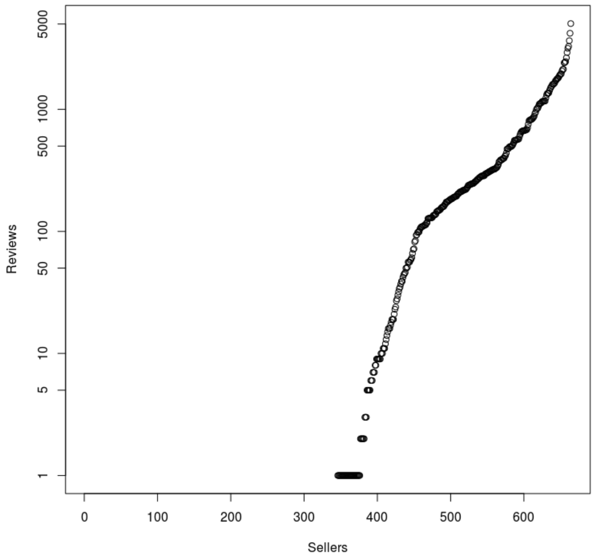

Figure 6.3: Number of reviews for vendors

7 Application

This chapter describes the application for investigating bitcoin address. The application consists of three parts. The scraping module, that downloads bitcoin blockchain and also scrape data from publicly available sites mention in section XXX. The computanional module, which imports data to the database and also transform data. so that searching in these data would be fast. The scraping, import and computanional modules are available for linux only. The GUI written in HTML/JS/CSS, that is connecting to neo4j database REST endpoint and provides visualisation of data. The GUI can be given a configuration string, to connect to neo4j REST API endpoint, so the gui can be viewed in broser from any device, as long as the server with neo4j data is reachable from that device.

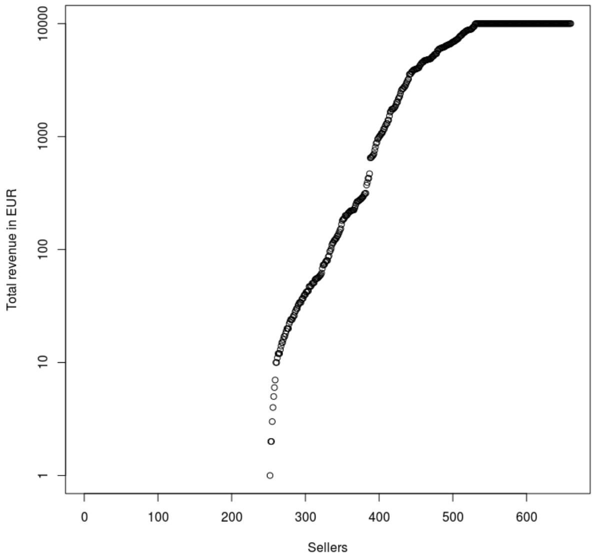

Figure 6.4: Total revenue of vendors

7.1 Implementation

The importing module is responsible for parsing bitcoin blockchain files and importing the data into neo4j database. The importing module take two parameters, the directory of .blk files, which store blockchain data and directory for creating neo4j graph database. The import module firstly parses the .blk files and save blockchain as multiple .csv files. This intermidiary step is useful for debugging and also simplifies importing to neo4j database.

The next importing script is scrape_identities.py script, which crawl popular forums and multiple websites and creates identities.csv. File identities.csv contains 3 collumns.

- Address - bitcoin address the identity is associated with
- Identity - String representing identity, usually username
- URL - Url where the Identity and Address were scraped

If the user has his own data about the owners of different bitcoin addresses, he can import it through the web GUI later.

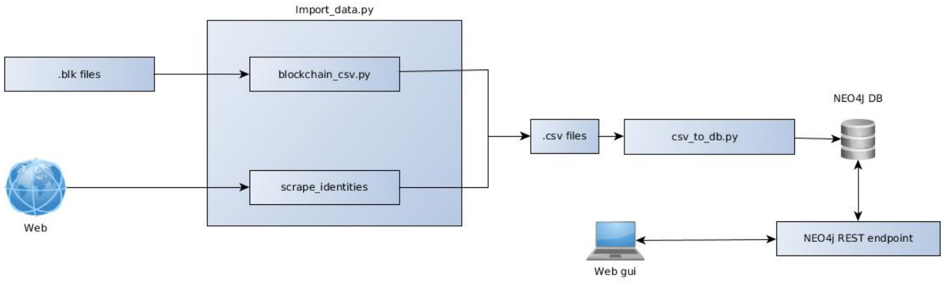

Figure 7.1: Neo4j database ER diagram

7.2 Usage

See the following command :

```
$ ./import_module ~/.blockchain/ ~/neo4j/graph.db
```

7.3 Future development possibilities

8 Testing and verification of the created tool

This chapter describes the way, the POC application was tested.

The testing were performed by sending bitcoins to drug markets and withdrawing them. Than marking the addresses from where the bitcoins were recieved as

8.1 Method of testing

8.2 results

9 Conclusion

Here you can insert the appendices of your thesis.gg

Bibliography

Borgman, Christine L. (2003). *From Gutenberg to the global information infrastructure. access to information in the networked world*. 1st ed. Cambridge (Mass): The MIT Press. xviii, 324. ISBN: 0-262-52345-0.

Fleder, Michael, Michael S Kester, and Sudeep Pillai (2015). "Bitcoin transaction graph analysis". In: *arXiv preprint arXiv:1502.01657*.

Greenberg, David (1998). "Camel drivers and gatecrashers. quality control in the digital research library". In: *The mirage of continuity. reconfiguring academic information resources for the 21st century*. Ed. by B.L Hawkins and P Battin. Washington (D.C.): Council on Library and Information Resources; Association of American Universities, pp. 105–116.

Hàn Thé, Thành (2001). "Micro-typographic extensions to the TEX typesetting system". PhD thesis. Brno: The Faculty of Informatics, Masaryk University. URL: `http://www.pragma-ade.nl/pdftex/thesis.pdf` (visited on 12/09/2016).

Masaryk University (1996–2009). URL: `https://www.muni.cz/en` (visited on 12/09/2016).

Nakamoto, Satoshi (2008). *Bitcoin: A peer-to-peer electronic cash system*.

Reid, Fergal and Martin Harrigan (2013). "An analysis of anonymity in the bitcoin system". In: *Security and privacy in social networks*. Springer, pp. 197–223.

Ron, Dorit and Adi Shamir (2013). "Quantitative analysis of the full bitcoin transaction graph". In: *International Conference on Financial Cryptography and Data Security*. Springer, pp. 6–24.

Spagnuolo, Michele, Federico Maggi, and Stefano Zanero (2014). "Bitiodine: Extracting intelligence from the bitcoin network". In: *International Conference on Financial Cryptography and Data Security*. Springer, pp. 457–468.