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

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 "Projects of Large Research, Development, and Innovations Infrastructures " programme (CESNET LM2015042) is greatly appreciated.

Abstract

This thesis has two goals. The first goal is to perform a quantitative statistical analysis of the Valhalla cryptomarket. We scraped the Valhalla cryptomarket website for information about vendors, listings and buyers and brought up a lot of interesting statistics about them.

The second 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 the dark web. The aim of this tool is to help investigators collect intelligence about entities related to these drug markets. Users and operators of these markets employ multiple means to prevent their deanonymization. Cryptomarkets are operated as hidden services, PGP encryption is often required to communicate between multiple parties and bitcoin is used as a way of paying for goods or services.

We scraped multiple publicly available social sites and websites related to bitcoin (Twitter, Bitcointalk, Reddit, blockchain.info...) and also Valhalla cryptomarket itself. We stored all the data in a Neo4j database. We created a tool collecting these data, importing them to the database, runnning multiple heuristics and providing user an interface for visualizing data and metadata of addresses and identities related to cryptomarkets and blockchain. The tool can find the nearest addresses related to drug markets for given bitcoin address and also find the nearest addresses that are mentioned in scraped websites.

To test the accuracy of this tool, we created two profiles on Valhalla cryptomarket and deposited and withdrew bitcoins multiple times. This way, we obtained multiple bitcoin addresses owned by Valhalla cryptomarket. We ran the tool and certified the tool's ability to identify, whether the addresses we obtained by these transactions were identified as belonging to the same identity (eq. Valhalla market).

Keywords

blockhain, bitcoin, darknet, drug market, Tor, cryptocurrency, anonymity, de-anonymization, Valhalla

Contents

List of Tables

List of Figures

1 Introduction

Relative anonymity of the Internet offers an incentive for criminal parties to use the Internet as a tool for their activities. Internet facilitates some forms of existing crimes (selling drugs, guns and counterfeits, running Ponzi schemes) and also enables many new types of frauds like hacking, phishing and identity theft.

Publicly available statistics prove that cybercriminals are much less likely to be discovered and persecuted, than criminals operating offline. In the USA in 2010, there were 5628 robberies and the loot was recovered in more than 20 % of cases [5]. In the same year, FBI received 303809 complaints related to cyber crime, resulting in just six convictions [1]. Criminals value their anonymity very highly and use various means to prevent being caught by police forces [27][30][3].

A big problem for criminals was getting the money obtained from criminal activity into their possession, as it required some form of physical presence or identification. Also, it was hard for two anonymous entities engaging in criminal activity to transfer value to each other, because it's difficult to set up an anonymous bank account and neither party could be sure about the origin of the money they have received.

For bitcoin, there is no central authority requiring bitcoin address (bitcoin equivalent of bank account number) to be linked to person's identity. The criminals can therefore just use their connection to the Internet to both receive and send bitcoins without disclosing their identity. However, all bitcoin transactions are publicly visible and so each bitcoin can be tracked through the whole transaction history.

Cryptomarkets are online marketplaces, where vendors offer illicit goods and services. Cryptomarkets are accessible via Tor network as a hidden service, and the users of cryptomarkets use PGP to communicate with each other and use bitcoin as a way to transfer value. These mechanisms make it possible for cryptomarkets to publicly operate, yet still be hard to reach by the law enforcement [8].

We scraped and examined data from the Valhalla market, to our best knowledge one of the currently most popular and well established operating drug markets (February 2018), in order to perform a statistical analysis of the scale of its operations.

All the bitcoin transactions are stored in a blockchain. The blockchain is publicly available, therefore anyone can access any bitcoin transactions data. Bitcoin transactions transfer bitcoins between bitcoin addresses. It is a common practice for users to generate new bitcoin address for each incoming transaction. When a user pays via bitcoin, he might use bitcoins only from some small subset of his bitcoin addresses, so the recipent doesn't know sender's addresses that were not used in the transaction. This mechanism helps the sender not to disclose, how many bitcoins he owns in total and also protects the receiver from finding all transactions that were done by sender in the past.

1.1 Goals

We created a toolchain for drug market users deanonymization by using publicly known data. The tool imports bitcoin blockchain into Neo4j graph database in order to create a graph of transactions. It then runs multiple heuristics over this graph, to cluster addresses that belong to the same user. The tool scrapes multiple public sources related to drug markets and bitcoin transactions in order to obtain some form of identification (eg. a username) linked to bitcoin address, and adds this information to database. It also provides a GUI, where the tool user can insert an address and the program finds addresses with linked identities that are few transactions away from the inserted address in the transaction graph. We also created multiple scripts for scraping information about trades from Valhalla market website and used these data for quantitative analysis of Valhalla market operations.

1.2 Structure of the thesis

Chapter *related terminology* starts with a quick introduction on how cryptomarkets work and the technology they use. It carries on describing the major technologies used by cryptomarkets – bitcoins, Tor and PGP. Chapter *Related work* gives an overview of previously published works on similar topics.

The chapter *Methods and tools of data retrieval and analysis* describes the process of collecting and storing the data from bitcoin blockchain, drug markets and publicly available forums and social networks.

The chapter called *Application for searching nearby identified addresses* describes the functionality, architecture and implementation of the application that was created as part of this work.

The chapter *Valhalla cryptomarket statistic* consists of various statistics about drug markets gathered by website scraping of the Valhalla market.

Chapter *Conclusions and future work* discusses the achieved goals, possible future extension and development of this work.

2 Related terminology

In this chapter, we explain the terms and technology related to cryptomarkets. Cryptomarkets use several technologies crucial to their anonymous operation. Bitcoin enables different parties to exchange value in an anonymous way. Tor allows users and administrators of a marketplace to hide from any third party performing packet sniffing on the network, that they are accessing drug marketplace. It also hides the location of drug marketplace web server from its users. PGP enables users and vendors to communicate between each other in an encrypted way, so that drug market administrators cannot eavesdrop on that communication.

2.1 Cryptomarkets

Illegal online markets have been around for more than 30 years [18]. On these markets, users can sell and buy drugs, weapons, hacking tools, stolen credit cards, counterfeit currency, forged documents and other illegal goods and services. Most markets forbid selling the most unambiguously harmful goods, such as child pornography or hitman services.

A new type of illegal online marketplace appeared in 2011, the so-called cryptomarket. A cryptomarket is an illegal online market accessible only via Tor network and using bitcoins as means of making payments. These two technologies provided a safer environment than previous markets hosted on forums and chatrooms.

Physical products, like drugs, are sent to a buyer via ordinary mail to the address provided by the buyer. The package is disguised as packages containing common goods sent by big online retailers [20].

Cryptomarkets are popular with vendors, as they offer them a high influx of customers and secure and anonymous environment for conducting their bussiness [28]. Cryptomarkets offer safer, more comfortable and more professional way of buying drugs, avoiding the need to meet face to face with dealers [6].

Nowadays multiple cryptomarkets exist, competing against one another. The risk of a failure of a deal is still high [32]. In order to protect buyers, cryptomarkets use Escrow and vendors' feedback to

identify scammers and minimize losses. Silkroad cryptomarket used tumbler services [23], which made it harder to analyze bitcoin transaction graph. Valhalla cryptomarket claims that they users can withdraw bitcoin from them using bitcoin tumbler too.

2.1.1 The process of ordering

Most of the cryptomarkets are publicly available and there is no fee for creating a user account. The act of buying drugs from them is considered user-friendly by buyers and is similar to the process of buying goods from popular lawful e-shops like Amazon.

The whole process consists roughly of these steps:

1. User creates an account on cryptomarket (unless he has one already)

2. He tops up his account by sending bitcoins to the bitcoin address generated for his deposits.

3. He then finds an offer he is interested in and buys it in a similar way as in any other e-shop.

4. Buyers' money is now locked by cryptomarket.

5. The buyer and vendor arrange the way of delivery.

6. If the buyer receives goods or services, he confirms it and money is unlocked to the vendor.

7. Buyer gives feedback to the vendor.

8. Vendor withdraws bitcoins from the market to his bitcoin address.

As vendors value their feedback ratings very highly, they encourage buyers to leave a positive feedback when the transaction goes well.

2.1.2 Valhalla cryptomarket

We selected Valhalla cryptomarket[1] based on three metrics. The first metric is its size.

We checked a list of active cryptomarkets mentioned on reddit community *r/darknetmarkets* and on the *https://www.deepdotweb.com* website. The site *https://www.deepdotweb.com* has alexa ranking 23500, which roughly corresponds to thousands visitors a day. The reddit community is the first link for google query "darkmarkets list" and has more than 160 000 readers. We therefore consider these sites to be relevant enough not to miss the biggest cryptomarkets in their lists. Cryptomarkets taken into consideration are in the table 2.1.

Valhalla market is a well known operating cryptomarket, with more than 20 000 active listings and 900 vendors, it's the second biggest currently operating cryptomarket.

Among these cryptomarkets, only Valhalla and Dream market have been operating since 2013. This represents an advantage for our analysis, as we can analyze matured cryptomarket with vendors, who have been selling for a longer time and have more reviews.

Among cryptomarkets mentioned above, the Valhalla cryptomarket provides the most information about vendors and buyers. Each feedback for given vendor consists of buyer's comment, date the feedbacks were given, the listing the user gave feedback to, the sum of money the buyer spent on market , the amount of trades the buyer did on the Valhalla market and first and last 2 characters of the buyer's nickname. (eq. for buyer with nickname "gopnik789" it shows "go***89")

The first and last 2 characters of buyer's username are really unique for Valhalla market, other markets offer first and last character of buyer's username at most. This will greatly help with recognizing the same buyer among multiple feedbacks. Apart from that, Valhalla is also the only market from the markets above, where the customer feedback contains the associated listing. This allows us to detect the most profitable and popular listings.

1. Available at http://Valhallaxmn3fydu.onion

Table 2.1: Visited cryptomarkets

Cryptomarket name	Number of listings
Dream Market	72830
Valhalla	25309
Wall street market	10566
Point market	9202

2.1.3 Vendor's feedback

Cryptomarkets usually employ reputation systems [22], where buyers can share their satisfaction with vendors. These systems are similar to systems used in popular e-commerce websites like Amazon or Ebay. Users can give feedback only to vendors they have traded with. On some cryptomarkets, it is only possible to upvote and downvote vendors, on some others, people can rate different parts of their interaction with the seller, like the easiness of communication, the speed of sending the goods and unsuspiciousness of packaging.

Feedback is not mandatory, but vendors encourage buyers to give them positive feedback [3][25], because high positive feedback ratio brings the vendor more customers over vendors with worse feedback ratio.

2.2 Bitcoin and blockchain

Bitcoin [19] is the first decentralized peer to peer cryptocurrency, created by the anonymous author(s) known by pseudonym Satoshi Nakamoto in 2008. Bitcoin transactions are not verified by central authority, they are processed by distributed peer to peer network of bitcoin nodes instead. The source code of bitcoin nodes is an open source and can be downloaded and run locally. The entire history of transactions is stored in a distributed public ledger called blockchain. Bitcoin combines multiple cryptography algorithms to achieve consensus among nodes on the state of the blockchain. The state of blockchain is protected by the computing power of bitcoin miners using cryptography to validate and add new transactions. State of blockchain should

only be modified by adding the new block of transactions to the end of the blockchain. A third party attacker could, in theory, modify the blockchain, if he had 51 % of the whole mining computing power. This has nearly happened once, the Ghash.io mining pool accounted for more than 42 % of mining power in 9.1.2014. In reaction to that event, some of miners under this mining pool left the pool in order to prevent this from happening. Also, it sometimes happens, that two miners add different new block to the blockchain at the same time and for some time, two versions of blockchain exist, with a different last block. These two versions exist until the next block is found for one version. Miners then accept the longer chain of the two versions and the shorter one is discarded. It is therefore recommended to wait multiple blocks after the transaction is added into blockchain to be sure it does not disappear by this forking.

2.2.1 Addresses, bitcoins and transactions

In order to receive and send bitcoins, a user needs to have a bitcoin address. A bitcoin address is a BASE58[2] encoded public key with 4 bytes added for checksum. Each address has its associated private key. In order to send bitcoin from bitcoin address, a user needs to have a private key associated with the given bitcoin address. Storage and usage of bitcoin addresses and associated private keys is automatically managed by a software called bitcoin wallet. Many third party software wallets currently exist.

All the transactions, bitcoins and addresses are stored in the blockchain. The transaction history with data about all transactions is publicly available. When someone pays by bitcoins, the recipent can see the recipent address and can search blockchain for all transactions the sender address was part of. In order to not see the whole history of transactions and balance of the address owner, the bitcoin wallets generate a new bitcoin address for each transaction where owner receives bitcoins. When spending bitcoins, the software uses subset of the previously generated addresses and can spend bitcoins from multiple addresses in one transaction. Therefore, when pairing the address to identity, we can directly obtain just the history of transactions related

2. https://en.wikipedia.org/wiki/Base58

to the given address, but not all transactions and balances of the user, as he is likely to own multiple bitcoin addresses.

Bitcoins in blockchain are represented as inputs and outputs of the transactions. Each transaction has some inputs and outputs. Input and output have the same data structure, they only differ in its relationship to the given transaction. Each input/output consists of its unique identifier, its value in bitcoins, bitcoin address that owns it and a script that descibes how the input/output can be used. The script allows a greater flexibility in spending bitcoins like multisignature wallets where two or more keys are required for spending bitcoins.

Every output can be used exactly once as an input of a new transaction, and therefore the owner of an output can not spend one output multiple times. He also can not directly spend just a portion of bitcoins from the output. In order to divide bitcoins in smaller parts, he needs to use "change" address. The change address mechanism is described later. When the sender sends a bitcoin to the recipient, he generates a transaction and sign it with the required private keys. A new transaction must comply with the following:

- The scripts of all used inputs trigger no failures and return True. (The sender is authorized to use inputs)
- No input has been used as an input in any other transaction. (The sender cannot spend one output multiple times)
- The sum of bitcoins of transaction inputs is equal to the sum of bitcoins of transaction outputs + fees

A new transaction must have 1 or more outputs. There can be multiple outputs in a transaction with different associated addresses and bitcoin values, however, there exists a common pattern. When the sender sends bitcoins to one recipient, the transaction typically contains two outputs. First output contains the recipient's address and the volume of bitcoins he receives. Second output is called *change output*. Since the sender usually doesn't own outputs with exactly the amount of bitcoins he wants to transfer, he adds a change output to the transaction. The change output's associated bitcoin address is owned by sender, and by it the sender gets back the bitcoins he doesn't want to transfer. This is the only way to split bitcoins into smaller parts. On

the figure 2.1 is a transaction scheme where sender spends output worth of 5 bitcoins, sends 2 bitcoins to recipent and get 3 bitcoins back as change.

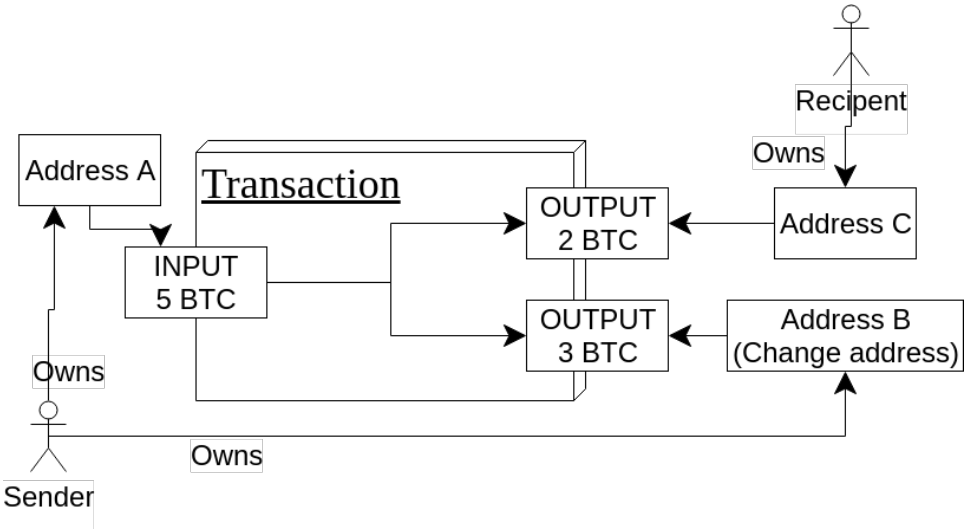

Figure 2.1: Mechanism of change address

When sending bitcoins from bitcoin wallet, wallet software creates the transaction data and signs it with required private keys. Then it sends them to one or multiple bitcoin nodes. Nodes collect transactions from users and broadcast them to other nodes on best effort basis. The validity of the transaction is later checked by miners and added to the newly generated block in blockchain.

2.2.2 Mining

The purpose of miners is to verify transactions and protect blockchain from being overwritten by a malicious third party. Miners are running bitcoin nodes and mining software, which enables them to create a new block of transactions, add it to blockchain and broadcast a new, longer version of blockchain to other nodes. Finding a new block of transactions is a hard problem from computational perspective. The SHA256 hash of newly generated block must be lower than a given number, called difficulty. The lower the difficulty, the harder it is to find the new block. Difficulty is adjusted every 2016 blocks in order

to match the computational power of miners, so that new blocks are generated on average once per 10 minutes.

Each block must contain the hash of the previous block in the blockchain, so it's impossible to precompute the problem for blocks that are to come further in the future. Each block contains, among other things, a list of transactions and some free space, where miners can insert random data to change the hash of the block. Miners try to find block that satisfy SHA256 requirement by brute-force by filling this space and when they find a solution, they are able to generate a new block of transactions and broadcast it to other miners. When a miner generates a new block, he can claim all of the fees of transactions included in that block, he is also able to create a special transaction called coinbase transaction, thus sending bitcoins from nowhere to his address. By these coinbase transactions, new bitcoins are emitted into the network. Blocks are limited by size and therefore may contain limited number of transactions. It is possible for new transactions to be waiting for being added to the blockchain for some time, because miners prefer to add transaction with higher fees to the next block.

2.2.3 Tumblers

Anyone can download blockchain and therefore obtain information about all bitcoin transactions that ever happened. Bitcoin transactions are seemingly anonymous, but when the user sends bitcoins to someone (exchange) who knows their identity, the recipient can pair the incoming bitcoin address to the identity of the sender. Although bitcoin users usually use multiple bitcoin addresses, their transactions and addresses are still susceptible to transaction graph analysis. This might identify other addresses belonging to the owner of the address we already know.

Bitcoin Tumblers exist in order to prevent, or at least obstruct such analysis. A user sends bitcoins to the tumbler service and the service mixes his bitcoins with other users' bitcoins by performing multiple transactions between its bitcoin addresses. [17]

The structure of these transactions differs for different tumbler services. The user sends their bitcoins to an address owned by the tumbler, then he generates a new, never used bitcoin address. He than receives bitcoins from tumbler service to his new bitcoin address.

Peer-to-peer tumblers also exist (CoinJoin, SharedCoin, Coinswap), that enable multiple users to directly create transactions to mix bitcoins among themselves. These transactions can be performed multiple times with different acters.

2.3 Tor - the onion routing

Communication between browser and web server is usually done via HTTPS protocol. This protocol uses asymmetric cryptography. The web server and browser exchange their public keys at the start of communication and encrypt the data using these keys. Decrypting the data is possible only by corresponding private keys, which the browser and web server keep locally. This protocol is susceptible 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 a web server for the user and as a user for the web server. To prevent these types of attack, a certification authority is needed. The certification authority is an institution that signs public keys, belonging to the web server. When the browser receives the public key, it automatically checks whether it is signed by any authority from its list of authorities and if not, it displays a warning or error message.

The HTTPS protocol encrypts data, but doesn't hide the identity of the user from the web server. The internet provider can also see, where is the user connecting. Tor aims to solve these issues.

Tor [10] is a free open source software that provides access to network of Tor nodes. The goal of the Tor project is to provide its users encrypted access to the Internet in order to prevent third parties from eavesdropping and analysing the transmitted data. While the Tor usage itself can be detected by a third party, the third party can not decrypt user's data that are transmitted via Tor. Some websites restrict access from Tor network due to many risks involved. Tor hides user's identity from the web server. Internet provider can also only detect that user is connecting to Tor, but can not identify the other side of communication and content of the transmitted data.

The simplest way for the user to use Tor network is to install the Tor browser bundle, consisting of a modified Firefox browser and onion

proxy. Onion proxy routes all traffic through Tor network. Tor network consists of more than 6000 publicly available Tor nodes (also called relays or routers). Anyone can run a Tor node, but there is a subset of speciallized servers called directory authorities. These directory authorities monitor relays and distribute the list of relays that are working and not under the same IP address. Tor nodes are divided into 3 groups, guard nodes, relay nodes and exit nodes. Client achieves anonymous communication by proxying his traffic through a chain of 3 Tor nodes, a so-called circuit, as it can be seen on picture 2.2. Client obtains public keys of these 3 Tor nodes from a directory authority and negotiate symmetric keys for encrypting the data with each node. When the client wants to send some data, it splits them into fixed size chunks (called Tor cells) and encrypt them with multiple layers of encryption by the symetric keys negotiated in previous step. You may find the encrypted Tor cell scheme on the 2.3 picture. Each node in the circuit decrypts the Tor cell and therefore removes one layer from it. The exit node decrypts the last layer of encryption, reads the original data of the client and sends the packet to destination address. The reponse is than again encrypted by symetric keys and sent back by the same circuit. This way, each node only knows its neighbours in the Tor circuit, but doesn't know the rest of the circuit. Tor employs multiple additional features for Tor relay selection and circuit building, to prevent attacks by behavioural and statistical analysis.

Cryptomarkets are running as hidden services on Tor network. Tor hidden services is a feature of Tor for providing a service (eg. a webserver, IRC server etc.) over Tor network without disclosing the server's IP address. The hidden services have special top level domain *.onion*.

When administrator of service wants to be accessed, he creates an RSA keypair. The public key (in base-32 encoding) is the service identifier, through which can the service can be accessed by accessing URL *identifier.onion*. The administrator chooses some Tor nodes as introduction points and establishes a circuit to each one of them. The administrator then uploads a descriptor of the service to 6 different hidden service directories (HSDir). HSDir is a Tor node with a special flag that was assigned to it by Tor authorities. The selection of HSDirs changes every 24 hours and can be deterministically computed from the timestamp and the hidden service public key. The descriptor

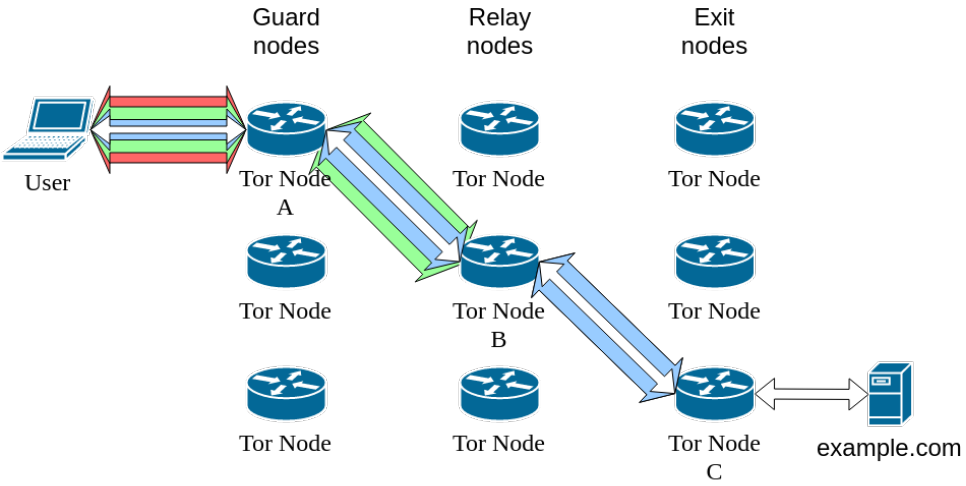

Figure 2.2: Tor routing schema

contains the hidden service public key and signed list of introduction points.

When a client starts connecting to the hidden service, he uses service's domain name to compute HSDirs with service descriptor. It then downloads a service descriptor from HSDir with list of introduction points. Client also generates one time secret, picks a random Tor node, asks it to act as randezvous-point and tells it the one time secret. The client then sends a message to the introduction point of the service (encrypted by service's public key) with the information about randezvous-point and the one time secret. The service receives the message, connects to randezvous point and uses one time secret to match with client. Randezvous-point then merely relays messages between the client and the hidden service. All the connections mentioned above can be seen on the **??** picture. Note that there is no direct connection between two servers during the hidden service set-up, all communication between servers is routed through regular Tor circuits. Randezvous-point just acts as a relay between two Tor circuits.

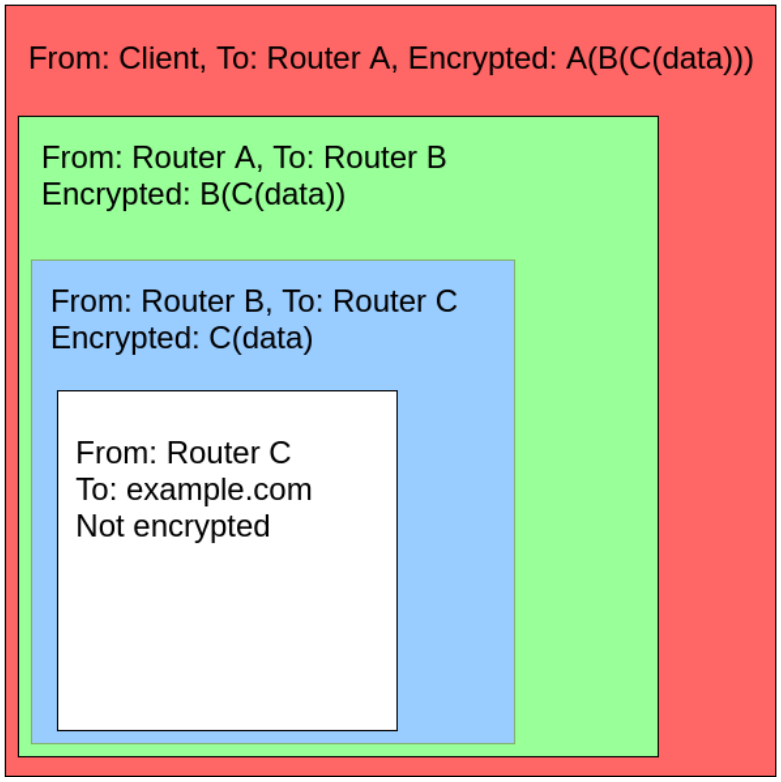

Figure 2.3: Tor packed encryption schem

2.4 PGP

PGP [33] 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 an open source, with the intention to provide an open, widely used standard for encrypted communication. Nowadays, the PGP program is not an open source anymore, but the standard is used by open source GPG software.

PGP uses public key cryptography. Unlike symmetric cryptography, public key cryptography uses two different keys for encrypting and decrypting. User generates a pair of keys, a public key for encrypting emails sent to him and a private key, which the user keeps for himself and uses for decrypting messages encrypted with the associ-

ated public key. The user also publishes his public key, so other users can send him encrypted messages.

PGP is used in the context of online drug markets as a means of secure communication between vendors and customers. Both vendor and customer have 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 users to keep their communication private also from the administrators of the marketplace.

3 Related work

3.1 Blockchain analysis and de-anonymization of bitcoin addresses

Multiple papers and tools were published regarding the blockchain analysis. Blockchain contains all bitcoin transactions and anyone can simply check the source and destination addresses of every transaction in the system. It is heavily encouraged for users of blockchain to use multiple bitcoin addresses and every major bitcoin wallet does so. We will focus on ways of detecting addresses belonging to the same user.

The authors of one [21] of the first research articles related to blockchain analysis parsed blockchain files to create a graph of bitcoin transactions, with vertices as addresses and edges between them representing bitcoin transactions. They created so-called user graph by clustering addresses belonging to the same user, using simple heuristics that the owner of all input addresses used in a transaction must be the same. The first version of this article was published in 2011 and dealt with a much smaller number of people using bitcoin and smaller transaction graph. Their analysis also focuses on deanonymization through multiple aspects of bitcoin protocol, while this thesis focuses on deanonymization from transaction graph and public data.

Androulaky [4] performed clustering using two heuristics. The first one is from the previously mentioneed article [21] – all inputs of transaction belong to the same user. The second heuristics is clustering some outputs of the transaction with its inputs. Most transactions have two outputs, one is owned by the transaction recipient, the other one (the aforementioned change output) is owned by the sender. The mechanism of change output is described in chapter 2.2.1 and the heuristics in chapter 4.4. If the transaction has two output addresses A and B, A has appeared in blockchain before and B has not, then it assumes B being the change address, as it's a common practise for the wallet software to generate new, never used addresses for change outputs. They also employed multiple clustering techniques based on the behaviour of users. They tested the success of their clustering techniques in their simulated bitcoin environment.

Advanced and similar work [26] was done by creators of Biodine application. They downloaded the blockchain, transformed it to a database and performed clustering to get a graph of transactions between users. Then they developed a tool for scraping data from multiple locations (Bitcointalk and bitcoin-OTC forum) to link off-chain data and identities to bitcoin addresses. They tested the tool on few popular transactions related to the seizure of Silkroad marketplace.

Similar work [13] to this thesis was done by Fleder. This paper use data from Bitcointalk, the most popular bitcoin forum. They apply a simple algorithm to group multiple bitcoin addresses belonging to one user together. Then they use the scraped data to show that some of the Bitcointalk users were using Silkroad marketplace or other popular services accepting bitcoin.

Ron and Shamir [24] focus on bringing interesting statistics about bitcoin transaction graph and provide a detailed analysis of really big bitcoin movements (more than 5000 BTC) through transactions in the network. In their other study [23], they analyzed transactions performed by Ross Ulbricht, who was the administrator of Silkroad marketplace. The FBI published their bitcoin address, which they used to collect all seized bitcoins from Ross Ulbricht. They took the size and frequency of transactions related to the seized bitcoins prior to the seizure and compared it to the estimated income of Silkroad. They found discrepancies between the relatively stable income of Silkroad marketplace and unstable balances in bitcoin addresses that were seized by FBI. They conclude that FBI seized around 22% of Ross Ulbrich bitcoins and found addresses possessing some of these bitcoins, that have not been used since Ross' arrest.

In contrast to previously mentioned papers, Meiklejohn [16] haven't only passively scanned the blockchain, he actively sent bitcoins to addresses of well-known services to track their bitcoins in the following transactions executed by the service. He also used the same two heuristics for clustering addresses as Androulaki [4]. He concluded that the network does not offer enough anonymity and large transactions can be traced.

All of the previously mentioned works had to deal with much smaller transaction graph, as the usage of bitcoin grew significantly over the last year. As far as I know, no previous work utilizes as many sources of data as this work. Also, the aim of our tool is to be able to

detect, whether some specific bitcoin address of investigator's interest was somehow connected to Valhalla drug market, or at least find nearby deanonymized addresses in the transaction graph to help with further investigation.

3.2 Behaviour of drug markets users and operators

Emerging cryptomarkets brought the attention of the scientific community and lots of articles have been published related to the phenomenon of drug trafficking via the Internet. Most of these papers were investigating the topic from the social and criminology perspective, performing qualitative analysis [3][6][14][12][29][31][15].

There are only a few articles focusing on statistically describing fully operating drug market and its vendors by collecting and analyzing data from a cryptomarket webpage. Short description of works like that follows.

Aldridge [2] scraped Silkroad in September 2013, the most popular cryptomarket of that time. He focuses on how the vendors and buyers perceive the risk of arrest and attempt to downsize it. He concludes that users of cryptomarket are aware of the risks both related to their physical and online activity and actively reduce their risk.

Decary [9] focuses on investigating, how loyal the buyers on cryptomarkets are to their vendors. As it seems, popular vendors successfully build their loyal customer base. These findings make sense, given the natural health risks related to using drugs, customers prefer vendors with high reputation and trust.

Broseus [7] restricted his research to vendors shipping from Canada and tracked their activity through multiple markets. His findings conclude, that the same vendors use the same usernames and sometimes PGP keys on multiple marketplaces because reputation is highly valued in these cryptomarkets, therefore vendors try to keep it when moving to the new marketplace. Also by his findings, some vendors are highly specialized in selling one category of drugs, while others offer a wide range of drugs.

Article by Doliver [11] is most similar to our work. They scraped and analyzed two popular cryptomarkets, Agora and Evolution, and quantitatively assessed the characteristics of vendors from both mar-

kets, focusing on the difference between different markets' vendor populations.

All previously mentioned works were analyzing cryptomarkets that are not operational nowadays, while this work focuses on describing the vendors from currently operating drug market Valhalla and examining, whether the behaviour of vendors or the nature of cryptomarkets has changed significantly. We also not only passively scraped cryptomarket's webpage, but created a user account on the marketplace as well and sent/received bitcoins from the marketplace in order to get data about the cryptomarket's flow of money.

4 Methods and tools of data retrieval and analysis

4.1 Valhalla cryptomarket webscraping

We scraped data from Valhalla cryptomarket during January and February 2018. We used official Tor daemon and software called privoxy to create a local proxy that will redirect all incoming traffic through Tor network. The privoxy was essential, as Tor daemon creates SOCKS proxy, which can not be used by wget. Thus, we created an HTTP proxy by privoxy, which redirected the traffic through Tor SOCKS proxy. We did not have to implement any login and captcha solving functionality for web scraping Valhalla market, because all the webpages about market listings, vendors and feedbacks are available without logging in.

Addresses of all market listings are in pattern `http://Valhallaxmn3fydu.onion/products/xxx` where `xxx` is a number incrementing with each new listing. The last listing had number 103770 and only 25309 numbers lead to the valid listing page. Rest of the numbers lead to 404 error. We believe these numbers refer to listings that were delisted by the vendor or administrator. Vendor profile pages were in format `http://Valhallaxmn3fydu.onion/xxx` and their reviews in the format `http://Valhallaxmn3fydu.onion/xxx/palautteet` where `xxx` is a vendor nickname. The description of format of the scraped data is in tables 4.1, 4.2 and 4.3.

We wrote several scripts for webscraping and parsing these data, their description is also in chapter 5. The scripts iterate through all of the listings and download them using wget command line tool. After downloading all the listing pages, the downloaded files are parsed using python and common Linux command line tools (cat, grep, cut, sed). We were not using python HTML parsing libraries (like beautifulsoup) for parsing downloaded webpages because HTML elements of Valhalla web pages don't have any unique identifiers and so these libraries bring us no significant advantage.

By scraping the listings webpages, we obtained unique vendor nicknames, which we used for downloading and scraping vendor's

Table 4.1: Data scraped from listings

Variable	data type
Vendor's nickname	string
Subcategory	string
Title of listing	string
Price	float

Table 4.2: Data scraped from vendor profile pages

Variable	data type
Vendor's nickname	string
count of positive feedbacks	integer
count of negative feedbacks	integer
revenue	integer
PGP key	string, not mandatory for all vendors
Country vendor ships from	string, not mandatory for all vendors

profile and feedback pages. The shortcoming of this method is that we were able to download and analyze only those vendors having at least one active listing at the time of data collection.

4.2 Valhalla cryptomarket metadata scraping and analysis

We thought that metadata from the photos of drugs, which vendors upload to show in their listing, might contain some information leading to identity disclosure. We downloaded 20 images from Valhalla market and examined EXIF metadata using `identify` command line tool.

Only metadata directly depending on image content, like the amount of red, green and blue colours were different for different images. Metadata that could potentionally help disclosing user identity, like date of creation nad modification, signature and name of software version

Table 4.3: Data scraped from feedback pages

Variable	data type
vendor nickname	string
rating	1-5
date	Timestamp, days resolution
buyer's nickname	string of length 4
money the buyer spent	int
trades the buyer did	int

were the same for all images. The software version contained exactly this string: `"ImageMagick 6.8.99 Q16 x86_64 20170731"`

ImageMagick is popular software library used for manipulating images, so it seems that market automatically rewrites EXIF metadata in uploaded images in order to protect privacy of users. To test this hypothesis, we created vendor account and uploaded an image with custom-made EXIF metadata. We than downloaded the uploaded image from webpage and saw that EXIF metadata were indeed overwritten.

We tested whether every transaction happening on drug market has its counter transaction in bitcoin blockchain. We deposited some Bitcoins to cryptomarket and bought a licit virtually deliverable product (Guide on weightlifting) and checked if anything happened to the deposited bitcoins. There was no follow up transaction happening for weeks after the deposit transaction was done. This means that market doesn't transfer bitcoins when service or goods are bought. All of bitcoin transactions that these drug markets do are accepting bitcoins deposits, sending bitcoins to withdrawing users and possibly bitcoin tumbler transactions, if they use such service.

We tried to scan ports of Valhalla cryptomarkets server and fingerprint their web server, in order to find any vectors for further information gathering. We scanned Valhalla server using netcat and found that the only opened ports are 80 (redirects to HTTPS) and 443 (HTTPS), which is used by web server. The webserver was popular software called nginx, as we detected both from HTTP headers and from webserver fingerprinting tool *httprecon*. The result of port scan

and web server fingerprinting doesn't indicate any new vectors for gathering data about cryptomarket.

Some vendors have published their PGP keys on their profile page. We scraped 150 PGP keys and tested them for ROCA vulnerability via python module roca-detect. None of these keys were vulnerable. All these PGP keys were searched for User-Id in metadata and found user-Ids were seached by Google. None of the searches for user-ids (both nicknames and mail addresses) returned usable results. Some of the usernames were just briefly mentioned in some posts on anonymous forums like 4chan.org.

4.3 Publicly available data scraping

In order to have some bitcoin addresses and bitcoins linked to identities, we searched the Internet for pages, where bitcoin addresses are tied to offline or virtual identities. The sites that we have decided to scrape were Bitcointalk forum, Reddit, Twitter and blockchain.info. The Bitcointalk is the most popular Internet forum related to cryptocurrencies. URL address of profile page on Bitcointalk contains the profile number, starting at 1 and being incremented by 1 for each new profile. It is therefore quite easy to iterate over all forum's profiles, including the ones with no posts, and check whether they have associated bitcoin address. The script bitcointalk-scraper.py visits profile pages of all profiles on the forum and scrape usernames and bitcoin addresses. Scripts for scraping Reddit and Twitter contain several hard-coded phrases like "Donate bitcoin" and "bitcoin address" and scrapes result of searches for such phrases. Blockchain.info is a webpage that serves primarily as bitcoin blockchain explorer. Secondary, it gathers multiple statistics about bitcoin blockchain and allows third parties to have their bitcoin address and identity listed on their webpage as well. The generated identities are stored as rows in csv files that have 3 columns:

1. bitcoin addres
2. URL where was the address scraped
3. nickname of the associated identity

4.4 Using own transactions to identify market wallets

The Valhalla cryptomarket generates unique deposit address for each new user. When the user wants to buy something, he is required to deposit bitcoins to deposit address, which will top up his account balance, which he can in turn use to buy the thing he wishes. Vendor or user can also request withdrawing their bitcoins from the cryptomarket. When he does so, the cryptomarket decreases its account balance and sends bitcoin to the user's address.

For identifying bitcoin addresses owned by the Valhalla cryptomarket, we deposited small amount of Bitcoins to our user accounts on Valhalla cryptomarket and performed multiple withdrawals by smaller amount than the deposited one, to receive bitcoins back. When we requested bitcoin withdrawals, we received bitcoins from different addresses than the deposit one. We consider the deposit address and all addresses we received bitcoins from during withdrawals as owned by cryptomarket.

Knowing these addresses alone wouldn't help us to decide whether someone sent or received bitcoins from the Valhalla cryptomarket. It is a common practise for any service accepting bitcoins to use multiple bitcoin addresses. For each Valhalla market account, a unique deposit address is required, so that the system can associate a bitcoin transaction with a Valhalla market account. In order to solve this issue, we used two heuristics used in the previous works [4][21] for clustering bitcoin addresses belonging to same user.

The first heuristic simply states that all inputs of one transaction belong to the same user. This is logical, since users generally don't share their private keys and collaborate on creating one transaction. When transaction has multiple inputs owned by different bitcoin addresses, we assume that all of these addresses are owned by one identity. The first heuristic's scheme can be found on figure 4.1.

The second heuristic focuses on detecting a change address described in chapter 2.2.1. The goal of the second heuristics is to detect change addresses in the transactions. When one transaction has two outputs with two different addresses, we assume one of them is a change address and is owned by the sender. We then search through the the blockchain for the first occurence of each output address of the transaction. If one of the output addresses had been used before the

transaction and the second has not, we can then safely assume, that the second one is a change address [4]. The second heuristic's scheme is on figure 4.2.

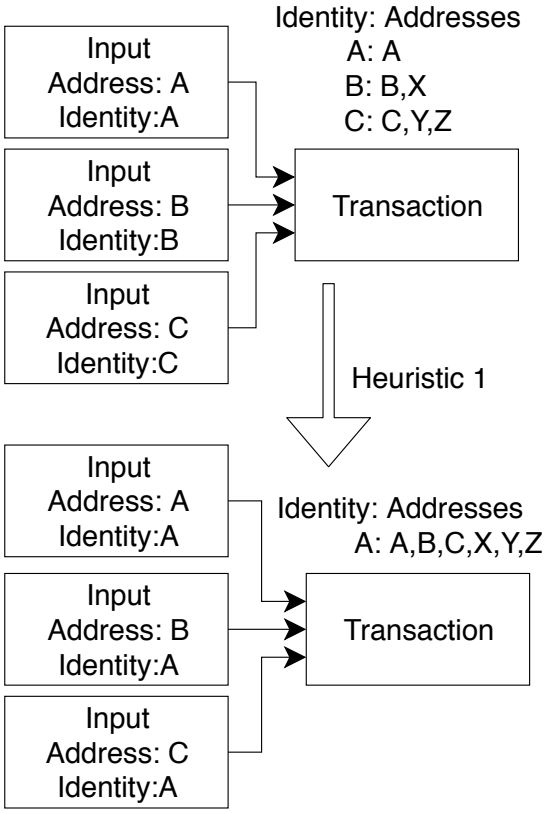

Figure 4.1: Schema how first heuristic works

Both of these heuristics are pretty strict and have just a slim chance of falsely merging addresses not belonging to the same user [4]. Although the concept of multisignature wallets and a few proposed anonymisation methods could make these heuristics misleading (like CoinJoin mixer and dark wallet), they only contribute to a negligible percentage of all transactions.

We measured the success of heuristic clustering in two ways. We had two accounts on Valhalla market, each performing one deposit of 0.011 BTC and immediately 10 withdrawals of 0.001 BTC, with roughly 0.001 BTC spent on cryptomarket fees. For the first account,

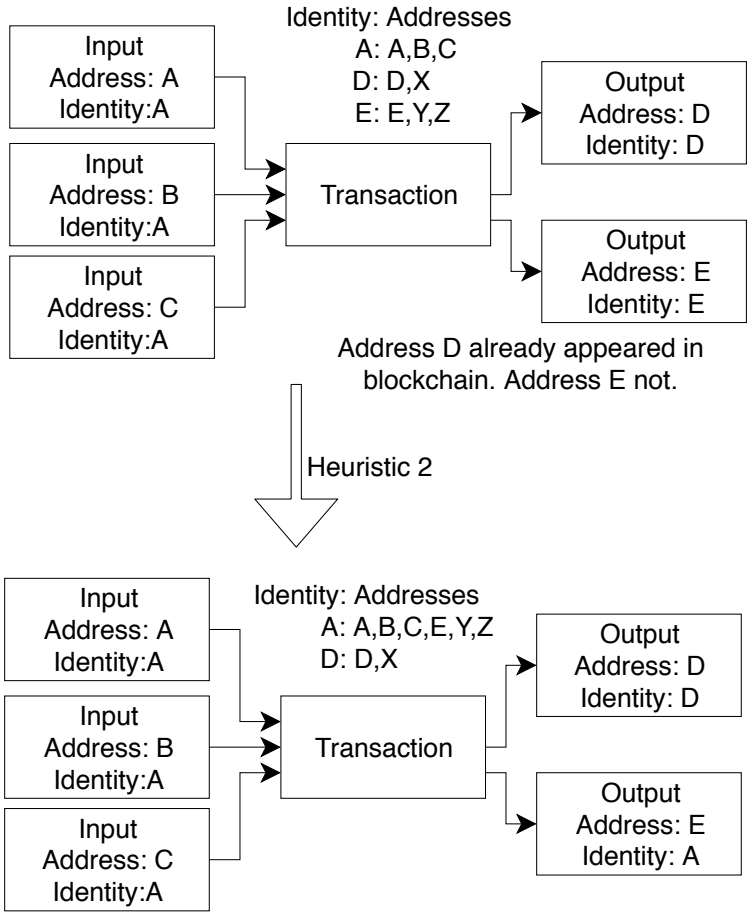

Figure 4.2: Schema how second heuristic works

we performed these transactions on January 11th 2018, for the second one on February 13th 2018.

We created a new bitcoin wallet via Mycelium mobile app, deposited bitcoins into it from bittrex exchange and used the wallet only with the first account. The second bitcoin wallet, created by coinomi mobile app, was used only by the latter account. Bitcoins were deposited to it from coinmate.cz bitcoin exchange. We installed both these wallets and generated a new, never used bitcoin address in each one, and also received bitcoins from two different exchanges in order to not affect our measurement by being the owners of both bitcoin addresses.

We then ran our application and looked into the neo4j database, storing blockchain transaction graph and associated identities. We checked the cryptomarket's addresses associated with our deposits and withdrawals, whether they were clustered as one identity. The table 4.4 shows, how many identities were associated with given deposits and withdrawals. The heuristics could not cluster our deposit addresses, because the deposit addresses just received bitcoins from us and not send them somewhere else till that time. Only 2 identities managed to contain addresses from both accounts.

Afterwards, we have checked, how much bitcoins have these identities received between 14. 1. and 14. 2. We scraped the Valhalla cryptomarket between 14. 2. and 15. 2. and received data from feedback pages, which we used to estimate the minimal amount of bitcoins users spent over the previous month on Valhalla market. We estimated in chapter 6, that users spent at least 527 730 EUR during that one month period. With average bitcoin price 8 219EUR it roughly equals to 64.2 bitcoins. The amount of bitcions these identities received is also in table 4.4. Our heuristics were not enough to cluster majority of market's transactions/addresses as one identity. Hovewer, with just two accounts, 2 deposits and 20 withdrawals, we were able to identify 5.71(9%) bitcoins out of estimated 64.2 bitcoins the market received within that one month timeframe.

Table 4.4: Mapping of found identities to addresses used in deposits and withdrawals

Identity	Deposits:withdrawals from first account	Deposits:withdrawals from second account	received bitcoins
1	1:0	0:0	0.011
2	0:0	1:0	0.011
3	0:4	0:1	2.15
4	0:2	0:0	0.7
5	0:1	0:1	1.5
6	0:1	0:0	0.15
7	0:1	0:0	0.1
8	0:1	0:0	0.1
9	0:0	0:3	1.22
10	0:0	0:2	0.12
10	0:0	0:2	0.37
11	0:0	0:1	0.3

5 Application for searching nearby identified addresses

This chapter describes our application for analyzing the gathered data. The application architecture is depicted in figure 5.1. It consists of multiple python scripts, divided into 5 categories.

- Scripts for scraping Valhalla market
- Scripts parsing bitcoin blockchain and data from publicly available sites mentioned in section 4.3.
- The script for importing data to a database, creating indexes and running heuristics mentioned in chapter 4.4.
- Webserver that handles GUI requests and retrieves data from database.
- The web GUI written in HTML/JS/CSS for sending requests to webserver, visualisations and searches in the retrieved data.

Scripts for scraping Valhalla market in order to get data used in next chapter are in statistics folder. Script *listing-scraper.sh* goes through Valhalla listings and downloads the HTML source code of these pages. Script *listing-parser.py* process the downloaded pages and generates the *listings.csv* file with information about listings. *Seller-scraper.py* opens *listings.csv* and downloads HTML source of vendors' profile pages. *Seller-parser.py* proccesses the downloaded vendor's profile pages and generates *sellers.csv* and *feedbacks.csv*. The variables and data types stored in these csv files are described in chapter 4.1 in tables 4.1, 4.2 and 4.3. We used these generated csv files for assembling Valhalla cryptomarket statistics contained in the last chapter.

The scripts for obtaining data for the application can be found in the *identity–scraping* folder. Each one of the scripts scrape webpage mentioned in its name and generates a csv file with 3 columns: bitcoin address, URL where the address was scraped and the associated identity (eg. nickname). Script blockchain_to_csv.py uses bitcoin-core API to iterate through bitcoin blockchain and generate several csv files

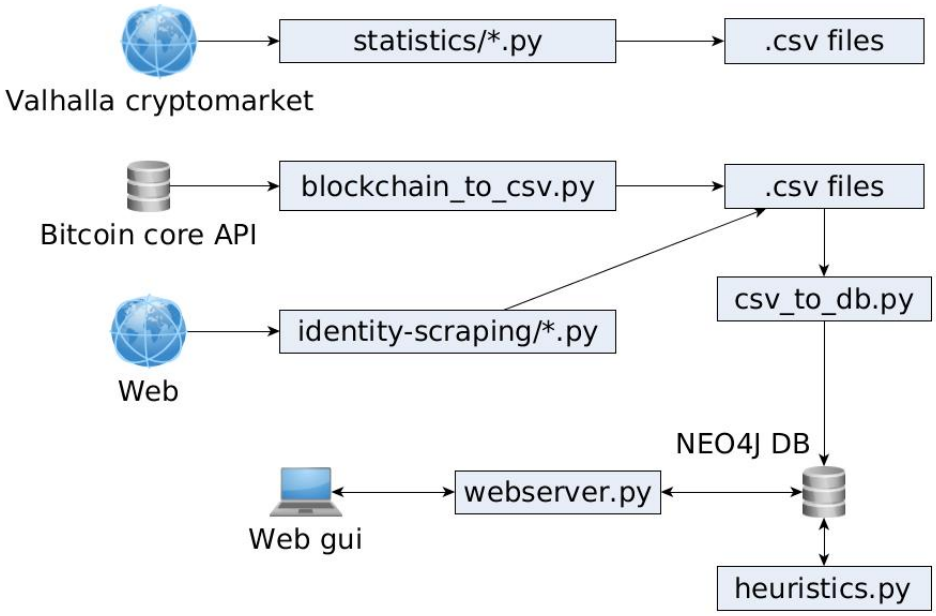

Figure 5.1: Architecture of application

with the transaction data and bitcoin blockchain addresses, which are later imported into neo4j database, as well as files with identities.

In order to create a tool for effectively searching and visualising both blockchain and web-scraped data, we need to store the blockchain and identity data locally in a way allowing an effective performance of the heuristic analysis. The natural representation of transactions happening between bitcoin addresses is a graph, so we decided to import all the data into Neo4j graph database. It's one of the most widely used graph databases, with native support for common graph algorithms. Simple representation with addresses as nodes and transactions between them as edges was not sufficient for our case, because we needed to represent outputs, inputs and time of transactions in the database in order to be able to compute previously mentioned heuristics. The graph schema is in figure 5.2. The entities in the schema are represented as vertices in the database and relationships between them are edges.

The importing script *csv_to_db.py* is responsible for parsing previously generated csv files and importing the data into neo4j database.

Having csv files as interminiary step between parsing blockchain and importing it to the database proved to be useful for debugging application, while it also provided an easy way to use the application with different blockchain or custom-generated data.

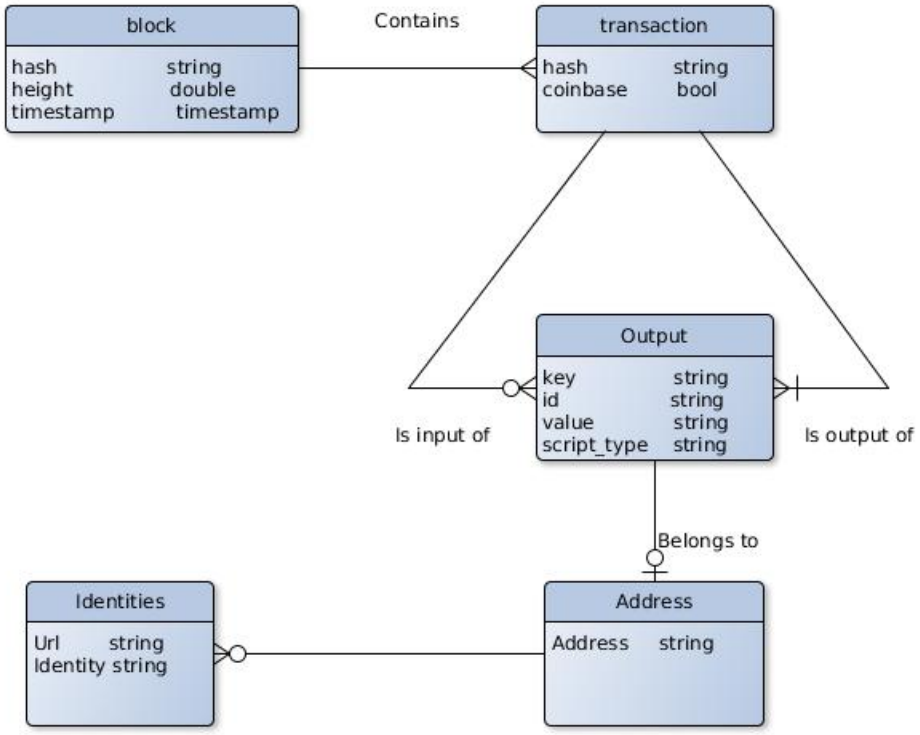

Figure 5.2: Neo4j database ER diagram

Script *heuristics.py* connects to database and runs the heuristics mentioned in the chapter 4.4. It then labels each address in our blockchain transaction graph with a unique ID, representing unique identity. The outputs/inputs owned by the each address are labeled with the same ID. Script then runs the first and then the second heuristic, each time rewriting IDs of merged address(and its outputs/inputs), so that addresses marked by the heuristic as belonging to the same identity have the same ID. The *webserver.py* connects to local neo4j database with the schema and data generated by previous scripts, and starts local flask webserver that provides a web GUI on port 5000.

The GUI has one input form, where a user of our application is supposed to write the bitcoin address he is interested in. The GUI then shows information about the inserted address, nearby addresses and found identities. It consists of 3 parts, input form, a table, and two pie charts. The screenshot of GUI can be seen on picture 5.3. The first part is a simple HTML form where user submits bitcoin address he is interested in. After submission, the application searches in the neo4j data and after a while, the other parts of GUI are updated with information about given address. The table in upper right shows a list of nearby addresses with known identities. The distance is number of transactions between the inserted address and the address found in our table. It also shows the URL where the identity was found and amount of bitcoins that were transferred between the searched address and the address in table. In the lower part, there are two pie charts. The first pie chart shows the amount of bitcoins that were received by the inserted address and what percentage of received bitcoins was the application able to link to some identity. The second pie chart shows the same information for outgouing bitcoins, the percentage of bitcoins that were sent from inserted address and ended up in address that we were able to find associated identity.

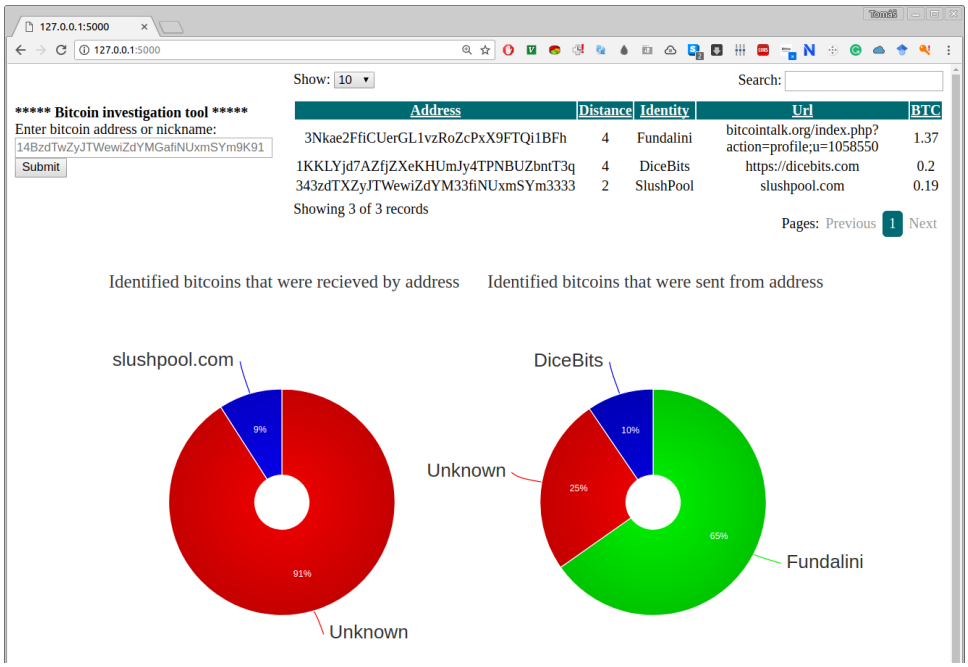

Figure 5.3: Screenshot of Gui

6 Valhalla cryptomarket statistics

We managed to scrape 25309 listings, 981 vendors and 6381 feedbacks. There were 17 314 (68 %) listings related to selling drug substances, the rest were related to ebooks, premium accounts, guns, fake IDs etc. Vendors had from 1 (90 vendors) to 1083 (1 vendor) active listings, with average of 25.69 (SD = 58.06) and median of 11.00 listings per vendor. Most vendors tend to have just few active listings, as can be seen on histogram 6.1. The histogram does not show 19 vendors with more than 200 listings.

The vendors were shipping drugs from 39 distinct countries. The occurence of countries as the shipping origin can be found in the table 6.1. The table 6.2 lists countries of vendors that have achieved 10 000EUR (150 vendors, 15%) or more in their revenue. Valhalla was originally estabilished as a local Finnish crypto market, which might be the reason for surprisingly many vendors shipping from Finland and also large percentage of Finns among high revenue vendors.

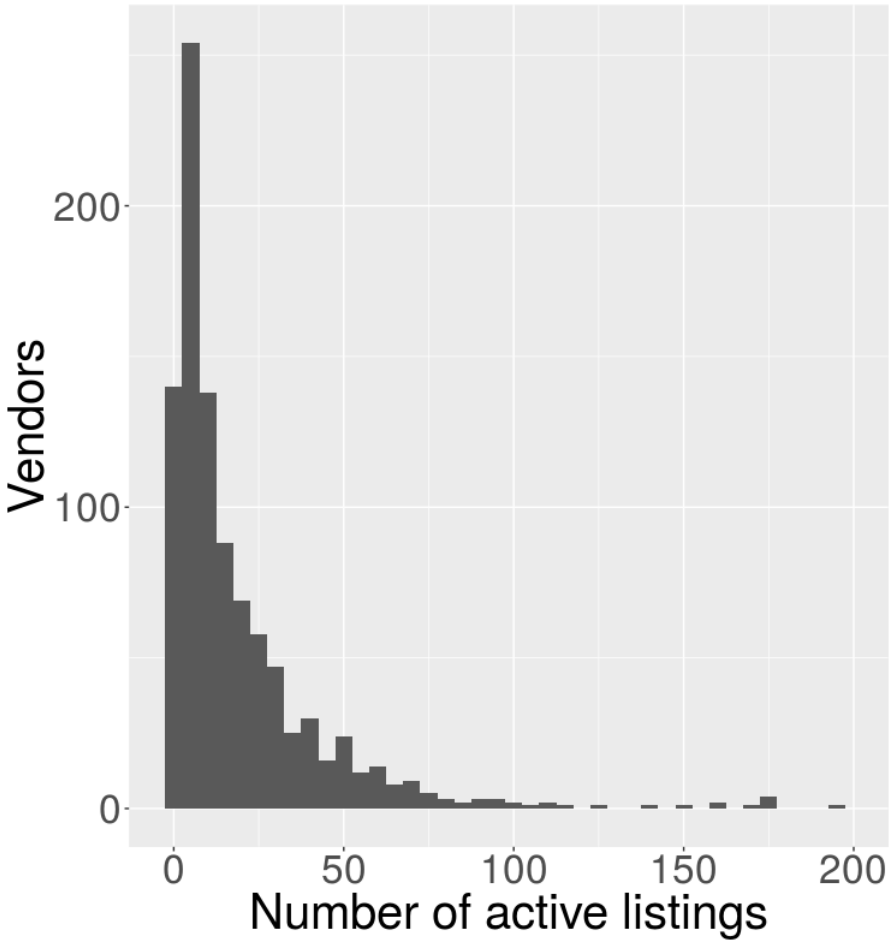

Figure 6.1: Histogram for number of active listings vendors have

They might be on the market since the time it was not internationally popular and had more time to generate revenue.

We looked at the total revenue of vendors, as depicted in figure 6.2a. The total revenue mentioned in the vendor's profile page is capped at 10 000 Eur. There are two major groups of vendors regarded to their revenue. Majority (598 vendors, 60%) of vendors have not had any positive or negative feedbacks and all of these vendors (except for 3) earned less than 300EUR on cryptomarket over their lifetime. Please note we were able to scrape just the vendors with active listings, so there might be much more vendors who stopped using Valhalla after

Table 6.1: Countries that vendors ship from

Country	Count of vendors
Finland	28
United Kingdom	23
United States	16
Germany	13
Netherlands	12
France	5
Norway	5
Spain	4
Canada	4
Australia	3
Poland	3
Others	31
Unknown	834

Table 6.2: Countries high revenue vendors ship from

Country	Count of vendors
Finland	24
UK	13
USA	10
Germany	6
Netherlands	6
Norway	4
Others	20
Unknown	69

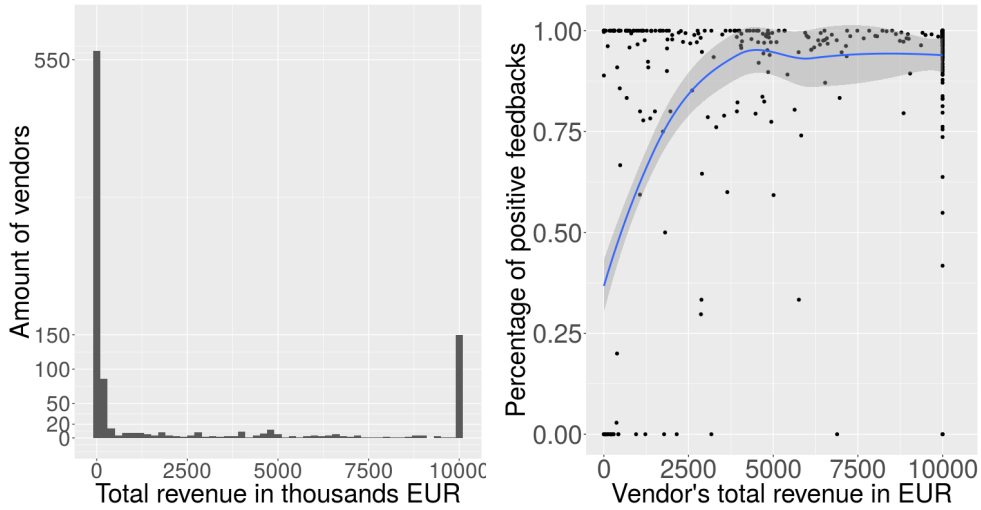

Figure 6.2: Distribution of vendors by their revenue

not being able to earn substantional amount of money. The second group is 150 vendors (15 %) who earned 10 000 EUR or more.

Vendors, who had high revenue, maintained high ratio of positive feedbacks. When counting only vendors with more than 10 reviews (304 vendors, 31%), the mean is 93.7% and median 98.3% of positive feedbacks. Vendors with 1 to 10 reviews had on average 27.2% of positive feedbacks. Figure 6.2b shows the the distribution of vendors with similar positive feedbacks ratio, their revenues and its LOESS curve. These findings indicate, that market is heavily competitive and having bad reputation during first few sales shuts down any opportunity for the vendor to continue selling his goods. It also concludes that majority of vendors don't succeed or use the market merely sporadically.

The figure 6.3a shows distribution of prices of all listings in the market (on the left graph) and the price distribution in feedbacks (on the right). Some listings might be counted multiple times or not at all in the right graph, because one listing might be bought several times or not at all and therefore generate multiple or no feedbacks. We expect that the price distribution in these feedbacks reflects more accuratelly the prices of the goods or services that are actually bought at market, because feedbacks can be given only by users who bought something from vendor. We calculated price of listings in EUR as price

in bitcoin multiplied by the average price of bitcoin for 30 days before the scraping (8219 EUR for 14.1.2018 - 14.2.2018, daily prices taken from blockchain.info).

The average price of listing is 336.2 Eur(1368.2 SD!) with median of 71.3 Eur, while feedbacks had average 82.7Eur(150.7 SD) with median of 55 Eur. There were 1533(6.05%) active listings with price greater than 1000 EUR but only 22(0.3%) feedbacks with price greater than 1000EUR. The price distributions in 6.3a show, that while there is a lot of free or very expensive listings, the majority(5937, 93%) of trades were between 0 and 200 EUR. There were 115 free listings on the market, but none of them was mentioned in the feedbacks. The free listings had titles like "CARDING SERVICE *FREE*", "Free Carding Tutorial 2017" and "FREE SAMPLE 84% MDMA CRYSTAL ROCKS" The titles of expensive listings don't indicate that these listings are somehow special, except for the higher amount of drugs offered.

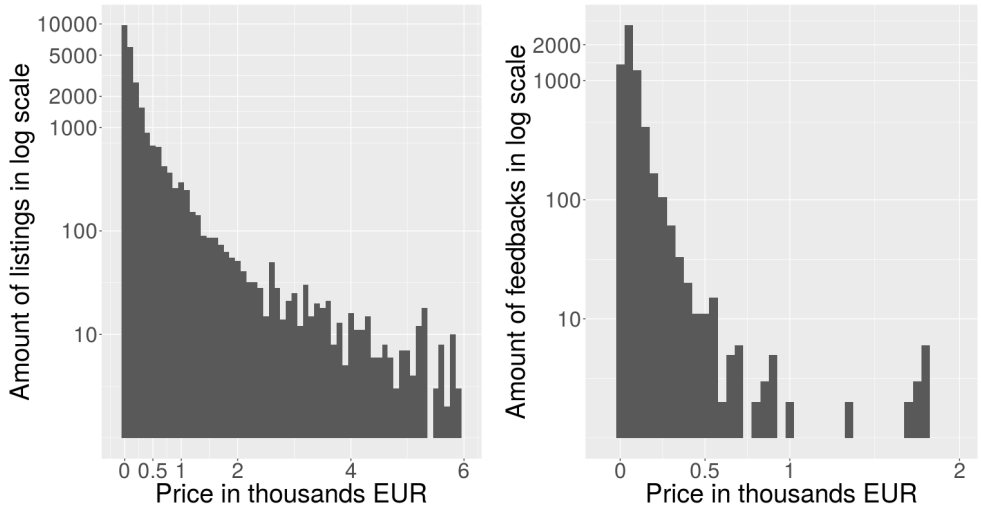

Figure 6.3: Distribution of prices in listings and feedbacks

We calculated the last month revenue for each category by summing up the prices over all feedbacks for listings in given category. This leads to minimal revenue, because feedbacks can be given only if the trade happened on cryptomarket, while trades happening with no given feedbacks are also possible. The total revenue over one month,

Table 6.3: Estimated monthly revenue for selected drug categories based on feedbacks

Category	listings	feedbacks	revenue in EUR	average price in EUR	market share
Cannabis	5139	1883	135693	72	27.6%
Stimulants	3493	1043	108157	103	22%
Opiates	1662	489	83135	170	16.9%
Pharmacy	2294	1104	62054	56	12.6%
Body building	679	402	31466	78	6.4%
Empathogens	2988	394	20344	51	4.1%
Other drugs	774	173	14350	82	2.9%
Psychedelics	1377	198	11796	59	2.4%
Other products	782	89	9106	102	1.8%
Self-defence	513	4	4635	1158	0.9%
Services	1662	35	4061	116	0.8%
Dissociatives	290	38	2317	60	0.5%
Classifieds	649	18	2182	121	0.4%

amount of listigs, feedbacks and market shares for given categories can be found in the table 6.3. Market share is a revenue of category divided by total revenue from all categories, which was 527 730 EUR. The average price of occured trades was similar in each category with two outliers. Opiates, which have a high price/dose ratio and self-defense, which only had 4 trades, 2 of them custom listings and 2 of them guns. Custom listings contain no special tag and no description of any goods nor services. They contain just titles like "custom listing for Paul", "custom 013" and so on. We found 300 listings with word "custom" in their title. We presume that vendors and buyers use these custom listings in order to be able to protect themselves by cryptomarket escrow service, in case of a pre-arranged deal.

Most of the vendors had listings in only a few categories. The graph 6.4 shows the amount of vendors with given revenue and number of

different subcategories they have active listings in. Most vendors sell goods only in a few categories, hovewer, vendors with 10 000EUR or more in revenue tend to have listings in more categories. The boxchart 6.5 shows the distribution of vendors revenue based on the category they were selling in. We have not counted vendors with 0 revenue for the boxchart. If a vendor was selling in multiple categories, we divided his revenue by the number of categories and counted him within all of them.

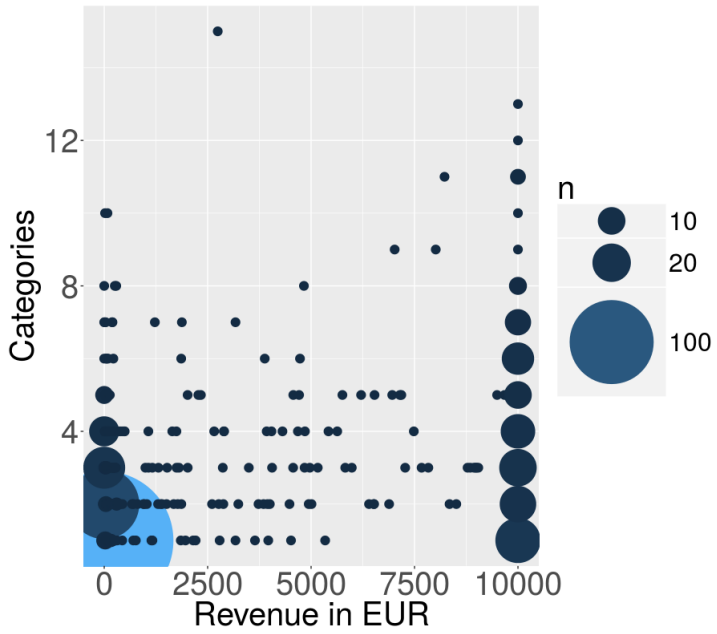

Figure 6.4: Amount of vendors with number of different categories they sell and their total revenue

Vendor pages contain counts of feedback gathered over lifetime, these were used in 6.2b. Vendors also have feedback pages, where feedbacks younger than 1 month contain first two and last two characters of buyers username, the rounded total number and price of trades performed by the buyer. We used these feedbacks to gather informations about buyers.

There are $36^4 = 1679616$ possible combinations of 4 alphanumerical characters. From our 6000 feedbacks, we gathered 2433 unique combinations of four letters from username, when added rounded

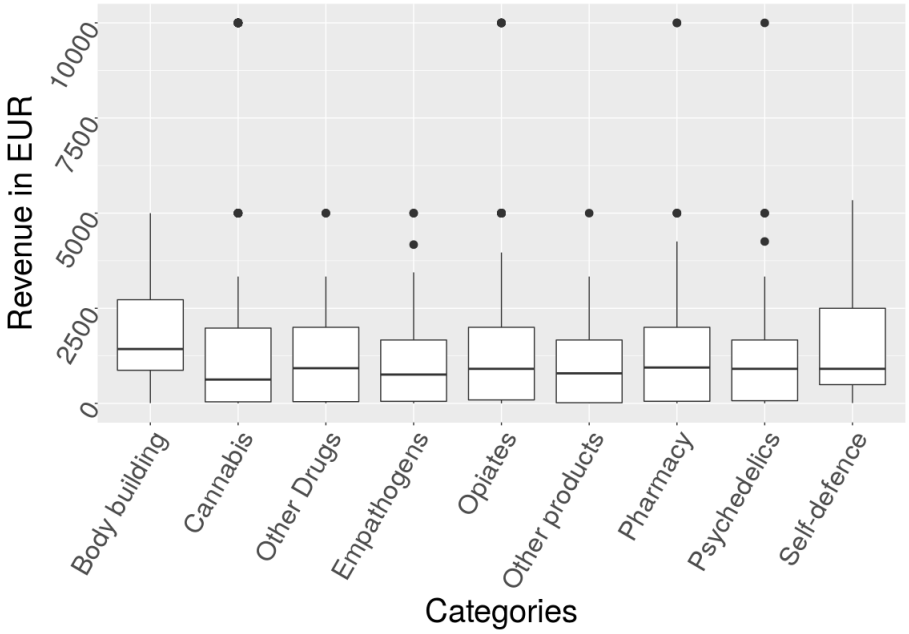

Figure 6.5: How much have Vendors earned based on category they sell

number of trades and buyer's total spending we got 2654. It is highly probable that some different buyers have the same username prefix, suffix, amount of trades and revenue, because when considering buyers total spending and trades, we managed to increase the number of identified buyers. Buyers total spending and total performed trades were rounded to only a few distinct values, so the following statistics are just estimates.

Each buyer have on average bought goods for 1176EUR through all his trades on Valhalla market. Just 129 (5%) of buyers had bought goods worth 10 000EUR or more, 1289 (50%) buyers spent 500EUR or less. This distribution is similar to revenue distribution of vendors, where vast majority of actors transfer just small amounts of value through Valhalla market and merely an extremely small percentge of actors trade on the market on a bigger scale. The average price of one trade (82.7 EUR) and the average of lifetime spendage (1176EUR) is way below the amount someone who actively resell drugs for profit

would buy. The distribution of buyers total spendage is on the figure 6.6a.

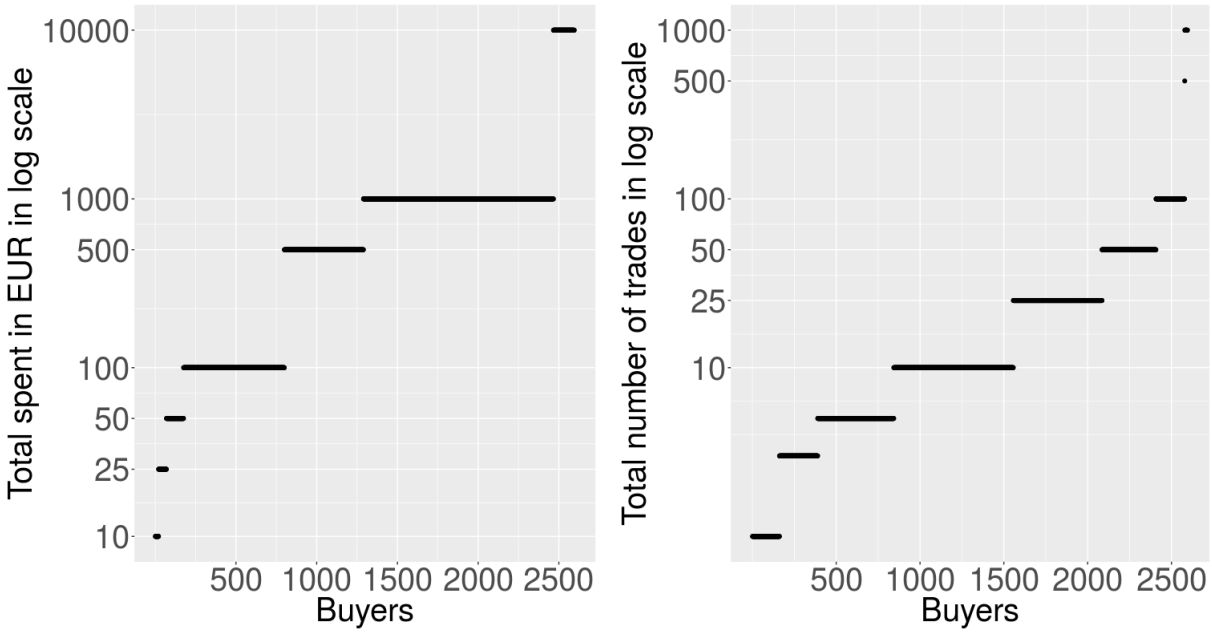

Figure 6.6: Distribution of buyers total spending and total performed trades

7 Conclusions and future work

The Valhalla cryptomarket scraping was successful and it brought up meaningful and intersting statistics about its vendors and users. Our statistical findings fit into the previously published desciptions of cryptomarkets, but we also found some statistics where Valhalla market is an outlier and came up with reasons why is it so. We scraped Valhalla market once, but the feedbacks data were related to the time frame of one month. Continously monitoring Valhalla cryptomarket for longer periods of time could enchance the statistical description with data about changes of various trends on this market over time.

Our application is fully working and production ready, it finds nearest addresses with associated identities for given address. The application could be extended in multiple ways. We used two heuristics that have the lowest risk of falsely clustering addresses belonging to different identities. The heuristics have not clustered all the addresses and transactions done within the cryptomarket, hovewer we found, that with just tens of deposits and withdrawals between us and the market, we were able to identify significant percentage of cryptomarket's cashflow over given month. The price that we paid in fees in these transactions was less than 100 dollars altogether. Perpetual depositing and withdrawals over a prolonged period of time seems financially feasible and might lead to disclosing majority of cryptomarkets addresses and transaction.

The application could be extended by adding more heuristics for clustering addresses from articles mentioned in chapter 3. These heuristics are not based merely on a bitcoin transactions graph, but also on expected behaviours of users and data obtainable by continously running one or multiple bitcoin nodes. These heuristics are less reliable, but offer more options to cluster addresses of the same owner.

The whole application backend can be run on consumer grade notebook. Backend database is indexed and utilizes memory well, the GUI requests take at most 30 seconds to process. This represents a great advantage of the application, hovewer the application set up takes a long time, for the simple reason that bitcoin blockchain consists of roughly 150GB of data and these must be inserted into database,

indexed and processed by heuristics. For setting the application up on server and having users access it remotely, different options viable for storing and retrieving data on server hardware can be considered (eg. a significant RAM extension to enable storing the whole database in RAM).

Bibliography

[1] *2010 Internet Crime Report*. https://pdf.ic3.gov/2010_IC3Report.pdf: Internet Crime Complaint Center, 2011.

[2] Judith Aldridge and Rebecca Askew. "Delivery dilemmas: How drug cryptomarket users identify and seek to reduce their risk of detection by law enforcement". In: *International Journal of Drug Policy* 41 (2017), pp. 101–109.

[3] Judith Aldridge and David Décary-Hétu. "Not an'Ebay for Drugs': the Cryptomarket'Silk Road'as a paradigm shifting criminal innovation". In: (2014).

[4] Elli Androulaki et al. "Evaluating user privacy in bitcoin". In: *International Conference on Financial Cryptography and Data Security*. Springer. 2013, pp. 34–51.

[5] *Bank Crime Statistics (BCS). Federal Insured Financial Institutions, January 1, 2010 – December 31, 2010*. Federal bureau of investigation, 2011.

[6] Monica J Barratt, Jason A Ferris, and Adam R Winstock. "Use of Silk Road, the online drug marketplace, in the United Kingdom, Australia and the United States". In: *Addiction* 109.5 (2014), pp. 774–783.

[7] Julian Broséus et al. "Studying illicit drug trafficking on Darknet markets: structure and organisation from a Canadian perspective". In: *Forensic science international* 264 (2016), pp. 7–14.

[8] Joseph Cox. "Staying in the shadows: The use of bitcoin and encryption in cryptomarkets". In: *Internet and drug markets, EMCDDA insights* (2016), pp. 41–47.

[9] David Décary-Hétu and Olivier Quessy-Doré. "Are repeat buyers in cryptomarkets loyal customers? Repeat business between dyads of cryptomarket vendors and users". In: *American Behavioral Scientist* 61.11 (2017), pp. 1341–1357.

[10] Roger Dingledine, Nick Mathewson, and Paul Syverson. *Tor: The second-generation onion router*. Tech. rep. Naval Research Lab Washington DC, 2004.

[11] Diana S Dolliver and Jennifer L Kenney. "Characteristics of drug vendors on the Tor network: a cryptomarket comparison". In: *Victims & Offenders* 11.4 (2016), pp. 600–620.

[12] Diana S Dolliver and Katherine L Love. "Criminogenic Asymmetries in Cyberspace: A Comparative Analysis of Two Tor Marketplaces". In: *Journal of Globalization Studies* 6.2 (2015), pp. 75–96.

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

[14] Nicolas Christin. "Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace". In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 213–224.

[15] James Martin. "Lost on the Silk Road: Online drug distribution and the 'cryptomarket'". In: *Criminology & Criminal Justice* 14.3 (2014), pp. 351–367.

[16] Sarah Meiklejohn et al. "A fistful of bitcoins: characterizing payments among men with no names". In: *Proceedings of the 2013 conference on Internet measurement conference*. ACM. 2013, pp. 127–140.

[17] Malte Moser, Rainer Bohme, and Dominic Breuker. "An inquiry into money laundering tools in the Bitcoin ecosystem". In: *eCrime Researchers Summit (eCRS), 2013*. IEEE. 2013, pp. 1–14.

[18] Marti Motoyama et al. "An analysis of underground forums". In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM. 2011, pp. 71–80.

[19] Satoshi Nakamoto. *Bitcoin: A peer-to-peer electronic cash system*. 2008.

[20] Masarah-Cynthia Paquet-Clouston. "Are Cryptomarkets the Future of Drug Dealing? Assessing the Structure of the Drug Market Hosted on Cryptomarkets". In: (2017).

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

[22] Paul Resnick et al. "Reputation systems". In: *Communications of the ACM* 43.12 (2000), pp. 45–48.

[23] Dorit Ron and Adi Shamir. "How did dread pirate roberts acquire and protect his bitcoin wealth?" In: *International Conference on Financial Cryptography and Data Security*. Springer. 2014, pp. 3–15.

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

[25] Kyle Soska and Nicolas Christin. "Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem." In: *USENIX Security Symposium*. 2015, pp. 33–48.

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

[27] Meropi Tzanetakis et al. "The transparency paradox. Building trust, resolving disputes and optimising logistics on conventional and online drugs markets". In: *International Journal of Drug Policy* 35 (2016), pp. 58–68.

[28] Marie Claire Van Hout and Tim Bingham. "Responsible vendors, intelligent consumers: Silk Road, the online revolution in drug trading". In: *International Journal of Drug Policy* 25.2 (2014), pp. 183–189.

[29] Marie Claire Van Hout and Tim Bingham. "'Silk Road', the virtual drug marketplace: A single case study of user experiences". In: *International Journal of Drug Policy* 24.5 (2013), pp. 385–391.

[30] Marie Claire Van Hout and Tim Bingham. "'Surfing the Silk Road': A study of users' experiences". In: *International Journal of Drug Policy* 24.6 (2013), pp. 524–529.

[31] Charlotte Walsh. "Drugs, the Internet and change". In: *Journal of psychoactive drugs* 43.1 (2011), pp. 55–63.

[32] Frank Wehinger. "The Dark Net: Self-regulation dynamics of illegal online markets for identities and related services". In: *Intelligence and Security Informatics Conference (EISIC), 2011 European*. IEEE. 2011, pp. 209–213.

[33] Philip R. Zimmermann. *The Official PGP User's Guide*. Cambridge, MA, USA: MIT Press, 1995. ISBN: 0-262-74017-6.