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 technica l 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 Infrastru cture MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures " (CESNET LM2015042), is greatly appreciated.

Abstract

This thesis has two goals. First goal is to perform quantitative statistical analysis of valhalla cryptomarket. We scraped 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 dark web. The aim of this tool is to hel p investigators with collecting intelligence on entities related to these drug markets. Users and operators of these mar kets employ multiple means to prevent their deanonymization. Cryptomarkets are operated ad TOR services, PGP encryption is often required to communicate between multiple parties and bitcoin is used as a way to pay for goods or servic es.

We scraped multiple publicly available social sites and websites related to bitcoin(twitter,bitcointalk, reddit, blockchain.info...) and drug markets thereself using python. We stored all these data into neo4j database, which is a graph database based on PostgreSQL. We created 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 n earest 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 trans actions to deposit and withdraw bitcoins.

Keywords

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

Contents

List of Tables

List of Figures

1 Introduction

The relative anonymity of internet offers an incentive for criminal parties to use internet as a tool for their activiti es. Internet facilitated some forms of existing crimes (selling drugs, guns and counterfeits, running Ponzi schemes) and also enabled many new types of frauds like hacking, phishing, carding and ident ity theft.

Publicly available statistics show, that cyber criminals are much less likely to be discovered and persecuted. In USA in 2010, there were 5628 robberies and the loot was recovered in more than 20% cases. [5] FBI recieved 303809 complaints related to cyber crime in 2010, resulting in just 6 convictions. [1] Criminals value their anonymity very high and use various means to prevent them from being caught by police forces. [24] [27] [3]

Big problem for criminals was getting the money they got from criminal activity to their possession, since that 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 hard to setup anonymous bank account and neither party could be sure about the origin of money they are recieving.

For bitcoin, there is no central authority requiring bitcoin address (bitcoin equivalent of bank account number) to be linked to person's identity and so criminals can use their anonymous connection to internet to both recieve and send bitcoins without disclosing their iden tity. Hovewer, all bitcoin transactions are publicly available and so each bitcoin can be tracked through the whole transactio Cryptomarkets are online marketplaces with listings offering illegal goods and services. Cryptomarkets are accesible via TOR network and users of cryptomarkets use PGP to communicate. n history. These mechanisms make it possible for cryptomarkets to publicly operate, yet be hard to reach by law enforcment. [8]

We scraped and examined data from Valhalla market, one of the most popular and well estabilished currently (February 2018) operating drug markets, in order to do statistical analysis of the scale of it's operations.

We collected data from multiple public sources related to drug markets and bitcoin transactions and explored possibilities to de-

anonymize drug market's users by linking publicly known identities to nearby bitcoin addresses. We also created a tool to visualize data obtained from these public sources and search for nearby bitcoin addresses.

1.1 Goals

The main goal of this thesis is to analyze Valhalla drug market and try multiple approaches to deanonymize users related to this drug market. The results of this work are data from cryptomarket and showing interesting statistics about the whole market as well as actors operating there. Also, we gathered addresses tied to some form of identification, like usernames, from social networks and publicly avai lable forums. Another goal of this work is to create a tool, that uses the data mentioned above to help investigator to disclose transactions , bitcoin addresses and identities related to online drug markets.

We managaed to partly fill our goals. We succesfully scraped data from cryptomarket and did an analysis. We scraped data about identities linked to bitcoin addresses and created a tool to visualize them. We were just partly succesfull with finding heuristics, that would cluster multiple bitcoin addresses belonging to the s ame owner.

1.2 Structure of thesis

XXX - mozna se zmeni lehce jmena/poradi kapitol, jinak povidani stejne

The following text describes individual chapters of this thesis. Chapter Related works gives overview of works already done on similar topics.

Chapter Technology and terms starts with a quick introduction to bitcoin and blockchain, which is used for paying on crypto markets. It also descibes how cryptomarkets work and tools that are used by cryptomarkets' users and administrators.

The chapter Methods and tools describes the proccess of collecting and storing the data from bitcoin blockchain, drug markets and pubicly available forums and social networks.

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

Chapter Statistics of drug markets consists of various statistics about drug markets, that were gathered during drug mar ket website scraping. It contains two parts, the first is focused on statistics related to cash flows, the second part is giving insight about non-money related statistics.

Chapter called Application describes the functionality, implementation, usage and possible future development of application for investigating bitcoin addresses, which was created as part of this thesis
.

Testing and verification is about the testing of the created application. Last chapter is discussion about achieved goals, problems of implementation and possible future improvements.

2 Related terminology

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 marketplace 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.

2.1 Bitcoin and blockchain

Bitcoin [18] is the first decentralized peer to peer cryptocurrency, created by 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 open source and can be downloaded and run locally. The entire history of transactions is stored in distributed public ledger called blockchain. Bitcoin combine multiple cryptography algorithms to achieve consensus among nodes on the state of blockchain. Anything once written in the blockchain can not be removed or modified. State of blockchain can be modified only in that way, that new block of transactions is added at the end of blockchain.

2.1.1 Addresses, bitcoins and transactions

In order to recieve and send bitcoins, user need to have a bitcoin address. Bitcoin address is simply a BASE58 encoded public key with 4 bytes added for checksum. Each address has it's associated private key. 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.

All the transactions, bitcoins and addresses are stored in blockchain, the balance of all addresses and all transactions 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. When spending bitcoins, it use o ne or more of the addresses the wallet generated previously. Therefore, when pairing the address to identity, we can directly obtain just the history of transactions related to the given address, but can not get all transactions and balance of the user, as he is likely to own multiple bitcoin addresses.

Bitcoins in blockchain are represented as inputs and outputs of transaction. Each transaction has some inputs and outputs. Input and output is the same data structure, it only differs in it's relationship to given transaction. Each input/output consists of it's unique identifier, it's value in bitcoins and it's owner address. Every output can be used exactly once as input of new transaction, and therefore the owner of output can not spend one output multiple times.

When sender sends bitcoin to recipent, he generates a transaction. The new transaction must satisfy:

- He owns address of the inputs = He can spend only bitcoins he own
- Each input has not been used as input in any other transaction = He can not spend one output multiple times
- The sum of bitcoins of transaction's inputs is equal as sum of bitcoins of transaction's output + fees

The new transaction must have 1 or more outputs. There can be multiple outputs in transaction with different associated addresses and bitcoin value, however, there happens to be a common pattern. When sender sends bitcoins to one recipent, the transaction contains two outputs. One output contains recipent's address and the volume of bitcoins he recieves. One output is called "change output". Because sender usually doesn't own outputs that sums to be equal to number of

bitcoins he wants to send, he adds a second output to the transaction. The second output has address he owns and amount of bitcoins, he will recieve from the transaction back. This is the only way to split bitcoins to smaller parts.

When sending transaction, wallet software creates transaction data and sign it with keys of addresses that are sending the bitcoins. Than it sends it to one or multiple bitcoin nodes. Nodes collect transactions from users and broadcast them to other nodes on best effort basis. The validity of transaction is later checked by miners and if everything is ok, they add it to the blockchain.

2.1.2 Mining

Miners are verifying transactions. Miners are running bitcoin nodes and mining software, which enables them to create a new block of transactions, add it to blockcha in and broadcast new, longer version, of blockchain to other nodes. Finding new block of transactions is a hard problem from computanional perspective. Miners looks for solution to the problem by bruteforce and when they find solution, the are able to generate new block of transaction. The difficulty of problem is adjusted every 2016 block, so that new block is generated on average every 10 minutes.

Some of the variables for the problem are dependent on the last block in blockchain, so it is impossible to precompute the problem for blocks that will come in the future. Miner and anyone else know the definition of the problem just for the block that will immidiately follow.

When miner generates 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 add ress. By these coinbase transactions, new bitcoins are emmitted into network. He also broadcast his new block, so other miners can update their blockchains.

2.2 TOR - the onion routing

Tor [10] is a 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 r outing technology. The usage of Tor can be detected by third party, but the third party can not decrypt 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 HTTPS protocol. This protocol use assymetric cryptography. The webserver and browser exchange their public keys at start of communicatio n 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 attack a certification authority is needed. Certification authority is an 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 autho rities 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 TO R, 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 n etwork and estabilish a circuit, as we can see on 2.1. The packet of data is encrypted with the each public key of the node in circuit, starting from last node as on 2.2. No TOR node knows the whole path of the packet, only his neighbour nodes on the path. The TOR node knows the previous node he recieved the packet from. It gets

also the address of next node by decrypting the packet and reading the added metadata.

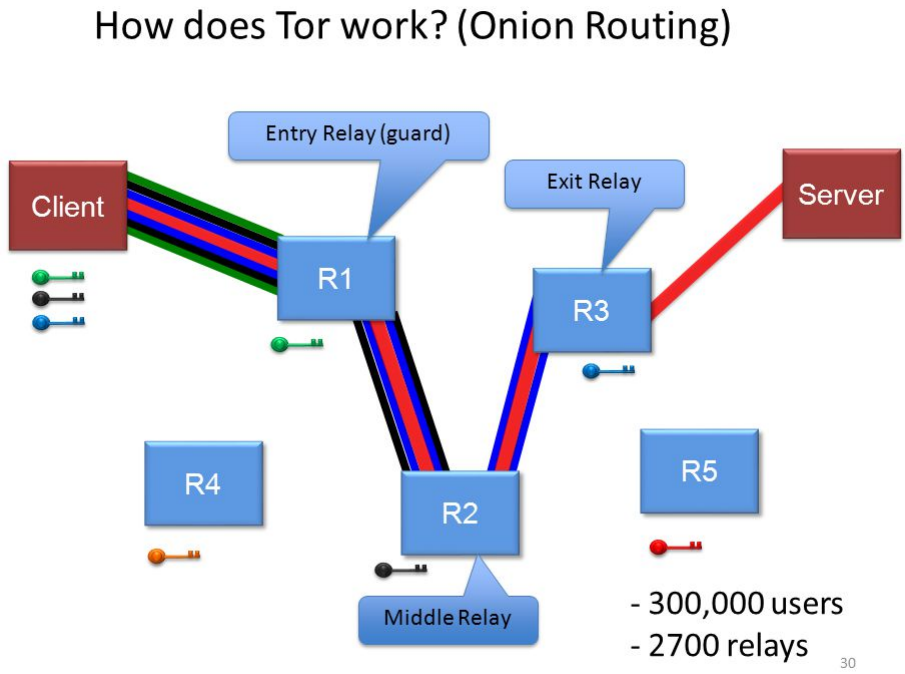

Figure 2.1: TOR routing schema

2.3 PGP

PGP [30] 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, public key cryptography uses two different keys for encrypting and decrypting. User generates a pair of keys, public key for encrypting mails sent to him and private key, which the user keeps for himself

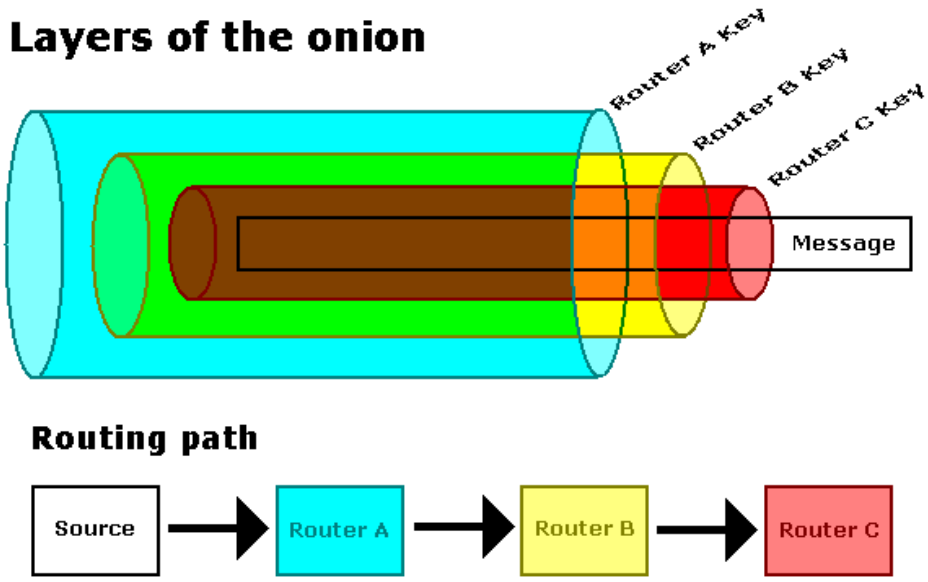

Figure 2.2: TOR packed encryption schem

and uses for decrypting messages encrypted with associated public key. The user also publish 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 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.

2.4 Cryptomarkets

Illegal online markets have been around for more than 30 years [17]. 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 unambigously harmful goods, such as child pornography or hitman services.

In 2011 appeared a new type of illegal online marketplace called cryptomarket. A cryptomarket is an illegal online market accesible only via TOR network and using bitcoins as a 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 buyer via ordinary mail to address provided by buyer. Package is disguised as packages containing common goods sent by big online retailers. [19]

Cryptomarkets are popular by vendors, because they offer high traffic, secure and anonymouse environment for conducting their bussiness[25]. Cryptomarkets offer safer, more comfortable and more professional way of buying drugs, avoiding the need to meet face toery competitive environment f face with dealers [6].

Nowadays, there exists multiple cryptomarkets competing against each other and the risk of a failure of a deal is still high [29]. In order to protect buyers, cryptomarkets use Escrow and vendors' feedback to identify scammers and minimize losses. Cryptomarkets also use Tumbler services, which makes it harder to detect and analyze bitcoin transactions related to these illegal activities.

2.4.1 Escrow

Escrow service is a service offered by cryptomarket for their customers. Escrow means, that the money transfered between two actors(vendor and seller) are held by third party(cryptomarket) and are released when both of the parties agree, that they are satisfied with the trade. Disputes are resolved by the operator of the cryptomarket.

Using the Escrow service is usually required. Some cryptomarkets allow vendors with long history of satisfied customers to bypass escrow service and recieve money im midiately.

2.4.2 Tumbler

moser2013inquiry Bitcoin transactions are publicly available, but it is not easy to identify their owners. It might seem, that bitcoin transactions are anonymoous, but when user send bitcoins to someone(exchange) who knows their identity, the recipent can pair the bitcoin address the bitcoins came from to identity of sender. Although bitcoin users usu-

ally use multiple bitcoin addresses, their transactions and addresses a re still suspectible to blockchain cashflow analysis, which might identify other addresses of the owner of addres we already know .

Bitcoin Tumblers exist in order to prevent such analysis. User sends bitcoins to the tumbler service, the service mix hi s bitcoins with bitcoins of other users by performing multiple transactions between it's bitcoin addresses. The structure of these transactions differs for different tumbler services. User generate new bitcoin address with no tie to his previous addresses and bitcoins are recieved from the service to th is address. There also exists peer-to-peer tumblers(CoinJoin,SharedCoin,coinswap), that enable multiple users to directly create transactions to mix bitcoins among themselves. The transaction can be performed multiple types with different actors.

2.4.3 Vendor's feedback

Cryptomarkets usually employs reputation systems, where buyers can publish their satisfaction with the vendor. These systems are similar to systems used in popular e-commerce websites like amazon or e-bay. Users can give feedback only to vendors with whom thay 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 seller, like the easyness of communication, speed of sending the goods and unsuspiciousness of packaging.

2.4.4 Valhalla cryptomarket

We selected Valhalla cryptomarket based on three considerations. The first consideration is it's size. Walhalla market has is well known operating cryptomarket. It has XXX listings and YYY vendors. This is the second most listing and vendor's, just behind dream market. Among significant cryptomarkets (dream market, Point market, Wall Street Market), Valhalla have been operating for longest time. This is to advantage of our analysis, as we can analyze vendors, who have been selling for longer time and their reviews. Third, it provides best information about vendors and buyers. When we look at feedback page of given vendor, for each given feedback we can see the comment, when the feedback was given, what was the listing the user gave feedback

for, what was the price and first and last 2 digits of buyer, the amount, how much the buyer bought overall on the Valhalla market and how many trades have buyer done. The 2 first and last digits 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 drastically help with granularity of recognizing different/same buyers and detecting, which listings brought the vendor most money.

3 Related works

3.1 Blockchain analysis and linking bitcoin addresses

Multiple papers and tools were published regarding analysis of blockchain. 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 (software, for recieving and sending bitcoins) do so. It is therefore a big challenge to cluster addresses belonging to same user.

The authors of first research article [20] parsed blockchain files to create graph of bitcoin transactions, with vertices as transactions and edges between them represented bitcoins flowing from one transaction to another. They created so called user graph by clustering addresses belonging to same user. They used simple heuristics, that the owner off all input addresses used in a transaction must be the same. First version of this article was published in 2011 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.

Androulaky [4] performed clustering using two heuristics. The first one is the same as [20] did, that all inputs of transaction belongs to same user. The second heuristics is clustering some outputs of transaction with it inputs. Most transactions have two outputs, one is owned by the transaction recipent, the other one is called change. The change is output of transaction, that is owned by the sender. The change output is needed, because that the only way to split bitcoin value of output is to use it as input for transaction. If the user owns 3BTC in one output and need to transfer 1 BTC, it generates a transaction with two outputs, one worth of 1 BTC with the recipent address and second output worth 2 BTC with the recipent address of sender. This way, sender can split his bitcoins for smaller transaction. They also employed multiple clustering techniques based on behaviour of users. They tested succes of their clustering techniques in their simulated bitcoin environment.

Advanced and similar work was done by [23]. 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-ch ain data and identities to bitcoin adresses. They tested the tool on few popular transactions related to seizure of silkroad marketplace.

Similar work to this thesis was done by [13]. 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.

Ron and Shamir [22] focus on bringing interesting statistics about bitcoin transaction graph and provided a detailed analysis of really big bitcoin movements (more than 5000 BTC) through transactions in the network. In their other study [21], they analyzed transactions performed by Ross Ulbricht, who was administrator of Silkroad marketplace. The FBI published their bitcoin address, which they used to collect all seized bitcoins from Ross Ulbrict. 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 discrepencies between the relatively stable income of Silkroad marketplace and unsstable balances in bitcoin addresses that were seized by FBI. They conclude, that FBI seized around 22% of Ross Ulbrich bitcoins and found addresses that posses a some of these bitcoins, which has not been used since Ross' arrest.

In contrast to previously mentioned papers, Meiklejohn [16] doesn't only passively scan blockchain, they actively send bitcoins to addresses of well known services to track their bitcoins in the following transactions executed by the service. They also used the same two heuristics for clustering addresses as Androulaki. [4] They concluded, that the network does not offer enought 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 expo nentionally over the last year. My work is unique in that way, that it utilize much more sources of data than the works previously 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.

3.2 Behaviour of drug markets users and operators

Emerging cryptomarkets brought attention of scientific community and lots of articles have been published related to the phenomena of drug trafficking via internet. Most of these papers were investigating the topic from the social and criminology perspective and performing qualitative analysis. [3] [6] [14] [12] [26] [28] [15]

There are only few articles focusing on statistically describing fully operating drug market and it's vendors by collecting and analyzing data from cryptomarket webpage. Short description of works like that follows. Aldridge [2] scraped Silkroad in September 2013, the most popular cryptomarket of that time . He focus on how the vendors and buyers percieve risk of arrest and attempt to limit them. 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] focus on answering the question, how loyal are buyers on cryptomarkets to vendors. It seems, that popular vendors succesfully build their loyal customer base. These findings make sense, given the natural health risk srelated to using drugs, customers prefer vendors with high reputation and trust.

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

Article by Doliver [11] is most similar to this work. They scraped and analyzed two popular cryptomarkets, agora nad evolution and quantitatively asses the characteristics of vendors from both markets, focusing on the difference between different markets' vendor populations.

All previous works were analyzing no longer existing cryptomarkets, This work focus on describing the vendors from currently operating drug market walhalla and see, if the behaviour of vendors or the nature of cryptomarkets has significantlly changed. We also not only passivelly scrape cryptomarket's webpage, but also create user account on marketplace and send/recieve bitcoins from marketplace in order to get data about the cryptomarket's money flow.

4 Methods and tools to get and analyze data

4.1 Valhalla cryptomarket webscraping

We scraped data from valhalla cryptomarket, one of the most popular drug markets available via TOR. The data was collected from walhalla drug market on 20.1.2018. 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 needed, because TOR daemon creates SOCKS proxy, which can not be used by wget. So we created a HTTP proxy by privoxy, which redirected the traffic through TOR SOCKS proxy.

For webscraping Walhalla market, we did not have to implement any login and captcha solving functionality, because valhalla listings and vendor pages are accessible and showing same data when accesing them without being logged in.

Addresses of all market listings are in pattern http://valhallaxmn3fydu.onion/products/ where xxx is number incrementing with each new listing. Only 1/3 of numbers let to valid listing page. Rest of the numbers led to 404 error. We believe, that these numbers refer to listings that were disabled by vendor or administrator.

From each listing, we parsed vendor's nickname, the subcategory where the listing was placed, title and price. We got 666 unique vendor names by scraping the available listings.

Vendor profile pages were in format http://valhallaxmn3fydu.onion/xxx and their reviews in format http://valhallaxmn3fydu.onion/xxx/palautteet where xxx is vendor username. We were therefore able to parse profile and feedback page of each vendor who had active listing.

From each vendor, we scraped his - nickname = string - number of positive and negative reviews - 2 integers - revenue = integet 0<x<10 000USD, for vendors with higher revenue it shows 10000+ - PGP key = string, is not mandatory for all vendors

We scraped his feedback page and for each review was scraped vendor nickname = string rating = 1-5 date = Timestamp, days resolution first and last 2 characters of buyers nickname = string of length 4 money the author of review spent on Valhalla market = int trades the author of review done = int

We wrote a small script in bash to iterate through all of the listings and download them using wget command line tool. After downloading all the pages of products, we parsed the downloaded files using python and common linux command line tools(cat,grep,cut,sed). We have not used python HTML parsing libraries(like beautifulsoup) for parsing downloaded webpages , as the HTML elements of valhalla webpages don't have any unique identifiers and so these libraries bring us no advantage.

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'.

4.2 Valhalla cryptomarket metadata scraping and analysis

We tested these keys, if they are vulnera ble to ROCA attack, via python module roca-detect. None of these 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 t he searches 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 ve rsion) 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. For both markets, there was no transaction happening for days after the transaction was done. This means, that markets d on't transfer bitcoins, when there is filled order, all the transactions that these drug markets do are just for depositing bitcoins on drug mar ket 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 transfere d and where the withdrawn bitcoins originated. These deposits and withdrawals are used to test the resulting application

4.3 Drug market server fingerprinting

We tested, if every transaction that is happening on drug market has its counter transaction in bitcoin blockchain. We se nt 0.05 bought a virtually deliverable legally service(Link to secret forum) and checked

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 use d by webserver. We used httprecon to fingerprint used HTTP server. The fingerprinting consists of sending multiple malformed HTTP reques ts 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 bi tcoin 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 "Don ate 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 row s 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.

4.5 Detecting wallets owned by drug markets

4.6 Using own transactions to get market wallets

5 Statistics of Walhalla cryptomarket

5.1 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 5.1.

Each circle in 5.2 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 5.1. 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 5.4

asfd asdf

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

Table 5.1: 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

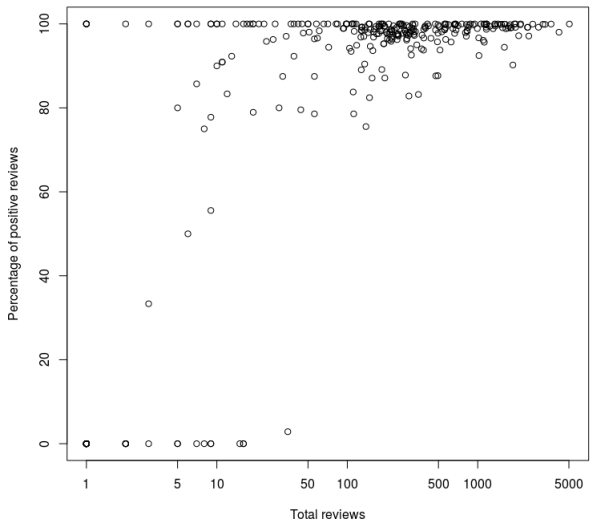

Figure 5.1: Positive reviews of vendors

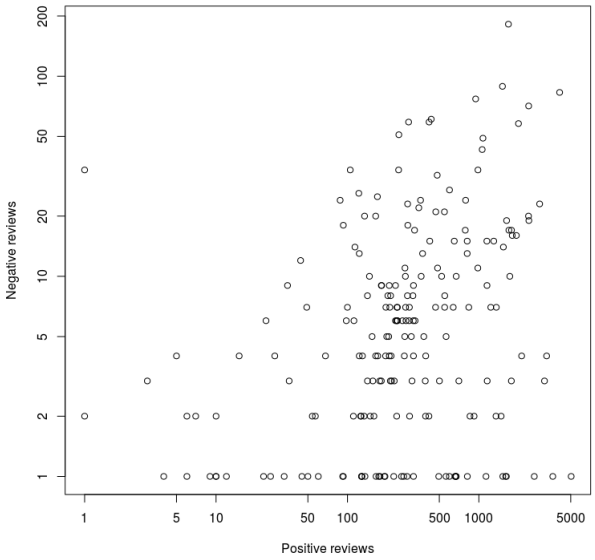

Figure 5.2: Positive/negative reviews of vendors

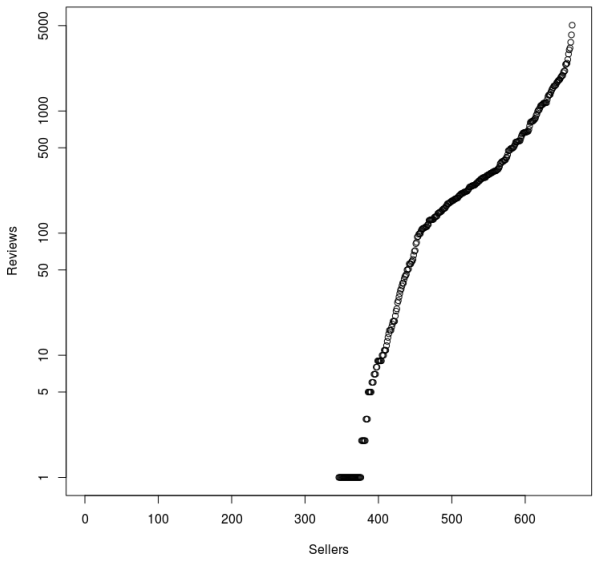

Figure 5.3: Number of reviews for vendors

6 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 sec tion XXX. The computanional module, which imports data to the database and also transform data. so that searching in these data wo uld 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 f rom any device, as long as the server with neo4j data is reachable from that device.

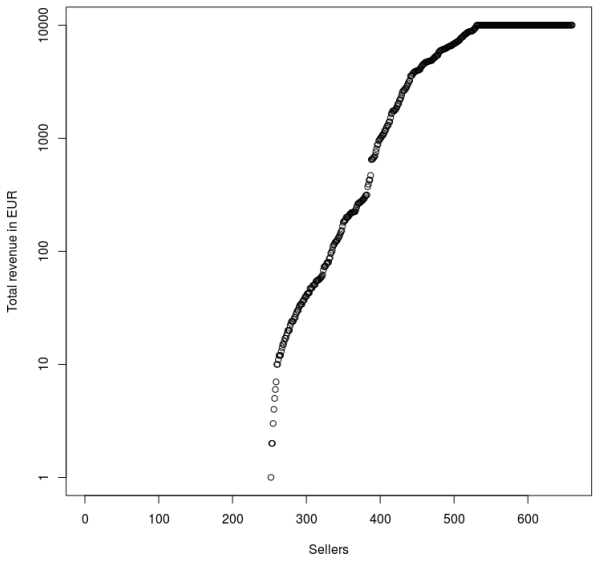

Figure 5.4: Total revenue of vendors

6.1 Retrieving,storing and analyzing blockchain data

In order to create a tool, that will effectively search and visualize blockchain data, we need to store the blockchain locally in that way, so that common graph algorithms can be effectively executed. We ran the official bitcoin daemon (further referenced as bitcoind), to obtain a copy of bitcoin blockchain. Bitcoind st ore 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 comp ute statistics and heuristics. All entities in the 6.1 are represented as graph nodes, the relationships between them are edges.

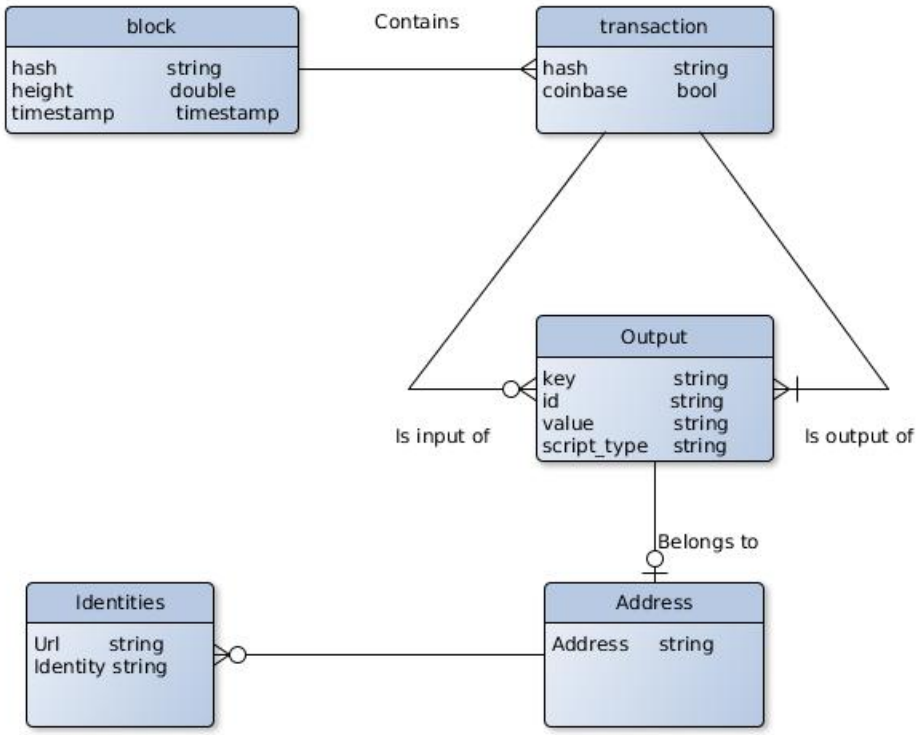

Figure 6.1: Neo4j database ER diagram

6.2 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 cre ating neo4j graph database. The import module firstly parses the .blk files and save blockchain as multiple .csv files. This intermidiary step is us eful 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

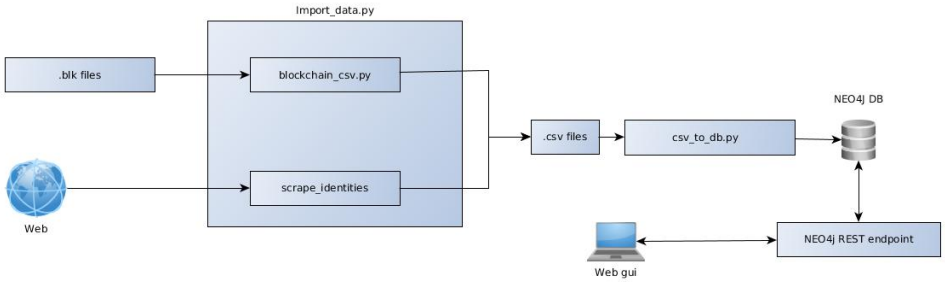

Figure 6.2: Neo4j database ER diagram

- 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 .

6.3 Usage

See the following command :

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

6.4 Future development possibilities

7 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

7.1 Method of testing

7.2 results

8 Conclusion

Here you can insert the appendices of your thesis.gg

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*. XXX: 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] 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.

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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