Masaryk University
Faculty of Informatics

OSINT: Correlation and inference of information from social media

Master's Thesis

Tomáš Šíma

Brno, Spring 2018

Masaryk University
Faculty of Informatics

OSINT: Correlation and inference of information from social media

Master's Thesis

Tomáš Šíma

Brno, Spring 2018

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

Declaration

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

Tomáš Šíma

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

Acknowledgements

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

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

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

Abstract

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

I scraped multiple publicly available social sites and websites related to bitcoin(twitter,bitcointalk, reddit, blockchain.info...) and drug markets thereself using python. I stored all these data into AgensGraph database, which is a graph database based on PostgreSQL. I 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 nearest adresses that are mentioned in scraped websites.

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

Keywords

blockhain, bitcoin, OSINT, darkweb, drug market, TOR, cryptocurrency, anonymity, metadata

Contents

List of Tables

List of Figures

1 Introduction

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

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

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

1.1 Goals

The main goal of this thesis is to map two currently available drug markets and try multiple approaches for deanonymization of identities related to these drug markets. Another outcome of this work is a proof of concept tool, that uses the data mentioned above to help investigator

to disclose transactions, addresses and identities related to online drug markets. The secondary outcome of this work is gathering data about the trades, which happened on the drug markets and gather interesting statistics about the whole market as well as actors operating there.

1.2 Structure of thesis

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

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

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

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

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

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

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

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

2 Related works

2.1 Pairing blockchain transactions with public data

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

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

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

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

2.2 Behaviour of drug markets users and operators

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

3 Technology and terms

In this chapter, I explain the terms, tools and related technology used for achieving the goal of this thesis.

3.1 Bitcoin and blockchain

Blockchain For a broader coverage related to this topic, see Blockchain.

Number of unspent transaction outputs The blockchain is a public ledger that records bitcoin transactions. A novel solution accomplishes this without any trusted central authority: the maintenance of the blockchain is performed by a network of communicating nodes running bitcoin software.Transactions of the form payer X sends Y bitcoins to payee Z are broadcast to this network using readily available software applications. nodes can validate transactions, add them to their copy of the ledger, and then broadcast these ledger additions to other nodes. The blockchain is a distributed database – to achieve independent verification of the chain of ownership of any and every bitcoin amount, each network node stores its own copy of the blockchain.[50] Approximately six times per hour, a new group of accepted transactions, a block, is created, added to the blockchain, and quickly published to all nodes. This allows bitcoin software to determine when a particular bitcoin amount has been spent, which is necessary in order to prevent double-spending in an environment without central oversight. Whereas a conventional ledger records the transfers of actual bills or promissory notes that exist apart from it, the blockchain is the only place that bitcoins can be said to exist in the form of unspent outputs of transactions.

Transactions

Number of bitcoin transactions per month (logarithmic scale) See also: Bitcoin network Transactions are defined using a Forth-like scripting language.Transactions consist of one or more inputs and one or more outputs. When a user sends bitcoins, the user designates each address and the amount of bitcoin being sent to that address in an output. To prevent double spending, each input must refer to a previous unspent output in the blockchain.[52] The use of multiple inputs corresponds to the use of multiple coins in a cash transaction. Since

transactions can have multiple outputs, users can send bitcoins to multiple recipients in one transaction. As in a cash transaction, the sum of inputs (coins used to pay) can exceed the intended sum of payments. In such a case, an additional output is used, returning the change back to the payer.[52] Any input satoshis not accounted for in the transaction outputs become the transaction fee.

Privacy Bitcoin is pseudonymous, meaning that funds are not tied to real-world entities but rather bitcoin addresses. Owners of bitcoin addresses are not explicitly identified, but all transactions on the blockchain are public. In addition, transactions can be linked to individuals and companies through "idioms of use" (e.g., transactions that spend coins from multiple inputs indicate that the inputs may have a common owner) and corroborating public transaction data with known information on owners of certain addresses. Additionally, bitcoin exchanges, where bitcoins are traded for traditional currencies, may be required by law to collect personal information.

To heighten financial privacy, a new bitcoin address can be generated for each transaction. For example, hierarchical deterministic wallets generate pseudorandom "rolling addresses" for every transaction from a single seed, while only requiring a single passphrase to be remembered to recover all corresponding private keys. Researchers at Stanford University and Concordia University have also shown that bitcoin exchanges and other entities can prove assets, liabilities, and solvency without revealing their addresses using zero-knowledge proofs. "Bulletproofs," a version of Confidential Transactions proposed by Greg Maxwell, have been tested by Professor Dan Boneh of Stanford. Other solutions such Merkelized Abstract Syntax Trees (MAST), pay-to-script-hash (P2SH) with MERKLE-BRANCH-VERIFY, and "Tail Call Execution Semantics, have also been proposed to support private smart contracts.

3.2 TOR - the onion routing

Tor is free software for enabling anonymous communication. The name is derived from an acronym for the original software project name "The Onion Router". Tor directs Internet traffic through a free, worldwide, volunteer overlay network consisting of more than seven

thousand relays[10] to conceal a user's location and usage from anyone conducting network surveillance or traffic analysis. Using Tor makes it more difficult to trace Internet activity to the user: this includes "visits to Web sites, online posts, instant messages, and other communication forms". The intent for Tor's use is to protect the personal privacy of its users, as well as their freedom and ability to conduct confidential communication by keeping their Internet activities from being monitored.

Tor does not prevent an online service from determining when it is being accessed through Tor. Tor protects a user's privacy, but does not hide the fact that someone is using Tor. Some websites restrict allowances through Tor. For example, the MediaWiki TorBlock extension automatically restricts edits made through Tor, although Wikipedia allows some limited editing in exceptional circumstances.

Onion routing is implemented by encryption in the application layer of a communication protocol stack, nested like the layers of an onion. Tor encrypts the data, including the next node destination IP address, multiple times and sends it through a virtual circuit comprising successive, random-selection Tor relays. Each relay decrypts a layer of encryption to reveal the next relay in the circuit to pass the remaining encrypted data on to it. The final relay decrypts the innermost layer of encryption and sends the original data to its destination without revealing or knowing the source IP address. Because the routing of the communication is partly concealed at every hop in the Tor circuit, this method eliminates any single point at which the communicating peers can be determined through network surveillance that relies upon knowing its source and destination.

3.3 Online drug marketplaces

A darknet market or cryptomarket is a commercial website on the web that operates via darknets such as Tor or I2P They function primarily as black markets, selling or brokering transactions involving drugs, cyber-arms, weapons, counterfeit currency, stolen credit card details,[4] forged documents, unlicensed pharmaceuticals, steroids and other illicit goods as well as the sale of legal products. In Decem-

ber 2014, a study by Gareth Owen from the University of Portsmouth suggested the second most popular sites on Tor were darknet markets.

Transactions typically use Bitcoin for payment, sometimes combined with tumblers for added anonymity and PGP to secure communications between buyers and vendors from being stored on the site itself. Many sites use Bitcoin multisig transactions to improve security and reduce dependency on the site's escrow. The Helix Bitcoin tumbler offers direct anonymized marketplace payment integrations.

On making a purchase, the buyer must transfer cryptocurrency into the site's escrow, after which a vendor dispatches their goods then claims the payment from the site. On receipt or non-receipt of the item users may leave feedback against the vendor's account. Buyers may "finalize early" (FE), releasing funds from escrow to the vendor prior to receiving their goods in order to expedite a transaction, but leave themselves vulnerable to fraud if they choose to do so.

Following Operation Onymous, there was a substantial increase in PGP support from vendors, with PGP use on two marketplaces near 90%. This suggests that law enforcement responses to cryptomarkets result in continued security innovations, thereby making markets more resilient to undercover law enforcement efforts.

4 Methods and tools to get and analyze data

4.1 Obtaining,storing and analyzing blockchain data

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

Than I imported these files into neo4j graph database, to have whole transaction graph in one place and be able to compute statistics and heuristics. All entities in the 6.1 are represented as graph nodes, the relationships between them are edges.

4.2 Drug markets web scraping and data collection

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

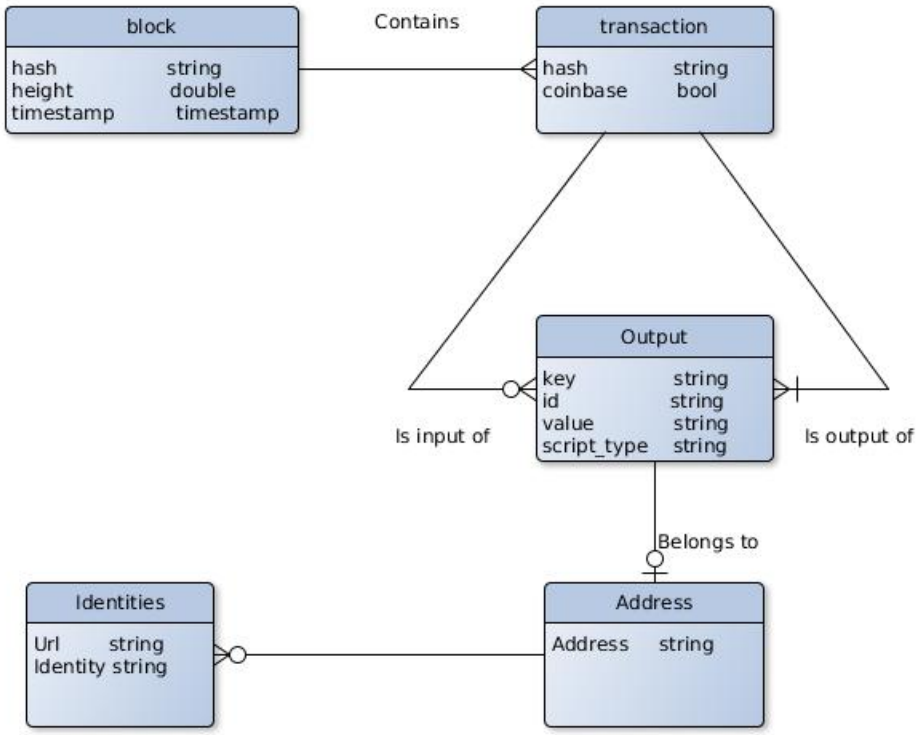

Figure 4.1: Neo4j database ER diagram

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

I thought that metadata from the photos of drugs, which are available on the drug markets might be useful. I downloaded hundreds of pictures both from walhalla and dream market. Only metadata directly dependending on image content(like amount of red, green and blue colors) differ, metadata that could potentionally help dislosing user identity(date of creation nad modification, signature, software version) were the same. The software version contained line: $ImageMagick 6.8.9-9 Q16 x86_64 2017-07-31 http://www.imagemagick.org$ I 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 I believe, that markets automatically scrape metadata from uploaded images in order to protect privacy of the users.

4.3 Drug market server fingerprinting

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

4.4 Publicly available data scraping

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

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

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

5 Deanonymization techniques

5.1 Detecting wallets owned by drug markets

5.2 Using own transactions to get market wallets

6 Statistics of drug markets

6.1 Financial flows in dream market

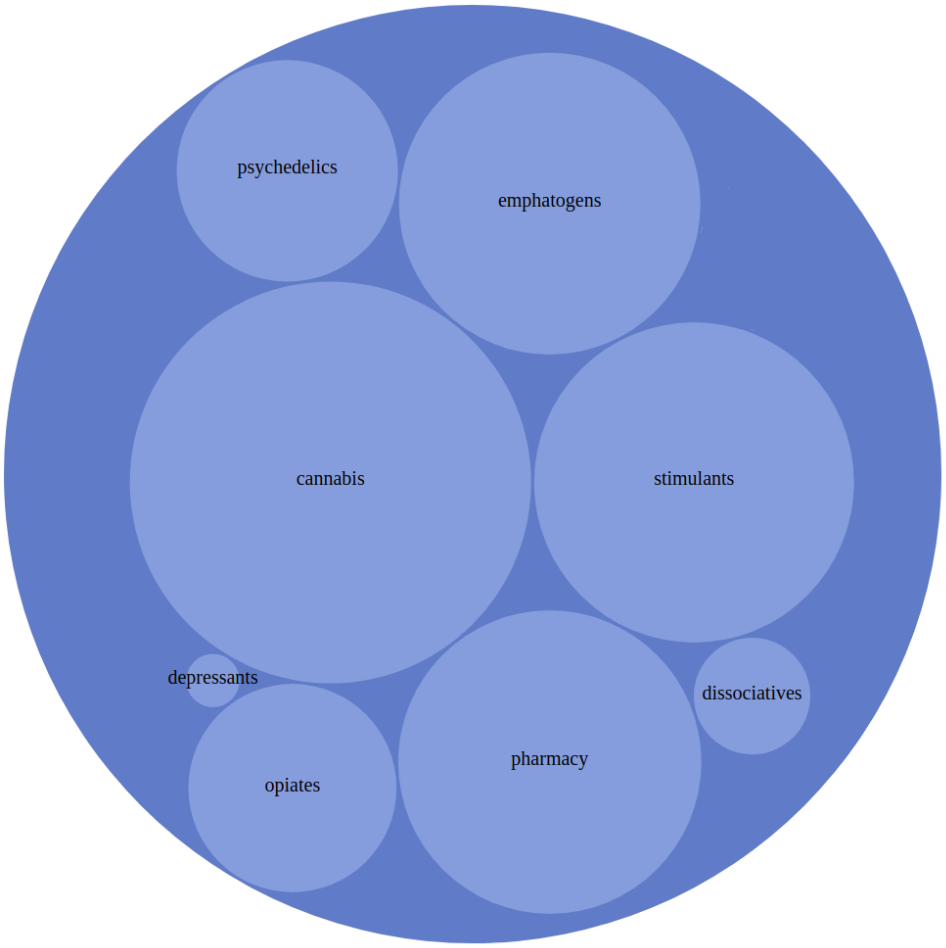

Figure 6.1: Neo4j database ER diagram

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

7 POC application

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

7.1 Implementation

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

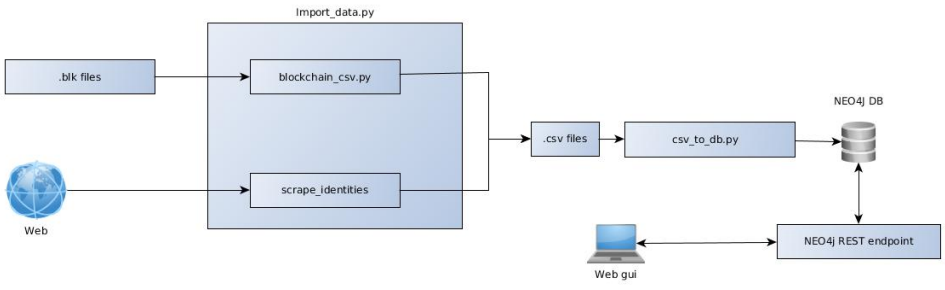

Figure 7.1: Neo4j database ER diagram

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

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

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

7.2 Usage

See the following command :

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

7.3 Future development possibilities

8 Testing and verification of the created tool

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

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

8.1 Method of testing

8.2 results

9 Conclusion

Here you can insert the appendices of your thesis.gg

Bibliography

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

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

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

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

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

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

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

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

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