Masaryk University
Faculty of Informatics

OSINT: Correlation and inference of information from social media

Master's Thesis

Tomáš Šíma

Brno, Fall 2017

Masaryk University
Faculty of Informatics

OSINT: Correlation and inference of information from social media

Master's Thesis

Tomáš Šíma

Brno, Fall 2017

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 sources of 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

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.

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.

1 Related works

1.1 Analysis of graph of blockchain transactions

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.

1.2 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.

2 Technology and terms

2.1 Bitcoin and blockchain

Bitcoin is a distributed, decentralized crypto-currency. The users of Bitcoin are called clients, each of whom can command accounts, known as addresses. A client can send Bitcoins to another client by forming a transaction and committing it into a global append-only log called the blockchain. The blockchain is maintained by a network of miners, which are compensated for their effort in Bitcoins. Bitcoin transactions are protected with cryptographic techniques that ensure only the rightful owner of a Bitcoin address can transfer funds from it. The miners are in charge of recording the transactions in the blockchain, which determines the ownership of Bitcoins. A client owns x Bitcoins at time t if, in the prefix of the blockchain up to time t, the aggregate of transactions involving that client's address amounts to x. Miners only accept transactions if the balance at the source is sufficient.

2.2 TOR

3 Methods and tools to get and analyze data

3.1 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 3.1 are represented as graph nodes, the relationships between them are edges.

3.2 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-

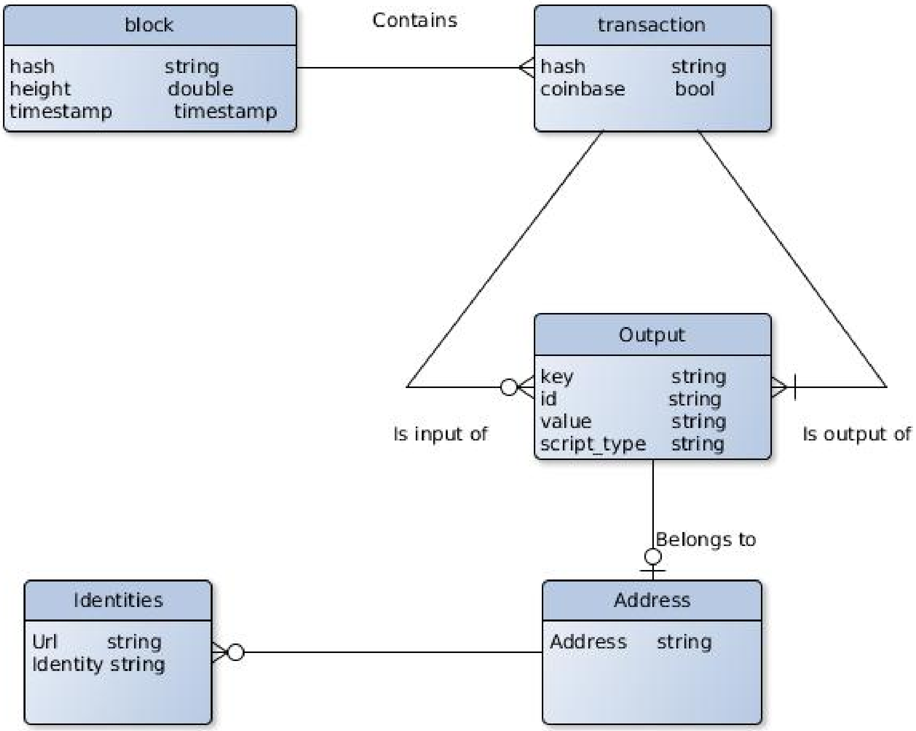

Figure 3.1: Neo4j database ER diagram

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 thesearches 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: $ImageMagick6.8.9-9Q16x86_642017-07-31http://www.imagemagick.org$ 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.

3.3 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.

3.4 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, secundarly, 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.

4 Data Analysis

4.1 Detecting wallets owned by drug markets

4.2 Wallets behaviour during forks

4.3 Using own transactions to get market wallets

5 POC application

5.1 Implementation

5.2 Usage

6 Testing and verification of the created tool

6.1 Method of testing

6.2 results

commands.

Paragraphs and

subparagraphs are available as well. Inside the text, you can also use unnumbered lists,

- such as
- this one
    - and they can be nested as well.
    - » You can even turn the bullets into something fancier,
    - § if you so desire.

twoside Numbered lists are

1. very
    (a) similar

and so are description lists:

Description list A list of terms with a description of each term

The spacing of these lists is geared towards paragraphs of text. For lists of words and phrases, the paralist package offers commands

- that
    - are
        - ∗ better
            - · suited

1. to
    (a) this
        i. kind of

A. content.

The amsthm package provides the commands necessary for the typesetting of mathematical definitions, theorems, lemmas and proofs.

Theorem 6.2.1. *This is a theorem that offers a profound insight into the mathematical sectioning commands.*

Theorem 6.2.2 (Another theorem)**.** *This is another theorem. Unlike the first one, this theorem has been endowed with a name.*

Lemma 6.2.3. *Let us suppose that $x^2 + y^2 = z^2$. Then*

$$\left\langle u \,\middle|\, \sum_{i=1}^{n} F(e_i, v) e_i \right\rangle = F\left(\sum_{i=1}^{n} \langle e_i | u \rangle e_i, v \right). \tag{6.1}$$

Proof. $\nabla^2 f(x,y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$. □

Corollary 6.2.4. *This is a corollary.*

Remark. This is a remark.

7 Floats and references

The logo of the Masaryk University is shown in Figure 7.1 and Figure 7.2 at pages 13 and 14. The weather forecast is shown in Table 7.1 at page 14. The following chapter is Chapter 8 and starts at page 15. Items 3, 3b, and 3(c)iv are starred in the following list:

1. some text
2. some other text
3. ⋆
   - (a) some text
   - (b) ⋆
   - (c) some other text
     - i. some text
     - ii. some other text
     - iii. yet another piece of text
     - iv. ⋆
   - (d) yet another piece of text
4. yet another piece of text

If your reference points to a place that has not yet been typeset, the `\ref` command will expand to **??** during the first run of `pdflatex fi-pdflatex.tex` and a second run is going to be needed for the references to resolve. With online services – such as Overleaf – this is performed automatically.

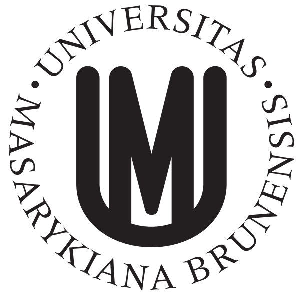

Figure 7.1: The logo of the Masaryk University at 40 mm

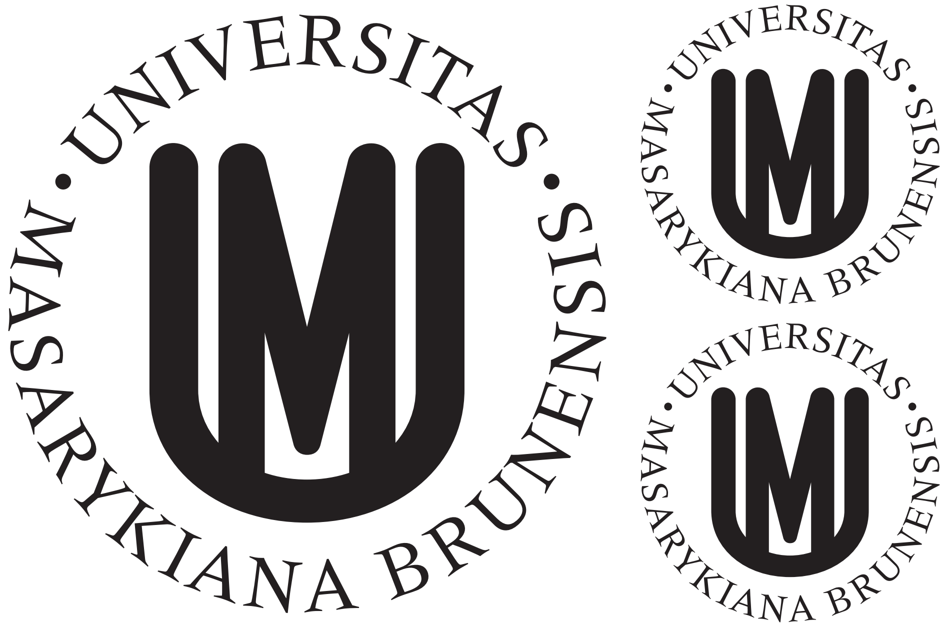
UNIVERSITAS · MASARYKIANA BRUNENSIS ·
UNIVERSITAS · MASARYKIANA BRUNENSIS ·
UNIVERSITAS · MASARYKIANA BRUNENSIS ·

Figure 7.2: The logo of the Masaryk University at $\frac{2}{3}$ and $\frac{1}{3}$ of text width

Table 7.1: A weather forecast

Day	Min Temp	Max Temp	Summary
Monday	13°C	21°C	A clear day with low wind and no adverse current advisories.
Tuesday	11°C	17°C	A trough of low pressure will come from the northwest.
Wednesday	10°C	21°C	Rain will spread to all parts during the morning.

8 Mathematical equations

TeX comes pre-packed with the ability to typeset inline equations, such as $\mathrm{e}^{ix} = \cos x + i \sin x$, and display equations, such as

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det(\mathbf{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

LaTeX defines the automatically numbered `equation` environment:

$$\gamma P x = P A x = P A P^{-1} P x. \tag{8.1}$$

The package amsmath provides several additional environments that can be used to typeset complex equations:

1. An equation can be spread over multiple lines using the `multline` environment:

$$\begin{multline} a + b + c + d + e + f + b + c + d + e + f + b + c + d + e + f \\ + f + g + h + i + j + k + l + m + n + o + p + q \tag{8.2} \end{multline}$$

2. Several aligned equations can be typeset using the `align` environment:

$$\begin{align} a + b &= c + d \tag{8.3} \\ u &= v + w + x \tag{8.4} \end{align}$$

$$i + j + k + l = m \tag{8.5}$$

3. The `alignat` environment is similar to `align`, but it doesn't insert horizontal spaces between the individual columns:

$$\begin{alignat}{2} a + b + c + d &\qquad&= 0 \tag{8.6} \\ e + f + g &&= 5 \tag{8.7} \end{alignat}$$

4. Much like chapter, sections, tables, figures, or list items, equations – such as (8.8) and (My equation) – can also be labeled and referenced:

$$\begin{alignat}{4} b_{11}x_1 &+ b_{12}x_2 &+ b_{13}x_3 & & &= y_1, \tag{8.8} \\ b_{21}x_1 &+ b_{22}x_2 & & &+ b_{24}x_4 &= y_2. \tag{My equation} \end{alignat}$$

5. The gather environment makes it possible to typeset several equations without any alignment:

$$\psi = \psi\psi, \tag{8.9}$$

$$\eta = \eta\eta\eta\eta\eta\eta, \tag{8.10}$$

$$\theta = \theta. \tag{8.11}$$

6. Several cases can be typeset using the cases environment:

$$|y| = \begin{cases} y & \text{if } z \geq 0, \\ -y & \text{otherwise.} \end{cases} \tag{8.12}$$

For the complete list of environments and commands, consult the amsmath package manual[1].

[1] See http://mirrors.ctan.org/macros/latex/required/amslatex/math/amsldoc.pdf. The \url command is provided by the package url.

9 We have `several` FONTS *at* **disposal**

The serified roman font is used for the main body of the text. *Italics are typically used to denote emphasis or quotations.* `The teletype font is typically used for source code listings.` The **bold**, SMALL-CAPS and sans-serif variants of the base roman font can be used to denote specific types of information.

We can also change the font size, although it is usually not necessary.

A wide variety of mathematical fonts is also available, such as:

$$\mathrm{ABC}, \mathcal{ABC}, \mathbf{ABC}, \mathsf{ABC}, \mathit{ABC}, \mathtt{ABC}$$

By loading the amsfonts packages, several additional fonts will become available:

$$\mathfrak{ABC}, \mathbb{ABC}$$

Many other mathematical fonts are available[1].

[1] See `http://tex.stackexchange.com/a/58124/70941`.

10 Inserting the bibliography

After linking a bibliography database files to the document using the `\thesissetup{bib={file1,file2, ... }}` command, you can start citing the entries. This is just dummy text (Borgman 2003) lightly sprinkled with citations (Greenberg 1998, p. 123). Several sources can be cited at once: Borgman 2003; Greenberg 1998; Hàn Thé 2001. "Camel drivers and gatecrashers" was written by Greenberg in 1998. We can also produce Greenberg (1998) or (Greenberg (1998), 1998). The full bibliographic citation is: *David Greenberg (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.* We can easily insert a bibliographic citation into the footnote[1].

The `\nocite` command will not generate any output, but it will insert its arguments into the bibliography. The `\nocite{*}` command will insert all the records in the bibliography database file into the bibliography. Try uncommenting the command and watch the bibliography section come apart at the seams.

When typesetting the document for the first time, citing a `work` will expand to [**work**] and the `\printbibliography` command will produce no output. It is now necessary to generate the bibliography by running `biber fi-pdflatex.bcf` from the command line and then by typesetting the document again twice. During the first run, the bibliography section and the citations will be typeset, and in the second run, the bibliography section will appear in the table of contents.

The `biber` command needs to be executed from within the directory, where the LaTeX source file is located. In Windows, the command line can be opened in a directory by holding down the Shift key and by clicking the right mouse button while hovering the cursor over a

[1] David Greenberg (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.

directory. Select the Open Command Window Here option in the context menu that opens shortly afterwards.

With online services – such as Overleaf – or when using an automatic tool – such as LaTeXMK – all commands are executed automatically. When you omit the `\printbibliography` command, its location will be decided by the template.

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.

11 Inserting the index

After using the `\makeindex` macro and loading the `makeidx` package that provides additional indexing commands, index entries can be created by issuing the `\index` command. It is possible to create ranged index entries, which will encompass a span of text. To insert complex typographic material – such as α or TeX – into the index, you need to specify a text string, which will determine how the entry will be sorted. It is also possible to create hierarchal entries.

After typesetting the document, it is necessary to generate the index by running

```
texindy -I latex -C utf8 -L ⟨locale⟩ fi-pdflatex.idx
```

from the command line, where ⟨*locale*⟩ corresponds to the main locale of your thesis – such as `english`, and then typesetting the document again.

The `texindy` command needs to be executed from within the directory, where the LaTeX source file is located. In Windows, the command line can be opened in a directory by holding down the Shift key and by clicking the right mouse button while hovering the cursor over a directory. Select the Open Command Window Here option in the context menu that opens shortly afterwards.

With online services – such as Overleaf – the commands are executed automatically, although the locale may be erroneously detected, or the `makeindex` tool (which is only able to sort entries that contain digits and letters of the English alphabet) may be used instead of `texindy`. In either case, the index will be ill-sorted.

Index

A An appendix

Here you can insert the appendices of your thesis.gg