Identifying and analyzing digital payment flows regarding illegal purposes on the Internet

In collaboration with CGI and the Swedish Financial Coalition

Master Thesis VT16
Jesper Asplund & Caroline Berggren

ISRN: LIU-IEI-TEK-A--16/02648—SE
Identifying and analyzing digital payment flows regarding illegal purposes on the Internet

Production Economics, Department of Management and Engineering, Linköping University

Jesper Asplund & Caroline Berggren

ISRN: LIU-IEI-TEK-A--16/02648—SE
Abstract
The aim of this study was to illustrate an unexplored illegal exploitation of legal businesses, with the purpose of limiting this market and especially the related transactions. The issue of transactions regarding illegal material executed with credit cards was solved through involving the companies who issues the credit cards, making the market more transparent and thus preventing this kind of transactions. The thesis will illustrate how cryptocurrencies, such as Bitcoin, are being exploited regarding illegal transactions and more specifically, transactions regarding selling and purchasing Child Abusive Material within file hosting services (cyberlockers). The analyzed data was gathered using a webcrawler and different methods for analyzing correlation were implemented on the data to find relationships between different data points. The data points were then clustered, using an algorithm to create a relationship network. The developed model analyzed the data to identify trends and patterns regarding the illegal transactions and the results can be used to find the most prominent users who are potential perpetrators that actively distributes illegal material. A deeper analysis is then performed on the, according to the model, most interesting users in an attempt to identify their underlying identity.

When cryptocurrencies are used by perpetrators to pay and get paid for illegal material, the transaction flows cannot immediately be connected to specific identities and therefore it is required to first identify potential perpetrators and track their transactions, to later compare them with the transactions that has already been identified as payments for illegal material. Apart from this model, a framework has been created to identify certain patterns and trends regarding the cyberlockers’ transaction flows. This was performed through analysis of the transaction flows connected to cyberlockers that were suspected to contain Child Abusive Material or other illegal material.

With the results from the first and second model, the most interesting cyberlockers for future investigations were discovered, according to the trends and patterns in their surrounding transaction flows. When that analysis was performed and the first model was implemented, potential perpetrators was identified through collaborations between the investigating unit, the Police, the cyberlockers in question and the relevant exchange services. Through this collaboration the identities of the perpetrators are revealed and the transaction flows can then be analyzed to limit further distribution of Child Abusive Material within cyberlockers and consequently limit the illegal transactions with cryptocurrencies.

Keywords: Bitcoin, Child Abusive Material, Cyberlockers, Illegal payments, Cryptocurrency, Webcrawler, Correlation, Relationship network.
Acknowledgements

It is a pleasure to thank those who made this thesis possible. Jörgen Blomvall has been a great help considering supervising this thesis and offering examples of topics of interest when proceeding with the work. The supervisors at CGI; Elin Swedlund, Johan Wadenholt, Robert Book, Magnus Lagercrantz and Leif Eddemo, has been of great support in the proceedings of the work considering advice and the presentations. With their help, the Power Point presentations have been clear to the public and interviews with relevant people have been possible. A special thank you goes out to the police department NOA and especially Per-Ake Wecksell and Christian Squalli, for taking the time to meet and provide with relevant information. Also, Thomas Andersson and Caroline Persson at ECPAT has been of great help in the gathering of information regarding the subject of Child Abusive Material. Without the help and the opportunity to receive a license to the tool Reactor by Chainalysis’ co-founder and CEO, Michael Grönager, the results of the thesis would not have been possible to generate. The tool has been of great help considering the spared time and energy that Reactor offers. The Steering Committee of the Swedish Financial Coalition, with Mats Odell at the head, deserves a big thank you considering letting the authors join the meetings and thereby receive greater insight to the amazing work that the coalition performs.
Abbreviations

CAM: Child Abusive Material
EC3: European Cybercrime Centre
ECB: European Central Bank
EFC: European Financial Coalition
I2P: Invisible Internet Project
ISP: Internet Service Provider
IWF: Internet Watch Foundation
KYC: Know Your Customer
P2P: Peer to Peer
SFC: Swedish Financial Coalition
TOR: The Onion Router
Table of Contents

1 Introduction 1
1.1 Purpose 6
1.2 Delimitations 6

2 Method of study 8
2.1 Literature study and source criticism 8
2.2 Tools 8
2.3 Time frame 9

3 Scientific methods 10
3.1 The fraud triangle 10
3.2 Time series analysis 10
3.2.1 Correlation 11
3.3 Controls 11
3.4 Data Collection: Webcrawler 11
3.5 Multidimensional Scaling 12
3.6 Agglomerative Hierarchical Clustering 13
3.7 Clustering 14

4 Theory 16
4.1 Payments 16
4.1.1 The conventional payment system 17
4.1.2 Centralized versus decentralized payment systems 18
4.2 Virtual Currency 18
4.2.1 Bitcoin 19
4.2.1.1 Exchange services 19
4.2.1.2 Encryption 19
4.2.1.3 Transactions 19
4.2.1.4 Block chain 19
4.2.1.5 Mining 20
4.3 Anonymity 20
4.3.1 TOR 21
4.3.2 Tumblers 21
4.3.3 CoinJoin 21
4.4 Cybercrime 21
4.4.1 Cybercriminals using financial solutions 22
4.4.2 Child Abusive Material 22
4.5 The Fraud triangle 22
4.6 Time series analysis 23
4.6.1 Analyzing trends 24
4.6.2 Analyzing seasonality 24
4.6.3 Correlation 25
4.7 Controls 25
4.7.1 Preventative 26
4.7.2 Detective 26
4.7.3 Corrective
4.7.4 Conclusion
4.8 Webcrawler
  4.8.1 Selection policy
  4.8.2 Re-visit policy
  4.8.3 Politeness policy
  4.8.4 Crawling the deep web and the darknet
4.9 Multidimensional Scaling
4.10 Agglomerative Hierarchical Clustering
4.11 Clustering

5 Chosen methods
5.1 The CAM triangle
  5.1.1 The first corner – the opportunity to buy, sell or produce CAM
  5.1.2 The second corner – the ability to rationalize the crime
  5.1.3 The third corner – the pressure on the individual.
  5.1.4 Summary
5.2 Primary method
  5.2.1 Webcrawler
  5.2.2 Multidimensional Scaling
  5.2.3 Agglomerative Hierarchical Clustering
  5.2.4 Summary
5.3 Secondary method
  5.3.1 Identifying and investigating cyberlockers
  5.3.2 Chainalysis

6 Results and analysis
6.1 Primary method
6.2 Secondary method
  6.2.1 Conclusion
6.3 Combination of the primary and secondary method
  6.3.1 Process flow

7 Discussion
7.1 Primary method
  7.1.1 Results from Moore and Rid’s experiment
7.2 Secondary method
7.3 Preventing payments
7.4 Risks
7.5 Ethical aspect
  7.5.1 Ideology
  7.5.2 Integrity and privacy
    7.5.2.1 Encryption
7.6 Further research
  7.6.1 Methods
  7.6.2 Machine learning
7.6.3 Collaborations 69
7.6.4 Legal investigations 69

8 Conclusions 70

Bibliography 71

Appendix 76
   A. Gantt-scheme of the time frame 76
   B. Table of the investigated cyberlockers. 77
   C. Table of examples of how the collected data from the webcrawler can look like. 78
   D. Dissimilarity/Similarity matrix 79
List of figures

Figure 1 Illustrates the file hosting service, cyberlocker. (G2 Web Services, 2015) 1
Figure 2 Types of Internet services exploited to host CAM URLs in 2014 compared to 2013 (IWF, 2014). 2
Figure 3 Illustration of the relevant payment flows to and from the cyberlocker. 3
Figure 4 Illustrates how a tumbler works. 5
Figure 5 Focuses on the users’ payment flows to and from the cyberlocker. The cyberlocker’s corporate account corresponds to both the virtual wallet as well as the traditional bank account. 6
Figure 6 Webcrawler used for data collection 12
Figure 7 The relationship between the user profiles. 12
Figure 8 Multidimensional Scaling 13
Figure 9 Agglomerative Hierarchical Clustering 14
Figure 10 Illustration of how the transaction flows could look like before clustering. 15
Figure 11 Illustration of how the transaction flows could look like after clustering. 15
Figure 12 The evolution of payments according to Leinonen (2008). 16
Figure 13 Life cycle of a non-cash payment (ECB, 2010) 17
Figure 14 The Fraud Triangle 23
Figure 15 An example of how a Shepard diagram can look like. 30
Figure 16 Illustrates \( C = \{c_1, \ldots, c_n\} \) and the distances between different clusters. 31
Figure 17 Illustrates the distances between the clusters through a dendrogram. (Lavrenko, 2014) 32
Figure 18 Illustrates two examples of how to cluster when performing AHC. 33
Figure 19 Primary method. 35
Figure 20 Secondary method. 35
Figure 21 Illustration of the primary method. 37
Figure 22 Example of the connections between different forums regarding time and date of post, which username that is used and which link that is shared. 38
Figure 23 Resulting dendrogram from the example. 42
Figure 24 Illustrates the transactions from wallet A to B and from B to C (the cyberlocker’s corporate wallet). 45
Figure 25 Illustrates some transactions to and from wallet C. 46
Figure 26 Example from manual data gathering. 48
Figure 27 Example from converted data 49
Figure 28 Sample from output matrix. 49
Figure 29 The generated dendrogram illustrating the relationships between the data points in the example. 50
Figure 30 Illustrates the surroundings of cyberlocker D (Bitstamp). 53
Figure 31 Illustrates the surroundings of cyberlocker E’s wallet, were B is the temporary address. 54
Figure 32 The surroundings of cyberlocker F and G. 55
Figure 33 Illustrates cyberlocker H’s Bitcoin connections. 56
Figure 34 Illustration of the combined result from the two methods. 60
List of tables

Table 1 Explanation and examples of the different payment methods. .................................................. 3
Table 2 Illustration of the time frame. ...................................................................................................... 9
Table 3 A money matrix (Henning and Nordin, 2014)(ECB, 2012). ....................................................... 18
Table 4 The identified parameters represented as numerical values. .................................................... 40
Table 5 Illustrates two examples of the categorization by different parameters of the file hosting services. .......................................................................................................................... 45
Table 6 Illustrates the transaction to and from the different services. .................................................. 46
Table 7 Illustrates the information behind the most relevant data points from the experiment. .......... 51
Table 8 Part of the result from the initial investigation of the 155 cyberlockers. ................................. 52
Table 9 Illustrates the considered parameters when generating the framework. ................................. 56
Table 10 illustrates the total cash flow to and from the three cyberlockers. ......................................... 57
Table 11 Illustrates the amounts and shares connected to a TOR market, CoinJoin and Tumbler. ...... 57
Table 12 Explains the different classifications. ....................................................................................... 64
Table 13 Illustrates the result from the webcrawling session. ................................................................. 65
1 Introduction
There has been a clear shift from traditional credit card payments to solutions that provide a higher degree of anonymity regarding trade of illegal material on the Internet (EFC, 2015). Virtual currencies, such as Bitcoin, are therefore getting more popular among perpetrators for the distribution of Child Abusive Material (CAM) within legitimate file hosting services, hereby described as cyberlockers. The European Financial Coalition (EFC) advocates further investigation of these cyberlockers to gain knowledge of the payment methods that are used by offenders (EFC, 2015). Hence, this thesis will investigate and analyze the transactions regarding illegal material and identify patterns and trends. The transactions regarding illegal material incorporate the cash flow to and from users of a cyberlocker with the purpose to upload or download CAM.

As previously mentioned, a cyberlocker is a designation of a file hosting service, where a user can host files of different types, for example images, videos and compressed archives etc. It allows users to upload files on the cyberlocker’s cloud server and thereby make it possible for the user to share the file with other members of the service through a shared link, according to Figure 1. A cyberlocker can hide the user’s identities, due to that the IP addresses of members are kept anonymous from each other and are known only to the cyberlocker’s operator (Chow et al., 2015). This type of business is usually legitimate and there are many companies offering file hosting services for companies and private persons to store large files. What separates the exploited cyberlockers from this set of legitimate businesses is that the cyberlocker’s different business models give the users incentive to upload files that others desire and hence increasing the frequency of downloads. The user who uploaded the file gets paid according to how desired the file was and consequently how many times it was downloaded.

Asplund, Berggren, 2016
upload, in some cases, are hidden from free users in a “premium segment” of the cyberlocker which makes it impossible to find those files without a premium account. This further complicates it for the authorities to find and delete the files that contains illegal material, which are stored in this segment.

During 2014 there was a distinct increase in the distribution of CAM for financial gain, which was the consequence of abuse of pay-for-premium services such as cyberlockers. As seen in Figure 2 there has been an extreme expansion in the exploitation of these kinds of services over the past years. The figure illustrates how many URLs to these kind of exploited services, are posted on different types of sites, where there has been an increase in the exploitation of image hosts, file hosts and forums of different sorts. Both the exploitation of legitimate image hosting services and file hosting services increased more than 300 % from 2013 to 2014. This increase is one of the reasons the banks of Sweden and the Swedish Financial Coalition (SFC) has expressed a vision to prevent this type of exploitation by trying to understand the cash flows and increase the knowledge about how to prevent these payments in the future.

A link that is connected to a file within a cyberlocker is typically distributed through third-party sites, such as web forums. This allows followers that already have a premium account, to easily download the relevant file by clicking on the link. One possible way of gathering data is to investigate the forums in which links with connected files within cyberlockers are marketed and shared. Chow et al. (2015) have developed a method that collects and sorts this data, which is further described in Chapter 3.4.

Figure 2 Types of Internet services exploited to host CAM URLs in 2014 compared to 2013 (IWF, 2014).
To clarify which payment flows this thesis will focus on, see Figure 3. The main focus lies on the users of the services and not the owners/administrators of the cyberlocker. There are several different ways to pay for a premium membership and Figure 3 illustrates the different scenarios for transactions regarding the cyberlocker. The user pays for a premium account and gets in return paid for uploading popular files within the service, according to Figure 5. The payment from the service to users differs between virtual currencies, real economy money and also, in some cases, credits per download which later can be converted into real economy money or used to extend the premium account period. After an initial investigation of a few cyberlockers, two types of payment methods have been identified as the most common: credit card and virtual currencies (mainly Bitcoin). See the explanations and examples of the different payment methods in Table 1.

Table 1 Explanation and examples of the different payment methods.

<table>
<thead>
<tr>
<th>Real Economy money (stored digitally in a bank account)</th>
<th>Virtual money (stored in a virtual wallet)</th>
<th>Payment service (stored in a digital account)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USD</td>
<td>Bitcoin</td>
<td>PayPal</td>
</tr>
<tr>
<td>EUR</td>
<td>Litecoin</td>
<td>WyWallet</td>
</tr>
<tr>
<td>SEK</td>
<td>Dogecoin</td>
<td>Klarna</td>
</tr>
</tbody>
</table>

Transactions to/from Cyberlocker

![Diagram of payment flows to and from the cyberlocker.](image)

Figure 3 Illustration of the relevant payment flows to and from the cyberlocker.
In Figure 3 “other services” are illustrated, which refers to solutions like Paysafe card and Call2Pay which basically is prepaid cards that is bought in a retail store and these transactions are consequently relatively anonymous. If these solutions are to be investigated, information such as surveillance footage and transaction history, need to be collected from the stores which requires warrants for the police.

The real economy money is easy for the police to track through the transactions to and from a traditional bank account or a digital wallet which has led to a shift in what payment methods the perpetrators normally use. The banks and payment services have cooperated with other operators to prevent that their services are connected to CAM and other illegal activity. If the user pays for the premium account through a credit card, a simple cash flow of the digital money can be identified and the transactions from the user’s bank account to the cyberlocker’s bank account can be tracked and charted. Digital real economy money is not tangible like a bill but accounted for and transferred using computers and can be turned into physical money through ATMs (Investopedia, 2016). The digital money is stored in a personal traditional bank account or in a personal digital account, for example within a service such as PayPal, and the two types of storages is hereby denounced as digital wallets. The virtual money, on the other hand, is stored in a virtual wallet which is connected to a specific ID number and is therefore not necessarily related to you personally. A lot of companies offers safe virtual wallets for a fee, and consequently the Bitcoin are better protected from hackers and other offenders. The cyberlocker is listed as a normal company and therefore holds a corporate bank account if real economy money is involved within the service. The problem arises when investigating the virtual money’s transaction flows.

Cryptocurrencies is a subset of virtual currencies, which means that it inherits all attributes that a virtual currency has, but also has some unique features that virtual currencies do not have. The main difference lies in the encryption of transactions and the complete anonymity of the users, while the transactions still are completely public. This will be described further in this chapter and in Chapter 4.2.1.

Bitcoin is the most commonly used cryptocurrency within cyberlockers; therefore, the main focus will lie on this. Like digital and physical wallets, virtual wallets exist to store virtual money. Bitcoin is a decentralized and a peer-to-peer based virtual currency that was publicly launched in 2009. There are different ways to access Bitcoin; through online currency exchanges, mining and sales of products or services, given an already existing Bitcoin wallet. If the Bitcoin are exchanged through an online currency exchange, one of the parts in the transaction will transfer money from a bank account and the other part will be the initial holder of the Bitcoin. In order to keep track of all Bitcoin in the network, a ledger file is kept and agreed upon which contains all transactions ever made in the network (Nakamoto, 2008). The phenomenon is called the block chain and if a Bitcoin transaction is executed between the cyberlocker’s account and the user’s, the money can be tracked through this. For further information about the block chain see Chapter 4.2.1.4. The tracking process would then consist of manually identifying the cyberlockers’ virtual wallet ID by looking at the users’ payments for premium accounts and then it would be possible to track payments from the cyberlocker to users who have uploaded files desired by other users.

Another difficulty arises when measures are taken to further anonymize transactions. A distributed network of computers, in a process called mining, can create brand new Bitcoin for a reward through solving complicated math problems, for further information about mining see Chapter 4.2.1.5. The

---

1 Swedish Financial Police. Interview 2016-02-11.
process demands high capacity so therefore people join together in a mining pool to generate new Bitcoin and consequently share the reward. The members of the mining pool are therefore relatively anonymous depending on how many that take part in the pool. Another way to increase the anonymity of the Bitcoin transactions is to scramble different peoples’ coins with others, which basically means that you send your Bitcoin to a laundry service where the coins are scrambled with others’. When the scrambling is completed, the corresponding amount is sent to the predetermined receiver without revelation of who the Bitcoin was derived from, see Figure 4 (Bitlaunder, 2013). The degree of anonymity depends on how many transactions that are scrambled at the same time within the laundry service. There are some registered companies that offers the service to scramble and anonymize your Bitcoin for a fee. CoinJoin is another anonymization method for Bitcoin transactions. The idea for a payer is to find someone else who also wants to make a transaction and make a joint payment together. In these joint payments it will be hard to relate input and output in one Bitcoin transaction and thus the exact direction of money movement will remain unknown to a third party. In these different anonymization cases, investigators cannot track transactions as easily. The green arrows in Figure 5 represent where the anonymization of Bitcoin can occur.

Figure 4 Illustrates how a tumbler works.

The answer to why virtual cash flows is much harder to analyze than normal cash flows is that everyone involved in a payment are anonymous and replaced by the ID of their virtual wallets, but the flow of the currency is public. This means that if someone wants to trace their Bitcoin back and check who the previous owner was, the only thing they will find is the ID of the previous owner and the route the currency has traveled. There are certain webpages that keeps track of all the worlds’ Bitcoin and every transaction ever made with a Bitcoin. With the help of these pages and the tool Reactor (further explained in Chapter 2.2), it is easier to determine the virtual cash flow of a certain ID number regarding both in- and outflows.
Asplund, Berggren, 2016

1.1 Purpose
The purpose of this study is to develop methods to investigate accessible data to identify and analyze digital payment flows regarding illegal material and consequently prevent further transactions involving Child Abusive Material on the Internet.

1.2 Delimitations
The area of illegal material on the Internet is broad. Because this thesis is elaborated with the Swedish Financial Coalition (SFC), the focus will lie on CAM. It can be difficult to differentiate CAM from a dataset that contains different sorts of illegal material and if that was the case, all of the material was analyzed.

Figure 5 Focuses on the users' payment flows to and from the cyberlocker. The cyberlocker's corporate account corresponds to both the virtual wallet as well as the traditional bank account.

The purpose of this thesis is not to defile cyberlockers or virtual currencies. These phenomena were not created to deal or be connected with illegal activity. Therefore, careful revising and delimitations need to be executed when investigating the area so that false accusations are not distributed. It is important though, to highlight the flaws with the services so the awareness in society increases and preventative measures can be taken to reduce the amount of illegal material online.
There are many different ways for perpetrators to share and upload CAM on the Internet but, as mentioned earlier, cyberlockers are the most growing method. The thesis focuses on these legitimate hosting services that are abused by offenders. Within the services only digital material can be uploaded and shared.

Since the perpetrators can choose between different payment methods, the most used and advocated kind was identified and hence was the primary focus of this thesis, while the other solutions had secondary focus and was analyzed if time allowed it.

The administrators and owners of the cyberlockers are often highly competent within the sophisticated technique required. Consequently, the transactions to and from the users of the cyberlocker was prioritized, see Figure 3. Within the group we call “users”, both sharers, followers and downloaders are included. These three types of users can switch roles and therefore no distinction was drawn between them.

Bitcoin is the most prominent virtual currency, accounting for 80% of the market capitalization (European Central Bank, 2015). Therefore, when considering virtual currencies within the cash flow to and from a cyberlocker, Bitcoin was assumed to be the most relevant.

Services such as Call2pay provide payment solutions that are not connected to the payer’s identity and was therefore out of scope in this thesis due to the operational measurements that is needed.
2 Method of study
When conducting this thesis, a combination of quantitative and qualitative data was collected and analyzed. The gathered data was collected from interviews with representatives of companies who support the work of the Swedish Financial Coalition (SFC), such as ECPAT, the Swedish Police, CGI and many more. The companies have valuable insights about the preventive measures necessary and the proactive work regarding prevention of CAM on the Internet. The opportunity has been given to participate at SFCs steering group meetings as observers, with the possibility to ask questions directly and also be provided with other relevant contacts.

2.1 Literature study and source criticism
The literature that has been used to perform this thesis is from a wide range of different sources. The main source being books and scientific papers, which have been collected from the University of Linköping's online library and Google Scholar. If a source has been found on the University of Linköping's online library, it has been cross-checked with Google Scholar to verify the source, and vice versa. Other online sources of scientific papers and books have also been used to verify the sources and as well to find new ones. Since this subject is quite uncharted, similar subjects have been used and parallels have been drawn to make the literature fit the thesis. Fraud and fraudulent activity is one example of that, where similarities with the subject of CAM exists, both in the behavior of the actors and the criminal conduct.

There is not much literature on the specific subject of this thesis but there are a lot of existing electronic sources available on related subjects, which have been used to support presented claims and theories. Since these sources are harder to verify, and even if they are verified, it is difficult to conclude that the presented information is legit, one has to be cautious when using them. Some sources could not be verified, because the information presented was only found at one electronic source, and therefore they could not be used. All electronic sources used in this thesis have therefore been verified using this method of cross-checking several similar sources and extracting the verified and legit information.

2.2 Tools
Michael Grönager, CEO and co-founder of Chainalysis, has offered a license for the tool Reactor which visually illustrates different Bitcoin transactions. The tool was of great help when investigating transactions to and from cyberlockers, tumblers and joint payments.
2.3 Time frame

The provisional timetable is represented in Table 2 below and as a GANT-scheme in appendix A. The dates have changed a bit during the process and the initial plan was to present the master thesis in May.

Table 2 Illustration of the time frame.

<table>
<thead>
<tr>
<th>Task Name</th>
<th>Duration</th>
<th>Start</th>
<th>Finish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature collection</td>
<td>9 days</td>
<td>Thu 16-01-14</td>
<td>Tue 16-01-26</td>
</tr>
<tr>
<td>Write planning report</td>
<td>24 days</td>
<td>Mon 16-01-11</td>
<td>Fri 16-02-29</td>
</tr>
<tr>
<td>Planning report seminar</td>
<td>1 day</td>
<td>Wed 16-02-19</td>
<td>Wed 16-02-19</td>
</tr>
<tr>
<td>Write mid-term report</td>
<td>35 days</td>
<td>Thu 16-02-15</td>
<td>Mon 16-04-01</td>
</tr>
<tr>
<td>Establish contacts with authorities</td>
<td>1 day</td>
<td>Wed 16-01-20</td>
<td>Thu 16-01-20</td>
</tr>
<tr>
<td>Establish contacts with the SFC (banks, police, ECPAT etc.)</td>
<td>1 day</td>
<td>Wed 16-01-20</td>
<td>Wed 16-01-20</td>
</tr>
<tr>
<td>Investigate cyberlockers</td>
<td>7 days</td>
<td>Fri 16-02-04</td>
<td>Fri 16-02-14</td>
</tr>
<tr>
<td>Data collection on file sharing sites, through interviews and such.</td>
<td>30 days</td>
<td>Mon 16-02-01</td>
<td>Wed 16-03-11</td>
</tr>
<tr>
<td>Mid-term report seminar</td>
<td>1 day</td>
<td>Mon 16-04-08</td>
<td>Mon 16-04-08</td>
</tr>
<tr>
<td>Analyze data and identify patterns and trends</td>
<td>15 days</td>
<td>Thu 16-03-14</td>
<td>Thu 16-04-01</td>
</tr>
<tr>
<td>Investigate different payment methods and contact banks and payment institutions</td>
<td>7 days</td>
<td>Fri 16-04-04</td>
<td>Fri 16-04-12</td>
</tr>
<tr>
<td>Draw conclusions and write report</td>
<td>40 days</td>
<td>Mon 16-04-04</td>
<td>Thu 16-05-27</td>
</tr>
<tr>
<td>Present the thesis (SFC)</td>
<td>1 day</td>
<td>Mon 16-05-12</td>
<td>Mon 16-05-12</td>
</tr>
<tr>
<td>Present the thesis (school)</td>
<td>1 day</td>
<td>Fri 16-06-17</td>
<td>Fri 16-06-17</td>
</tr>
<tr>
<td>Oppose</td>
<td>1 day</td>
<td>Fri 16-06-17</td>
<td>Fri 16-06-17</td>
</tr>
</tbody>
</table>
3 Scientific methods

The first question that needed to be answered was: which payment method is the most common regarding the transactions to and from the cyberlocker? The hypothesis was that Bitcoin is popular due to the currency’s high level of anonymity. When the most popular payment method was revealed, the cash flows could be investigated. One way of collecting data was through web forums and through services, such as blockchain.info and Reactor, where Bitcoin block chains are registered and consequently the history of each Bitcoin could be studied. To make the investigation of the perpetrators easier, a collaboration with the cybercrime center and the financial division within the Swedish Police was of interest, to hopefully receive access to a joint database of usernames and other information about perpetrators. A literature study has been conducted and some methods for analyzing payment flows within forensic accounting have been identified, mostly regarding fraud and money laundering. The aim was to adapt and shape these methods for application on transactions regarding CAM.

3.1 The fraud triangle

According to Cressey (1973) the fraud triangle is composed of three parts; pressure, opportunity and rationalization which all together can make a “normal” person commit fraud. Pressure can evolve through a personal financial crisis and hence he/she can use the opportunity to abuse the position of trust to solve the problem. The offender has to rationalize the crime afterwards considering that he/she might not have a criminal past and therefore come up with excuses such as: the money was taken as a loan. (Cressey, 1973) The fraud triangle was modified to fit the problem of CAM and will be of use in the prevention of future offences through a better understanding of the perpetrators.

The main thought was to create a model which is based on the fraud triangle but is more relevant to people who are buying, selling and/or producing CAM. The shape of the model remained a triangle and the adapted method was composed of the three corners; pressure, opportunity and rationalization. The main difference is that the three corners have a more specific definition since this is a crime that cannot be “accidentally” committed.

Input: The fraud triangle; Behaviors of offenders committing fraud.
Output: The CAM triangle; Behaviors of offenders that are connected to Child Abusive Material.
Purpose: To better understand the offenders and their behaviors. The framework is meant as a tool for the discussion of this thesis.

3.2 Time series analysis

The method is constructed through historical data to predict future values considering seasonal trends. Basically, time series analysis is a more sophisticated form of regression analysis. Using time series analysis in forensic analytics is a relatively new method. The method is mostly applied on cases regarding detection of fraud, where deviations from the benchmark easily can be detected. For example, if a restaurant’s sales numbers deviate from the seasonal norm this is a warning sign for the investigators. (Nigrini, 2011)

Input: Historical data on Bitcoin transactions to and from infected cyberlockers. Derived from Reactor.
Output: Patterns and trends regarding timeframes, amounts, senders/recipients etc. in the transactions.
Purpose: To investigate and predict future developments of transactions regarding CAM and consequently find preventative measures to prevent future trade of illegal digital images.
3.2.1 Correlation
The idea was to export data on transactions between known infected cyberlockers and different tumblers, CoinJoins and other complicated solutions used to further anonymize Bitcoin transactions. This data was generated through Chainalysis’ tool Reactor. When the exportation of data was completed, the correlation between different factors was investigated. An example from the investigation is how many of the infected cyberlockers that have a connection to a tumbler or a CoinJoin. The different parameters that was investigated is explained in more detail in Chapter 5.3.2.

**Input:** Data from Reactor regarding transactions between infected cyberlockers and tumblers, CoinJoins etc.  
**Output:** Common denominators in the Bitcoin transactions between the different cyberlockers and the anonymization services.  
**Purpose:** Distinguish correlation between infected cyberlockers and suspicious services to create a framework. This framework will be of use for the police when investigating suspicious cyberlockers.

3.3 Controls
According to the Institute of Internal Auditors (2005) different controls of fraud can be classified as preventive, detective and corrective. Preventive controls are focused in preventing errors and other security incidents from occurring in the first place. These controls are embedded in company’s database system. Detective controls comprise detection of already existing errors and corrective controls focuses on correcting these defaults. The corrective controls are an attempt to reduce further losses. (Institute of Internal Auditors, 2005)

The different controls could be of use in this study to prevent further transactions regarding CAM. The preventive control, which should be embedded in the cyberlocker’s system, would block Child Abusive Material from the site and thereby eliminate the problem. This incorporates that all images need to pass a filter when first introduced to the cyberlocker and CAM is not allowed through the filter. The detective control would be the second barrier in the system. If offenders can avoid getting rejected by the filter, an application would be, for example, PhotoDNA that warns the system when CAM is present within the service. PhotoDNA is a technique developed by Microsoft to detect illegal digital images online (Microsoft, 2016). The corrective control is useful regarding removing illegal digital images to reduce the interest in the material.

**Input:** The framework of the three controls is applied in all cyberlockers.  
**Output:** No Child Abusive Material will be stored within the cyberlockers and hence no transactions regarding illegal material exists in connection to the services.  
**Purpose:** To eliminate CAM within cyberlockers and hence stop the illegal transactions.

3.4 Data Collection: Webcrawler
A link that is connected to a cyberlocker is typically distributed through third-party sites, such as web forums, which allows visitors to easily download the relevant file through the link if the person already has a premium account connected to the cyberlocker. To collect data and information about the operators within the web forums that is connected to cyberlockers, crawlers are used. A web crawler is typically a program that methodically browses the World Wide Web and creates copies of the visited pages for later processing (ScienceDaily, 2015). In this way, up-to-date data can be collected. According to Figure 6, the relevant area for the crawlers in Chow et al.’s experiment, was the web forums where links to files within cyberlockers are posted and shared.
**Input:** The input to this model is hard to define, because the writer of the webcrawler defines what data the webcrawler is supposed to gather, and preferably a website as a starting point of the search.

**Output:** In this thesis the output is data on four different parameters; forum, link, timestamp and username, preferably in a table.

**Purpose:** To gather the data needed to conduct this master thesis.

![Data Collection](image)

**Figure 6** Webcrawler used for data collection

When the crawlers have collected the requested data, the user profiles are constructed and segmented into downloaders, followers and sharers. A person who uploads and shares links to files within a cyberlocker is called a sharer and someone that only responds in the forum is called a follower. The definition of a downloader is a person that visits the posted links that corresponds to the connected file, a downloader can also respond in the forum and could potentially upload files as well. This means, according to Figure 7 that a user can be a categorized as a downloader, sharer and a follower at the same time. (Chow *et al.*, 2015)

![Relationship between user profiles](image)

**Figure 7** The relationship between the user profiles.

### 3.5 Multidimensional Scaling

By applying the technique Multidimensional Scaling (MDS) on the represented data, connections and patterns between operators can be deduced. For example, if the users are defined as objects and an object A is in close proximity to an object B but far away from an object C. Then object A and object B have a strong relationship while a weak or no relationship exists with object C. Basically, MDS is a sophisticated correlation analysis. (Chow *et al.*, 2015)
In this thesis the main focus lies on the users of the cyberlocker and not the owners/administrators. This means that prior to the MDS analysis, a method that generate relevant data on users need to be implemented, tentatively a webcrawler solution. The output from the data collection would then be compiled and used in the MDS, as seen in Figure 8. The relevant data in this case is the relationship between the parameters; usernames, forums, timestamp and links, which will be further explained in Chapter 5.2.2.

Through the data collection performed prior to this method, relevant data on sharers and followers can be collected, but data regarding the downloaders is more difficult to collect. One way to identify downloaders is to contact the site administrator and get permission to intercept the IP addresses of the users who actively click on the download link. This method requires the site administrators' permission which can be hard to get, since it could be argued that this is a violation of the users’ privacy. Observations of cyberlockers in general has shown that they often write in their privacy statement about actively cooperating with the relevant authorities if suspicion of criminal activity exists. This should mean that if the police find out about criminal activity from a user on one of these sites, they can demand to get the users’ IP address from the site administrator.

**Input:** The data generated in the data collection, preferably conducted using a webcrawler.

**Output:** A four dimensional coordinate space containing all gathered data, where the distances between the points in the space corresponds to the relationship between them.

**Purpose:** To get a statistically reliable way of mapping the relationships between different users. Also, to get a more reliable way of gathering data which can be implemented to find producers and distributors of illegal material, who otherwise are very hard to find.

---

**3.6 Agglomerative Hierarchical Clustering**

After a multidimensional scaling analysis is performed, the users are represented as points in a lower dimensional map, incorporating the relationships between each other. Agglomerative Hierarchical Clustering (AHC) is then performed on the coordinates of the resulting points to identify potential clusters in which users have similar behavior, as seen in Figure 9. When the methods are completed the result can be evaluated through stress tests to determine the quality of the fit. (Chow et al., 2015) The two methods combined illustrate a chart of activity, which could be of use by the police to track the users.
behind the aliases found online. The method can also be used in a financial purpose to identify payment flows and patterns, instead of tracking user identities, which is Chow et al.’s purpose. More information regarding MDS can be found in Chapter 5.2.2.

**Input:** The space generated with the MDS model.  
**Output:** A clustered view of the space generated with the MDS model, which will be easier to analyze and interpret results from.  
**Purpose:** To make the data, that is used and rearranged with the MDS model, easier to interpret and analyze. Also, to receive an output which later can be viewed and analyzed by third parties, such as the police.

![Diagram of Multidimensional Scaling](image)

**Figure 9 Agglomerative Hierarchical Clustering**

### 3.7 Clustering

Data clustering can be used to group together a set of data objects to a bigger entity, which simplifies the viewing of the data. Instead of looking at thousands of nodes with similar properties, clustering simplifies the viewing by clustering together these nodes and hence minimizes the number of entities. One assumption made in this thesis is that a majority of the users who exploit cyberlockers will use some sort of laundry-service for their Bitcoin used in transactions to and from the cyberlocker. Since these users could be using a range of different Bitcoin-laundering services, the transaction flows will be hard to investigate. Clustering analysis could then be used to cluster the different laundry services into one entity, which would make the transaction flows easier to follow and then later analyze, see Figure 10 and Figure 11. The clustering of the laundry services facilitates and saves time in the future work when investigating the input/output from the services to/from users and the cyberlocker account. Basically, it will be easier to investigate Figure 11 than to investigate Figure 10.

**Input:** A number of laundry services connected to a cyberlockers.  
**Output:** One entity corresponding to all connected laundry services.  
**Purpose:** To simplify the investigation of the Bitcoin transaction flows and thereby save time. Time that the police can use in a more efficient way to further prevent CAM.
Figure 10 Illustration of how the transaction flows could look like before clustering.

Figure 11 Illustration of how the transaction flows could look like after clustering.
4 Theory
This chapter explains the theories behind the presented methods in Chapter 3 and also the basic background concepts to clarify the concluding results and analysis of the investigation of the payments connected to Child Abusive Material. The theoretical background is based on eleven sections and starts off explaining the basic concepts needed to understand the analysis and discussion of the thesis. Then, the theories of the presented methods are explained. First a presentation of the fraud triangle, which will be of use in the discussion of the thesis. Thereafter the theory behind the method of collecting data is presented and the theoretical background of the methods used to analyze the collected data. Some extensive methods are presented as, for example, preventative measures and are hence not connected to the collected data. These methods are examples of different measures taken in other cybercrime situations and could be shaped to fit the problem of the distribution of CAM.

4.1 Payments
According to Leinonen (2008) payments are basically fund transfer services and the end result is that the payer’s account is debited and the payee’s account credited. The concept “money” is a diffuse subject. It all started with a trade of goods, i.e. bartering. The transformation from bartering to cash payments took several centuries and since then, new payment methods have developed exponentially, according to Figure 12. Paper-based transfers developed through cash payments when it was discovered to be more efficient than physically moving cash. Now the paper-based system is replaced by electronic account transfers, due to the increase in efficiency and that integrated e-payments are getting more popular. (Leinonen, 2008)

This development of payments indicates that customers demand a higher level of speed, security, simplicity and privacy at a lower cost for their transactions. New Payment Service Providers (PSPs), such as PayPal, and other non-traditional payment methods are competing with the conventional system to assist the market, due to the demand of speed and the growing use of innovative technology in payments. (Bank for International Settlements, 2012) (Leinonen, 2008)

Figure 12 The evolution of payments according to Leinonen (2008).
4.1.1 The conventional payment system

In every economy a large number of transactions take place each day, involving trade of goods, services or financial assets. When a transaction occurs, a payer is defined as someone that wants to pay money for a good, service or financial asset and a payee is the recipient of the amount of money. Regarding the payments of this trading system, banks and other entities play an important role. Financial institutions compete with each other to provide services to customers, but still have to collaborate when dealing with transactions between them. Consequently, banks may join common systems that facilitates the transaction process. (ECB, 2010)

The most common ways to transfer funds from a payer to a payee is represented by cash and non-cash payment instruments. Cash payments (payments with bills and coins) are usually associated with face to face transactions between individuals or between an individual and a merchant. If the parties do not exchange information about their identity, a cash payment is relatively anonymous. This is a payment solution that generate a fast and secure transaction. Identification measures are taken when a large sum of cash is moved, with the purpose to handle money laundering and the financing of terrorism. (ECB, 2010)

Non-cash payments involve the transfer of funds between accounts. The institutions role in the transaction is to eliminate the frictions in the transaction for a fee. In the process of a transaction, the steps in Figure 13 are crucial. The first part corresponds to authorization and submission of a payment, which means that the payer gives the bank authorization to the transfer of funds. The processing involves payment instructions to the exchange between concerned banks and accounts. A clearinghouse is the third party in the processing step and registers the transaction in an administrative and legal sense, for a fee. The clearinghouse takes responsibility for the contracts completed by the counterparty and consequently reducing the counterparty- and systematic risk. The settlement takes place when the payer’s bank has to compensate the payee’s through a third party and the final settlement of debts and claims between the two institutions is performed, typically by the national central bank. Because the central banks play an important role in the transactions, a great confidence lie in them to maintain the price stability, meaning the value of the stock of the currency. If the payer and payee hold accounts within the same bank a simplified procedure is performed internally, without involvement of other parties. Due to the complexity of an external transaction, a payment system needs to be supported by a sound legal basis. (ECB, 2010)

“Safe, reliable and efficient market infrastructure for payments, securities and derivatives is crucial to the maintenance of stability in the banking sector and the financial system in general.” (ECB, 2010)
4.1.2 Centralized versus decentralized payment systems
The conventional payment system is centralized and thereby highly regulated by laws and external operators. The centralized structure relies on one operator to make decisions and provide directions for the future procedure. Most of the newly developed PSPs that handle real economy money are centralized and regulated. The consequences of the conventional system are that it takes time and money to transfer funds. With a centralized system, the regulation entails identification of consumers and it is very hard for users to remain anonymous. Out of the complications of the centralized systems, decentralized systems evolved.

The decentralized system corresponds to leaving the responsibility to the individuals of the community and thereby making the transaction faster without any costs. The transactions are approved by other members of the community and no identification measures are taken, which contributes risk. Criminals abuse the decentralized system to make anonymous transactions, which puts a higher pressure on law enforcement agents (Financial Action Task Force, 2010). Hence, there is a need for regulations that fit this new technological environment (European Central Bank, 2011).

4.2 Virtual Currency
A virtual currency is a type of unregulated digital money, which is issued and often controlled by the developers of the currency (a decentralized system). It is used and accepted among specific virtual communities. (European Central Bank, 2012) Table 3 illustrates the different money formats and corresponding legal status.

Table 3 A money matrix (Henning and Nordin, 2014)(ECB, 2012).

<table>
<thead>
<tr>
<th>Legal status</th>
<th>Physical</th>
<th>Digital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unregulated</td>
<td>Certain types of local currency</td>
<td>Virtual currencies</td>
</tr>
<tr>
<td>Regulated</td>
<td>Bills and coins</td>
<td>E-money</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Commercial bank money (deposits)</td>
</tr>
</tbody>
</table>

The first virtual currency, Flooz, was introduced online in February 1999 and was an attempt to establish a unique currency for internet merchants. By 2001 the FBI notified the company Flooz.com about their currency being used in a money-laundering scheme by a Russian organized crime syndicate and the co-founder of Flooz stated in 2001 that 19% of purchases made with Flooz were fraudulent (Tedeschi, 2001). Later in 2001 the company announced its closure and all unused Flooz became worthless and no way of refunding them existed, which led to an exhaustion of about 35-50 million USD in venture capital (Aune, 2010).

There are many virtual currencies and today’s biggest and most used virtual currency, Bitcoin, was introduced in early 2009 by someone called Satoshi Nakatomo. He stated that he had solved the issue of “double-spending”. By using a peer-to-peer network (P2P-network) the currency was completely decentralized and had no servers or central authority (Bitcoin, 2016). A P2P-network is a non-hierarchal network of connected nodes, or computers, which does not communicate through the standard client-server model. The computers within the P2P-network is not assigned specific roles in the communication and therefore no one has any privileges relative to the other nodes.
4.2.1 Bitcoin

The issue of double-spending is a classic problem within systems handling different virtual currencies. A digital “coin” can easily be copied, which gives the copier opportunity to use the coin an infinite amount of times and/or share it with others. The easiest way to handle this problem has earlier been to use a central authority which regulates and keeps track of what has been spent etc. What separates Bitcoin from other virtual currencies is the fact that it’s completely decentralized. Bitcoin is built on peer to peer technology and no central authority is therefore needed. (Nakamoto, 2008)

4.2.1.1 Exchange services

Different exchange services offer the possibility to exchange real economy money through bank transfers, Internet banking, swish or other payment solutions into Bitcoin. There are different security measurements taken within these exchanges but most of them require a personal identification of the person buying Bitcoin. An example is Coinbase that demands that you register your driving license or another identification document. (Bitcoin, 2016) It exists some more anonymous ways to purchase Bitcoin through cash transactions. Some services offer Bitcoins for cash sent in an envelope through mail. There are even brokers that match buyers and sellers for a fee, so the exchange can be realized through cash deposits anonymously (Bitcoin-Brokers, 2016).

4.2.1.2 Encryption

Bitcoin uses asymmetric encryption which in short means that two different encryption keys, one public and one private, are used to encrypt and decrypt. What this means is that the public key, which is used to encrypt, can be shared. This encryption key can then be used by all people who got it, to encrypt messages, while only the person holding the private decryption key, can decrypt the specific messages. Within Bitcoin, the public part of this key-pair is called a Bitcoin address or the Bitcoin wallet ID. This means that a public address has an account balance and anyone can send Bitcoin to it, similar to an account number. To send money from an address, the private key has to be known. (Nakamoto, 2008)

4.2.1.3 Transactions

When Bitcoin is to be transferred between two different addresses, a transaction is created by the owner of the sending address. The transaction is then signed with the private key corresponding with the sending address and then the transaction is made public to the network. When the transaction is made public, the network of nodes (other users) knows about the sending and receiving address and also the amount of the transaction. This makes it impossible for the owner of the transaction to send the same transaction to multiple addresses thus making double-spending impossible. To make sure that the correct receiver of the transaction gets his/her payment, then network agrees upon which receiving address that is correct. An unwritten rule, which is somewhat arbitrary, is that at least six separate confirmations has to be made to finalize the payment. (Nakamoto, 2008)

4.2.1.4 Block chain

Since Bitcoin is a decentralized virtual currency, no central authority exists who can regulate and approve transactions, there exists a public ledger. This ledger is a distributed database called the block chain, where all transactions that occur in the Bitcoin network are stored. For a transaction to be verified, it needs to exist in a block in the block chain. To create a new block in the chain, one has to find a smaller hash code than the current smallest that exists, more in Chapter 4.2.1.5. A hash function is used to represent data of arbitrary size with data of fixed size. The data is given a hash value, which depends on how the hash function is constructed and this makes the database and table lookup faster. A block is composed by the transactions that are supposed to be included, a reference to the previous block and also an arbitrary number denounced “nonce”. These attributes are then hashed with SHA-256 hash protocol and if the
value is lower than the current difficulty-level, a new block is created and added to the block chain. The difficulty-level is defined as a measure of how difficult it is to find a hash value below a given target. The Bitcoin network has a set global block difficulty, which is regulated every two weeks to limit the created number of blocks to one new block every ten minutes as of now. If a new block is created, the creator is paid 25 Bitcoin as of now, a number which is bisected approximately every four years and the maximal amount of Bitcoin will over time be approximately 21 million BTC. (Nakamoto, 2008)

4.2.1.5 Mining
Miners solve complicated math problems to create new blocks and thereby new Bitcoin. Mining is designed to demand high capacity and it is difficult to limit the number of blocks found each day. To be deemed valid, each individual block must contain a proof of work which is verified by other Bitcoin nodes each time the nodes receive a block. To create a new block is a “trial-and-error”-work where the goal is to find a hash value that is small enough and consequently this means that the more hash values a single user is able to create during a certain time, the higher the chance that one of these values can be used to create a new block. If a new block is generated, a reward will be paid to the creator of the block. When Bitcoin were introduced in 2009, anyone could verify transactions and thereby create new Bitcoin, using their own computer. The Central Processing Unit (CPU) of the computer did the work and the more power the CPU had, the higher the chances of creating a new block. Because of the extreme CPU-power needed, the time to find a hash value that is small enough to create a new block can be very long, which created the need for miners to join together in groups, or pools, to increase their joint CPU-power and then later share the reward if a block is created. (Bitcoin, 2016)

4.3 Anonymity
Anonymity means “without a name”, or “nameless”. The definition as nameless does not capture the complete context of what anonymity is (Pavliček, 2005). Someone that is anonymous, is without a known identity and is unreachable, untraceable and/or not possible to identify. Anonymity today has transformed from being a concept to being a technique of modern privacy.

The concept of anonymity is not something that was brought to the world by the Internet. Nowadays it is so closely connected to the internet though, because the phenomenon made it simpler to be anonymous. It is more difficult to prove your identity on the Internet, than it is to hide it (Pavliček, 2005). Depending on the level of technical competence, anonymity can be fully achieved or only partially achieved. If the user is of high technological competence, and does what is necessary, they can be completely anonymous (Goddyn, 2001). It is difficult to achieve that complete anonymity and there is a big difference between complete anonymity and perceived anonymity, where perceived anonymity is when a user has gone through some processes to become anonymous and is satisfied, while hackers with a high degree of technical competence can reveal their identity in a matter of seconds.

Anonymity creates more freedom of expression and less accountability (Berglund and Palme, 2004). The internet was not created and designed to help users be anonymous. On the internet, every computer has a IP address which is an address that is used to access resources on the internet and can thereby be used to identify the computer and hence connect it to regional information. To address this anonymity flaw of IP tracking, users tend to use anonymizing techniques such as The Onion Router (TOR), which is further described in 4.3.1, and Invisible Internet Project (I2P). Basically, what these services does is rerouting the packets that are sent by the user’s computer through a network of other computers before sending it to the destination address. The most important thing about this rerouting is that the only thing that is known to the other computers that the packets are sent to, is the previous address and the next address. This means that both the origin of the packets and the final destination is never known by any computer.
in the network. (TOR project, 2016)

4.3.1 TOR
The Onion Router is a network with volunteer-operated servers that helps people increase their privacy and security on the internet, by becoming more anonymous. Users who connect to the TOR network allows routing of traffic through their servers and computers. This routing is done through a series of computers, making it harder to track the origin of the data. Users do not connect directly to the network, but instead connects through a series of virtual tunnels, thus further anonymizing their identity. TOR can also be used to connect to sites that are blocked by their Internet Service Provider (ISP) and/or for socially sensitive communication such as searching for specific syndromes of illnesses, going on a forum for rape victims and/or searching for illegal material. The TOR project is also advocating smart surfing and the TOR browser to further anonymize the users’ identities thus making it harder for authorities or other investigators to identify the users online. Hence, TOR helps repressed people regain their freedom of speech in countries that deliberately blocks certain websites. (TOR project, 2016)

4.3.2 Tumblers
A virtual currency tumbler is a service that certain companies offer to mix identifiable virtual currency funds with others, to increase the difficulty of tracing them back to the original owner. This can be done with or without criminal activity involved, some users just want their online funds to be anonymized, while criminals use it to confuse legal authorities in their attempts to find evidence of illegal activity. Compared to the traditional financial system, using a tumbler is similar to moving funds through banks located in countries with bank-secrecy laws that does not allow the banks to mediate any information to international authorities. Like those banks, tumblers take a fee for using their service, which usually is a small percentage of the total transaction (1-3%). There are people who works with financial crimes who advocates criminalizing tumbling services, such as Jeffrey Robinson who is a financial crimes author. The reason why they want to criminalize the tumblers is because of the incentive for criminal activity which it gives and also of the potential for criminals to use this to further increase their anonymity. (Allison, 2015)

4.3.3 CoinJoin
The idea behind CoinJoin and other similar methods is to make joint payments with other outside parties who are also going to make a payment of similar size. When the two parties merge and make a joint payment there is no way to connect the input and output of the payments for a third party, which means that authorities will have a hard time trying to relate the payments to the correct identity. This method increases the privacy for all parties, even those who are not using the method. This is because when a transaction arrives in a single wallet, it is easy to trace the money back to the previous owner, but it will be harder to trace the money back to its original owner if someone on the way has made a joint payment. (Maxwell, 2013)

4.4 Cybercrime
Cybercrime is a fast growing segment of the total crime sector. Offenders of the internet is exploiting the speed and anonymity that is offered, generating victims worldwide. Law enforcement agents typically divides internet related crimes in to two sections; advanced cybercrime and cyber-enabled crime. The first corresponds to sophisticated attacks against computer hardware and software while cyber-enabled crimes are “traditional” crimes that have adapted to the new technological environment. Involving for example exploitation of children, money laundering and terrorism. Today, criminal networks use the Internet to facilitate their business for commercial purposes. Individual perpetrators are getting more educated in the sophisticated techniques required to securely commit crimes, which put a higher pressure on the law enforcement. (Interpol, 2016)
4.4.1 Cybercriminals using financial solutions

Cybercrime is the second most reported crime within the economic sector (PwC, 2016). When referring to cybercrime within finance, a connection to hackers of the bank system (advanced cybercrime) are assumed. In this section, the offenders that have adapted to the technological landscape (cyber-enabled crime) is mapped, hence not professional hackers. This corresponds to people committing “traditional” crimes online by using the different financial systems, for example Bitcoin- or cash transactions.

The Internet is divided into two parts; the deep web and the surface web. According to Brightplane (2014) the surface web is anything that a search engine can find while the deep web is anything that a search engine cannot find. “Normal” people only visits the surface web during a life time on the Internet. The dark web is classified as a portion of the deep web that has intentionally been hidden and is inaccessible through standard web browsers. The most famous content that resides on the dark web is found in the TOR network. This is the part of the Internet most connected to illegal activity due to the anonymity associated with TOR. (Brightplane, 2014) Silkroad was a popular black market trading place within the dark web that offered visitors all kinds of illegal goods and services. The payments for these illegal products were made through Bitcoin transactions. The decentralized and relatively anonymous payment solution was thereby misused by criminals to purchase illegal goods. This is just one example of how the virtual currency is abused by offenders.

4.4.2 Child Abusive Material

Viewing, sharing, downloading or in other ways being connected to Child Abusive Material is against the law in Sweden. The definition from the European Union law explains that child abusive material corresponds to “any material that visually depicts any person appearing to be a child engaged in real or simulated sexually explicit conduct or any depiction of the sexual organs of any person appearing to be a child, for primarily sexual purposes” (Eur-Lex, 2011).

With the development of the Internet follows an efficiency in sharing illegal images. Instead of trading/buying/sharing physical images Peer to Peer (P2P) the offenders have discovered that it is much easier and safer to exchange/buy/share material online. In this way no physical evidence may be used against the criminals if they have an advanced knowledge of the required anonymization techniques. As mentioned earlier, the dark web is used to trade illegal images, and much more, but there are other solutions that are available on the surface web that are getting more popular, such as cyberlockers (Köhler, 2015).

To remain anonymous, the perpetrators are increasing their knowledge about the technological aspects required, which makes it more difficult for the law enforcement to catch them. A clear example of this development is the shift from card payments in the purpose to purchase CAM to more sophisticated solutions, such as Bitcoin transactions, which requires a higher degree of technical competence.

4.5 The Fraud triangle

The Fraud Triangle is a model which originates from the theories of Donald Cressey, an American criminologist who studied organized crime, sociology of criminal law, white-collar crime and criminology (Cressey and Sutherland, 1978). The model is created with the purpose of explaining the underlying factors that can cause someone to commit fraud. The model consists of three parts, which together can

---

lead to fraudulent behavior. The three “pillars”, or corners, in this model are; Pressure, Opportunity and Rationalization, which can be seen in Figure 14. According to Cressey et al. (1978) all three factors must be present at the same time, in order, for an ordinary person to commit fraud.

**Figure 14 The Fraud Triangle**

The first corner of the fraud triangle is represented by *pressure*. This is the underlying motivation of the person to commit fraud. For example, if the person has financial problems that he/she is unable to solve through legitimate measures, illegitimate means becomes more considerable. This can range from stealing cash and robbery, to fraud and embezzlement. The problematic financial situation of the person can be either personal, i.e. personal debt, or financial, with pressure from a person higher up in the hierarchy as an example. Other pressures that can be an underlying reason for an ordinary person to commit fraud include; inability to pay one’s bills, drug addiction and desire for status symbols. (Cressey and Sutherland, 1978)

The second corner of the model is called *opportunity* and is defined as the approach to the crime by the person and hence the method by which the crime is committed. The person must see some way to solve his/her financial problems with a low risk of getting caught. The most important thing for these fraud perpetrators, is that it is possible to commit this crime in secret, without their family, friends and/or coworkers finding out. White-collar crime is often committed for the sole purpose of maintaining social status, while a secret drug problem which the person no longer can pay for can make a person steal to be able to conceal the addiction while still being able to get drugs. The problem that the perpetrator is trying to conceal is often something that can hurt his/her social status and if the perpetrator is caught it will most likely hurt the social status at least as much as the underlying problem would have. This means that apart from the fact that the perpetrator needs to be able to steal funds, he/she also needs to complete the act in such a way that it minimizes the chances of getting caught. (Cressey and Sutherland, 1978)

The third and last corner is *rationalization*. A majority of the people committing fraud are first-time offenders without a criminal record, which means that neither they nor the society views them as criminals. Their image is often ordinary, honest people and this is also their own view of themselves. A vast majority of the perpetrators that are first-time offenders believe that the crime is legitimized by them just being caught in a bad set of circumstances. Consequently, the perpetrator justifies the crime for himself/herself to make the act more acceptable. (Cressey and Sutherland, 1978)

4.6 Time series analysis

Time series analysis has two main purposes; identifying the nature of the phenomenon represented by the sequence of observations and forecasting. The analysis is based on sequences of measurements that
follows a non-random order and that the data points are presented with equally spaced time intervals. To achieve the purpose of the analysis the pattern of observed data is identified to thereafter be interpreted and integrated with other data, for example seasonal price changes. Later the identified patterns can be extrapolated to predict future events. (Dell, 2015a)

It is assumed that the time series data consists of pattern and a random noise, which complicates the identification of the pattern. Therefore, most time series analysis involves some sort of filtering to limit the noise. Most patterns can be divided into two components, trend and seasonality. Trends represents a general systematic linear or nonlinear component that changes over time and does not repeat or at least does not repeat within the time range captured by the collected data. Seasonality may have a similar nature but it repeats itself in systematic intervals over time. The two general classes of time series components can coexist in real life data. For example, a company’s sales can consistently grow over 20 years but it still follows seasonal patterns, like the sales numbers are always spiking in December every year. (Box and Jenkins, 1976)

### 4.6.1 Analyzing trends

If the time series data points clearly are consistently increasing or decreasing the analysis is quite simple. If the data series is disturbed by considerable errors, different measures can be taken. Smoothing is a technique that involves local averaging of data. The most common type is moving average, which replaces each element of the series by either the simple or weighted average of $n$ surrounding elements, where $n$ is the width of the smoothing window. (Box and Jenkins, 1976) (Velleman and Hoaglin, 1981)

When the measurement error is very large, the distance weighted least squares smoothing or negative exponentially weighted smoothing techniques can be used. The methods will filter the noise and convert the data into a smooth curve that is relatively unbiased by outliers. Series with relatively few and systematically distributed points can be smoothed with cubic splines. (Dell, 2015a)

Many monotonous time series data can be adequately approximated by a linear function; if there is a clear monotonous nonlinear component. The data need to be transformed to remove the nonlinearity, which can be performed through logarithmic, exponential or polynomial functions. (Dell, 2015a)

### 4.6.2 Analyzing seasonality

Seasonality is defined as correlational dependency of order $k$ between each $i$th element of the series and the $(i - k)$th element and is measured by autocorrelation, where $k$ is the lag. If the measurement error is not too distinct, seasonality can be visually identified in the series as a pattern that repeats every $k$ element. (Kendall, 1976)

Seasonal patterns of time series can be examined via auto correlograms. The technique displays graphically and numerically the Autocorrelation Function (ACF) for consecutive lags in a specified range of lags. The autocorrelation function is based on serial correlation coefficients and their standard errors. Typically, the size of the auto correlation is of higher interest than its reliability, because the relevant ones are the strongest, and consequently the most significant. When examining correlograms, it is important to keep in mind that autocorrelations for consecutive lags are dependent. For example, if the first element is closely related to the second, and the second to the third, then the first element must also be related to the third one etc. This implies that the pattern of serial dependencies can change considerably after removing the first order auto correlation. (Dell, 2015a)

Another useful method to examine serial dependencies is to examine the partial autocorrelation function (PACF) which is an extension of autocorrelation, where the dependence on the intermediate elements is
removed. Hence, if a lag of 1 is specified then the partial autocorrelation is equivalent to autocorrelation. The partial autocorrelation provides a clearer picture of serial dependencies for individual lags. (Box and Jenkins, 1976) (Hay et al., 1980)

Serial dependency for a particular lag of \( k \) can be removed by differencing the series, which means converting each \( i \)th element of the series into its difference from the \( (i - k) \)th element. Thereby the hidden nature of seasonal dependencies in the series can be identified and the seasonal dependencies is removed to make the series stationary, which is necessary for some other techniques. (Dell, 2015a)

### 4.6.3 Correlation

Correlation is a statistical technique that can show whether, and how strongly, pairs of variables are related. Although some correlation is fairly obvious, the data may contain unsuspected relationships and a correlation analysis can lead to a greater understanding of the data. There are several different correlation techniques, but the most common type is called the Pearson product-moment correlation, which is best performed on linear relationships. The method calculates the covariance between the two variables to later divide the result by the two’s standard deviations, according to

\[
\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}
\]

(4.6.3.1)

where \( \rho_{XY} \) is the correlation between two stochastic variables, \( X \) and \( Y \), with the expected values, \( \mu_X \) and \( \mu_Y \), and the standard deviations of \( \sigma_X \) and \( \sigma_Y \). (The Survey System, 2012)

Correlation is applied on quantifiable data and the result of a calculation is called the correlation coefficient, \( \rho_{XY} \), that has a range from \(-1 \) to \(+1 \). The closer \( \rho_{XY} \) is to \(-1 \) or \(+1 \), the more closely the two variables are related. If \( \rho_{XY} \) is close to \( 0 \), it means that there is no relationship between the variables. If \( \rho_{XY} \) is positive, it means that as one variable gets larger the other gets larger. If \( \rho_{XY} \) is negative, this corresponds to that as one variable gets larger, the other gets smaller. This phenomenon is denounced as an inverse correlation. (The Survey System, 2012)

While the correlation coefficient is presented as \(-1 \leq \rho_{XY} \leq +1 \), squaring the result makes the phenomenon easier to understand. The square of the coefficient, \( \rho_{XY}^2 \), is equal to the percentage of the variation in one variable that is related to the variation in the other. Hence, if \( \rho_{XY} = 0.5 \) this corresponds to that 25% of the variation is related. (The Survey System, 2012)

### 4.7 Controls

A vulnerability is a defect in a process, system, application or other asset that creates the potential for loss or harm. Vulnerabilities are measured primarily through the identification of defects or weaknesses to determine a system’s or process’ propensity for failure. In the surrounding environment there are external controls that protect companies from certain risks. The administrative control is based on laws, regulations, policies, practices and guidelines. The logical control is based on application of technical controls such as firewalls, anti-virus software and encryption. The physical control such as video surveillance systems, gates and barricades, serves the purpose to govern access to an office or other important buildings. These three controls are critical to the creation of an effective control environment. However, these elements do not provide clear guidance on measuring the degree to which the controls mitigate the risk. To further minimize the risk of exposure within the organization, three more controls are important to consider: preventative, detective and corrective. (Information System Handbook, 2016)
4.7.1 Preventative
The preventive controls are procedures that prevent the loss or harm from occurring. For example, a control that enforces segregation of responsibilities (one person can submit a payment request, but a second person must authorize it), minimizes the chance that an employee can issue fraudulent payments. (Information System Handbook, 2016) Preventive controls can be as simple as locks and access codes to sensitive areas of a building or passwords for confidential information (CFO Career Planning Tools, 2013).

4.7.2 Detective
These controls monitor activity to identify instances where practices or procedures were not followed. For example, a business might reconcile the general ledger or review payment request audit logs to identify fraudulent payments. (Information System Handbook, 2016) A security camera is a good example of a detective control. An access log and an alert system can quickly detect and notify management of attempts by employees or outsiders to access unauthorized information or parts of a building. (CFO Career Planning Tools, 2013)

4.7.3 Corrective
Corrective controls restore the system or process back to the state prior to a harmful event. Coupled with preventive and detective controls, corrective controls help mitigate damage once a risk has materialized. An organization can document its policies and procedures, enforcing them by means of warnings and employee termination when appropriate. When managers back up data they can restore a functioning system in the event of a crash. If a disaster strikes, business recovery can take place when an effective continuity and disaster management plan is in place and followed. (CFO Career Planning Tools, 2013)

4.7.4 Conclusion
Of the three types of controls, preventative controls are clearly the best, since they minimize the possibility of loss by preventing the event from occurring. Corrective controls are next in line, since they minimize the impact of the loss by restoring the system to the point before the event. However, the restoration procedure may result in some degree of loss, since the restoration procedure may lead to the unavailability of systems and applications along with possible lost productivity, customer dissatisfaction, etc. The least effective form of control, but the one most frequently used, is detective controls - identifying events after they have happened. Depending on how soon the detective control is invoked after an event, a business may uncover a loss long after there is any opportunity to limit the amount of damages. (Information System Handbook, 2016)

One other valuable distinction to be made with controls is whether they are manual or automated. A business can implement manual controls to minimize the chance of fraudulent payments, such as requiring an administrator and a manager to manually sign the applicable paperwork to indicate that the transaction was authorized and approved. As an alternative, the business could automate these controls by introducing a computer program with logical access, segregation of duties and checker controls. (Information System Handbook, 2016)

4.8 Webcrawler
A webcrawler is a program that is meant to imitate human behavior on the internet. The webcrawler is created to systematically browse the World Wide Web, with the purpose of indexing the web, or to collect specific information from the web. Web search engines, such as Google, uses webcrawlers or other spidering software for indexing the World Wide Web to let the users’ search results be more relevant and also to be able to display sites that are currently offline by presenting an older offline-version of the site in question. There exist tools for websites to hide from webcrawlers to make their website not crawlable. It
is almost impossible to make a complete index of the internet, since the number of pages that exists is extremely high.

A webcrawler works along a given protocol as well as dynamic and/or static functions. It starts with a given list of URLs to visit, often called the seeds of a webcrawler. When the crawler visits the seeds it starts with identifying all of the hyperlinks presented on the given URL and adds them to the list of URLs to visit next, called the crawl frontier. These sites that are listed in the crawl frontier are then recursively visited according to the policies set up in the webcrawler. Depending on what type of webcrawler, the mission can differ. If the purpose of the crawler is to perform an archiving of websites, the webcrawler will copy and save the information found on each site. For an archiving crawler, the information found will be saved as “snapshots” from the time of crawling the site, and can later be viewed and read in “offline-mode”. The number of possible URLs crawled is generated by the server-side and this makes it difficult for webcrawlers to avoid retrieving duplicate content. There exists an endless number of possible URL-based parameters, but only a small selection of this will return unique content. (Masanès, 2007)

This creates a mathematical problem for crawlers since they must sort through these seemingly endless combinations to find unique content. For example, if a website containing downloadable files offer three options to the user and there exists four different ways of sorting the files, three choices of preview size, two different file formats and an option to disable content that other users has uploaded, the same set of content can be accessed with 48 different URLs, which all give the same unique content. (Masanès, 2007) This puts pressure on the developers of webcrawlers to make them more efficient and scalable to the maintain quality of the crawl and creates a demand of carefulness when choosing the URLs for the crawl frontier (Edwards et al., 2001).

The earlier mentioned protocol and policies that defines the behavior of a webcrawler is the outcome of a combination of different policies:

- a selection policy which defines which sites to visit and which pages to download,
- a re-visit policy that states how often to check if a website has changed since the last download,
- a politeness policy that defines how overloading of websites is avoided, and
- a parallelization policy which states how to coordinate distributed webcrawlers. (Castillo, 2004)

4.8.1 Selection policy

Even the largest search engines only cover a portion of the surface web, also known as the publicly available web. According to Gulli and Signorini (2005), the large-scale search engines successfully index about 40-70% of the surface web. Since a webcrawler only downloads a fraction of a web page, it is very desirable that this fraction contains the most relevant pages and not just random samples from the website. This creates a need for a metric of prioritizing content and pages by importance, which corresponds to a function of its quality, its number of visits and the amount of links to the website. A good selection policy must also be designed to work with only partial information, since a complete set of web pages is not known during crawling. (Cho and Garcia-Molina, 2000)

Crawlers may be created with the intention to only crawl content once, something that can be done with the use of URL normalization. This refers to the process of modifying and normalizing a URL in a dependable manner. The normalization can be done by, for example, converting all URLs to lowercase, to remove all punctuations etc. (Menczer et al., 2004) The prioritization of importance done by a crawler can be expressed as a function of the similarity of a page to a given query. If the importance function is written in this matter, and the webcrawler consequently only searches for pages similar to the given query, the method is called focused crawling.
4.8.2 Re-visit policy
As mentioned in Chapter 4.8.1, the internet is extremely large and the web also has a very dynamic nature. To successfully crawl even a fraction of the surface web can take several weeks, which means that when the crawling is done, new websites could have been created and some deleted. Plus, the content will most likely have changed on some of the websites. For the major search engines, an instance of not detecting an event implicates a cost, since they will have an outdated copy of the page until crawled again. The cost-functions that are relevant here are called freshness and age (Cho and Carcia-Molina, 2000).

Freshness is a binary measure which indicates if the local copy is still accurate or if there exists a more accurate and updated version on the surface web. The freshness, $F_p(t)$ is defined as

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases},$$

(4.8.2.1)

where $p$ is the page in question and $t$ is the time of the observation.

Age is defined as an indication of how outdated the local copy is and will therefore be relevant mainly when the freshness has the value zero. The age $A_p(t)$ is defined as

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified at time } t \\ (t - \text{modification time of } p) & \text{otherwise} \end{cases},$$

(4.8.2.2)

where $p$ is the age of the page in the repository and $t$ is the time of the observation. (Cho and Carcia-Molina, 2000)

4.8.3 Politeness policy
The reason for using a webcrawler is that the automated work performed by a crawler is much quicker and reaches a greater depth than human searches would have. The negative part about it is that the crawler can have a crippling impact on the website's performance. If multiple crawlers were to crawl the same site at once, and every single crawler were to perform multiple requests and downloads of files from the site every second, the server would be overloaded. Webcrawlers demand high bandwidth and a works with a high degree of parallelism which can lead to server overload if the frequency of accesses to a given server is too high. If a crawler is poorly produced, it can crash servers.

To keep this problem at bay, a solution called the robots exclusion protocol was made. This is also known as the "robots.txt"-protocol that is now a standard for site administrators to write, to indicate which parts of their servers that should not be accessed by crawlers (Koster, 1996). Limiting the interval of crawler visits to a certain website is the most effective way of avoiding server overload, but this is not part of the robot's exclusion protocol. It is possible, and used by the largest search engines, to implement a crawl delay in the robots.txt-protocol. The first proposed delay between successful page loads was 60 seconds, but this posed problem when trying to crawl a great set of websites, since each website can consist of hundreds of thousands of pages. (Koster, 1993) Which delay that the developers implement varies, but a common use is one second.

4.8.4 Crawling the deep web and the darknet
A great amount of pages lies hidden from the public surface web and exists on something called the deep web. This data can for example be data only accessible by submitting a query to a database and this is something that regular crawlers are not able to do. Google's Sitemaps is an example of a protocol intended to discover these deep web resources.
In 2016, a report called Cryptopolitik and the darknet by Moore and Rid was published, where a crawler has been implemented to crawl the darknet. The results that they were looking for was a categorization of all websites found on the darknet. To be able to get these results they developed a webcrawler for indexing pages and websites on the darknet and combined the output of the crawler with a machine learning software called Support Vector Machine document classifier. When crawling the darknet, one has to be careful of illegal material. Moore and Rid (2016) avoided this problem by only harvesting material in plain text. All other material that was not textual content was immediately discarded. (Moore and Rid, 2016) The results from Moore and Rid’s study is presented in Chapter 7.1.1.

4.9 Multidimensional Scaling
Multidimensional scaling is a method that can be used to analyze almost any kind of distance or similarity matrix. The similarities in the matrix can be represented by almost any arbitrary factor, depending on what is supposed to be analyzed. An example input to a MDS analysis could be all the major cities of Sweden. (Dell, 2015b) If we would limit the analysis to two dimensions, the graphical representation of the input would be a two-dimensional map corresponding to the normal geographical map of Sweden. If MDS then were implemented it would be possible to explain these different distances between the cities in terms of different underlying dimensions, for example how many people that travels between certain cities.

In similarity to Principal Component Analysis (PCA), the actual orientation of the axes is in the end arbitrary. This means that the resulting coordinate system can be rotated as desired, and still display the same results. The final orientation of the axes is therefore just a result of the researcher’s choice of representation. Because of these factors, the conclusion can be drawn that MDS is not an exact procedure, but rather a method to rearrange objects efficiently and give the researcher a better configuration that better approximates the observed distance between the coordinates. What the MDS method does in practice is to evaluate the distances between the coordinates and then control how well the coordinate system can be reproduced with the new configuration of axis. This is done with a function minimizing algorithm to maximize the goodness-of-fit. (Dell, 2015b)

In combination with the MDS method, it is necessary to evaluate the goodness-of-fit, by for example using stress measures. The value used for this evaluation is \(\Phi(i)\), which is described as the raw stress value and holds the definition

\[
\Phi(i) = \sum (d_{ij} - f(d_{ij}))^2,
\]

(4.9.1)

where \(d_{ij}\) corresponds to the distance between points \(i\) and \(j\) in the reproduced coordinate system, \(d_{ij}\) stands for the observed input data and \(f(d_{ij})\) is a monotone transformation of the input data. This stress measure will attempt to produce the general ordering of distance between the analyzed objects and this transformation has the output \(d\). This method is a variation of the well-known Minimum Mean Square Error (MMSE) method. (Dell, 2015b)

The input to the MDS method is a matrix containing the distances between the data points. This matrix is typically constructed through the Euclidean distance method, but not necessarily. The Euclidean distance between two points is defined as the line segment connecting them. If \(p = (p_1, p_2, \ldots, p_n)\) and \(q = (q_1, q_2, \ldots, q_n)\) are two points in an Euclidean n-space, then the distance, \(D\), from \(p\) to \(q\) is given by the definition of Pythagorean formula:
\[ D(p, q) = D(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}, \quad (4.9.2) \]

The output of the transformation, \( \hat{d} \), can be observed when constructing a so-called Shepard diagram, consisting of both the observed input data and also the reproduced distances. The Shepard diagram is a scatterplot which uses the input data on the horizontal (X) axis and the reproduced data on the vertical (Y) axis, as illustrated in Figure 15. The plot will show a step-function which corresponds to the \( \hat{d} \)-values from transformation of the input data. Deviations from the step-line indicate a lack of fit and a perfectly reproduced model will have all the coordinates on this step-line and therefore have minimal Phi-value. (Dell, 2015b)

\begin{figure}[h]
\centering
\includegraphics[width=\linewidth]{shepard_diagram.png}
\caption{An example of how a Shepard diagram can look like.}
\end{figure}

4.10 Agglomerative Hierarchical Clustering

The basic idea of Agglomerative Hierarchical Clustering (AHC) is to ensure that nearby (shortest physical distance, calculated through the Pythagorean theorem) points end up in the same cluster (Lavrenko, 2014). All \( n \) clusters are included in the collection \( C \), according to Figure 16. The procedure is repeated until one entity is created with internal connections between all clusters (see Figure 17 for an illustration of the cluster network). To begin, each cluster is represented by one data point, \( c_i = \{x_i\} \).
The next step corresponds to identifying the pair of closest clusters, \( \min_{i,j} D(c_i, c_j) \), where \( D \) is the physical distance between clusters, to later merge these two into a new cluster \( c_{n+k} \). \( c_i \) and \( c_j \) are then removed from the collection \( C \) and replaced by \( c_{n+k} \), \( k = 1, 2, 3 \ldots l \). When referring to Figure 16, if \( n = 7 \), then \( c_1 \) would be merged with \( c_5 \) and the two individual clusters would be removed from the collection \( C \) to be replaced by \( c_8 \). This procedure is iterative \( n \) times to measure all distances between clusters until an entity is created (Lavrenko, 2014).

When the iteration is complete a dendrogram is created. A dendrogram illustrates the distances between the different clusters within the entity, according to Figure 17. A distance threshold can be implemented to restrict the number of clusters desired (see the red line in Figure 17). (Lavrenko, 2014)
For example, if it is desired to perform AHC on the 2-dimension analysis generated from the MDS of the map of Sweden. Then the input to the AHC model is the standard geographical map of Sweden. The different cities are the data points defined as individual clusters. If you zoom in on the surrounding cities of Stockholm, you can distinguish different distances between them, according to Figure 18. When performing AHC on these cities a cluster is derived from Uppsala and Västerås, which are the closest ones. Thereafter, Stockholm is the closest cluster to the cluster $A$ in Figure 18. So Uppsala, Västerås and Stockholm now exist within the same cluster, if desired. Örebro is the next closest cluster and thereafter Nyköping. So basically, a decision is required to choose how many clusters that are desired. The alternatives in Figure 18 are five, four, three, two or one cluster(s). If AHC is performed in the whole of Sweden one example might be to cluster the cities that lie within the different landscapes, which generates 24 clusters. Or maybe three clusters that represents the cities in the north, middle and south part of Sweden.
4.11 Clustering

Cluster analysis is meant to be used on a dataset to group together sets of objects in a way that objects in the same group, or cluster, are more similar than objects in other clusters. It is a common technique for statistical data analysis and can be used within machine learning, pattern recognition and image analysis etc. Cluster analysis is not one set algorithm, but rather the concept on which specific cluster analysis algorithms rely. To create a cluster analysis algorithm, the developer of the algorithm needs to define what constitutes a cluster and how to efficiently find them. A cluster can include groups with small distances between the group members, dense areas of the data space and particular statistical distributions, which means that clustering can be viewed as a multi-objective optimization problem. The formulation of the algorithm depends on what parameter settings are used, how the data set is structured and the intended use of the output of the clustering analysis. Because of this, clustering is not an automated process, but rather an iterative one of trial and error. (Kumar et al., 2006)

The fact that it is very hard to define exactly what a cluster is, is the main reason there exists so many clustering algorithms. The common denominator though, is that a cluster always is a group of data objects. Some typical clustering models are listed below:

- Connectivity models, such as hierarchical clustering models based on distance connectivity. Hierarchical clustering models have the property that the objects to a child cluster also belong to the parent cluster. For more information regarding hierarchical clustering, see Chapter 4.10.
- Distribution models, where clusters are modeled using statistical distributions such as multivariate normal distributions.
- Group models, where the models do not provide a model for the results and just provide the grouping information.

Figure 18 Illustrates two examples of how to cluster when performing AHC.
5 Chosen methods

This segment describes the chosen solution method applied to fulfill the purpose presented in section 1.1 of this thesis. Each chosen method is not always implemented together, but the authors have combined the methods to make them fit the purpose of preventing CAM on the internet and to identify and analyze the payment flows regarding this illegal activity. Two different methods to solve the problem has been implemented and is called the primary and secondary method of implementation. The primary method is divided into five parts; data collection, rearranging data, clustering, comparing results and tracing the flow of money. The first part of the solution, called data collection is based on Chapter 3.4 where an introduction to webcrawlers are given and then how to produce one, which is described in section 4.8. The relevant data output from the webcrawler will be presented with the parameters; forum, link, username and timestamp, which will be the input to the next model, Multidimensional Scaling (MDS).

MDS, as introduced in 3.5 and further described in 4.9, is a way to rearrange the data and find relationships between the data points. The output of this model will be a four-dimensional space where the parameters are given numerical values to be able to map them on an axis. The different usernames will for example be called $1, \ldots, n$, where $n$ is the last username found. To understand the relationships and represent this data in a way that the identification of relationships becomes easier, Agglomerative Hierarchical Clustering (AHC) will be performed on the output of the MDS. AHC is introduced in section 3.6 and a more detailed view of the model is given in Chapter 4.10. The idea of this model is to cluster together points with a short distance between them and iterate this process until all data is presented in one unit. The relationships can then be viewed in a dendrogram, as seen in Figure 17. The output from this model can then be compared with a database that the police possess, to give more credibility to suspicions against potential perpetrators and offer the police reasons and evidence to start a preliminary investigation against the perpetrators.

Since the purpose of the thesis is to investigate and analyze payment flows, which is very hard to do when the payment flows mainly consists of virtual currencies, a collaboration with the police is necessary. When the police have enough evidence to start a preliminary investigation of a given suspect, they can then start to trace the money, with the help of currency exchanges and cyberlocker administrators. Given the information presented by these companies, it is likely that they will cooperate with the police if there are legitimate reasons. The police will have to identify the suspects IP-address and also turn to the banks to see which currency exchange the perpetrator used to trade virtual currency. The next step is to collaborate with the currency exchange in question and trace what payments the suspect has done to the cyberlocker and vice versa. Then a collaboration with the cyberlocker will let the police find out what material the suspect has posted on the website and also what material the suspect has downloaded. This will work as evidence against the suspect and the police can then proceed with the investigation. This will be further described in Chapter 5.2.
The secondary method will consist of creating a framework with which it would be easy to say if a file hosting service is believed to conduct their business with knowledge of the illegal exploitation from perpetrators. The framework would mainly be of use to the police and the banks when they are performing their analysis of different file hosting services. The first step was to identify and investigate existing cyberlockers which is described in 5.3.1. The investigated cyberlockers are investigated and categorized by what payment solutions they offer, if they have an option for premium accounts, if the users get payments for uploading successful material etc. The next step is to conduct an analysis of the transaction flows to and from a predefined set of cyberlockers that is proven to contain illegal material. This analysis was performed using Chainalysis tool Reactor and is described in Chapter 5.3.2. After this analysis was performed, the framework was completed with regards to what services are closely connected to the cyberlockers corporate virtual wallet. The thought was that tumbling services, TOR markets and CoinJoins etc, would be closely connected to the corporate wallet. Correlation between what services the cyberlockers use and in what magnitude is stated and implemented in the framework. The framework can then be used on other cyberlockers to establish if it is believed that the viewed cyberlocker is exploited by perpetrators to share and download illegal material, seen from a transaction perspective.
5.1 The CAM triangle
The development of the Fraud Triangle into the CAM triangle is described in this chapter. The method is mainly of use in the discussion chapter to substantiate the arguments.

5.1.1 The first corner – the opportunity to buy, sell or produce CAM
It is a known fact that if the government completely lifts the regulations on alcohol and drugs, more people would try it and some would get stuck and become addicts. When it comes to people trading CAM, a similar pattern can be expected. If child abusive material gets easier to come by, more people would view it out of curiosity. If more people view it, the number of perpetrators who start to regularly download, sell and/or produce CAM will grow since the viewing of the material could awaken a latent pedophilic tendency. One consequence is that, after viewing material of this sort and doing so more regularly, the chances of committing physical abuse on children increases. (Svensson, 2012)

5.1.2 The second corner – the ability to rationalize the crime
This part is a lot different from the original model, since the rationalization performed by the criminal depends on if they are a distributor or a consumer of the CAM. This thesis will, as mentioned earlier, mainly focus on the consumers of CAM. The rationalization is a cognitive way to justify the crime and adjust his/her perception of the act so that it fits his/her moral compass. The major part of the CAM consumers is believed to have pedophilic tendencies and therefore the rationalization will in most cases be that the society’s rules do not fit their sexuality and therefore the society is also at fault. Pedophilia is a psychological disorder, which makes the pedophile in most cases unable to emotionally and sexually bond with adults and therefore seeks the relationship with children. (Svensson, 2012)

5.1.3 The third corner – the pressure on the individual.
The pressure on the individual in this case is the psychological disorder’s pressure on the affected individual to perform these acts, which in many cases is against their will but the pressure is so strong that they cannot resist. The producers of CAM on the other hand have a monetary interest in the matter and the pressure can, for example, be from higher up in their criminal organization for them to produce this type of material so that the organization can make a profit.

5.1.4 Summary
The adapted fraud triangle will work as a guideline to understand the offender’s behavior and to consequently track the transactions regarding CAM. To conclude, this adaption of the fraud triangle will be of use in the future work of the thesis, mostly regarding the consequences of anonymity. If it is easy and safe for offenders to access CAM they will most likely abuse the opportunity and rationalize the behavior. When the addiction is strong and no more images can be accessed, the offender will be more careless and try to purchase images in a not so secure way and hopefully be detected and arrested. The pressure from the society that pedophilia is stamped worse than murder, combined with the fact that the offender’s worst fear is that their relatives and loved ones know the truth about their secret, has given them incentives to increase their knowledge about security and anonymity online. Even though the offenders operate more anonymously, the perpetrators that are the easiest for the police to catch are the ones that are desperate for material and therefore they are more likely to make mistakes which gets them caught. The offenders that are spreading CAM for financial gain are not as desperate for material and hence harder to catch.
5.2 Primary method

In this chapter the primary method is described and the steps are illustrated in Figure 21.

5.2.1 Webcrawler

Complete webcrawlers exists online, which individuals can adapt to generate the preferred output. In this thesis two complete web crawlers were implemented to validate that they work and generate the desired output, which in this case was all posted links within a discussion segment on a web forum. These crawlers were not specified to create the exact output that was desired in this thesis, and were merely implemented to try the concept of data collection through web crawling.

In Sweden it is illegal to actively search for CAM. The legislation incorporates the search of keywords connected to CAM without entering any link or webpage. Due to the legislation combined with the fact that the authors does not wish to accidentally view any CAM, this thesis does not implement webcrawlers in connection to CAM and validation of the method is done on a similar subject instead of CAM. The recommendation is to let the police crawl the surface web, and eventually also the darknet, to gather information about the posted links connected to cyberlockers. To manually validate if a webcrawler
solution is relevant as the data collection for this thesis, a manual search for links to a copyrighted movie hosted within a cyberlocker were conducted instead. The idea was to create a method to identify perpetrators based on illegally shared movies, which can be adapted and implemented with regards to CAM.

When searching for the movie “The Hateful Eight” on Google and filtrating the result by the recent week, several hits included cyberlockers and forums. When entering the identified forum, a link was presented to a cyberlocker containing the illegally distributed movie. The link was not entered, but it most likely contained “The Hateful Eight” in a compressed format to ease downloading. The name of the movie was not necessarily texted in the URL but it was mentioned in connection to the link. When visiting a forum containing a posted link, a username was identified. This link may be posted in another forum by the same or a different username, according to Figure 22. In this way, a person who uses one or two usernames in two different forums can be identified. If the username is the same in the different forums, the conclusion can be drawn that this is the same person that is posting the link. If the username differs from the different forums, the timestamp for the different posts can be identified and if the time difference between the posts is small, a conclusion can be drawn that the user who posted the links is the same underlying person, or that the persons behind the usernames have a close relationship. If no conclusions can be drawn, it is reasonable to assume that further research needs to be implemented, for example according to the MDS method in Chapter 5.2.2. The police have the right to check the two user’s IP addresses and in this way identify the person behind the aliases if a preliminary investigation has been initiated. The problem arises when anonymity measures, such as TOR, are taken by the users. The police also have the authority to follow the link and access the file to validate if the suspicious material indeed is illegal.

The constructed crawler scans the surface web and presents the links that are posted. These are sometimes connected to cyberlockers, which was validated by manually searching for the movie “The Hateful Eight” on Google. The links to the cyberlocker containing the rar/zip file was not entered and therefore no proof exists that the link actually contained illegal material. The idea is to crawl the whole surface web and hence identify all posted links. The list with links is then compared with the identified cyberlockers in Appendix B and the links containing the name of a cyberlocker is marked as interesting. Whether the link is connected to illegal material is still unknown at this stage. When the list of links containing cyberlocker names in the URL is presented, the list should be completed with the corresponding forum in which the link is posted. Basically, the list of forums, links, timestamp and usernames is the output of the crawling session.

The belief is that a webcrawler can be constructed to find, filtrate and organize the data as specified above. The writers lack the sophisticated programming skill to construct such a webcrawler. Due to this, the legal situation and that the area lies out of scope for this thesis, a recommendation is made to
continue the work through a collaboration between the police and professional programmers.

The next step is to investigate the data with the help of the MDS method to identify patterns, trends and relationships between forums, links, the time of posting and usernames, see Chapter 5.2.2. In the long run, after the MDS step, the connections are mapped and offenders can be identified by the police and hence these people cannot continue distributing/downloading CAM. Pedophiles are known for being in contact with each other through chats or other platforms. Hence, when catching one perpetrator, many others can be identified and stopped, resulting in a decrease of CAM on the online market, this will be discussed further in the discussion chapter. Because a complete webcrawler was not constructed, a manual search for forums containing shared links to cyberlockers was conducted. The result is illustrated in Appendix C and illustrates all points in four dimensions; forum, link, username and timestamp. This table is the basis when validating the MDS and AHC method.

5.2.2 Multidimensional Scaling
The observed input data in this thesis will be the output generated from the webcrawler. The MDS will be performed on data which will be categorized in four dimensions; forum, file/link, timestamp and username. The input data is presented in a table with four categories corresponding to the four parameters, according to Appendix C. Each row of the table will have information in each category, i.e. name of forum, full URL to link, what time it was posted and by which user. The reason for choosing these categories is, for example, to discover if the same person uses different usernames within different forums. A way to prove this is to look at the timestamp. If the link is posted close in time in two different forums by two different usernames, the probability is high that the two usernames are represented by the same person. Another example is if one username is active within a specific forum, sharing different links connected to CAM within different cyberlockers. Then a further investigation of the cyberlocker and the username is of great support for the police.

The table in Appendix C was then converted into numerical values, according to Table 4, where the first forum presented will be converted to the numerical value of one, and the second presented will be converted in to the number two and so on. For example, in Table 4 the three first rows are three different shared links within the same forum, by the same username and the timestamp differs. The date and time was converted into numerical numbers through configuration in Excel. This will be done for all four categories which will make it possible to work with the data using statistical and mathematical models.
Table 4 The identified parameters represented as numerical values.

<table>
<thead>
<tr>
<th>Forum</th>
<th>File/link</th>
<th>Username</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>42480,33</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>42450,46</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>42481,50</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>42487,46</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>42480,27</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4</td>
<td>42481,00</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>42486,25</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>3</td>
<td>42487,24</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>5</td>
<td>41689,10</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>6</td>
<td>42089,99</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>7</td>
<td>42488,36</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>8</td>
<td>42487,71</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>9</td>
<td>41025,27</td>
</tr>
</tbody>
</table>

When the data has been organized in the manner presented in Table 4 it will be converted in such a way that the built in MDS method in Matlab can handle the data correctly. If the data were to be used in the way presented in Table 4, it would for example appear as if the data represented by the first row is more similar to row 4 than to row 12. The distance from row 1 and row 12 would appear to be farther than between row 1 and row 4, which is not the case. Row 4 and row 12 are equally dissimilar from row 1 since neither of their respective parameter matches with the parameter values of row 1. To avoid this dissimilarity issue, the data was instead evaluated individually against each row, and a penalty function was implemented to perform the evaluation. In this case there were four parameters in the dataset and three of them were weighted equally. This means that when this study was performed there were none of these three parameters that were believed to be more important than the other. The parameter that was weighted differently than the others were “timestamp” because the potential similarities found regarding this parameter in this example is purely coincidental. For larger data sets the parameter Timestamp needs to be handled in a more sophisticated way, as seen in Chapter 6.1. The values assigned is then evaluated and the lower the value, the smaller the distance, which consequently means that the points have a higher probability of having a relation between them. In the matrix created, see Appendix D, there is a diagonal with zeroes, which corresponds to when the currently investigated row is compared to itself. It is a zero because no dissimilarity can be found and therefore the penalty function will give each of the parameters the value zero.

As stated earlier, there are four parameters in this dataset and three of them are weighted equally. The fourth parameter, timestamp, has been given a very small weight and will have almost no impact in this analysis. The other three parameters have been given the weight of 0,25 each, to make it possible to increase the weight of the timestamp-parameter up to 0,25 if necessary. If the parameter in the investigated row differs from the one that it is currently being compared with, it will be given the value of 0,25 and if it is equal it will be given the value of zero. This is done for the three equally weighted parameters. The timestamp-parameter is evaluated, in this case, with the following formula:

\[
\frac{\text{investigated row timestamp}}{\text{current compared row timestamp}} - 1 \right) / 4
\]

(5.2.2.1)
which will result in a small impact on the result of the comparison. This can be seen in the matrix in Appendix D as the deviations from 0.25, 0.5 and 0.75. If a row has similarities with another row, the value will appear as close to 0.25 or 0.5 and if no similarities can be found, the value will be close to 0.75.

The MDS method will rearrange the objects in a more efficient manner and from that, it will be possible to identify connections between the forums and the different users if the same link has been posted on different forums. For example, if a link is posted at a certain time by a specific user, and just moments later posted at a different forum by a user with a different username, there is a relationship between these situations that would not be possible to see if the timestamp was not a recognized parameter.

When practically performing the Multidimensional Scaling on the example, Matlab was used as a tool. First of all, the table of similarities in Appendix D was imported through the function `xlsread` from Excel to Matlab. The distances between the data points in the similarities matrix was generated through the function `pdist` in Matlab. A vector containing distances, D, was hence created and functioned as an input to the MDS function `cmdscale` in Matlab.

### 5.2.3 Agglomerative Hierarchal Clustering

Agglomerative Hierarchal Clustering (AHC) is an interpretation tool for the output generated from the MDS. The output from MDS is specific parameters (forum, link/file, username and timestamp) presented as numerical values. Hence, the distances between clusters of points with similar behaviors, can be structurally mapped and segmented. Through a dendrogram the data can be divided into a preferred number of clusters that matches the designated results.

Initially all data points are defined as clusters and the distances between them need to be measured, through the equation 4.10.1 in Chapter 4.10, to find the pair that is physically closest in the four dimensional space. The two closest clusters are then joint together into a new cluster. Thereafter the distance between the next two closest clusters are investigated iteratively until all are clustered to one unit according to Chapter 4.10. Through a dendrogram, a visualization of the distances between the different clusters are presented and a decision can be made of how many clusters that are desired. Hence, the number of clusters desired is hard to decide with no finished dendrogram to view.

A webcrawler that generates all links within a forum was found online and implemented. A more sophisticated one needs to be created to fit this thesis, but the writers lack the sophisticated programming skills. Hence, the methodology was instead verified by performing examples imitating the future performance of the desired webcrawler. When performing the Agglomerative Hierarchal Clustering practically on the generated example of data points, the Matlab function `linkage` was used on the distance vector, D. The output was then the input when creating the dendrogram in Matlab through the function `dendrogram`.
The output from the MDS and AHC method is represented as a dendrogram in Figure 23. When dealing with only 13 data points, as in this example, it is quite simple to see the connections between data points visually in Appendix C. This method is meant for a bigger data sample where the dendrogram will be of use to visually detect the connections between the points. The connections of great interest in the example study are the ones between five-eight and one-two. These are the closest linked data points, which in this example means, for example, that the points are represented in the same forum and by the same username. The conclusion from the dendrogram is to further investigate the four points’ common denominators to build a case against the username or the cyberlocker.

5.2.4 Summary
The first choice of action would be to establish a collaboration with the identified cyberlockers that are infected and from there extract the relevant user’s information. If the user does not use an anonymization service and/or proxy server, the IP address of the user’s computer can be identified. The IP address can then be compared with transaction history to the identified cyberlocker containing the specific link that the user posted and the specific Bitcoin address that is used in that case. Thereby, there exist two alternatives for future investigation; further investigation of the cyberlocker that is included in the shared URL that contains CAM and to follow that individual’s Bitcoin cash flow through Reactor to discover if he/she has exploited other cyberlockers. If the individual’s Bitcoin wallet is connected to an exchange service, then the banks can be warned that the incoming money originates from a bad source.

The investigators best choice of action is to contact the relevant internet service provider (ISP) and collect the information from there. This will need a court order and it is therefore of interest to involve the police in the procedure. When or if the IP-address has been found and the person behind it has been identified the researcher will have to hand over the investigation to the police, because the next step of the investigation will include extracting information from the banks and this will require a certain warrant since this information is classified by the banks secrecy act.

Figure 23 Resulting dendrogram from the example.
When the transaction history has been extracted from the banks, the data need to be analyzed and hopefully some transactions to and from a virtual currency exchange can be found. The currency exchanges will more often than not work with the police if a court orders it or if a valid warrant is submitted. This means that the transactions to and from the cyberlockers can be identified and also the previously hidden information on what files the now identified user has posted and downloaded on the website. The patterns in the transactions to and from the cyberlockers can give an indication on whether or not there is illegal material present on the site. The hypothesis is that the files that are most sought after by other users, tend to be illegal material. To make sure that the hypothesis is valid, the researcher will have to investigate the cyberlocker further, but the methodology will make the investigations easier to handle and far quicker.

There exists an international collaboration between many countries through Europol and organizations such as INHOPE, where ECPAT is included. For example, pictures that has been discovered in Latvia that portrait a situation in Sweden is sent by the Latvian police to the Swedish. Basically, the collaboration results in that the material that is produced in respective country ends up in that country’s police’s hands. Hence, the primary method is recommended to be used by Europol and not the national police, because this is an international problem and due to the collaboration it is more effective to perform the method on a higher level.
5.3 Secondary method

The secondary method is based on investigating historical data on Bitcoin transactions to and from cyberlockers to see if connections to suspicious services, such as Tumblers, exists. Basically, the correlation between different services was investigated to create a template of comparison for the police. Another part of the framework will be to determine how big the transaction flows to/from the different cyberlockers are, thereby determining how much money that flows in and out from their individual corporate wallet. The hypothesis is that if the cyberlocker has relatively large amounts of money flowing in and out of their business, the probability that the cyberlocker is getting exploited by illegal businesses for illegal purposes is higher.

5.3.1 Identifying and investigating cyberlockers

To achieve a list of classified file hosting services, a manual investigation of 372 cyberlockers was performed. Lists of cyberlockers were discovered online and out of the 372 cyberlockers, only 155 still exists. These 155 file hosting services were investigated based on certain parameters, according to Table 5. The information was collected through manually investigating the web pages based on the registered parameters. In some cases, it was necessary to create an account to access the information regarding the cyberlocker’s payment methods. A temporary email address was created due to the writers’ unwillingness to receive spam emails after the thesis is completed. With this said, no personal identification is required when creating an account within these services.

As stated, Table 5 illustrates two examples of the categorization made of the different identified cyberlockers. The first four columns describe the different payment methods that can be seen within the cyberlockers. Column four, other payment solutions, can for example be Call2Pay, which is a telephone based payment method where the user calls a service number and the costs will be received on the next phone bill. Columns five and six are just descriptive to how the cyberlocker works with their own payment model, and defines if there is a possibility to buy premium accounts on the cyberlocker to receive certain privileges and column six describes if the users can receive payment for successful files. Column seven corresponds to if the site administrator, or anyone from the company that owns the cyberlocker, has their contact information listed on the website. More often than not, this will have the answer “No” and this does often mean that there is a public email-address, similar to: “contact@cyberlockernameno.info”. The last column of the table is advocated security, and this is just the intuitive feeling of the authors about the sites security. If the cyberlocker presents what security they have and for example how anonymous a user can expect to be when using their service, this will provide them with a higher number in this column. For example, Cyberlocker 2 in Table 5 brags about their cryptology and further promises the user that he/she will be anonymous within the service. Therefore, this cyberlocker received the highest value, five, on the scale of advocated security. Cyberlocker 1, on the other hand, does not even mention the user’s privacy or cryptology at all and hence receives a value of one in the scale. The values in between, are harder to define specifically and here is where the intuition is used. If the cryptology of the service is mentioned on the start page within the service and nowhere else, then this should receive a value of two. The value of three should be registered if the anonymization measures are mentioned in the start page of the cyberlocker. The value of four is registered if the anonymization of the users is mention here and there within the service, but not everywhere and not underlined.
Table 5 Illustrates two examples of the categorization by different parameters of the file hosting services.

<table>
<thead>
<tr>
<th>Cyberlocker</th>
<th>Bitcoin</th>
<th>Card</th>
<th>Solutions like Paypal, Klarna, WyWallet etc.</th>
<th>Other payment solutions, such as Call2Pay</th>
<th>Premium account</th>
<th>Paid to upload</th>
<th>Contact information to a private person</th>
<th>Advocated security (1-5, where 1 is low)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyberlocker 1</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>1</td>
</tr>
<tr>
<td>Cyberlocker 2</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>5</td>
</tr>
</tbody>
</table>

### 5.3.2 Chainalysis

The idea was to identify the cyberlocker's Bitcoin wallet ID by just looking within the service. As discovered, the cyberlocker does not reveal the ID of their Bitcoin wallet until you make a payment for a premium account. Therefore, 0.13 Bitcoin (BTC), worth 50 Euro at the time, was bought at the exchange Coinbase and a Bitcoin Wallet, hereby denounced A, was created. After that, a premium account was purchased with Bitcoin (0.049 BTC) at the cyberlocker which hereby will be called C. When a meeting took place with Thomas Andersson at ECPAT it was revealed that cyberlocker C has been of interest to their hotline several times recently. Therefore, to create a framework, the research started with an investigation of this cyberlocker.

It took about 1 hour from the purchase until the transfer took place, due to the necessary verifications from other Bitcoin users. The 0.049 BTC was transferred to cyberlocker C’s posted Bitcoin address, hereby denounced B. After a few hours it was possible to view the transaction in the tool Reactor according to Figure 24. Wallet B only contained received funds from various addresses and about 24 hours later, the total received amount was sent to a bigger wallet, the wallet called C, which is believed to be the cyberlocker’s corporate wallet.

![Figure 24 Illustrates the transactions from wallet A to B and from B to C (the cyberlocker's corporate wallet).](image)

When investigating the transactions to and from the cyberlocker’s corporate wallet, C, it was discovered to have about 170 000 transactions in total and several connections to suspicious services. Figure 25 illustrates some connections to tumblers, CoinJoins, mining pools and a TOR market, which is a marketplace on the Darknet.
It is large amounts that are transferred to and from the cyberlocker. The largest amount received corresponds to approximately 4 400 000 SEK (1 BTC = 3716 SEK. 2016-05-13) and the largest amount sent is about 4 078 000 SEK. Table 6 illustrates the largest transactions to and from the related services.

Table 6 Illustrates the transaction to and from the different services.

<table>
<thead>
<tr>
<th>(1 BTC = 3716 SEK. 2016-05-13)</th>
<th>Largest amount sent (SEK)</th>
<th>Largest amount received (SEK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyberlocker C</td>
<td>4 389 528</td>
<td>4 077 600</td>
</tr>
<tr>
<td>Cash solution</td>
<td>25 876 904</td>
<td>20 242 971</td>
</tr>
<tr>
<td>Tumbler</td>
<td>4 856 007</td>
<td>15 293 697</td>
</tr>
<tr>
<td>TOR market</td>
<td>22 527 663</td>
<td>8 651 041</td>
</tr>
<tr>
<td>CoinJoin</td>
<td>251 830 667</td>
<td>100 457 462</td>
</tr>
<tr>
<td>Mining pool</td>
<td>1 621 646 186</td>
<td>667 402 952</td>
</tr>
</tbody>
</table>

To find and define transactions connected to the cyberlocker, Reactor is used to manually track a specific Bitcoin. Chainalysis' tool Reactor visually illustrates a Bitcoin's pathway, which simplifies the investigations for the law enforcement and other institutions.

The idea was to receive the 16 cyberlockers posted in the G2 Web Services’ report and investigate their transaction patterns in Reactor to identify common patterns. These 16 cyberlockers were identified as containing illegal material and therefore their behavior will work as a template of comparison for future investigation of the 51 cyberlockers in Appendix B. If the cyberlockers in appendix B have the same behavior parameters as the registered patterns of the 16 identified cyberlockers, they are marked as interesting for future investigation for the police. The investigated parameters are listed in Table 9. The investigation was preformed manually through Reactor. The same procedure was expected, as earlier described, to identify respective cyberlocker’s corporate Bitcoin wallet ID, to later be able to view the transactions to and from the file hosting service.
6 Results and analysis

In this chapter the results from the two different methods are presented and analyzed. The primary method is a composition of several different methods and models. The output of these composed methods and models will be a relationship network of nodes, where it will be easy to see the connections between different users. By using this method, it is possible to locate the uploading source of specific material and also reveal who the underlying person behind the alias is. In the latter parts of the method, a collaboration between the police, cyberlockers and exchange services are needed to be able to identify these perpetrators and consequently prevent further material to be uploaded within cyberlockers. This also means that the Bitcoin transactions regarding this matter will be limited and in the end completely prevented.

The results from the secondary method is a framework for categorizing and classifying cyberlockers based on their transaction flow history. The framework was created by first investigating and analyzing cyberlockers that were proven to consequently contain illegal material. Trends and patterns in these cyberlocker’s transaction flows were identified and used as input to the framework. The created framework can be used to categorize and classify new cyberlockers by comparing their transaction flows to the framework. In the end of the chapter, the results from a combination of the two methods is presented. The different methods complement and support each other in the purpose to stop payments regarding illegal material online and finding offenders.

6.1 Primary method

From the webcrawler session, data was derived. MDS was used to sort this data and AHC to further understand the data. Hence, the result from the primary method is the relationship between different data points. The result will be of use for the police to compare with their database, which according to a source within the Swedish Police exists and contains information about known groomers and perpetrators that have been connected to CAM in the past. This database contains usernames that the perpetrators have been using and also other information that differs from person to person, i.e. IP-address, phone number, home address etc. The police want to cooperate to prevent people from producing and uploading CAM, as well as downloading and viewing the material, and it is therefore possible to establish a cooperation in which the police will cross-check usernames found through the author’s method with their database to start building a case against the perpetrators.

For example, if a frequently used username appears in different forums sharing the same link, this username can be crosschecked in the police’s database to further validate if the person behind the alias has been investigated before or in other ways been connected to suspicious activity. If the username also is registered in the police’s database, a stronger case against the person is built.

The most prominent card providers and the banks united in the past to stop the trade of illegal material, which led to a decrease of credit card payments for illegal purposes. This was a result of thorough investigations of the corporations that wanted to have credit cards as a payment method for their customers online. If the card providers detected illegal activity, this business was blocked in collaboration with the police. This resulted in that these kind of illegal payments by credit cards is not as used as before, and the authors’ primary method in collaboration with the police and exchange services, can eventually

---

3 Cybercrime Division within the Swedish Police. Interview 2016-03-29.
lead to the same result regarding Bitcoin transactions. If the police have identified a specific wallet that is connected to illegal material, the exchange service that has been involved with the identified wallet, can block the wallet from using its’ services. If all exchanges do this consistently, the use of Bitcoin for illegal purposes will decrease, in similarity to credit card payments. If, for some reason, a number of exchange services do not want to cooperate to prevent the distribution of CAM, the assumption is that the concentration of illegal material increases in connection to these services. This focuses the investigation for the police to the exchange services that do not want to collaborate and a narrower area of the problem is created.

The goal of the primary method is to create a network of relationships between perpetrators to make it easier to pursue both the uploaders and the downloaders of CAM. The primary method consists of creating a sophisticated webcrawler for extensive gathering of data which then is analyzed using Multidimensional Scaling to easier understand the correlations between the different data points. When this new representation of the data is completed, the data is processed with the Agglomerative Hierarchical Clustering method to cluster together the points that have high correlation, which further makes it easier to understand the relationships between the data points. This process has been validated using a small dataset in Chapter 5.2.1 – 5.2.3. To make the primary method more realistic and applicable to this field of study, a larger example containing more parameters was conducted. In this example, the handling of the different parameters aims to imitate the handling of parameters when performing this method on the intended field of study, CAM. As seen in the earlier validation, illegally shared copyrighted movies have a lot of similarities with the illegal distribution of CAM and are therefore used in this larger example as well.

The gathering of data was performed manually, but with the aim to imitate how the future webcrawler will perform. This means that when an illegally shared movie was found and the data regarding that movie had been gathered, the links connected to the uploader’s and the followers’ user profiles were followed to make sure that all data connected to them on that forum was found. The initial search was also meant to imitate a webcrawler, which meant that one illegally shared movie at one specific forum by one specific uploader was chosen and then the manual search was conducted from there. The search resulted in a total of 14 different illegally shared movies within 12 different forums, uploaded by 23 unique usernames. The followers are seen as users, which means that someone who is merely commenting on one forum post, could potentially be the uploader of another forum post. When accounting for the unique users who were followers, the grand total of 148 usernames was collected and the total amount of data points were 36. An example from the data gathering session can be seen in Figure 26.

To make it possible to process the data through the MDS method, a numerical representation of the data was performed. This was done individually for each parameter. For example, the parameter Forum was transformed to the numerical values of 1 to 12, where the first forum in the matrix of gathered data was given a value of 1 and the last occurring unique forum was given the value of 12. The parameter users, both uploaders and followers, was given the values of 1 through 148 and the parameter links was given values between 1 through 14. As seen in Figure 26, the parameter Timestamp is converted in excel to a
numerical value and how this parameter is handled will be explained later. An example of the converted data can be seen in Figure 27.

![Figure 27 Example from converted data](image)

In this larger example there were five unique parameters which were weighted equally when implementing the penalty function, as seen in Chapter 5.2.1 - 5.2.3. This meant that the weights of each parameter was 1/5 as maximum, since the largest achievable value, if no parameters were equal, was one. As for the three parameters; **Forum**, **Link** and **User (uploader)**, the conversion was done by simply checking if the values were equal and if they were, the penalty given was zero and if they were not, they were given the penalty of 0.2. The handling of the parameter **Timestamp** was in this case performed by checking if any parameters matched apart from this parameter and if there were no matches, the timestamp was not considered relevant and hence were given the penalty of 0.2. If there were one or more parameters who were equal, the timestamp-parameter were analyzed further by looking at how far apart in time uploads were done. If the time between them was less than two hours the penalty became lower and for example if the time between uploads was one hour, the penalty given was 0.1.

As for the followers, the penalty was constructed by finding similar followers or if a follower was the uploader of a different post. If the row currently analyzed has 10 followers and the row that it is being compared to have two followers who are the same, the penalty value given will be $\left(\frac{8}{10}\right) \times 0.2$. If all followers are the same, the penalty would be zero and if no followers are the same the value will be 0.2. Lastly the penalties for the five parameters are summarized together to create the final penalty value, which have a maximum value of one.

The conversion to the relationship-matrix was done with a script in Matlab. The output matrix has the dimension of \((\text{number of data points}) \times (\text{number of data points})\), since each data point is compared to all other data points. This means that each data point in this example is compared 36 times and this is done for all 36 data points, which creates an output matrix with the dimension of 36*36. This also means that the output matrix will have a diagonal of zeros since when comparing a data point with itself, no differences will be found. A small sample from the output matrix that shows how the matrix will look like after the conversion can be seen below in Figure 28. Figure 28 illustrates the first six data points but only the relation to the first ten data points. When the value of a cell is 0.8 it means that one parameter is equal and when it is lower it means that more than one parameter is equal.

![Figure 28 Sample from output matrix.](image)

The output matrix consists, as earlier stated, by the sum of the penalties for each parameter for a data
point against other data points. This means that, the lower the penalty value, the higher the probability of a strong relation between the data points. This matrix is converted with the function `pdist` in Matlab to create the distances between the data points in the output matrix, that later is processed through the `cmdscale` method in Matlab, which is the MDS method. The output is a new representation of the distances which makes it easier to understand the relations. This output from the MDS method can then be clustered with the `linkage` function in Matlab, which corresponds to the AHC method, to cluster together data points with the lowest distance between them. This process is iterated until all data points are clustered together. The clusters will be represented in a matrix which is the main result generated from these methods. To be able to easier understand this matrix, the function `dendrogram` was used to visually illustrate the clusters, as seen in Figure 29. This function will show which points that are clustered together and at what level. The horizontal lines between data points represents connections between the points and the lowest horizontal line represents the cluster with the highest probability of a strong relation between these data points, which in this example is the data points 22 and 23 in Figure 29.

![Figure 29 The generated dendrogram illustrating the relationships between the data points in the example.](image)

From the data, which is visually presented in Figure 29 above, it is easy to see which data points, and hence which usernames, that are most interesting to investigate first. Data point 22 and 23 are the two closest data points, distance wise, due to the fact that the same link was shared in the same forum by two different usernames at almost the exact same time, according to Table 7. In this example, the investigators should start their investigation with these two data points and if Link 1 is proven by the police to contain CAM, the identity behind the usernames “A” and “B” should be explored. The probability is high that these two usernames, corresponding to point 22 and 23, communicate in some way or is the same person due to the close time span regarding when the link was shared (within 15 min). The next closest pair of nodes are 17 and 19 that are closely linked to the data points 16 and 18. The reason for node 17 to be closest to 19 is that the comparison with all of the other data points is performed from the top and down. Otherwise, the naturally closest nodes are 18 and 19 that have 2 common parameters and are closer in time than between 19 and 17, according to Table 7. Anyway, the usernames of interest derived from these points are “C” and “D”.

Asplund, Berggren, 2016
The next step for the police is to contact the relevant cyberlockers (X, Y, Z and T) to retrieve all the information about the four usernames (A, B, C and D). Both the IP-addresses and the users’ Bitcoin wallets can be extracted. If the user uses an anonymization service such as TOR, the IP-address cannot be revealed and the person behind the alias remains unknown. In that case, the cash flow to/from the user’s Bitcoin wallet can be mapped and hopefully one transaction between the user’s wallet and an exchange service can be found. If that is the case, the police can extract information from the exchange regarding the username’s underlying identity.

Table 7 Illustrates the information behind the most relevant data points from the experiment.

<table>
<thead>
<tr>
<th>Data point</th>
<th>Forum</th>
<th>Link</th>
<th>User (uploader)</th>
<th>Timestamp</th>
<th>Followers</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>X</td>
<td>Link 1</td>
<td>A</td>
<td>42514,85220</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>X</td>
<td>Link 1</td>
<td>B</td>
<td>42514,83930</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Y</td>
<td>Link 2</td>
<td>C</td>
<td>42514,81257</td>
<td>92, 93, 94, 95, 96, 97, 57, 98, 99, 100, 101, 102, 103, 104, 105, 106</td>
</tr>
<tr>
<td>18</td>
<td>T</td>
<td>Link 2</td>
<td>C</td>
<td>42514,82155</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>T</td>
<td>Link 2</td>
<td>D</td>
<td>42514,83664</td>
<td></td>
</tr>
</tbody>
</table>

If/when the police verify that the link contains CAM, the connected cyberlocker consequently gets marked as a service that contains illegal material. The marked cyberlocker can be affected negatively and when the business wants to exchange their Bitcoin to real economy money, the exchange service can refuse to serve the business. The banks do not want to be connected with money that is derived from trade with CAM. Therefore, the exchange services need to be watchful so that no connection exist with money connected to CAM. Therefore, it is in the cyberlocker’s and the exchange service’s interest to prevent CAM within their services to be able to have a good relationship with the banks. This is the goal of the primary method and as presented, it is possible to find the points with the highest probability of a strong relation between them and hence find the perpetrators who most actively share and produce illegal material.
6.2 Secondary method
ECPAT supported this thesis with a few file hosting services (cyberlocker C, D and E) that has appeared in their hotline. Also, three cyberlockers (F, G and H) were extracted from the initial investigation of cyberlockers, displayed in Appendix B, to further validate the hypothesis regarding the surroundings of these corporate wallets. The six cyberlockers were investigated to create the framework regarding the transactions to and from the services, according to the parameters in Table 9. Cyberlocker F, G and H are interesting to investigate considering the ability to get paid for uploaded files within cyberlocker G and H and the fiercely advocated safety in cyberlocker F, according to Table 8. The reason for the absence of cyberlocker D and E in Table 8 is that these were not included in the primary investigation of the different cyberlockers. The secondary method was implemented to illustrate what a framework of larger size could look like and how this framework could be of use when investigating new cyberlockers. A larger sample size is needed for the framework to be legitimate.

Table 8 Part of the result from the initial investigation of the 155 cyberlockers.

<table>
<thead>
<tr>
<th>Cyberlocker</th>
<th>Bitcoin</th>
<th>Card</th>
<th>Solutions like Paypal, Klarna, WyWallet etc.</th>
<th>Other payment solutions, such as Call2Pay</th>
<th>Premium account</th>
<th>Paid to upload</th>
<th>Contact information to a private person</th>
<th>Advocated security (1-5, where 1 is low)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyberlocker C</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>Cyberlocker F</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>5</td>
</tr>
<tr>
<td>Cyberlocker G</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>4</td>
</tr>
<tr>
<td>Cyberlocker H</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>3</td>
</tr>
</tbody>
</table>

Cyberlocker C’s surroundings is illustrated in Chapter 5.3.2 and in Table 8, 9 and 10 more information is presented regarding this file hosting service. According to ECPAT’s hotline, cyberlocker C is currently the most popular service among perpetrators and should thereby be of top priority for the police in future investigations. Considering the surroundings of cyberlocker C’s corporate wallet, all suspicious services are connected as listed in Table 9. The amounts sent and received connected to the wallet lies around 270 million SEK. As illustrated in Table 10, this amount is fairly similar to the other cyberlockers’ cash flows except for the exchange service Bitstamp. With the gathered information, the conclusion can be drawn that cyberlocker C is of high interest in future investigations regarding the distribution of CAM due to its connections to a TOR market, a CoinJoin and a Tumbler even though the service does not offer the possibility to get paid for uploading popular files within the service. For the offenders, the payment can, instead of money, be the possibility to watch other people’s uploaded illegal images. Also, if the user uploads a popular file, he/she could get payed through a higher score or other internal rewards within the cyberlocker.
Cyberlocker D does not offer any premium account but advocates donations instead. Therefore, a donation was made of 0.005 BTC to the designated address, which proved to be an exchange called Bitstamp. Hence, the hypothesis is that the donated Bitcoin are transferred to an exchange to directly be converted into real economy money. Some suspicious services surround this exchange, such as tumblers and TOR markets which is illustrated in Figure 30. The amount sent and received to/from Bitstamp is approximately 40 billion SEK, which highly differs from the amounts regarding the corporate wallets of cyberlockers in Table 10. This is expected due to the fact that this is an exchange service and not a normal Bitcoin wallet. Exchange services such as Bitstamp and Bittylicious (illustrated in Figure 33) are of high relevance considering that these services should be the first to be contacted regarding the preventative measures, further explained in chapter 6.3, that need to be taken regarding CAM. The fact that Bitstamp is connected to services such as CoinJoins and TOR markets further validates its suspicious activity. The yellow circle in Figure 30, BitPay, has occurred in three (cyberlocker F, G and H) of the six transactions made to the different cyberlockers. This is an outsourcing service for Bitcoin similar to Klarna regarding real economy money. Different cyberlockers offers the ability to pay with Bitcoin, but when the payment takes place the payer is redirected to a BitPay page where all of the information regarding the transaction is listed. BitPay allows the company to receive payments in Bitcoin and get the amount in real economy money deposited to a bank account.

Figure 30 Illustrates the surroundings of cyberlocker D (Bitstamp).
A premium account was purchased within cyberlocker E as well. According to Figure 31, the transaction made from wallet A turned out to be to a temporary address, B. This address was investigated and was under surveillance until the Bitcoin later were transferred from this address. The day after the surveillance was initiated, a large amount was sent from this temporary address to a Bitcoin exchange service. Surrounding this temporary address and the exchange service that the temporary address sent Bitcoin to, was a large number of suspicious services such as tumblers, TOR markets, CoinJoins etc according to Figure 31. In the figure, BitPay occurs again as a yellow circle.

Figure 31 Illustrates the surroundings of cyberlocker E’s wallet, were B is the temporary address.
When investigating cyberlocker F’s and G’s corporate wallets, it appears that they are identical. The two wallets have different wallet IDs but exactly the same transaction history. The answer to why this is, could be two different Bitcoin addresses within the same wallet. This would explain the exact same transactions and information within the different Bitcoin addressess. Hence, the connecting services are illustrated for the two cyberlockers in Figure 32. The two file hosting services have connections to services such as CoinJoins and gambling services, but no connection to a tumbler or a TOR market which is quite unexpected. If this is a large wallet containing several addressees, you could anticipate many connecting suspicious services, but this is not the case.

Figure 32 The surroundings of cyberlocker F and G.
In the cases of cyberlocker F and G, the payment goes to a temporary address and in cyberlocker H’s case, the payment goes directly to the corporate wallet. In Figure 32 and Figure 33 the scam service MMMGlobal is illustrated in connection to the corporate wallets of cyberlocker F, G and H. This is a service that survives on defrauding individuals or groups by exploiting the classical Ponzi scheme. It can be seen that transactions between the corporate wallet and the scam service are mutual. To what purpose is hard to say, maybe the owner of the cyberlocker also is the owner of MMMGlobal. If the scam service is disregarded, cyberlocker H only has one other connection of interest, to a CoinJoin. The joint payment service SharedCoin appears throughout the investigation of different cyberlockers. The service’s website seems professional and an attempt to collaborate with this service in the future regarding CAM, would be of relevance.

Figure 33 Illustrates cyberlocker H’s Bitcoin connections.

This sample of six different cyberlockers is too small to be a basis of a reliable framework, but it gives an indication of what kind of trends and patterns in the transaction flows of cyberlockers that can be found. It is easy to assume that if large amounts are sent from a cyberlockers corporate account through tumbling services and CoinJoins, this is a warning sign that illegal activity surrounds this cyberlocker. This assumption can be drawn since the services charges a percentage fee to anonymize the Bitcoin transactions, which means that the cyberlockers who uses these services, sacrifices a certain percentage of their profit to erase the trace on their money, which is highly suspicious.

Table 9 Illustrates the considered parameters when generating the framework.

<table>
<thead>
<tr>
<th>Cyberlocker</th>
<th>Connection to a TOR market (YES/NO)</th>
<th>Connection to a CoinJoin (YES/NO)</th>
<th>Connection to a Tumbler (YES/NO)</th>
<th>Connection to a mining pool (YES/NO)</th>
<th>Connection to other suspicious service (YES/NO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyberlocker C</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Cyberlocker D</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Cyberlocker E</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Cyberlocker F</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Cyberlocker G</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Cyberlocker H</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>

When investigating the total cash flows to and from the six cyberlockers, large amounts is found, according to Table 10. The exchange service Bitstamp is of particular interest due to the suspicious...
connections that the service has and the large transactions it sends and receives. The reason cyberlocker D’s transaction flow is distinctly larger than the other services is because of its starting date and due to that it is an exchange service and not a regular Bitcoin wallet. One deviation is that cyberlocker E has been active for about two more years than cyberlocker C, F, G and H and yet the cash flow reaches about a third of their total transactions to and from the wallets. The conclusion can be drawn that business is not great for cyberlocker E and could be a result of not containing popular illegal material, but that is just speculations.

Table 10 illustrates the total cash flow to and from the three cyberlockers.

<table>
<thead>
<tr>
<th>Cyberlocker</th>
<th>1 BTC = 3 737 SEK (2016-05-25)</th>
<th>Active since</th>
<th>Sent (SEK)</th>
<th>Received (SEK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyberlocker C</td>
<td>2016-01-27</td>
<td>271 112 491</td>
<td>271 163 582</td>
<td></td>
</tr>
<tr>
<td>Cyberlocker D (Bitstamp)</td>
<td>2011-08-17</td>
<td>39 799 148 870</td>
<td>39 804 202 068</td>
<td></td>
</tr>
<tr>
<td>Cyberlocker E</td>
<td>2014-07-10</td>
<td>136 435 638</td>
<td>136 536 291</td>
<td></td>
</tr>
<tr>
<td>Cyberlocker F</td>
<td>2016-03-29</td>
<td>339 296 938</td>
<td>339 396 118</td>
<td></td>
</tr>
<tr>
<td>Cyberlocker G</td>
<td>2016-03-29</td>
<td>339 296 938</td>
<td>339 396 118</td>
<td></td>
</tr>
<tr>
<td>Cyberlocker H</td>
<td>2016-03-29</td>
<td>327 565 963</td>
<td>330 638 524</td>
<td></td>
</tr>
</tbody>
</table>

Table 11 illustrates the percentages of the total amount sent and received to/from the cyberlocker to the services; TOR market, CoinJoin and Tumbler. Note that these percentages are only represented by transactions to and from individual services. For example, cyberlocker C’s connection to a TOR market is only represented by the total transactions to/from AlphaBay market, which is a merchant place within the darknet. Considering Table 11, the distinct most worrying scenarios are cyberlocker D’s relations to TOR markets and cyberlocker C’s transactions from CoinJoins. 0,5 % of the total sent amount is to a TOR market from cyberlocker D which corresponds to approximately 200 million SEK. 1,36 % received from a TOR market to Bitstamp corresponds to about 540 million SEK. Considering that the exchange service has been active for about five years, this is an amount of about 100 million SEK each year that is received to the exchange service Bitstamp and converted into real economy money. As many know, and after what has been explained in this thesis, not many legal products/services exist within the darknet. Bitstamp is not the only one connected to worrying percentages, almost 0,9 % of the total amount received to cyberlocker C originates from a CoinJoin. This corresponds to about 2,5 million SEK and the file hosting service has only been active for about four months. The results of the investigation emphasize that cyberlocker C and D should be further investigated by the police.

Table 11 Illustrates the amounts and shares connected to a TOR market, CoinJoin and Tumbler.

<table>
<thead>
<tr>
<th>Cyberlocker</th>
<th>TOR market</th>
<th>CoinJoin</th>
<th>Tumbler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyberlocker C</td>
<td>Sent (SEK)</td>
<td>Received (SEK)</td>
<td>Sent (SEK)</td>
</tr>
<tr>
<td>Cyberlocker D (Bitstamp)</td>
<td>Sent (SEK)</td>
<td>Received (SEK)</td>
<td>Sent (SEK)</td>
</tr>
<tr>
<td>Cyberlocker E</td>
<td>Sent (SEK)</td>
<td>Received (SEK)</td>
<td>Sent (SEK)</td>
</tr>
<tr>
<td>Cyberlocker F, G and H</td>
<td>Sent (SEK)</td>
<td>Received (SEK)</td>
<td>Sent (SEK)</td>
</tr>
</tbody>
</table>

When conducting the investigation of the percentages of cash flow to/from suspected services, it was discovered that cyberlocker F, G and H generated the same result. It was earlier established that F and G
were identical and probably two addresses within the same wallet. Now the suspicion arises that also cyberlocker H is a part of that same wallet when considering the results from the investigation and that the three wallets originated on the same date (2016-03-29). These three are different individual cyberlockers, but they are probably gathered under the same central organization which explains the similarities between them. That would explain that the three services all uses BitPay as an outsourcing service. If that fact is disregarded, the three cyberlockers are “only” connected to a CoinJoin with relatively low shares which implies a low chance of containing illegal material, in comparison to the other cyberlockers. In the beginning of this thesis, the hypothesis was that tumbling services were the most popular way of “washing” Bitcoin. The result from the secondary method implies that this is not the case. CoinJoins seems to be the most popular form of further anonymizing Bitcoin transactions and many external businesses offer the service for a fee. This fee could be represented by a percentage of the amount sent/received or as a fixed charge. In either way, the users of these services are willing to pay a certain amount to anonymize their Bitcoin transactions.

Every cyberlocker has made at least 200 000 transactions. Hence, all of the individual transactions have not been investigated whether or not the sent and received address is a CoinJoin, tumbler, TOR market etc. When investigating the connecting services, the amounts of the individual transactions were sorted so that the highest amount sent and received was listed on the top in Reactor. This simplified the work and the significant transaction volumes could be further investigated. If smaller amounts were received/sent to a suspicious service, they could thereby have been missed.

Parameters such as mining pools and scams are not included in Table 11 due to that these are not considered to be relevant for the problem of CAM. Scams are a worldwide problem and it affects people that willingly pay money without any deeper research of the organization receiving the amount. Mining pools are not illegal, but only a way to earn money, create new Bitcoin and to create new blocks in the block chain. Gambling services are not considered either, due to its irrelevance regarding the subject of CAM.

6.2.1 Conclusion

The secondary method resulted in a template of comparison for the police when investigating suspicious cyberlockers in the future. The police need to use a larger sample size for the framework to be efficient and correct but this method was performed to illustrate what kind of patterns and trends that can be visualized by a framework of this type. If patterns regarding CoinJoins used by cyberlockers can be found, i.e. if a majority of the exploited cyberlockers uses joint payment services, the conclusion can be drawn that a majority of the investigated cyberlockers that uses this anonymization service are getting exploited for illegal purposes. Considering that CoinJoins charges a percentage fee of the received amount, the party that pays for the service must have incentive to do so, either because they know that the money is tainted and still want to use it, or that they know about it but can’t do anything to stop it. There are few logical reasons to why a company that only deals with legal businesses and legal purposes would pay a percentage fee of their profit to anonymize their Bitcoin transactions. One reason could be transactions for totally legal services/products that are of sensitive nature. The reason why it is relevant to include connections between cyberlocker’s corporate wallets and different TOR markets is because of the kind of business that is being conducted on these markets. According to the classification of the darknet done by Moore and Rid and presented in Chapter 7.1.1 in Table 13, more than half of all the active pages on the darknet are of illicit nature. This means that when transactions are made to/from different TOR markets, the probability that the money is connected to illegal activity is high.

Because the framework was created based on a small sample size, a comparison with the 51 identified cyberlockers was not performed. It would have been a waste of time and money on a comparison with a
framework that is not statistically correct, due to the money it would cost for buying premium accounts within every service.

6.3 Combination of the primary and secondary method
The primary method results in a network of relations between the gathered data points. The closest connections are prioritized and the first to be examined due to the common denominators regarding CAM. The secondary method results in a framework that illustrates suspicious behaviors within cyberlockers. The framework can be handed to the police for future investigation. When the cyberlocker is validated to contain CAM by the police, further explained under Chapter 5.

Figure 34 illustrates where in the process the two methods can be of use. Both methods are necessary in the pursuit of perpetrators and to prevent future Bitcoin transactions regarding CAM. For example, one part of the result from the primary method could be a link that is proven by the police to contain CAM, which has been shared by the same user in different forums and during a time period of less than 2 hours. If the shared link contains a cyberlocker name, this cyberlocker can be further investigated manually according to the secondary method. If 30% of this cyberlocker’s total amount of files are proven by the police to contain illegal material, the assumption is made that at least 30% of the outgoing transactions are classified as illegal money. The idea is to further track the illegal moneys’ pathway in Reactor and later investigate the percentage of illegal money in each wallet. The most interesting wallets to investigate in this thesis are the infected cyberlockers’ corporate wallets and the wallets directly connected to the infected cyberlocker. The cash flow from the suspected cyberlocker to an individual wallet is an indication that the person has uploaded a popular file within the cyberlocker and hence is getting paid for the number of downloads. The wallets of the perpetrators found in the primary method are further investigated and if transactions are discovered to be connected to an exchange service, the police can demand personal information about the persons behind the wallets. This can only be done if a preliminary investigation has been approved by a prosecutor. Hopefully the information from the primary and secondary method combined with the police’s database, is enough for the law enforcement to initiate an investigation of a private person.
The gathered information contributes to building a case against individual offenders. The exchange services play an important role when Bitcoin is exchanged into real economy money and integrated with the traditional financial system. The banks that are responsible for the accounts where the exchanged Bitcoin end up, do not want to be connected to illegal activity. This was a clear message when the banks joined together to stop trade with illegal material via credit card payments. The same attitude remains regarding illegal transactions with Bitcoin. If an exchange service is known to be connected to infected cyberlockers, banks could refuse to accept the exchanged amount to an account within their organization. Therefore, it lies in the exchange’s interest to prevent business with perpetrators and infected cyberlockers. The primary method can identify individuals and the secondary method generates potentially infected cyberlockers. These cyberlockers’ and offenders’ Bitcoin wallets should be marked with a warning signal when trying to exchange Bitcoin into real economy money through an exchange service. In this way, it is up to the individual exchange service to decide if illegally earned Bitcoin can be exchanged into real economy money.

Another way to diminish infected cyberlockers is to post a warning on the start page within the service, where it is stated that the cyberlocker has been proven by the police to contain illegal material. Thereafter, the visitor has a choice to use another cyberlocker and dismiss the infected one. In this way, it is also in the cyberlocker’s interest to actively prevent the illegal distribution of CAM and collaborate with the police. The administrators of the cyberlocker has access to the Bitcoin address that are used in individual cases and with a warrant, the police can gather the location of certain computers through the IP-address. If services such as TOR are used by the offender, Reactor can be helpful to manually track the perpetrators transaction flows through the listed Bitcoin address. Hopefully it will lead to an exchange service, where another attempt is made to identify the person behind the alias. Collaborations are needed between the police, cyberlockers and the exchange services to succeed in the elimination of Bitcoin transactions regarding CAM.

Figure 34 Illustration of the combined result from the two methods.
6.3.1 Process flow

The idea is to begin the future investigations with identifying and map as many cyberlockers as possible by considering the result from the secondary method. Consequently, a list with classified cyberlockers is derived. This list is matched with the parameter “link” in the primary method to find connections. If a cyberlocker is defined as a service likely to contain illegal material (or proven to contain CAM by the police) considering the result from the secondary method, the links containing that cyberlocker’s name should receive a higher weight. If the parameter receives a higher weight, deviations will be clearer since the penalty will be higher, which means that similarities between data points will be distinguished and intensified focus lies on these. The output from the crawling session in the primary method is, as earlier stated, a list with several columns with different parameters, where the number of columns is equal to the number of parameters. For example, one row can contain the name of the cyberlocker, username, link and timestamp. Hence, when suspected cyberlockers with connected usernames are derived, the list is sent to the police to be crosschecked with their database containing known online offenders. If the police discover correlations of interest regarding a username, this alias is further investigated through the information derived in the result from the primary method. This username will then be a way to distinguish a cluster that is connected to the alias in some way and different used usernames can be detected to have the same underlying person behind it, while other cyberlockers could be discovered to be of interest for the police. To conclude, the result from the primary method, in combination with the secondary method, is a set of data points and the relationship between them, which results in an easier way to discover nodes of interest and the results are also useful in other parts of the process flow, which can be seen in Figure 34.

In the proceeding of the investigation, the result from the primary method is analyzed and the most interesting usernames within different cyberlockers is highlighted. The next step is for the police to contact these cyberlockers and demand all information about the aliases. This includes the users’ IP-addresses, which files that have been downloaded/uploaded and consequently information about the user’s transactions which will give the investigators information about the users Bitcoin wallet. As earlier mentioned, most of the cyberlocker that has been manually investigated (155 cyberlockers) states that the service is willing to cooperate with the law enforcement if crime is suspected. At this stage, the police can verify if the username indeed is an offender regarding the distribution/possession of CAM through the shared link. If the user does not use services such as TOR, the location of the perpetrator can be revealed. If the user use services such as TOR, another way of retrieving the information of the offender is implemented. Since the user have made transactions to/from the cyberlocker, the administrators know the address of the user’s Bitcoin wallet. In Reactor, the police can track the cash flow to/from the wallet, which hopefully eventually reveals a transaction to/from an exchange service. Now the police can demand the user’s information from the exchange service which includes name, address and bank information etc. In this way the offender behind the aliases is revealed.

If a collaboration between the different parties is established, it would be possible to convince the exchange services to create a warning signal if/when the perpetrator tries to exchange Bitcoin and also if the connected cyberlocker to exchange their Bitcoin. If several links within different cyberlockers are proven through the primary and the secondary method to contain CAM, an assumption is made that a share of their total transactions is represented by illegal money. This would be enough proof for the exchange services to register a warning signal when these services wants to exchange their Bitcoin into real economy money. These are preventative measures that will protect the banks and also cyberlockers and exchange services with legitimate purposes.
7 Discussion
Regarding the result and analysis of the primary method, some discussion about the police’s procedures is necessary. From the available data, relevant parameters and relationships have been extracted, but the risk exist that this is not enough for a prosecutor to approve an initial investigation. Also, IP-addresses are bound to a specific computer within a specific country. The transactions made from this IP address will therefore be originated from this country, which means that this specific countries police force and the national banks should be involved. With this said, the international cooperation is important to combat the issue of CAM and the recommendation is to implement the primary method on an international level.

The offenders that are desperate for material are easier to catch compared to the perpetrators that are connected to CAM for financial gain. The desperate ones more often make mistakes in their attempt to receive material, hence the police can easier catch them. The primary method is a tool for the police to further map the people that are distributing/downloading CAM for financial purposes, which is clear due to the Bitcoin cash flow network.

The perpetrators who are not downloading and uploading material for financial gain, but instead are doing it because they are desperate for material, are often ashamed that they have this need. The last thing that they want is for the act to become public, partly because they don’t want to get caught and partly because they don’t want their families to find out. Despite this fact, they feel the pressure to find new material because of their pedophilic tendencies. If they are presented with an opportunity to get new material, they will do almost anything to do so. While in the act of finding new material or producing it, they rationalize the crime in their head, to be able to live with themselves and afterwards they feel this earlier mentioned shame. If they were not to rationalize the crime, a majority of the perpetrators would not be able to go through with it.

7.1 Primary method
In the larger example under Chapter 6.1 a more thorough testing of the different methods is implemented, with five different parameters; Forum, link, username (uploader), timestamp and username (follower). This was a result of the desire to investigate the relevant parameters for sharing movies through links to cyberlockers and to find a way to imitate the real implementation regarding CAM. When implementing the methods to suit the problem of CAM, more parameters could potentially be added. For example, an implementation of PhotoDNA to identify if the exact same image is being shared several times. This would not have been useful when investigating the illegal distribution of movies since it would not have generated any results in that example. Regarding the parameter timestamp it is important to try different approaches to find the best way to handle this parameter. In the larger implemented example containing 36 data points, an approach that could be implemented regarding CAM is tested and performed. This approach is based on a time interval of two hours and is only considered if one or more parameters appear equal. If one or more parameters appear equal, the time between the posts decide the size of the penalty in a way that two hours equal to maximum penalty, one hour equals to half of the maximum penalty and no time difference gives zero penalty. This is one way to handle the parameter, to only give it a significant weight when other parameters are correlated, as performed in the implementation in Chapter 6.1.

Depending on which parameter that is considered, different weights can be assigned. In the example implemented in Chapter 5.2 a lesser weight was assigned to the parameter timestamp, while in the larger example in Chapter 6.1 all parameters were weighted equally. In the same way the weights should be adapted when investigating CAM. For example, if the same picture is discovered to be shared multiple times, the parameter PhotoDNA should receive a higher maximum weight, so that the focus temporarily
can be put on this parameter. Using the same logic, the results from the secondary method explained in Chapter 6.2, the weights of the parameter Link could be relevant to increase for the cyberlockers that have been put on the “warning list”. This would ease the continued investigation, and together with the results from the Primary method, this would create a more definitive path regarding what users to investigate further.

The primary method would have been conducted in a better way if the names of the popular forums amongst perpetrators and the names of the proven infected cyberlockers was known for the authors. Consequently, greater weights could have been assigned to these forums and cyberlockers and the result would clarify which usernames to focus on. The authors did not have the authority to receive these names according to Europol and the legislation. As a complementary solution, the secondary method should be performed before the primary method so that suspicious cyberlockers are matched with the names of the cyberlocker in the output matrix generated from the webcrawling session. At least then, suspicious cyberlockers are highlighted and focused on.

7.1.1 Results from Moore and Rid’s experiment

Moore and Rid (2016) performed an in-depth, lengthy webcrawl for every web based hidden service, accessible by an individual seeking content within the darknet. They took website lists from two such services; onion.city and ahmia. The two sites yielded 5615 unique onion addresses, which was enough for the experiment. The entire list of websites was crawled during a two-month period (January – March, 2015). To avoid illegal material, Moore and Rid (2016) only harvested textual content automatically.

The analysis was divided into two steps. The first incorporated included a manual categorization of websites. They then randomly sampled the harvested data and assembled a collection of webpages that was used to train the automatic Support Vector Machine content classifier. The classifier is a statistical classification algorithm used in machine learning to categorize data. The result was a taxonomy of twelve categories, see Table 12. The second phase corresponded to assign the trained classifier to categorize the rest of the documents into the twelve categories. The results were randomly inspected for potential errors to validate the classification.
Table 12 Explains the different classifications.

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arms</td>
<td>Trading of rearms and weapons.</td>
</tr>
<tr>
<td>Drugs</td>
<td>Trade or manufacture of illegal drugs, including illegally obtained prescription medicine.</td>
</tr>
<tr>
<td>Extremism</td>
<td>Content espousing extremist ideologies, including ideological texts, expressions of support for terrorist violence, militant how-to guides and extremist community forums.</td>
</tr>
<tr>
<td>Finance</td>
<td>Money laundering, counterfeit bills, trade in stolen credit cards or accounts.</td>
</tr>
<tr>
<td>Hacking</td>
<td>Hackers for hire, trade or distribution of malware or DDoS45 capabilities.</td>
</tr>
<tr>
<td>Illegitimate pornography</td>
<td>Pornographic material involving children, violence, animals or materials obtained without participants’ consent.</td>
</tr>
<tr>
<td>Nexus</td>
<td>Websites primarily focused on linking to other illicit websites and resources within the darknet.</td>
</tr>
<tr>
<td>Other illicit</td>
<td>Materials that did not easily fit into the other categories but remain problematic, such as trade of other illegal goods and fake passports or IDs.</td>
</tr>
<tr>
<td>Social</td>
<td>Online communities for sharing illicit material in the form of forums, social networks and other message boards.</td>
</tr>
<tr>
<td>Violence</td>
<td>Hitmen for hire, and instructional material on conducting violent attacks.</td>
</tr>
<tr>
<td>Other</td>
<td>Non-illicit content, such as ideological or political content, secure drop sites, information repositories, legitimate services.</td>
</tr>
<tr>
<td>None</td>
<td>Websites which were either completely inaccessible or otherwise had no visible content, including websites which hosted only placeholder text, indicating that their operator had yet to generate indicative content. Indicates a lack of content, and was thus counted as neither licit nor illicit.</td>
</tr>
<tr>
<td>Unknown</td>
<td>Means that content was there but Moore and Rid (2016) were unable to determine its nature, as it was too sparse or illegible.</td>
</tr>
</tbody>
</table>

The database analysis enlightened the presence of illicit material on the TOR network. The scan generated 5205 live websites and out of these 2723 were successfully categorized within the different segments with a high confidence level, see Table 13. The outcome was once again inspected manually to ensure the accuracy. The crawler accessed about 300 000 addresses within the TOR hidden services network, generating 205 000 unique pages.

The financial category mainly consisted of three sub categories; Bitcoin based methods for money laundering, trade in illegally obtained credit cards and stolen accounts and trade in counterfeit currency. The pornographic content was the most disturbing, according to the authors, even though it was not represented by images but only in text. Websites dedicated to providing links to videos purporting to depict rape, bestiality and pedophilia were abundant. One example included a service that sold online video access to the vendor’s own family members.
Table 13 Illustrates the result from the webcrawling session.

<table>
<thead>
<tr>
<th>Category</th>
<th>Websites</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>2,482</td>
</tr>
<tr>
<td>Other</td>
<td>1,021</td>
</tr>
<tr>
<td>Drugs</td>
<td>423</td>
</tr>
<tr>
<td>Finance</td>
<td>327</td>
</tr>
<tr>
<td>Other illicit</td>
<td>198</td>
</tr>
<tr>
<td>Unknown</td>
<td>155</td>
</tr>
<tr>
<td>Extremism</td>
<td>140</td>
</tr>
<tr>
<td>Illegitimate pornography</td>
<td>122</td>
</tr>
<tr>
<td>Nexus</td>
<td>118</td>
</tr>
<tr>
<td>Hacking</td>
<td>96</td>
</tr>
<tr>
<td>Social</td>
<td>64</td>
</tr>
<tr>
<td>Arms</td>
<td>42</td>
</tr>
<tr>
<td>Violence</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>5,205</td>
</tr>
<tr>
<td>Total active</td>
<td>2,723</td>
</tr>
<tr>
<td>Total illicit</td>
<td>1,547</td>
</tr>
</tbody>
</table>

Moore and Rid’s example of crawling the darknet is proof that even the deep web can be investigated and illegal activity can be classified. This further scares the perpetrators, and hopefully, with enough scientific proof, some of the services can be regulated or shut down.

7.2 Secondary method
When creating a framework that will be used in future investigation of cyberlockers to make an initial classification based on their transaction flows, a number of problems can arise. The first thing that comes to mind is the number of cyberlockers included in the framework. The optimal way of creating such a framework would be to include all existing cyberlockers and use the framework to classify new cyberlockers that are created. This framework would both be extremely expensive and difficult to create, which leads to the more realistic number of cyberlockers that should be included to create a framework that is desirable and usable. For example, if there exists 1000 different active cyberlockers on the market, a realistic number of investigated cyberlockers to include in the framework could be 100, which would equal to 10% of the total amount. This would be enough to be able to find patterns and trends in the transaction flows. This method is quite risky to use though, since a non-complete framework can include trends and patterns that would not be present in the complete framework. This could potentially mean that some cyberlockers are put on the “warning list” even though they are not being exploited by offenders. It is important that the investigating unit collaborates with the police to validate concerns about the exploitation of cyberlockers, so that the problem with false accusations is handled correctly.

7.3 Preventing payments
During 2005 the Swedish National Criminal Police established a collaboration with Swedish ISPs to prevent CAM on the internet. Today, almost all Swedish ISPs are included in this collaboration and the work consists of blocking pages with CAM at the ISPs so that the pages cannot be viewed by the ISPs customers. Between the years of 2005 and 2011, over 40,000 pages were thoroughly investigated and over 100,000 tips were received from the public. Apart from this collaboration, there exists a collaboration between the Swedish National Criminal Police, The Swedish Financial Coalition and the banks of Sweden to prevent payments for illegal material through credit cards. This is a collaboration that was established
during 2009 with the sole purpose of preventing purchases of CAM with credit cards. The method that was created to prevent these payments was based on an American method, constructed by The Financial Coalition Against Child Pornography, an organization funded by the National Center for Missing and Exploited Children. The method relies heavily on the cooperation between the credit card companies, the police and the banks and initially the aim was to investigate which companies and services that uses the credit cards as a payment option for illegal material. A large portion of the identified illegal material purchased with credit cards was CAM and a majority of the material found, were found on file hosting services. The collaboration between these parties stopped the payments by identifying these companies and prevented them from using their credit cards as a payment option. This forced the uploaders of the illegal material to rethink their “business idea” and consequently the amount of illegal payments through this type of services was minimized.

The collaboration also led to an increased transparency between the bank, the police and the credit card companies, which meant that new services who tried to use credit cards as a payment option for illegal material, were stopped quickly. When the perpetrators were buying and selling illegal material with credit cards and this collaboration was established, a lot of personal information regarding the perpetrators became visible to the police which led to a heightened level of fear amongst the perpetrators, since this could lead to them getting arrested.

Now, the perpetrators have found a new way of buying and selling illegal material and that is with crypto currencies. Therefore, it is important to establish the same kind of collaboration regarding these payments, as was done with the issue of credit cards, to be able to prevent this activity in the future. If the primary method, as explained in Chapter 6.1, and the secondary method, explained in Chapter 6.2, were to be implemented, the police would have a lot of information regarding which services and perpetrators to initially focus on. This would also give them enough information to start preliminary investigations against these perpetrators and consequently lead them towards establishing collaborations with cyberlockers and exchange services. At that point the banks are involved in the process, to create the same kind of transparency as with the credit cards, but now with crypto currencies instead. This could potentially mean, that crypto currencies cannot as easily be used to buy and sell illegal material. This transparency would also help the banks avoid being connected to the money that is used by these illegal businesses.

7.4 Risks
The risk with the illegal trade with Bitcoin is that the currency will be removed, in similarity with the virtual currency Flooz. If the board of Bitcoin recognize that a large amount of the virtual currency is used for illegal purposes, they can decide to shut it down. Therefore, to keep Bitcoin alive, a collaboration need to be developed because the exchanges loose if Bitcoin is dismissed. It is in everybody’s interest that the currency is not defiled.

If the Bitcoin currency seizes to exist or starts to get regulated in some way, new payment solutions will be developed. As earlier mentioned, the evolution of payments is exponential and ideas does not require long time to implement into reality. Therefore, it is best for the exchange services and to some extent also the public that Bitcoin stays active, because everyone is starting to understand and grasp the technique behind the currency. The risk is that a new virtual currency is developed and abused by offenders. This is why you need to narrow down the problem to catching only the people committing crimes and not innocent people using Bitcoin for legal purposes. The solution presented in this thesis does not involve a regulation of the currency, but a preventative control for exchanges to protect the currency.
7.5 Ethical aspect
In the context of CAM, a lot of ethical aspects need to be considered. The problem of the distribution of illegal material could be prevented if the ethics of the children was considered. This thesis wants to emphasize the problem and what solutions that could be of use in the future but the methods could be problematic for the public to accept, due to the ethical aspects.

7.5.1 Ideology
Anonymity was not created on the internet and the concept of anonymity was not something that was developed for the internet. Anonymity and the phenomena of hiding one’s identity is something that always existed and has been used both for privacy reasons and for illegitimate reasons for thousands of years. Some of the methods that have been used for centuries are still being used today, and while some are still used, most of them have transformed into the new age. The method of anonymizing a letter by not signing it, is probably the oldest method in the book when it comes to anonymizing correspondence. In modern times, this is still done, but is mainly used with electronic mail instead. Anonymity on the internet has not existed for as long as the internet has, but was rather developed on online communities as a pan-internet culture with a refined system of values. The most unique feature of these systems is the utmost and deep respect for anonymity, and more specifically the disassociation from one’s online identity and their real life identity.

This phenomena of viewing anonymity and pseudonym as a political philosophy is, as earlier mentioned, something that has risen from discussions within online communities. Their main political standpoint is the belief that anonymity is a human right. With great anonymity comes great privacy, since one can hide their identity while performing various tasks, ranging from sending correspondence to downloading illegal files. The initial thought of using pseudonyms to achieve anonymity online was not to enable illegitimate business and actions but was rather a consequence of the phenomena.

The initial purpose of virtual currencies, such as Bitcoin, was to create a currency which was not controlled by any state or government, but rather only controlled by the public and completely non-regulated. The users of virtual currencies mainly use it to be anonymous and make anonymous transactions and payments. The initial user community of Bitcoin does not like the commercialization of the currency. This narrow group of people does not use Bitcoin for illegal purposes, but rather uses it instead of standard currency just because of the fact that Bitcoin and other virtual currencies offer anonymous transactions.

The people who uses other anonymization services, such as TOR and I2P, can be viewed as a subset of the people who uses virtual currencies. TOR and other similar services, was created to make it more difficult for authorities and other people to trace where data was sent from, and who sent the data. The data in this case can be transactions, email and forum posts etc. The sender of an email, for example, does not sign the email with his/her name, so the recipient starts to trace his/her IP-address instead. These services will reroute the data to multiple sources before sending it to the final source, which makes it much more difficult to trace back to the original source. Services like this can be used by people who are living under oppression in regime controlled countries, dictatorship countries and countries where government obstructs the people’s freedom of speech. In these countries the usage of internet is often highly regulated and most of the websites on the surface web is blocked centrally by the government. The people can then use these services to bypass this blockade and this helps them to spread word to the world about for example injustices in their respective country.

TOR, I2P and similar services can also be used by perpetrators to hide their identity and location while conducting illegal business. This is the backside of the “cloaking”-services. The question is whether the
bad outweighs the good, and if so if these kind of services should be forbidden. These services were initially created as a tool for oppressed to bypass online oppression and the fact that these tools could be used to achieve anonymity online was merely a consequence that was not intended in the first place. The exploitation of these services for illegal use was not intended either, and can also be viewed as a consequence. In this case, the good does outweigh the bad, since there are hundreds of millions of people who are oppressed and to give them a chance to avoid this is worth the bad that it brings, especially if the governments and authorities find ways to get to these perpetrators even if they use cloaking services. The illegal business conducted using TOR, I2P and such, would still exist and would still be conducted even if these services didn’t exist. The main difference being that other anonymization services would be used and therefore the good does outweigh the bad and these services are a necessary evil.

7.5.2 Integrity and privacy
There has been a lot of discussions regarding privacy online recently. Private persons do not want the government, or other actors, to filter their behaviors online to gain control over the public. Therefore, they use ISPs such as Bahnhof which advocates “Internet with Secrecy”. The problem then arises that private persons are violated. Victims that could have been protected if filtering of big data was implemented. By not letting the authorities filter the data streams, innocent people lose their integrity in different ways. You can state that CAM only is represented by images of children and no assault was committed, but these children have been forced to pose or in other way act in front of the camera without any clothes on. Their integrity is violated and abused. If the law enforcement can get access to big data, investigations of the illegal activities can be performed and perpetrators can be caught.

7.5.2.1 Encryption
Moore and Rid (2016) argue that more encryption is better most of the time, but not all of the time. They refer to the pros and cons with encryption for the public. On one hand, encryption and cryptography provides privacy, democracy, commerce and even cyber security for the public. On the other, law enforcement argues that the more cryptography they cannot break, the worse. They also emphasize that terrorists, criminals and hostile states use encrypted ways of communicating, which further complicates for the government to protect the public. The discussion is relatively new, but very important at this time of the digital evolution. There is even a name for the subject; crypto politics.

7.6 Further research
The following segment represents some examples of future research that would be of use in the scope of this thesis. Also, regarding the chosen methods, a choice was made to validate the primary method through a manual search for illegally distributed movies. This due to the writers’ unwillingness to view and actively search for CAM. To better fit the problem of CAM, some examples are given regarding parameters that should be added within the specific scope of CAM.

The willingness of future collaboration between different parties is also discussed and other ways of interpreting the huge amount of data that the webcrawler generates. Unfortunately, the legal field is not considered as much as the writers would have wanted. The area should be further investigated regarding the methods presented in this thesis.

7.6.1 Methods
The primary method, which is composed of data gathering, correlation analysis and mapping of networks etc., could be expanded with further parameters to widen the search and make the results even more statistically reliable and more precise. More information regarding adding more parameters to the data gathering session will be discussed in Chapter 7.
The initial thought of both the primary and secondary method, was to deliver a complete framework and complete package of models and methods to the police so that they could easily implement them and adjust them after their wants and needs. The obstacle to pursue this within the scope of this thesis was that the authors did not receive the 16 names of proved cyberlockers that contain CAM from Europol. Hence, only three cyberlockers was validated through ECPAT to contain CAM and the other three is suspected to contain illegal material due to its transaction behaviors, illustrated in the result generated by the secondary method. Another factor to why the comparison between the framework and the 51 cyberlockers in Appendix B was not conducted, was the large amount of money it would cost to buy premium accounts within all 51 cyberlockers. Instead a small framework was created with the intention to show the thought process behind creating such a framework, so that a framework of larger scale easily could be created. The primary method was performed on a small data sample to validate the method and then on a larger data sample to imitate the implementation on the intended field of study, CAM. This method is meant to be performed on a very large data set eventually. Therefore, if further research on this subject were to be conducted, a sophisticated webcrawler needs to be created and implemented. The webcrawler should scan the whole surface web and either use the parameters specified in this thesis, or expand the webcrawler with more parameters, as discussed in Chapter 7.

7.6.2 Machine learning
Moore and Kid (2016) used machine learning as a tool for categorization of data. This could be an easier solution than MDS and AHC and should definitely be further investigated in relation to the data generated in this thesis. The biggest advantage with machine learning is that the program used will learn to recognize trends and patterns in the data and if the trends and patterns found are deemed important, the program will consequently search even more after these. It is also possible to tweak the program continuously if new trends and patterns arise and if new parameters are needed. This program could then incorporate the three first parts of the primary method stated in this thesis; data gathering, correlation analysis and the mapping of networks.

7.6.3 Collaborations
A further research that would be interesting to investigate, is how willing the Bitcoin exchange services are to collaborate with the police regarding the prevention of distribution of illegal material. This would be very interesting for the future work with the problem. There also exists a need for a collaboration with the cyberlockers that emerge as exposed, which has to be contacted.

7.6.4 Legal investigations
Another thing that the authors have considered is if you statistically can decide when a prosecutor will approve an initial investigation for the police. This would help the future mathematical work if there is a clear line of when enough underlying information is gathered by the police for an investigation to start. This would also help when further developing the primary method since it would then be possible to import more parameters to increase the reliability of the results that the model produce. If an exact decision point is presented, it could possibly be enough to perform an investigation with the primary method or a further developed version of the primary method and gather enough material to start an official investigation of a potential perpetrator.
8 Conclusions

The purpose of this study was to develop methods to investigate accessible data to identify and analyze digital payment flows regarding illegal material and consequently prevent further transactions involving Child Abusive Material on the Internet. Considering the purpose to prevent further transactions regarding CAM, previous experience of transactions regarding CAM with credit cards has been analyzed and adjusted to fit the problem regarding payments with cryptocurrencies connected to illegal material. The idea of preventing the distribution of CAM through a warning on the web page that the site could contain illegal material, decreased the interest for the perpetrators and they had to find other solutions, such as Bitcoin. Now the same preventative measures need to be taken regarding Bitcoin transactions. It is in everyone’s interest to protect the currency from exploitation, so no official regulations limit the currency.

Another way of answering the purpose has been to find new methods that would fit the accessible data and return the desired output. The idea was to use already existing forensic methods regarding cyberterrorism, cybercrime and money laundering and adapt these to fit the problem of CAM. The chosen methods use the data collection method, webcrawling, to receive as much information as possible. The information is then handled through Multidimensional Scaling and Agglomerative Hierarchal Clustering to better understand the data. In this way, the police receive as much information as possible and an investigation can be initiated.

So, to investigate the accessible data a crawling session was implemented. Through investigations and analysis of the cyberlockers surrounding transactions, a small framework was created through manual searches in Chainalysis’ tool Reactor. To prevent further transactions regarding CAM, the method of stopping payments with cards regarding illegal material was implemented regarding Bitcoin transactions.

To make these methods successful, a collaboration is needed between different parties and preferably a team within the police department need to shoulder the responsibility of continuing and developing these methods. Through automated processes, less man power is needed and resources can be spared. Basically, the methods presented in this thesis are meant to relieve and streamline the processes within the law enforcement so time can be spared to physically catching the offenders.
Bibliography


European Financial Coalition, EFC. 2015. *A Strategic Assessment; Commercial Sexual Exploitation of Children Online.*


*Asplund, Berggren, 2016*


Kumar, V., Steinbach, M. and Tan, P-N. 2006. *Introduction to Data Mining*. Pearson. USA.


Menczer, F., Pant, G. and Srinivasan, P. 2004. *Crawling the web*. Department of Management Sciences,
Asplund, Berggren, 2016

School of Library and Information Science. The University of Iowa, Iowa City. USA.


Appendix
A. Gant-scheme of the time frame
B. Table of the investigated cyberlockers.

<table>
<thead>
<tr>
<th>Cyberlocker</th>
<th>Bitcoin</th>
<th>Card</th>
<th>Solutions like PayPal</th>
<th>Other payment solution</th>
<th>Premium account</th>
<th>Paid to upload</th>
<th>Contact to a private person</th>
<th>Advocated safety (1-5, where 1 is low)</th>
</tr>
</thead>
<tbody>
<tr>
<td>depositfiles</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>5</td>
</tr>
<tr>
<td>BigFile</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>4</td>
</tr>
<tr>
<td>Plunder</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>Share-Online.biz</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>ifichier.com</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>cloudsix.me</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>doraupload.com</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>easybytez.com</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>fileboom.me</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>filespace.com</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>florenfile</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>hugefiles.net</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>interfile.net</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>keep2share.cc</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>limefile.com</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>mega.co.nz</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>movshare.net</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>nitroshare.com</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>novafile.com</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>rapidgator.net</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>rarefile.net</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>salefiles.com</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>speedyshare.com</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>tusfiles.net</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>upload.cd</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>userscloud.com</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>3</td>
</tr>
<tr>
<td>Egoshare.com</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>turbobit.net</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>Uploaded.to</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>alfille.net</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>datafile.com</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>defile.com</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>dfiles.eu</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>Filejoker.net</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>filerio.in</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>filesmonster.com</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>hitfile.net</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>hotbytez.com</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>rapidrar.com</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>rockfile.eu</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>streamcloud.eu</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>subyshare.com</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>upstore.net</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>uptobox.com</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>2</td>
</tr>
<tr>
<td>Filefactory</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>1</td>
</tr>
<tr>
<td>Ziddu</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dataport.cz</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>1</td>
</tr>
<tr>
<td>letitbit.net</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>1</td>
</tr>
<tr>
<td>MultiShare.cz</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>1</td>
</tr>
<tr>
<td>QuickShare.cz</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>1</td>
</tr>
<tr>
<td>RGhost</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>1</td>
</tr>
</tbody>
</table>
### Table of examples of how the collected data from the webcrawler can look like.

<table>
<thead>
<tr>
<th>Year</th>
<th>Excerpt</th>
<th>Timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>Example 1</td>
<td>2016-01-01</td>
</tr>
<tr>
<td></td>
<td>Example 2</td>
<td>2016-01-02</td>
</tr>
<tr>
<td></td>
<td>Example 3</td>
<td>2016-01-03</td>
</tr>
<tr>
<td></td>
<td>Example 4</td>
<td>2016-01-04</td>
</tr>
<tr>
<td></td>
<td>Example 5</td>
<td>2016-01-05</td>
</tr>
</tbody>
</table>

*Source: Asplund, Berggren, 2016*
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.250175928</td>
<td>0.249993155</td>
<td>0.749958639</td>
<td>0.5000006351</td>
<td>0.74996077</td>
<td>0.749962197</td>
<td>0.49999338</td>
<td>0.74744428</td>
<td>0.752338587</td>
<td>6.749552789</td>
<td>0.749956589</td>
<td>0.758866863</td>
</tr>
<tr>
<td>2</td>
<td>0.249824196</td>
<td>0</td>
<td>0.249817315</td>
<td>0.749782264</td>
<td>0.49982547</td>
<td>0.749830275</td>
<td>0.749789417</td>
<td>0.499783542</td>
<td>0.74565867</td>
<td>0.75214072</td>
<td>0.749777818</td>
<td>0.747788815</td>
<td>0.75864823</td>
</tr>
<tr>
<td>3</td>
<td>0.35006845</td>
<td>0.25017278</td>
<td>0</td>
<td>0.49964483</td>
<td>0.500007197</td>
<td>0.75802922</td>
<td>0.74972042</td>
<td>0.49966183</td>
<td>6.74751104</td>
<td>0.752325416</td>
<td>0.749596344</td>
<td>0.749633344</td>
<td>0.758873551</td>
</tr>
<tr>
<td>4</td>
<td>0.75004968</td>
<td>0.750219762</td>
<td>0.50035122</td>
<td>0</td>
<td>0.750043319</td>
<td>0.75803044</td>
<td>0.50000716</td>
<td>0.75800129</td>
<td>0.752360844</td>
<td>0.74999449</td>
<td>0.74995849</td>
<td>0.758910319</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.749999609</td>
<td>0.8001357</td>
<td>0.49992883</td>
<td>0.749958788</td>
<td>0</td>
<td>0.74995725</td>
<td>0.49964846</td>
<td>0.24998987</td>
<td>0.7474447</td>
<td>0.75238152</td>
<td>0.749982438</td>
<td>0.74956237</td>
<td>0.758866499</td>
</tr>
<tr>
<td>6</td>
<td>0.750003923</td>
<td>0.750179854</td>
<td>0.74997078</td>
<td>0.74996162</td>
<td>0.75004275</td>
<td>0</td>
<td>0.74996912</td>
<td>0.74996261</td>
<td>0.754748126</td>
<td>0.75232447</td>
<td>0.749556712</td>
<td>0.74969511</td>
<td>0.758870925</td>
</tr>
<tr>
<td>7</td>
<td>0.750034808</td>
<td>0.75017076</td>
<td>0.75002796</td>
<td>0.49999341</td>
<td>0.50033515</td>
<td>0.75803884</td>
<td>0</td>
<td>0.4999414</td>
<td>0.75478097</td>
<td>0.752335618</td>
<td>0.749878591</td>
<td>0.74999139</td>
<td>0.758905905</td>
</tr>
<tr>
<td>8</td>
<td>0.500046668</td>
<td>0.580216625</td>
<td>0.50003582</td>
<td>0.749997071</td>
<td>0.2504102</td>
<td>0.75803674</td>
<td>0.5000056</td>
<td>0</td>
<td>0.754776269</td>
<td>0.752359553</td>
<td>0.74999345</td>
<td>0.7499725</td>
<td>0.758908973</td>
</tr>
<tr>
<td>9</td>
<td>0.74534548</td>
<td>0.745531699</td>
<td>0.74533683</td>
<td>0.74532569</td>
<td>0.74534399</td>
<td>0.74531903</td>
<td>0.74530644</td>
<td>0</td>
<td>0.747688871</td>
<td>0.754297217</td>
<td>0.745308945</td>
<td>0.75405258</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.747702797</td>
<td>0.747877399</td>
<td>0.747696015</td>
<td>0.747661222</td>
<td>0.747701164</td>
<td>0.747668434</td>
<td>0.747662369</td>
<td>0.752410426</td>
<td>0</td>
<td>0.74766402</td>
<td>0.74765785</td>
<td>0.754848184</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.75004722</td>
<td>0.750233181</td>
<td>0.750048373</td>
<td>0.75005251</td>
<td>0.750047871</td>
<td>0.750043295</td>
<td>0.75401242</td>
<td>0.75000615</td>
<td>0.74791244</td>
<td>0.752336115</td>
<td>0</td>
<td>0.5000383</td>
<td>0.758917877</td>
</tr>
<tr>
<td>12</td>
<td>0.75004449</td>
<td>0.750219377</td>
<td>0.750036782</td>
<td>0.750045185</td>
<td>0.750043777</td>
<td>0.750045959</td>
<td>0.750008761</td>
<td>0.7520027</td>
<td>0.747879971</td>
<td>0.752363238</td>
<td>0.4999962</td>
<td>0</td>
<td>0.758911821</td>
</tr>
</tbody>
</table>
| 13  | 0.74143685 | 0.741646752 | 0.74143024 | 0.741396327 | 0.74143719 | 0.74143061 | 0.74140324 | 0.741397581 | 0.746019156 | 0.74675943 | 0.741391257 | 0     | 0.741394296 | 0