Behavioral Profiling of Darknet Marketplace Vendors

Sylvester Shan
u6049249

Supervised by
Ramesh Sankaranarayana

A thesis submitted for the degree of
Bachelor of Advance Computing (Honours)
The Australian National University

June 2020

Except where otherwise indicated, this thesis is my own original work.

Sylvester Shan
June 12, 2020

I am dedicating this thesis to my family and dear friends. With out the support and help from everyone, I would not have made it this far.
Thank you

Acknowledgments

I want to thank the support that I received from my parents and sister.

I would like to acknowledge my supervisor Ramesh, and thank him for his patience, guidance and support throughout my research.

And thanks to all my friends who support me.

Abstract

The usage and number of darknet users has increased rapidly in recent years. A key reason is that the darknet allows users to be fully anonymous when browsing on the darknet. Though such privacy is needed for some users, others decide to abuse darknets by selling or buying illicit goods off the darknet marketplace without being arrested or punished. Despite the hidden nature of darknet marketplaces, they often shut down due to reasons such as law enforcement activities or exit scams. As a result, the average life span of a darknet marketplace tends to be around 8 months. This leads to an important question: *If a vendor has built up a good reputation before a darknet was shutdown, does that mean he will start over again from scratch?* Not likely. A vendor would most likely use their username as a brand, in order to be recognizable on a different darknet marketplace when others shuts down.

This thesis states and explores the hypothesis: *Accounts that belong to the same individual are likely to have similar usernames, which are being used as a "Brand" by the vendor*. To verify this hypothesis, we first devise a method to correlate the accounts in a darknet marketplace data set using their PGP keys, thus linking multiple accounts to a single user. We then devise a method for determining username similarity, and check if the correlated accounts have a username similarity above a certain threshold. These experiments are done both internally within the datasets for the *Evolution* marketplace and the *SilkRoad2* marketplace, and also between the two datasets.

From the experiments, four behaviors were identified and they were used to verify and strengthen the hypothesis. Most importantly, we find that two accounts that belong to the same user are likely to have similar usernames if the accounts belong to different marketplaces, but not if the accounts belong to the same marketplace. We thus conclude a modified version of our initial hypothesis: *Accounts that belongs to the same individual, but are on different marketplaces, are likely to have similar usernames, which are being used as a "Brand" by the vendor*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Section 1.1 The Darknet: Introduces the clear web, deep web, darknet, the anonymity property of the dark net and how the dark web is used.
Section 1.2 The Tor Browser: Introduce the software Tor browser, which is required to access the darknet.
Section 1.3 The Darknet Marketplace: Introduces what the darknet marketplace is, how vendors and buyers operate on the marketplace and how the marketplace operates.
Section 1.4 Feedback is Everything: Address the importance of feedback to vendors and buyers.
Section 1.5 Returning buyer looking for vendor: This section introduces why it is likely that vendors uses their username as a "brand" on different marketplaces.
Section 1.6 Motivation: Provides the inspiration behind this thesis.
Section 1.7 Thesis Statement: The hypothesis this thesis verifies.
Section 1.8 Key Assumptions: Addressing important assumptions that were made and used in this thesis.
Section 1.9 Chapter Summary: Summarizes this chapter and introduces the next chapter.
Section 1.10 Thesis Outline: Outline of the rest of the thesis.

1.1 The Darknet

The internet can be divided into three sections: the surface web (clear web), deep web and the darknet (dark web), where the darknet is a subset of the deep web. It's estimated that the surface web takes up 5% to 10% of the internet and the deep web takes up 90% to 95%.

When it comes to an average internet user using the internet to look up a piece of information, they will likely use search engines such as Google, Bing, Yahoo, DuckDuckGo, etc. Search engines use crawlers to index websites, which is later used to return a list of indexed websites depending on the search query. All the indexed websites that can be reached or found using these search engines belongs to the surface web.

An average internet user that uses the internet on a daily basis is also likely to use the deep web. Deep web refers to web pages or websites that are intentionally designed in a way such that it cannot be indexed by a standard search engine. For example, the content of an individual's google drive, email or social media, data that is stored in databases by organizations etc. Various methods exists to prevent web pages from being indexed. For example, websites that requires username or password, a website that can only be accessed using certain software such as Tor.

The darknet is a subset of the deep web. Many illegal activities takes place on the darknet, such as malware service, hacker forum, selling illicit goods on different marketplaces, such as drugs, weapons, stolen information etc. It is unlikely for an average internet user to accidentally access the content on the darknet, as certain software are required to access it, such as Tor, I2P etc. By using such software, it allows users to access the content on the darknet with anonymity. In most countries it is not illegal to access the darknet, and by using the anonymity property of the darknet, it is near impossible for the user's identity to be lifted by the police, government or legal agencies. This is the major reason why the darknet roams with illegal activities, as users' identities are hidden away by the darknet.

The darknet is not criminogenic, as it is a tool which can be used in both good and bad ways [Mirea et al., 2018]. The anonymity property of the darknet brings convenience for users who abuses it for illegal activities, but it also protects the privacy of other users without such intention. For example, whistleblowers uses the darknet to share sensitive information from companies or organizations without revealing their identity, journalists use the darknet to collect sensitive information and avoid censorship etc. The dark side of the darknet often attracts medias' attention, while the bright side is often forgotten. [Mirea et al., 2018]

In this thesis, we will be focusing on the dark side of the darknet: the marketplaces.

1.2 The Tor Browser

The Tor Browser most commonly used software to access the darknet. To ensure the anonymity of a user when browsing the darknet, many servers called Tor nodes are needed, which are distributed around the world. When accessing a website on the darknet, the Tor Browser will connect users to a random Tor node, which then routes through other random Tor nodes, after exiting the final Tor node, it connects the user to the website. The path of the traffic is completely random and encrypted, assuring it is impossible to backtrack to identify a user. This process repeats every time when the user wishes to access a different web page.

Therefore, it is near impossible to shut down the darknet, as this will requires to shut down all the Tor nodes and hidden servers that hosts websites at the same time across the world.

1.3 The Darknet Marketplace

Darknet marketplaces are similar in many ways to the marketplaces on the clear web, such as Amazon and Ebay. The key difference between the darknet marketplace and clear net marketplace is that users are allowed to buy or sell goods with anonymity. Therefore by abusing such property, vendors are able to sell illicit items such as drugs, weapons, stolen information etc [Decary-Hetu et al., 2016]. The majority of items that are listed on darknet marketplaces are drug related. According to Europol [2017], with respect to several popular darknet marketplaces such as AlphaBay, Dream Market etc., around 62% of the listed items on these marketplace are drugs or drug-related chemicals.

1.3.1 Drug Dealers and Buyers Going Online

Before darknet marketplaces exists on the darknet, illegal drugs were sold by drug dealers on the streets. There are multiple risks for both drug dealers and buyers when involved in the process of selling or purchasing drugs, which encourages both vendors and buyers to sell or buy illicit goods online [Buxton and Bingham, 2015].

Drug dealers faces many risks other than the risk of getting arrested. They also face the risk of violence from other competitors and customers. According to Buxton and Bingham [2015], to reduce such risk drug dealers build reputation based on violence and fairness. With a violent reputation, it's less likely for the drug dealer to be threatened by other competitors or buyers. With a reputation of fairness, it helps maintain long-term relationship with buyers and employees. Therefore to be successful, it's important to build and maintain a good reputation. Though drug dealers might face the risk of violence from buyers, the reverse also holds true. Buyers could also face the risk of violence from drug dealers [Buxton and Bingham, 2015].

It is safer for drug dealers and buyers to sell or buy drugs on the darknet marketplace. As vendors are able to use the anonymity property of the darknet to protect their true identities, avoid being arrested by law enforcement agencies. This also helps avoidance of violence from other competitors and buyers. Selling or buying illicit goods or drugs on the darknet also help avoid putting the vendor and buyer at the risk of violence. The amount of risk vendors or buyers need to take on the darknet marketplace is substantially lower than the traditional face-to-face transaction method [Buxton and Bingham, 2015].

1.3.2 Communication: PGP Encryption

On the marketplace on the clear web, it uncommon for customers to ask questions about the product that is listed on the marketplace. This also applies to the darknet marketplaces, buyers might need to communicate with the vendor or admin of the marketplace. To maintain the anonymity property of the darknet, communications need to be encrypted. The most common method is using PGP encryption to encrypt messages [Afilipoaie and Shortis, 2015a; Buxton and Bingham, 2015].

1.3.3 Cryptocurrencies

Real world currencies cannot be used on darknet marketplaces, as transactions leave records behind, making it traceable [Buxton and Bingham, 2015]. Cryptocurrencies with anonymity properties are used instead, where Bitcoin is the most commonly used cryptocurrency across different darknet marketplaces Afilipoaie and Shortis [2015a].

1.3.4 Marketplace Income

Similar to the marketplace on the clear web, the darknet marketplace operators makes money by charging vendors fees and commissions. Meland et al. [2020] For example, operators on the darknet marketplace Silk Road makes approximately USD 92,000 per month in commissions. Christin [2012] Marketplace Agora charges fees for new vendors and also charge fees for each order. Refer to Figure F.1 for more details.

1.3.5 Marketplace Rules

Though the darknet provides protection to users identities, darknet marketplace operators still don't want to attract law enforcement agencies attention. Hence each marketplace has their own set of rules to keep a low profile.

For example, many darknet marketplaces such as Agora, Dream Market frowns upon selling goods related to child exploitation material (CEM) [Broadhurst, 2019; Branwen et al., 2015]. Agora also forbidden selling weapons of mass destruction, poison or services that does harm to other individuals such as hitman hiring etc. Figure F.2 includes the market rules for Agora.

Marketplace also have their own policy and rules to avoid scammers creating accounts to scam inexperienced buyers [Afilipoaie and Shortis, 2015a]. For example, Agora asks new vendors to deposit 1.5 BTC first before Agora grants them a vendor account. At the time this information was collected (around September 2017), 1.5 BTC worth approximately 5,500 Australian Dollars. The money could be returned if a vendor decides to leave Agora, but this still requires scammers to first collect 1.5 BTC before setting up a fake vendors account. For more information about the rules, please refer to F.3.

Various marketplaces offers escrow services, for example SilkRoad [Christin, 2012], Apollon [Ball et al., 2019] etc. This helps prevents new buyers from being scammed, also helps new vendors to build their reputation.

1.3.6 The End of a Marketplace

The average life span of a darknet marketplace is 8 months Europol [2017]. A marketplace could shut down due to law enforcement activities, exit scam, voluntarily exit or competitors DDoS attack the marketplace. Where an exit scam refers to marketplace operators stealing the escrowed bitcoins and shutting down the marketplace without any notice.

1.4 Feedback Is Everything

As mentioned above, successful drug dealers needs to maintain a good reputation. This is also similar for vendors on darknet marketplaces, as they need to maintain a good reputation by gaining positive or neutral feedback [Afilipoaie and Shortis, 2015a].

Positive and neutral feedback are crucial for vendors. For a vendor to be successful on the marketplace, the vendor needs to be trust worthy and have a good reputation. For a buyer, the general method to judge if a vendor is trust worthy or not is by looking at the feedback, which are given by other buyers Afilipoaie and Shortis [2015a]. With many positive and neutral feedback, it increases the chance of a first time buyer to purchase goods from the vendor. If buyer is satisfied with the goods, it's likely they will become a returning customer. Therefore vendors takes their orders seriously, from product quality to packaging, in order for the buyer to leave a positive feedback Redman [2016]. Vendors will try their best to offer solutions to avoid negative feedback, or be very specific about their terms and conditions in their product descriptions. Some vendors will explicitly ask buyers to contact them before leaving negative feedback. An item description example can be found in the appendix F.4.

Positive, neutral and negative feedback helps buyers to avoid scammers Afilipoaie and Shortis [2015a]; Buxton and Bingham [2015]. Though majority vendors have the intention of making money, there do exists a minority of scammers in marketplaces. Due to the anonymity property of the darknet, and purchasing illicit items from the darknet marketplace being illegal, it's near impossible to take legal actions when buyers are scammed. Though this is an advantage for scammers, it is still hard for them to scam inexperienced buyers. As discussed above, scammers first need to create accounts. This is hard for popular marketplaces, as each marketplace will have their own policy when creating new accounts. Using Agora as an example, individuals who wants to create a new account first needs to collect and deposit 1.5 bitcoins, or show evidence that they are legitimate vendors from other marketplaces. For more details about the rules and policy for Agora, refer to Figure F.3 in the appendix.

Buyers tend to believe a vendor is trust worthy if the vendor has many neutral or positive feedback for their goods. While buyers tend to avoid vendors with no or few feedback, as it's an indication that the vendor is possibly a scammer. It is likely that when scammers are reported to the marketplace, their account will be banned or removed.

Therefore from the reasoning above, we made the assumption $\mathcal{A}1$, such that there are no scammer accounts in the used data sets for this thesis.

1.5 Returning buyer looking for vendor

It's common for vendors to create accounts on different marketplaces. Darknet marketplaces has an average life span of 8 months [Europol, 2017]. Meaning when a marketplace shuts down, if a vendor wants to continue their business, they will need to create a new account on a different marketplace. In fact, vendors can have multiple benefits for creating accounts on different marketplaces before a market closes down. Many marketplaces includes the feature of importing feedback from other marketplaces. This can help avoid the "fresh" looks of an account and increase the chance of a buyer to purchase items. Having accounts on different marketplace platforms can help expand the customer base and increase the number of orders. It helps the vendors business to continue when a marketplace shuts down, avoiding multiple awkward situations. For example, if a marketplace shuts down and doesn't have accounts on a separate marketplace, and they want to continue business, they will then create a new account, but they won't be able to import feedback from the shutdown marketplace, forced to start with a "fresh" look and build their reputation from scratch again. After a marketplace shuts down, vendors with an account on a separate marketplace can continue their business with some customer base. Therefore from the discussion above, we make the assumption $\mathcal{A}2$: Assume vendors have different accounts across different marketplaces.

To ensure returning customers are able to find the same vendor on different marketplaces or when a new account is made, the vendor needs some something that stands out from other vendors. Thus, vendors are likely to use their usernames as a "Brand", to help returning customers to identify that they are the same vendor from other closed marketplaces. Furthermore, vendors are likely to use similar phrasing and structures in their item descriptions. This could be due to their natural style of writing, due to quickly copy and pasting the new listings from a previous account, or again to help identify themselves to returning customers. But what will instantly stand out will be the vendors username, as a buyer will generally see the vendor's username first before the item description.

1.6 Motivation

In real life, vendors often market themselves using a brand, which is recognizable and trusted by consumers. We do not know if this behavior exists among vendors in the darknet marketplace. Due to this, we explore whether this behavior exists by conducting experiments and analysis.

At the same time, we also profile other emergent behaviors of darknet vendors. We believe these behaviors could potentially give use insight, which could be used to make informed assumptions when building data sets.

For darknet marketplace data sets, we don't have the ground truth due to the anonymity property of the darknet marketplace, making it hard to study the darknet marketplace data set.

It is still possible to create data set that are close to the ground truth, such as how Zhang et al. [2019] constructed a data set by splitting a piece of text written by a vendor into two and match it to itself to create a positive data point, then match it to a piece of text that was written by a different vendor, in order to create a negative data point.

We proposed a vendor correlation method, where we match all the PGP key used by the two accounts from the darknet marketplace, if there is a match, we could nearly guarantee the two accounts belongs to the same individual, as PGP are long randomly generated strings. Hence we obtained a data set that is very close to ground truth.

Instead of using this data set to train machine learning models to correlate vendors or for other tasks, we decided to use this data set to put ourselves in the vendors' shoes, so that we could make observations of their behaviors on the darknet marketplace. And we believe we could use their behavior as assumptions to create larger data sets. For example, if we concluded a behavior such that darknet marketplace vendors uses the same username on each account on different darknet marketplace, then we could correlate vendors just based on measuring their username's similarity. Making it easier to create a data set that is closer to ground truth and used for other purposes such as machine learning.

1.7 Thesis Statement

From the discussion in the sections above, we believe that a behavior exists among the darknet vendors: they will use their username as a "brand", such that returning customers will be able to find them. Therefore, in this thesis we made and explore the hypothesis: *Accounts that belong to the same individual are likely to have similar usernames, which are being used as a "Brand" by the vendor.* Which is a behavioral of darknet marketplace vendors.

Thus we proceed in two steps: first we correlate the vendors account based on attributes other than the username. Then we compute the username similarity of the two correlated accounts, which is then used to calculate the percentage of correlated accounts with a username similarity greater than a threshold γ. While conducting experiment, attempt to conclude other interesting darknet vendor behavior.

1.8 Key Assumptions

Due to the anonymity property of the darknet, we make the following informed assumptions about its nature and to simplify problems:

$\mathcal{A}1$: Assume there are no scammers in the used data set. 1.4

$\mathcal{A}2$: Assume vendors have different accounts across different marketplaces.

$\mathcal{A}3$: Before any evidence, assume each account in all marketplaces belongs to a unique vendor.

$\mathcal{A}4$: Assume vendors operates as individuals on marketplaces.

$\mathcal{A}5$: If two PGP keys matches, then they belong to the same individual.

$\mathcal{A}6$: Vendors are likely to use usernames that differs to other usernames.

$\mathcal{A}7$: Assume usernames are unique in the same marketplace.

Assumption $\mathcal{A}3$ was made to formalize the problem. We initially treat each account belongs to one single individual and classify two or multiple accounts belongs to the same individual based on evidence. For example, we classify two accounts belongs to the same vendor when the PGP key matches. Assumption $\mathcal{A}4$ was made to simplify the problem. We cannot exclude the possibility of several individuals working together on the darknet marketplace using different accounts. Assumption $\mathcal{A}5$ was made as it is near impossible to generate two identical PGP keys. For a returning buyer, it will be hard to find the vendor they bought from if the vendor uses exact same usernames as other vendors. Therefore me made $\mathcal{A}6$.

Based on the assumptions above, in this thesis we attempt to figure out which accounts belong to the same individual, and use the correlated vendors summarize their different behaviors.

All assumptions made in this thesis can be found in Appendix E.

1.9 Chapter Summary

This chapter introduced the darknet marketplace and it's important anonymity property, which is abused by individuals who sells illicit goods online. Then we introduced how the darknet marketplace operates, the roles vendors and buyers play in the darknet economy system. Then we stated the motivation of this chapter: by using a ground truth data set, put ourselves into the vendor's shoes to conclude darknet marketplace vendor behaviors. Which could later be used as assumptions to create a larger data set. Finally the thesis statement was addressed and we introduced the key assumptions that was made in this thesis. In the next chapter, we introduce the work that was done in the past related to this thesis.

1.10 Thesis Outline

There are 5 more chapters in the thesis. Chapter 2 talks about the related work that has been done. Chapter 3 proposes the methods that we used to preprocess the raw darknet marketplace data set. Chapter 4 proposes the methods that was used to correlate vendors, how to calculate username similarities of two accounts and the procedure to verify the hypothesis Chapter 5 introduces different experiments that was conducted, their results and the behaviors that was concluded from the results.

Chapter 6 is the conclusion of this thesis, addressing what this thesis achieved, the limitations and future work.

Chapter 2

Related Work

This chapter introduces various research paper that has studied the darknet, darknet marketplace, darknet vendor and buyers.
Section 2.1 Drug dealers to Vendors: Discuss about the paper that address how the darknet operates and how the vendors and buyers sell or purchase goods.
Section 2.2 Correlation Problem: Discuss about the common correlation problem that many paper had.
Section 2.3 Correlation Methods: Discussed about the different correlation problems that was proposed.

Many different kinds of research was done related to the darknet and darknet marketplace. The i

2.1 Drug dealers to Vendors

There are multiple paper suggesting drug dealers are moving online, such as Afilipoaie and Shortis [2015a] and Buxton and Bingham [2015]. Afilipoaie and Shortis [2015a] stated it is trivial for a drug dealer to become a vendor online with limited understand of Tor, Bitcoin and PGP. This paper also introduces the common practices conducted by darknet vendors when selling illicit goods, the common practices conducted by a buyer when purchasing illicit goods. It also goes into the details of how to acquire bitcoins and set up stores for vendors. For buyers, it talks about the methods of purchasing the goods. Finally, Afilipoaie and Shortis [2015a] talks about how the escrow and payment system works to prevent scammers. The paper could also be treated as a guide to achieve all the topics it covered briefly. Afilipoaie and Shortis [2015a] addressed the risks a drug dealer needs to take. A drug dealer faces multiple risks at the same time: the risk of being arrested and exposure to violence. While for a darknet vendor, they can use the anonymity property fo the darknet to hide away from all those risks. To be a successful drug dealer, they need to build a violent and fair reputation. Such that competitors and customers will likely to leave them alone, at the same time maintaining a long-term good relation ship with their customers. For a vendor to be successful, from Afilipoaie and Shortis [2015a], they focus on their feedback. A vendor's reputation is built upon the feedback they get after selling goods. Comparing to the difficultly of keeping a successful business, is likely being a vendor is less difficult, as darknet vendors don't need to worry about being arrested, threatened or attack by other competitors or customers, as their true identity is protected by the marketplace. From Afilipoaie and Shortis [2015a], successful drug dealers adapts to the environment and they "should be considered active agents". With how simple it is to setup a store on the darknet marketplace, there is a non-zero chance sometime in the future all drug dealers will become darknet marketplace vendors.

2.2 Correlation Problem

Numerous quantitative and qualitative research was done to find the characteristics of darknet marketplace and darknet vendors. Christin [2012] was able to mind many interesting patterns from Silk Road 2, but the patterns cannot be applied to every other marketplace. Dolliver and Kenney [2016] was able to find few facts between the characteristics of vendors between marketplace Evolution and Agora, but the facts and patters that was found was not able to be applied to other marketplace. These were majorly due to poor vendor correlation method. Many chose to do a 1:1 comparison between the usernames of two accounts, if they match, then they classify the two accounts belongs to the same person. Many papers have difficulty to correlate vendors due to the anonymity property of the darknet, which prevents them from getting the ground truth data set to do such correlation.

2.3 Correlation Methods

A huge variety of correlation method was explored. The most common correlation method is string correlation of the username. Christen [2006] introduced a wide variety of methods to match real names. Though majority technique might not work well with usernames, it introduces techniques to quickly filter out names that doesn't match. Wang et al. [2016] introduced computing username similarity by constructing a username feature vector, then apply the main idea of IDF to each entry of the vector to create a self information vector. To compute the username similarity, the cosine similarity of two self information vectors were computed. Though the approach was interesting, the method was not suitable if the data set is consists of usernames from a darknet marketplace, as we won't know how well the method performs since we don't know the ground truth. The most interesting work was done by Zhang et al. [2019], they constructed a attribute heterogeneous information network, which correlates two vendors via meta paths. Many attributes were used as a vendors identification, such as the writing style, photography style, the drugs a vendors sells.

2.4 Summary

This chapter explored the paper that introduced how the darknet and darknet marketplace worked, the vendor correlation problem and solutions to the vendor correlation problem.

Chapter 3

Data Set Preprocessing

This chapter mainly focuses on how to preprocess raw data sets that was collected by a web-crawler. The nature and characteristics of the used raw data sets are addressed first. Then addresses the importance of studying the raw data sets and the unique ID for each data point before constructing the data set. Then introduces the attributes of the integrated data points, which will be used later in 4.

Section 3.1 Raw Data Sets: Provides justification for the chosen raw data sets and their nature.
Section 3.2 Information Scraper: States the preparations before implementing an information scraper for a raw data sets, and the attributes information scraper should be collecting.
Section 3.3 Data Integration: Introduces the motivation of data integration and the attributes of a data point after integration.
Section 3.4 Summary: The summary of this chapter.

3.1 Raw Data Sets

For this project, creating new raw data sets from scratch would be difficult and time consuming. To create a raw data set for one darknet marketplace, a web-crawler will need to be implemented. Ideally, this web-crawler will be able to capture and produce a static version of the marketplace whenever it is used. Hence to have a sufficient raw data set for one marketplace, a web-crawler needs to be run on a daily or regular basis, which could take up to months to do so. At the same time each marketplace's layout is different, therefore a separate web-crawler is needed for each marketplace. Yet we are still disregarding other problems and unknown factors, such as an error was found in one of the scrapers, scrapers need modification due to the change of layout of a marketplace, a marketplace shutdown before a sufficient amount of data was collected.

Therefore, instead of creating our own data set, we decided to use data sets from a public data dump: Darknet Market Archives. [Branwen et al., 2015] Though there are many preprocessed data sets that are available and could be used immediately, many did not include information such as the PGP key used by vendor, item description (which also could include the PGP key) or vendor's username. We have decided to create our own data sets using the raw data sets for marketplace Evolution and Silk Road 2. Where Evolution was active between 14th January 2014 [Afilipoaie and Shortis, 2015b] and mid March 2015, Silk Road 2 was active between 6th November 2013 and 6th November 2014 [Greenberg, 2014]. To visualize timeline, refer to Figure 3.1 below.

The dates when the two data sets where scraped overlaps. Evolution has been scraped between dates of 21st January 2014 and 17th March 2015. Silk Road 2 has been scraped between dates of 20th December 2013 and 6th November 2014. To visualize timeline, refer to Figure 3.1 below. The reason why no extra data was collected afterwards is due to both marketplace shutting down. Silk Road 2 was shut down due to an law enforcement operation called Operation Onymous. [Afilipoaie and Shortis, 2015b] Evolution was shutdown due to exit scam, where the marketplace operators stole all the escrowed bitcoins and shutdown the marketplace without any notice. [Greenberg, 2017] After Silk Road 2 shuts down, it is likely that vendors will move to Evolution, which was one of the most popular marketplace on the darknet. [Afilipoaie and Shortis, 2015b]

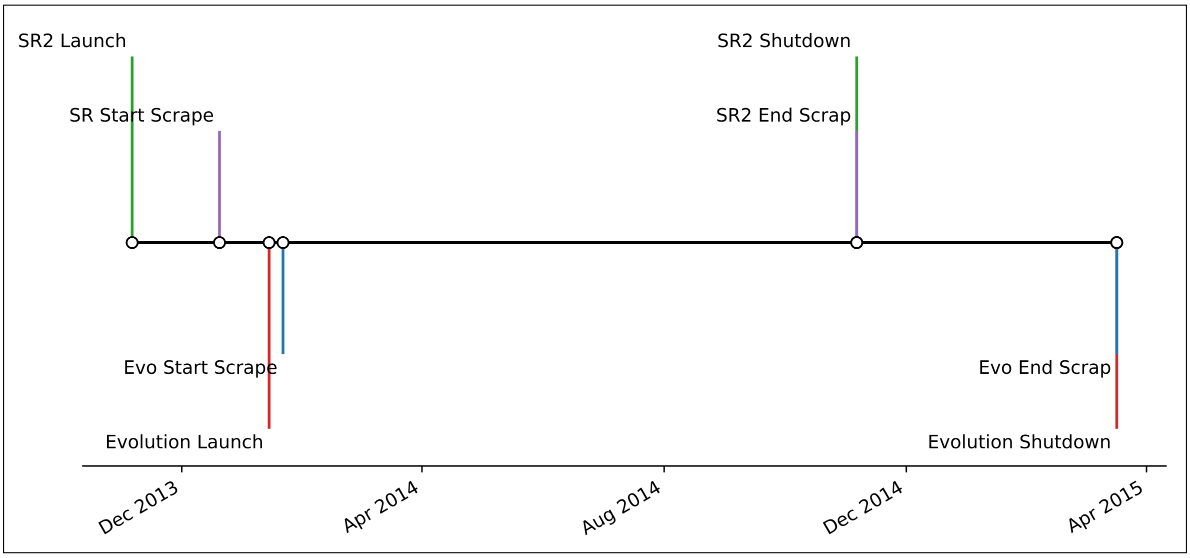


Figure 3.1: Time line of when Evolution and SilkRoad are active and when the data for both market are scraped

Both raw data sets contain a directory of folders, where each folder name is the date it was scraped. In each date, it consists different folders such as `profile` which contains the HTML files of vendors on the marketplace, `items` which contains the HTML files for all items that was listed on the marketplace etc. The layout of each raw data set are different, and was studied before implementing the information scraper.

3.2 Information Scraper: Collecting the Attributes

For each raw data set, an information scraper needs to be implemented to collect different attributes, which will then be used in the project (or other attributes that could potentially be useful in the future). For each account in the raw data set, the information scraper would create a data point that contains scraped attributes for the corresponding account and add it to the new data set *D*. For example, following is a data point that was returned by the information scraper for Evolution. The details of each attribute will be explained in the following sections.

$$\text{Evolution Data Point} \begin{cases} \text{Username} \\ \text{Date (the date the HTML files related this account was created)} \\ \text{PGP keys} \\ \text{Item names} \\ \text{Item description} \\ \text{Ship From} \\ \text{Ship to} \\ \text{Market} \end{cases}$$

After scraping the attributes from the raw data sets for Evolution and Silk Road 2, we would have two data sets. We denote the new data set for Evolution and Silk Road 2 as *D_evo* and *D_sr*2 respectively.

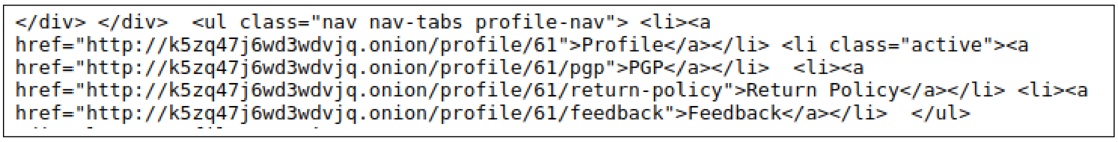
3.2.1 Before Implementing Scraper

Before implementing an information scraper for each marketplace, the HTML files in each were observed and studied. Since every marketplace has a different layout, hence different marketplaces could contain mutual and exclusive information. This means each raw data set needs it's own information scraper. It is important to fully understand how a marketplace is laid out, their nature and the available information that could be collected.

For example, Evolution and Silk Road 2 contains mutual information such as usernames, PGP keys details, item listing etc. For each account in Silk Road 2, they have a sections called "Vendor Description", where the account owner uses to advertise themselves, such as deals, packaging, shipping methods etc, while accounts in Evolution does not have such section.

3.2.2 Deciding Unique ID

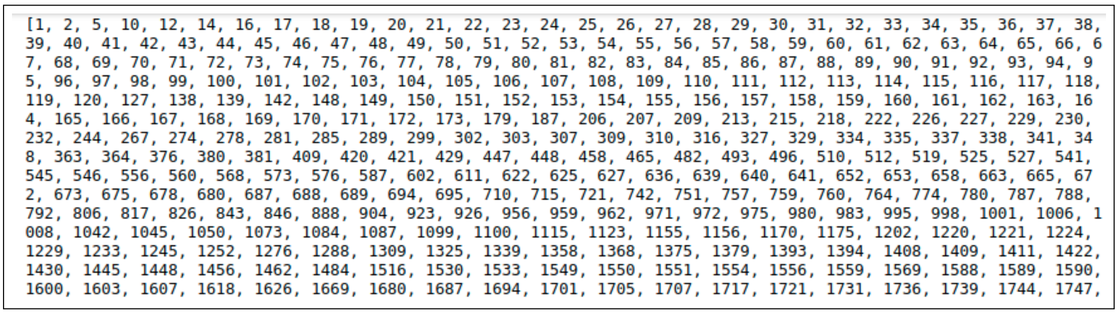
After studying what is displayed for each account's HTML file, a variable should be decided and used as a unique ID for each account, such that it is possible to differentiate itself from other accounts. This is also used later to reduce the size of data sets *D_evo* and *D_sr*2.



```
</div> </div>  <ul class="nav nav-tabs profile-nav"> <li><a
href="http://k5zq47j6wd3wdvjq.onion/profile/61">Profile</a></li> <li class="active"><a
href="http://k5zq47j6wd3wdvjq.onion/profile/61/pgp">PGP</a></li>  <li><a
href="http://k5zq47j6wd3wdvjq.onion/profile/61/return-policy">Return Policy</a></li> <li><a
href="http://k5zq47j6wd3wdvjq.onion/profile/61/feedback">Feedback</a></li>  </ul>
```

Figure 3.2: Hyperlinks found in a HTML file for an account from Evolution

For Evolution, an ID number in each account was used as the unique ID. When observing a HTML file of an account for Evolution, multiple hyperlinks was found, which can be seen in Figure 3.2. As can be observed, all the hyperlinks includes the number 61. A hypothesis was made: such number in the hyperlink for each account is a unique number assigned by the marketplace. For clarity, we refer to these numbers as "IDs". This was verified by first selecting a random date in the raw data set of Evolution, then collect all the IDs in the hyperlink for each account and put them in list L. If we convert L to a set S, the number of elements in L equals to the number of elements in S, indicating there are no duplicating IDs. If we sort the IDs in incrementing order, we would get the results in Figure 3.3. By observation, the smallest ID number starts from 1 and increments inconsistently.



```
[1, 2, 5, 10, 12, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38,
39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 6
7, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 9
5, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118,
119, 120, 127, 138, 139, 142, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 16
4, 165, 166, 167, 168, 169, 170, 171, 172, 173, 179, 187, 206, 207, 209, 213, 215, 218, 222, 226, 227, 229, 230,
232, 244, 267, 274, 278, 281, 285, 289, 299, 302, 303, 307, 309, 310, 316, 327, 329, 334, 335, 337, 338, 341, 34
8, 363, 364, 376, 380, 381, 409, 420, 421, 429, 447, 448, 458, 465, 482, 493, 496, 510, 512, 519, 525, 527, 541,
545, 546, 556, 560, 568, 573, 576, 587, 602, 611, 622, 625, 627, 636, 639, 640, 641, 652, 653, 658, 663, 665, 67
2, 673, 675, 678, 680, 687, 688, 689, 694, 695, 710, 715, 721, 742, 751, 757, 759, 760, 764, 774, 780, 787, 788,
792, 806, 817, 826, 843, 846, 888, 904, 923, 926, 956, 959, 962, 971, 972, 975, 980, 983, 995, 998, 1001, 1006, 1
008, 1042, 1045, 1050, 1073, 1084, 1087, 1099, 1100, 1115, 1123, 1155, 1156, 1170, 1175, 1202, 1220, 1221, 1224,
1229, 1233, 1245, 1252, 1276, 1288, 1309, 1325, 1339, 1358, 1368, 1375, 1379, 1393, 1394, 1408, 1409, 1411, 1422,
1430, 1445, 1448, 1456, 1462, 1484, 1516, 1530, 1533, 1549, 1550, 1551, 1554, 1556, 1559, 1569, 1588, 1589, 1590,
1600, 1603, 1607, 1618, 1626, 1669, 1680, 1687, 1694, 1701, 1705, 1707, 1717, 1721, 1731, 1736, 1739, 1744, 1747,
```

Figure 3.3: Ordering the IDs in increment order

We created a list L_d that consists of the difference between two consecutive IDs that is greater than 1, then we counted the total number of time each number appeared in the list. For example, the difference between "2" and "5" is 3, we found 3 appeared 139 times in L_d. Then we were able to obtain plot in Figure 3.4.

From Figure 3.4, we could observe the graph is skewed to the left, meaning the "gap" between two ID numbers are mostly low. This indicates the ID numbers are not generated randomly, but most likely to be assigned to new accounts incrementing order. For the accounts with the missing ID number, they are most likely banned or removed by the marketplace for violating the market rules. As discussed in Section 1.3.5, many marketplace has their own set of rules to avoid scammers and keeping a low profile to not attract law enforcement's attention. For example, many forbidden providing services that harms other individuals, selling mass destruction weapons, selling goods related to CEM etc.

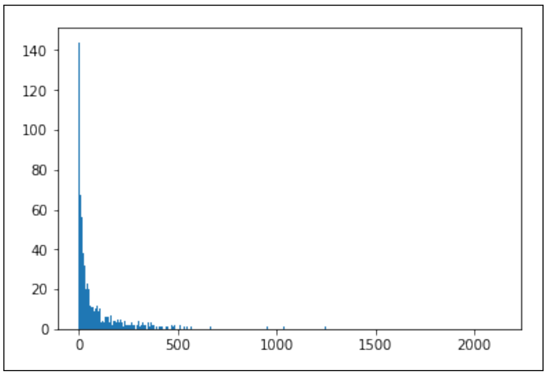


Figure 3.4: Counting the differences between IDs

Therefore, we treat the numbers in the hyperlinks for each account as a unique ID that was assigned by the marketplace.

For Silk Road 2, the username was used as the unique ID. From assumption $\mathcal{A}1$, we assumed that there does not exist any scammer in the marketplace, meaning there won't be accounts that pretends to be other vendors on the marketplace. From assumption and $\mathcal{A}6$, we assumed that vendors will intentionally use different usernames, indicating they will be unique and different. Therefore by the two assumptions, we could then assume usernames are unique and could be used as a unique ID.

3.2.3 Common Attributes in a Data Point

For each scraper, it should collect the following attributes from the raw data set. Following are the common attributes for each data point should contain when scraping through the HTML files of Evolution and Silk Road 2:

- Username: Username of each account.
- Date: The date when the web-crawler scraped the account's HTML file.
- PGP keys: As discussed in Section 1.3.2, PGP encryption are used such that vendors and buyers could communicate without revealing their true identity. It has been observed that even if a marketplace offers a specific section to put the PGP key, vendors might ignore it and put it in the item description. Therefore, when scraping the PGP key, we also check if the PGP key is in the item description.
- Item names: The names of listed items made by this account.
- Item Descriptions: The item descriptions for each listed items.
- Ship from: Where the vendors ship their items from.

- Ship to: Locations where the vendors are able to ship to.
- Market: The marketplace where this data point was created from.

3.2.4 Uncommon Attributes in a Data Point

As discussed before, due to the different layout of each marketplace, different information could be displayed. A information scraper should also collect attributes that could also be used in the future. Following attributes are exclusive to a data point created from Evolution:

- ID: The number in the hyperlinks in the account's HTML file.

Following attributes are exclusive to a data point created from Silk Road 2:

- Vendor description: For accounts in SilkRoad 2, there is a section in the vendor′s profile allowing vendors to introduce themselves, such as what deals they offer, the range of locations they ship to etc. Do note vendors also uses this attribute to display their PGP keys.

3.3 Data Integration

After scraping through the raw data set for Evolution and Silk Road 2, we have data set *D_evo* and *D_sr*2. It is hard to access certain attributes in the two data sets. For example, if we were to look for all the PGP key used by a username in *D_evo*, then we need to go through all the data points in *D_evo*. This becomes time consuming and expensive to compute. Therefore we could reduce the two data sets by integrating data points with the same unique ID separately, and obtain the reduced data set *DR_evo* for Evolution and *DR_sr*2 for Silk Road 2. The data points in the reduced data sets are more compact and the attributes will be summarized and easier to access.

After integrating the data points in both data set *D_evo* and *D_sr*2, both data sets' data point contains following common attribute:

- `pgp_history`: A list of PGP keys associated with the list of dates when it was used. From this attribute, we could obtain information such as how often and when does the vendor changes their PGP key.
- `listing_history`: A list of item names that was listed by the vendor, associated with all versions of description for that item and the dates when the item was listed. From this attribute we could obtain multiple information, such as how long will an item be listed, how often does the description of an item change etc.
- `ship_from`: A list of countries where the vendors postage from.
- `ship_to`: A list of countries that the vendor can postage to.

- `index`: The location of a data point in a data set.
- `market`: Which market this data point was created from.

The following attributes are exclusive to the data points in *DR_evo*:

- `ID`: The ID number for the data point, which was used as unique ID.
- `Uname_history`: A list of usernames that has been used by the account associated with a list of dates when it was used. From this attribute we could observe how often and when a username is modified.

The following attributes are exclusive to the data points in *DR_sr*2:

- `username`: The username of the account, which is used as unique ID.
- `vendor_description`: A list of vendor descriptions that belongs to the same username.
- `dates`: A list of dates that the account has been observed by the web-crawler. Note that the integrated data points in Evolution also contains such information, except it was combined with the usernames.

In this thesis, not all the attributes were used. The common attribute `pgp_history`, `index` and `market` were used. From *DR_evo*, attributes `ID` and `Uname_history` were used. From *DR_sr*2, attributes `username`, `vendor_description` and `dates` were used. As these attributes were enough to verify the hypothesis.

3.4 Summary

This chapter introduced the two raw data sets that was used for this paper: Evolution and Silk Road 2. Then introduced the two core concepts before implementing an information scraper, which is studying the HTML file to find the appropriate attributes to collect, and deciding the unique ID when constructing a data point. Finally, this chapter introduced the attributes of integrated data points for each data set, which will be used to correlate accounts in Chapter 4 and further analyzed in Chapter 5.

Chapter 4

Methodology

This chapter introduces the methods that was used to verify the hypothesis stated in the thesis statement 1.7. Two problems needed to be solved in order to verify the hypothesis. The first problem is the vendor correlation problem, which is how to classify if two accounts belongs to the same vendor. The second problem is how to compute the username similarity when given two usernames. Finally in this chapter, we introduce the procedure to verify the hypothesis after solving the two problems.

Section 4.1 Problem Statement: Detail description of the two problems that is needed to be solved in order to verify the hypothesis.
Section 4.2 PGP Correlation: Introduced the method that was used to correlated accounts.
Section 4.3 Username Feature Vector: Introduced how a username feature vector is created for the username of an account.
Section 4.4 Procedure to Verify Hypothesis : Introduced the procedure of verifying the hypothesis.
Section 4.5 Summary: The summary of the chapter.

4.1 Problem Statement

In Section 1.7, we stated the hypothesis: *Accounts that belong to the same individual are likely to have similar usernames, which are being used as a "Brand" by the vendor.* To verify this hypothesis, two problems needed to be solved: the vendor correlation problem and the username similarity problem.

Vendor Correlation Problem Statement

Given two accounts acc_1 and acc_2, where each account has n attributes associated with them. We want to determine if acc_1 and acc_2 belongs to the same individual by using some or all of the attributes.

$$Corr(acc_1, acc_2) = \begin{cases} 1 & acc_1 \text{ and } acc_2 \text{ belongs to the same individual} \\ 0 & \text{otherwise} \end{cases}$$

We propose a new method called PGP correlation to address this problem. Details of PGP correlation is discussed below in Section 4.2

Username Similarity Problem

Given two usernames u_1 and u_2, compute the similarity of the two usernames. We treat two usernames as the same if there similarity score is above a certain threshold γ.

To do this, we constructed a username feature vector for u_1 and u_2, then computed the similar similarity of the two feature vector using cosine similarity. The details of this method is discussed below in Section 4.3

4.2 PGP Correlation

For this method, we would be using the data set *DR_evo* and *DR_sr*2 from Chapter 3, where each data point in both data sets represents an account, associated with other attributes. In this method, we focus on three key attributes that are, "pgp_history", "market" and "index". Note that the "pgp_history" records all the used PGP keys of an account every time it has been seen by the information scraper. "Market" refers to which data set the data point is from. "Index" refers to the location of a data point inside a data set.

Method PGP correlation determines if two accounts acc_1 and acc_2 belongs to the same individual based on the PGP keys used by acc_1 and acc_2. Let's denote the PGP keys used by acc_1 and acc_2 respectively as PGP_1 and PGP_2. If any PGP keys in PGP_1 and PGP_2 matches, then we consider the the two accounts acc_1 and acc_2 belongs to the same individual and returns a tuple, which contains information which data set the accounts came from and their location in the data set. Hence we could get:

$$Corr(acc_1, acc_2) = \begin{cases} (m1, m2, r1, r2) & \exists p_1 \in PGP_1, \exists p_2 \in PGP_2 \text{ such that } p_1 \equiv p_2 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Where $m1$ and $m2$ indicates which data set acc_1 and acc_2 came from. $r1$ and $r2$ refers to the location (row index) of acc_1 and acc_2 in the data set of $m1$ and $m2$ respectively.

Note that we don't combine *DR_evo* and *DR_sr*2 at any time, as the structure of both data sets are slightly different and it will be easier to access a specific row of a specific data set.

4.3 Username Feature Vector

Methods to measure username similarity has been explored by Wang et al. [2016]. They created multiple feature vectors by using the username, then combined the feature vectors together to create the username feature vector [Wang et al., 2016].

This method was inspired by Wang et al. Similar to their work, we constructed our username feature vector (F_vec) with two feature vectors: gram frequency-inverse username frequency (GF-IUF) feature vector (G_vec) and basic feature vector (B_vec) [Wang et al., 2016].

4.3.1 Reusing TF-IDF: GF-IUF

Gram frequency-inverse username frequency is conceptually the same as TF-IDF [Manning et al., 2008]. The formula for TF-IDF is shown below:

$$tf\text{-}idf(t,d,D) = tf(t,d) \cdot idf(t,D) = f_{t,d} \cdot log\frac{\mid D \mid}{\mid \{d \in D : t \in d\} \mid} \tag{4.2}$$

Where t is a term, d is a document and D is the corpus. $tf(t,d)$ counts the number of times term t appears in document d. The value $idf(t,D)$ returns indicates if the term is common or rare among all the documents in D. For example, if $idf(t,D)$ returns a high value around 1, then this indicates the term is very rare in the corpus. If the value returned is close to 0, then this means the term is very common. Hence the value returned by $tf\text{-}idf(t,d,D)$ indicates how important the term is t is to the document d with respect to the corpus D.

With respect to the equation 4.2, instead of using terms t, document d and corpus D, we used grams g, username u and all usernames U from both data set DR_evo and DR_sr2.

$$gf\text{-}iuf(g,u,U) = tf(g,u) \cdot idf(g,U) \tag{4.3}$$

4.3.2 GF-IUF feature vector

Preprocessing Usernames

Let U denote all usernames in all data sets. Before constructing the GF-IUF feature vector for each username u in U, we preprocess each username u, to obtain U_p.

The preprocessing was done by simple removing all the usernames that is not a letter or digit, then converting all uppercase to lowercase. On the clear net, most websites' usernames are not case sensitive and limit the use of special characters such as "$", "%" etc. In fact, the characters that were used in usernames and are not in the alphabet or digit are "-" and "_". The two characters often acts as delimiters in the username. [Wang et al., 2016].

Constructing GF-IUF feature vector

Given a username u, we preprocess it and obtain u_p. First we create a list that contains all possible grams of size n, with respect to the string that contains all the alphabet letter and digits, i.e "abcefghijklmnopqrstuvwxyz0123456789". In this thesis, with an n value greater than 2 will create too many grams and the GF-IUF feature vector will be too large. Therefore we have chosen $n = 2$. The Cartesian Product of the string was used to create the list of all possible 2-grams. Which is demonstrated in Python below:

```
from itertools import product
lis = product("abcefghijklmnopqrstuvwxyz0123456789", repeat=2)
lis = [tup[0] + tup[1] for tup in lis]
# lis = ['aa', 'ab', 'ac', ..., "98", "99"]
```

Therefore, with respect to the list "lis" in code above, we construct the GF-IUF feature vector G_vec_u for preprocessed username u_p:

$$G\_vec_u = \langle gf\text{-}iuf(aa, u_p, U_p),\ gf\text{-}iuf(ab, u_p, U_p),\ \dots\ ,\ gf\text{-}iuf(98, u_p, U_p),\ gf\text{-}iuf(99, u_p, U_p)\rangle$$

4.3.3 Basic Feature Vector

As name suggests, the basic feature vectors composes of multiple basic features. When constructing this feature vector, we do not preprocess the usernames.

For each username u in U we return a vector B, where each entry of B corresponds to the following features:

$$B = \langle length,\ letter,\ dig,\ sum\_dig,\ special,\ upper,\ lower\rangle$$

Where:

length : The length of username u.

letter : The total number of letters in the username u.

dig : Total number of digits in username u.

sum_dig : The sum of digits in username u. This was used because if the username included some numbers, regardless if it was leet encoding or the numbers that represents a date, the higher the sum of digits of a username is, the more likely it will be unique, which can be observed in Figure 4.1 below. The larger the sum of digit is, the more unique it is, making it stand out compared to the total sum of other digits that has a low value.

special : Total number of special characters used in the username u.

upper : Total number of uppercase used in username u.

lower : Total number of lowercase used in username u.

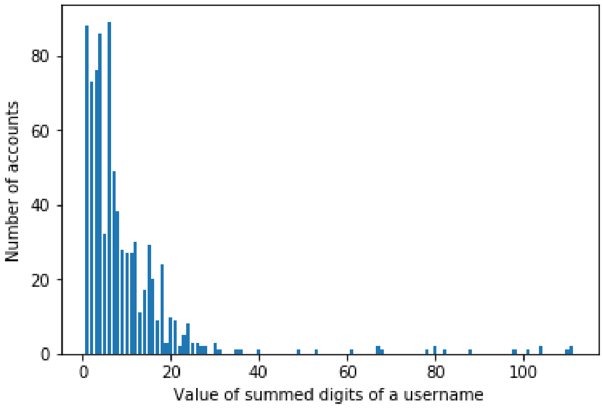


Figure 4.1: Number of usernames with different sum of digits in username (number of accounts with sum of 0 not included)

After creating the B vector for each username, we want to normalize each element in vector B to the range between 0 to 1. To do this, we find the max and minimum value for all the features above and get: $length_{max}$, $length_{min}$, $letter_{max}$, $letter_{min}$, dig_{max}, dig_{min}, sum_dig_{max}, sum_dig_{min}, $special_{max}$, $special_{min}$, $upper_{max}$, $upper_{min}$, $lower_{max}$ and $lower_{min}$.

Then we are able to construct a basic vector B_vec_u for username u in U by normalizing each element:

$$B\_vec_u = \langle \frac{length - length_{min}}{length_{max} - length_{min}}, \frac{letter - letter_{min}}{letter_{max} - letter_{min}}, \ldots \frac{lower - lower_{min}}{lower_{max} - lower_{min}}, \rangle$$

4.3.4 Updating the data points

For username u in U, after constructing the GF-IUF feature vector G_vec_u and basic vector B_vec_u, we could construct the username feature vector F_vec_u by concatenating vector G_vec_u and B_vec_u:

$$F\_vec_u = G\_vec_u + B\_vec_u$$

Note that for username u in U, we know which data point (account) u is from and where the data point is located. Hence we could update the data point which username u is from. Lets assume that u is from data point acc. Then we update the data point simply by creating a new attribute, which holds the username feature vector F_vec_u:

$$acc \begin{cases} Username \\ UsernameFeaturevector = F\_vec_u \\ Dates \\ PGPKeys \\ \vdots \end{cases}$$

We then update all the data points with their corresponding username feature vector.

4.3.5 Computing Similarity

After updating all the accounts by adding their corresponding feature vector, we are able to compute the username similarity when given two data points (accounts).

Given two data points acc_1 and acc_2, we are able to extract their usernames feature vectors F_vec_1 and F_vec_2. To measure the username similarity of acc_1 and acc_2, we simply compute the cosine similarity of username feature vectors F_vec_1 and F_vec_2:

$$Sim(acc_1, acc_2) = \frac{F\_vec_1 \cdot F\_vec_2}{\|F\_vec_1\| \times \|F\_vec_2\|} \tag{4.4}$$

4.4 Procedure to Verify Hypothesis

To verify the hypothesis, we first find the correlated accounts using PGP correlation from Section 4.2. Then for each matched accounts, we compute similarity score of the matched accounts' username. Finally we find the percentage of matched accounts that has a username similarity score greater or equal to a defined threshold γ.

Before doing any correlation, we would need to construct username feature vectors for all usernames in all data sets. This is done by using methods in Section 4.3.

From Section 4.2, by using the PGP correlation method, we are able to get a list of matched accounts $Match\_Accs$. Assume that we found n matches, then we have:

$$Match\_Accs = [(m_n,\ m'_n,\ r_n,\ r'_n) \mid m_n \neq m'_n, \quad r_n \neq r'_n, \quad n \in [1, \ldots, n]]$$

Where $Match\_Accs$ is a list of n tuples, where each tuple is of the form $(m,\ m',\ r,\ r')$. m and m' indicates which marketplace data set account acc_1 and acc_2 belongs to. r and r' indicates where acc_1 and acc_2 is located in marketplace data set m and m'. Note that we filter out tuples where $m = m'$ and $n = n'$, meaning it points to the same data point in marketplace m.

Succinctly, we could use m and r to extract the corresponding account acc_1 (data point) from data set m. This also applies for the second account acc_2 using m' and and r'. By doing so, we could update the tuples in $Match\_Accs$ by replacing m and r

with acc_1, m' and r' with acc_2. Therefore we could get:

$$Match\_Accs = [(acc_{1n},\ acc_{2n}) \mid acc_{1n} \neq acc_{2n}, \quad n \in [1, \ldots, n]]$$

By using equation (4.4), we could calculate the similarity score for each tuple in *Match_Accs*. Then we could update each tuples in *Match_Accs*, by simply adding the username similarity score to the tuple. Hence we get:

$$Match\_Accs = [(acc_{1n},\ acc_{2n},\ Sim(acc_{1n},\ acc_{2n})) \mid acc_{1n} \neq acc_{2n}, \quad n \in [1, \ldots, n]]$$

Following is the equation to calculate the percentage of matched accounts that has a username similarity score higher or greater than a defined threshold γ, where $\gamma \in [0, 1]$:

$$\text{percentage} = \frac{\mid (acc_{1n},\ acc_{2n},\ Sim(acc_{1n},\ acc_{2n})) \mid Sim(acc_{1n},\ acc_{2n}) \geq \gamma \mid}{\mid Match\_Accs \mid}$$

We decide if the hypothesis holds true based on the value of *percentage*.

4.5 Summary

In this chapter we addressed the procedure how to verify the hypothesis from Section 1.7. To verify the hypothesis, we first need to classify if two accounts belongs to the same vendor by using the PGP correlation method. After correlating two accounts, we introduced the methods to construct username similarity for each username, which is used to measure the username similarity of the two accounts. Lastly, by using the two methods, we are able to compute the username similarity of two correlated accounts, which then later used to calculate the percentage of correlated accounts with a username similarity score greater or equal to a defined threshold γ.

In the next chapter, we conducted three experiments to verify the hypothesis. Firstly we correlated the accounts within Evolution, then calculated the percentage of correlated accounts that has a username similarity score above multiple different threshold. Similar experiment was done for Silk Road 2. Last experiment was conducted by correlating the accounts of Evolution and Silk Road 2 then calculating the percentage.

Chapter 5

Vendor Behaviors and Observation

This chapter conducts experiments to find behaviors of darknet marketplace vendors. Four behaviors was concluded from the experiments, verified and strengthened our hypothesis from Section 1.7.

Section 5.1 Before Experiment: States the details of what needs to be done before conducting any experiment.
Section 5.2 Vendors use usernames as "brands": Three experiments was conducted, used the results to conclude three vendor behaviors, which is later used to verify and strengthen the hypothesis.
Section 5.3 Username Modification: An experiment conducted on data set *DR_evo* to see how often vendors change their username.
Section 5.4 Summary: The summary of this chapter.

5.1 Before Experiment

After preprocessing and integrating the raw data set of Evolution and Silk Road 2 using proposed methods from Chapter 3, we obtain *DR_evo* and *DR_sr*2. It is found that data sets *DR_sr*2, *DR_sr*2 had 4367 and 1226 unique accounts respectively.

Then we constructed the username feature vector for each account in both data set *DR_evo* and *DR_sr*2 by using method proposed in Section 4.3.

5.2 Vendors use usernames as "brands"

The motivation behind this experiment is to verify the hypothesis which was stated in Section 1.7: *Accounts that belong to the same individual are likely to have similar usernames, which are being used as a "Brand" by the vendor*.

5.2.1 Experiments

Note that it is possible for a vendor to create multiple accounts on the same marketplace, hence it will be interesting to observe if the hypothesis holds true for correlated accounts that are within the same marketplace.

By using the proposed method in Section 4.4, we conducted three experiments. For the first experiment, we correlated the accounts within Evolution (Evo & Evo). Then we calculated the percentage of correlated accounts with similarity score greater or equal to different values of γ. This experiment was repeated with matched accounts within Silk Road 2 (SR2 & SR2) and matched accounts between Evolution and Silk Road 2 (Evo & SR2).

5.2.2 Results

By correlating the accounts between Evolution and Evolution, Silk Road 2 and Silk Road 2, Evolution and Silk Road 2, we obtained the following Table 5.1:

	Evo & Evo	SR2 & SR2	Evo & SR2
Number of unique accounts	4367	1226	$5593 = 4367 + 1226$
Number of correlated accounts	123	8	358
Percentage of vendors with multiple accounts	2.81%	0.65%	6.40%

Table 5.1: Correlated vendors statistics

After computing the percentage of matched vendors from the three experiments with different thresholds γ, we got the following results presented in Table 5.2 below:

Threshold γ	Percentage for Evo & Evo	Percentage for SR2 & SR2	Percentage for Evo & SR2
1.00	6.5%	0.0%	74.58%
0.95	6.5%	0.0%	75.7%
0.90	8.13%	0.0%	78.21%
0.85	12.2%	12.5%	80.17%
0.80	15.45%	12.5%	82.4%
0.75	16.26%	12.5%	82.68%
0.70	16.26%	12.5%	84.64%
0.65	17.07%	12.5%	87.15%
0.60	17.07%	25.0%	87.43%

Table 5.2: Percentage of matched accounts with score greater or equal to different γ

5.2.3 Discussion

Behavior 1: For vendors who has two accounts on two different marketplaces, they are likely to use the exact same username.

From Table 5.1 and 5.2, we could see that 74.58% vendors who have accounts on both Silk Road and Evolution uses the exact same username. Note that the way humans judge if two usernames are the same is different to how a computer judge it.

Table 5.3 consists of some usernames of correlated accounts from Evolution and Silk Road 2, associated with their similarity score. For humans, with the knowledge that these usernames were created by the same individual, we could then judge if they are the "same" by looking at the structure of the two usernames, the semantics of letters in each username or detect abbreviations easily. For example, in the fourth row of Table 5.3, the computer gave a similarity score of 0.527, which is fairly low. For humans, based on the knowledge that we know these two usernames are created by the same individual, we could tell that the "C" in "CDreams" have a high chance of representing "california", allowing use to make the conclusion that these usernames are the "same". Another example, in the fifth row of Table 5.3, we could tell that "Chem" in "ChemBrothersAU" is short for "chemical". The "AU" in "ChemBrothersAU" probably refers to Australia, which could be disregarded. Hence we say "ChemBrothers" and "chemicalbrothers" are the same.

	Username1	Username2	Similarity Score
1	SC_Connect	socal-connect	0.640
2	SaltnPepper	salt-pepper	0.825
3	brownjames	jamesbrown	0.889
4	CDreams	californiadreams	0.527
5	ChemBrothersAU	chemicalbrothers	0.694
6	OrderOfPhoenix	orderofthephoenix	0.832

Table 5.3: Usernames that are the "same" to humans but not to computers

By judging from the similarity scores in Table 5.3, it would be fair to say that if given two correlated accounts with username similarity score greater or equal to 0.80, then for humans, we consider the username for both accounts are the "same".

Hence, with respect to Table 5.2, we could conclude that 82/68% of vendors who has accounts on Evolution and Silk Road 2, their username for both accounts will be the same. Therefore we could conclude **Behavior 1**.

Behavior 2: Vendors are highly likely to create new accounts on different marketplaces than creating another account on the same marketplace.

This behavior can be simply concluded from Table 5.1. In Evolution and Silk Road 2, only 123 and 8 vendors respectively has multiple accounts in the same marketplace. While 358 vendors has accounts in both marketplace. 358 is not a large number compared to the total number of unique accounts in both marketplaces 5593, but note that there are other darknet marketplace which exists at then same time when these data sets are collected, meaning a vendor could have an account in Evolution and Agora, but doesn't have an account in Silk Road 2. Selecting which markets to create an account is completely up to the vendor themselves, as each darknet marketplace has different policies and offers. By comparing the number of correlated vendors between Evolution and Silk Road 2 to the number of correlated vendors within Evolution and with in Silk Road 2, we could conclude **Behavior 2**.

Behavior 3: When a vendor has multiple accounts in the same marketplace, it is likely that these usernames will not be similar.

Out of the 123 correlated accounts within Evolution itself, $0.065 \cdot 123 \approx 8$ vendors decided to use the exact same username. Out of the 8 correlated accounts within Silk Road 2 none used the exact same usernames. Therefore we could conclude that if a vendor has multiple accounts in the same marketplace, it is likely that these usernames will have a low similarity score.

On the side note, many of the matched accounts from the same marketplace has usernames where one is the sub-string of the other. For example, for username "Nodnowinaus" and "nodnow", their similarity score is 0.667, but the latter username is a sub-string of the first username. Hence we added an extra step before computing the username similarity: if the shorter username is a sub-string of the longer username, then the similarity score is 1. Else continue with the original method by computing the cosine similarity of the two username features. Hence we could get:

$$Sim(acc_1, acc_2) = \begin{cases} 1 & u_1 \subseteq u_2 \vee u_2 \subseteq u_1 \\ \dfrac{F\_vec_1 \cdot F\_vec_2}{\|F\_vec_1\| \times \|F\_vec_2\|} & \text{otherwise} \end{cases}$$

Where u_1, u_2 are the usernames for account acc_1 and acc_2. By adding in this extract step, we could get Table 5.4:

Threshold γ	Percentage Evo & Evo	Percentage SR2 & SR2	Percentage Evo & SR2
1.00	13.82%	25.0%	83.24%
0.95	13.82%	25.0%	83.24%
0.90	13.82%	25.0%	83.52%
0.85	14.63%	25.0%	84.64%
0.80	16.26%	25.0%	85.75%
0.75	17.07%	25.0%	85.75%
0.70	17.07%	25.0%	86.87%
0.65	17.07%	25.0%	88.55%
0.60	17.07%	37.5%	88.83%

Table 5.4: Percentage of matched accounts with score greater or equal to different γ with extra step to calculate username similarity

Though the percentage was increased in each column, but the number of vendors with multiple accounts in the same marketplace with similarity score of 1 is still low. To be specific, $0.1382 \cdot 123 \approx 17$ vendors from Evolution and $8 \cdot 0.25 = 2$ from Silk Road 2. Therefore the concluded **Behavior 3** is still valid.

5.2.4 Hypothesis Conclusion

From **Behavior 1** concluded in Section 5.2.3, we know that if an individual has two accounts on two different marketplaces, then the usernames of the two accounts are highly likely to be the "same". Then by using **Behavior 1**, we could show our hypothesis *"accounts that belong to the same individual are likely to have similar usernames, which are being used as a "Brand" by the vendor"* holds true, but it's a weak statement. If two accounts are from the same marketplace, then by **Behavior 3**, the usernames of these two account will not be similar in any way.

From **Behavior 2**, we know that vendors are highly likely to create a new account on a different market. Then by using this behavior, we could strengthen our hypothesis by rewording it to: *Accounts that belongs to the same individual, but are on different marketplaces, are likely to have similar usernames, which are being used as a "Brand" by the vendor*.

5.3 Username Modification

With respect to data set *D_evo* (before integrating), the unique ID for each data point is the data point attribute `ID`. When integrating the data points in *D_evo* with respect to `ID`, it is possible to have a data point in *DR_evo* that has multiple usernames, meaning the vendor has changed the account's.

5.3.1 Experiment

The experiment is simply done by collecting all the data points that has more than two usernames in `Uname_history`. We extract the username and compute the cosine similarity.

5.3.2 Results

After finishing the procedure above, we got following Table 5.5:

Index	Username1	Username2	Similarity Score
1	ThunderWiz	thunderwiz	1.0
2	NOT_UKWHITE	UKWHITE	0.816
3	NOT_utopic	utopic	0.791
4	Phaethon	Addyshack	0.134
5	only	only_bak	0.707
6	Luxor	ShadyTom	0.0

Table 5.5: Accounts that used more than 1 username in Evolution

Note that for all the accounts in data set *DR_evo*, only 6 accounts modified their username, and each account has only modified once.

5.3.3 Discussion

From Table 5.1, we know that *DR_evo* has 4367 unique accounts, which means only $6 \div 4367 \approx 0.14\%$ users in Evolution changed their usernames. With respect to the timeline in Figure 3.1, we could see that the raw data set includes nearly the entire history of Evolution. We know that only 0.14% (6) vendors changed their account's username and each only once, we therefore could conclude **Behavior 4: Darknet marketplace vendors are unlikely to change their account's username.**

By observing the 6 data points, nothing was standing out or useful. For the accounts in row 2 and 3, the account owner might of been an scammer, who pretends to be someone else. This might be later reported to the marketplace operator, resulting the marketplace to add "NOT" to the front of their username.

5.4 Summary

This chapter conducted multiple experiments and was able to conclude four behaviors of vendors from darknet marketplace. By using Behavior 1, we were showed that the initial hypothesis holds true. And then we used Behavior 2 to strength our hypothesis and obtained: *Accounts that belongs to the same individual, but are on different marketplaces, are likely to have similar usernames, which are being used as a "Brand" by the vendor.*

Chapter 6

Conclusion

In this thesis, we wanted to verify if the behavioral hypothesis holds true: *Accounts that belong to the same individual are likely to have similar usernames, which are being used as a "Brand" by the vendor.*

First we introduced the two core concepts for constructing a useful data set for our purposes from the raw data of the darknet marketplace. These were *1.* Understand the structure of the data set and the available attributes so that information can be extracted, and *2.* Decide the unique ID attributes for an account that do not change over time, so that the information of a single account over the lifetime of the data set can be stored as a single data point within the data set.

Then we proposed the PGP correlation method to create our ground truth labels (which accounts belong to the same user), the method to construct username feature vectors (which determines the similarity score), and the experiment procedure to verify the hypothesis. In total four experiments was conducted and four vendor behaviors was concluded:

Behavior 1 : For vendors who has two accounts on two different marketplaces, they are likely to use the exact same username.

Behavior 2 : Vendors are highly likely to create new accounts on different marketplaces than creating another account on the same marketplace.

Behavior 3 : When a vendor has multiple accounts in the same marketplace, it is likely that these usernames will not be similar.

Behavior 4 : Darknet marketplace vendors are unlikely to change their account's username.

From the four behaviors, we are able to use **Behavior 1** to show our hypothesis holds true. However we also find that the hypothesis is not always right by **Behavior 3**, although this is the minority case according to **Behavior 2**. From these findings, we refine our hypothesis and obtain: *Accounts that belongs to the same individual, but are on different marketplaces, are likely to have similar usernames, which are being used as a "Brand" by the vendor.* Finally, **Behavior 4** shows us that the first three behaviors are useful findings, as usernames rarely get changed.

By using the refined hypothesis statement, it is now possible to predict which accounts between different marketplaces belong to the same vendor. Given a darknet marketplace data set, we would construct the username feature vector for all username and classify two accounts belongs to the same vendor if their similarity score is greater than a defined threshold γ, where in this thesis we decided upon $\gamma = 0.8$. We hope such findings will be useful for law enforcement agencies and further research upon darket marketplace behavior.

6.1 Future Work

Many of the methods proposed in this thesis could be improved. In Section 1.8, assumption $\mathcal{A}6$ and $\mathcal{A}7$ are strong assumptions and should not be used. As a result there will likely be cases where two accounts belongs to two different individual, but they have very similar usernames. Thus when predicting duplicate vendors, after we obtain a list of correlated accounts based on the stronger hypothesis, methods should be implemented to determine if the two correlated accounts belongs to different accounts based on other information in the dataset. For example we could compare the writing style of the two accounts in their item description. Such additional methods would further increase accuracy in our predictions.

From discussion in Section 1.4, we made the assumption $\mathcal{A}1$ that it is very unlikely for scammers to exists and make profit off the darknet marketplace, but it is not guaranteed. Methods should be implemented to identify scammers in the data set. Detecting scammers will be harder than correlating vendors, as scammers can always switch usernames consistently or abandon accounts when necessary. To implement such methods, we would need to put ourselves into the darknet buyers and scammers' shoes. We should investigate and summarise behaviors of buyers and scammers, such as what kind of feedback does scammers have, how long was an account active for or how does a vendor determine if a vendor is a scammer or not, and then develop a method based off these behaviors.

A big limitation for this project is that we don't have a live data base to record and update the results. Such a live data base should be constructed in a specific way, such that after scraping a static version of a darknet marketplace and formatting the newly collected raw data set, we could pass this raw data into the live data base and it will automatically correlate accounts and update information in the accounts. For example, perhaps currently account acc_1 and account acc_2 are not labeled as being owned by the same vendor as no common PGP key has been found. However after passing in today's scraped data, we find that acc_1 used a new PGP key and it is the same one that acc_2 used before, leading us to now determine that they are owned by the same vendor. Such live data bases could be used to keep track of darknet vendors movement, and could also potentially become useful evidence for law enforcement.

Appendix A

Final Project Description

The project is observe a behavior of vendors' username from darknet marketplace, and verify the hypothesis: vendors uses their username as a "brand". The hypothesis was verified by first correlating accounts across two data sets of marketplace Evolution and Silk Road 2. Then we compute the username similarity of the correlated accounts, and see the percentage of matched accounts that has a username similarity above a certain threshold. Then use the data sets to find interesting behaviors of darknet marketplace vendors.

Following are the detailed tasks:

Create data sets : By using the raw data set of Evolution and Silk Road 2 from Darknet Marketplace Archives Branwen et al. [2015], create the two new data sets which contains the attributes collected from the raw data sets.

Vendor Correlation : Implement correlation methods using the attributes from the two new data sets, which returns all the matched vendors in two marketplaces

Username feature vectors : Implement and create feature vectors for usernames, which will be used to compute username similarity.

Verify hypothesis : Using the found correlated accounts, compute the similarity between the usernames of the correlated accounts. Then see how many correlated accounts has a username similarity above certain threshold. For example, if we consider usernames with a similarity above 0.85 as the same, then what's the percentage of correlated accounts that has a similarity score above 0.85.

Finding other behaviours of Vendors : The verified hypothesis is a behavior of darknet marketplace vendors. By using the two data sets, find and conclude other interesting vendor behaviors, such as vendor's movement when a marketplace shuts down etc.

Appendix B

Study Contract

Australian National University

INDEPENDENT STUDY CONTRACT SPECIAL TOPICS

Note: Enrolment is subject to approval by the course convenor

SECTION A (Students and Supervisors)

UniID: u6049249

SURNAME: SHAN________________ FIRST NAMES: Sylvester_______________

TOPIC SUPERVISOR (*may be external*): Prof Ramesh Sankaranarayana

FORMAL SUPERVISOR (*if different, must be an RSSCS academic*): Prof Weifa Liang

COURSE CODE, TITLE AND UNITS: COMP4560 Advanced Computing Project

SEMESTER ☒ S1 ☒ S2 YEAR: 2019 & 2020

TOPIC TITLE:

Vendor Correlation & Understanding Colloquialisms on Darknet Marketplaces

LEARNING OBJECTIVES:

1. Gain knowledge in developing methods to solve real world problem with real world data.
2. Identify potential points of correlation between users across multiple darknet marketplaces
3. Understand and enhance data collection & analysis methods in the field of darknet markets.
4. Design & implement practices for classifying drug products based on known & unknown colloquialisms

DESCRIPTION:

The ANU Cybercrime Observatory collects data on the behaviours of buying & selling illicit products on darknet marketplaces. While hypothesised that vendors may operate over multiple marketplaces, it has been difficult to confirm this using existing practices. This project aims to research and develop a methodology to determine to a degree of confidence if a vendor from one darknet market is the same on other darknet marketplaces based on natural language analysis techniques.

The challenge of this research project will be that the dataset is not well formed, and no training set can be provided. As each darknet marketplace is unique, the available data regarding a vendor will differ from each website. Different methods will be developed and be used to solve such problem using the provided dataset, such as the similarity between vendors name across different darknet market, the similarity of grammar errors in the description of the drug and the similarity of the products different vendor sells across different market.

As this problem is a natural language problem, it is possible to also further understand the use of colloquialisms on darknet marketplaces to describe drug products. This will involve classifying the existing products into abstract clusters based on the type of drug it may be and perform predictions on unknown colloquialisms based on other product data. A tool such as this not only benefits the understanding of a vendor's breadth and depth of the understanding of the language, but assists in all data analysis methods pertaining to the use of data relating to drug products sold.

Summary: Develop methods according to existing techniques to correlate vendors across different darknet marketplaces using data provided by ANU Cybercrime Observatory.

Australian National University

ASSESSMENT (evaluated by the Topic Supervisor, unless stated otherwise here)

Assessed project components:	% of mark	Due date	Evaluated by
Report	60		
Artefact	30		
Presentation	10		

MEETING DATES (IF KNOWN):

STUDENT DECLARATION: I agree to fulfil the above defined contract:

Sylvester ……………………………………………………..   16/8/2019 ………………………..

Signature   Date

SECTION B (Supervisor):

I am willing to supervise and support this proposal. I have checked the student's academic record and believe the student can fulfil this contract. If I have nominated an examiner above, I have obtained their consent (via signature below or attached email)

……………………………………………………..   16/08/2019 ………………………..

Signature   Date

Examiner:
Name: Eric McCreath ………………………………………   Signature ……………………

REQUIRED DEPARTMENT RESOURCES:

SECTION C (Course convenor approval)

……………………………………………………..   ………………………..

Signature   Date

Appendix C

Description of Software

With respect to the artefact layout, the folder name "4-Code" contains all the code that was used for this project. All the code in "4-Code" were written by the student from scratch. (Except for imported modules)

It is hard to test the correctness of the entire code, as the code was implemented to verify a hypothesis and generate statistics that was later analyzed. No testing programs is available to test all the code, but testing code included between cells or commented out. One or multiple small test code is commented at the end of the cell for many functions. Majority functions are all well documented, where each function indicates what data type it should take in etc.

Following are the hardware that was used:

- CPU: amd ryzen 9 3900x
- GPU: ASUS ROG Strix GeForce RTX 2080 Ti OC 11GB
- RAM: 32 GB
- OS: Ubuntu 18.04.4 LTS 64-bit

Following are the software and important packages that were used:

- Python 3.7.7
- Pandas 0.24.2
- Pool from multiprocessing
- tqdm 4.31.1
- pandarallel 1.4.6
- numpy 1.16.2
- conda 4.8.3

Data set used from website https://www.gwern.net/index, use the following command to download the two data sets. If the following command does not work, please refer to the website above

- Evolution: rsync –verbose rsync://78.46.86.149:873/dnmarchives/evolution.tar.xz ./
- Silk Road 2: rsync –verbose rsync://78.46.86.149:873/dnmarchives/silkroad2.tar.xz ./

Appendix D

Readme

D.1 Download the data sets

The data sets that was used in this project were from https://www.gwern.net/index.

Use the following command in terminal to download the data sets (3.6GB each, might take some time):

Evolution rsync –verbose rsync://78.46.86.149:873/dnmarchives/evolution.tar.xz ./
SilkRoad2 rsync –verbose rsync://78.46.86.149:873/dnmarchives/silkroad2.tar.xz ./

After downloading the two data sets, extract them to desired location.

D.2 Files to modify:

1. Inside folder 4-Code, open the file paths_variables.py, follow the instructions and modify the paths and some variables.
2. Inside folder 4-Code/1_extract_DNM, open file paths_variables_old.py. Follow the instructions, they are the same variable you used in paths_variables.py.

D.3 Running the progrmas

1. Open folder 4-Code/1_extract_DNM, run all the cells in evolution_part1.ipynb, evolution_part2.ipynb and evolution_part3.ipynb one after the other. This creates the dataframe that we will be used later and is stored in the "storge_path" location.

2. Open folder 4-Code/1_extract_DNM, run all the cells in silkroad2.ipynb.

   This creates the dataframe for silkroad 2.

3. Open folder 4-Code/2_text_and_writing_style_Extraction. Run all the file 1_writing_style_extraction.ipynb. The content of this file was not included in the thesis. I think it changes the data structure of dataframes which i encounter for later on, but not sure. Running it just in case.

4. Open folder 4-Code/3_username_feature: run the cells of 1_username_feature.ipynb until you see a block of hastags #. That's when to stop. The following functions are included in a different python file "classes_functions.py".

5. Open folder 4-Code/4_Classification: run all the cells of file 1_pgp_classification.ipynb. This files classifies the accounts that correlates depending on the PGP.

6. Open folder 4-Code/5_analysis: run all the cells of file 1_pgp_analysis.ipynb. This file conducts all the three experiments in the thesis which is used to verify the hypothesis. The results are printed out.

D.3.0.1 Optional

7. Open folder 4-Code/6_social_networka: codes in this constructs social network graphs using the match tuples. It's still in experiment stage. Though it does return some grpahs, but they are not interesting graphs.

8. Open 4-Code/7_Verifying_facts: verifying_facts.ipynb, this file contains blocks of code used to create plot graph in thesis and some experiments.

Appendix E

Assumptions

The project makes the following assumptions.

$\mathcal{A}1$: Assume there are no scammers in the data set. (1.4)

$\mathcal{A}2$: Assume vendors have different accounts across different marketplaces. (1.8)

$\mathcal{A}3$: Before any evidence, assume each account in all marketplaces belongs to a unique vendor. (1.8)

$\mathcal{A}4$: Assume vendors operates as individuals on marketplaces. (1.8)

$\mathcal{A}5$: If two PGP keys matches, then they belong to the same individual. (1.8)

$\mathcal{A}6$: Vendors are likely to use usernames that differs to other usernames. (1.8)

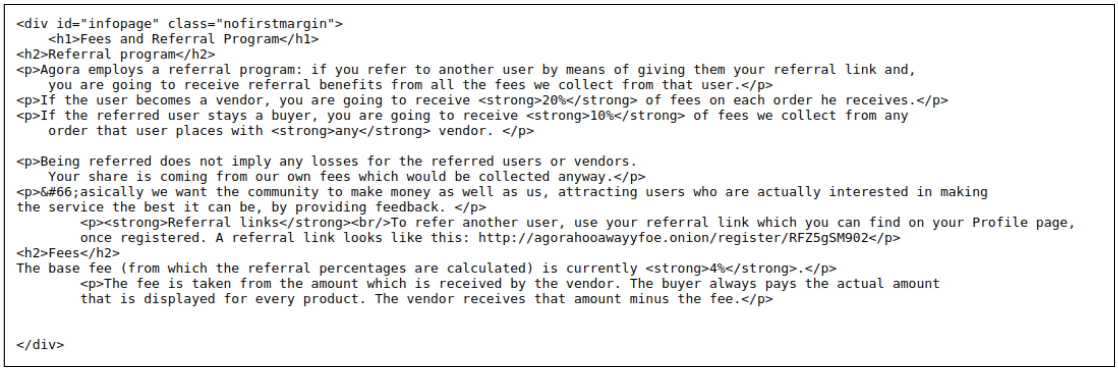
$\mathcal{A}7$: Assume usernames are unique in the same marketplace. (1.8)

Appendix F

Snippets of HTML files

to simplify the problem. This is due to the problem This appendix includes snippets of HTML files that was scrapped from different darknet marketplaces. For readability, the corresponding text in the HTML file is shown below each image.

More details or delete this line?



```
<div id="infopage" class="nofirstmargin">
    <h1>Fees and Referral Program</h1>
<h2>Referral program</h2>
<p>Agora employs a referral program: if you refer to another user by means of giving them your referral link and,
    you are going to receive referral benefits from all the fees we collect from that user.</p>
<p>If the user becomes a vendor, you are going to receive <strong>20%</strong> of fees on each order he receives.</p>
<p>If the referred user stays a buyer, you are going to receive <strong>10%</strong> of fees we collect from any
    order that user places with <strong>any</strong> vendor. </p>

<p>Being referred does not imply any losses for the referred users or vendors.
    Your share is coming from our own fees which would be collected anyway.</p>
<p>&#66;asically we want the community to make money as well as us, attracting users who are actually interested in making
the service the best it can be, by providing feedback. </p>
        <p><strong>Referral links</strong><br/>To refer another user, use your referral link which you can find on your Profile page,
        once registered. A referral link looks like this: http://agorahooawayyfoe.onion/register/RFZ5gSM902</p>
<h2>Fees</h2>
The base fee (from which the referral percentages are calculated) is currently <strong>4%</strong>.</p>
        <p>The fee is taken from the amount which is received by the vendor. The buyer always pays the actual amount
        that is displayed for every product. The vendor receives that amount minus the fee.</p>


</div>
```

Figure F.1: Agora’s referral information and fees

make it into a code block?

Following are the text from F.1:

Referral program

Agora employs a referral program: if you refer to another user by means of giving them your referral link and, you are going to receive referral benefits from all the fees we collect from that user.

If the user becomes a vendor, you are going to receive 20% of fees on each order he receives.

If the referred user stays a buyer, you are going to receive 10% of fees we collect from any order that user places with any vendor.

Being referred does not imply any losses for the referred users or vendors. Your share is coming from our own fees which would be collected anyway.

Basically we want the community to make money as well as us, attracting users who are actually interested in making the service the best it can be, by providing feedback.

Referral links

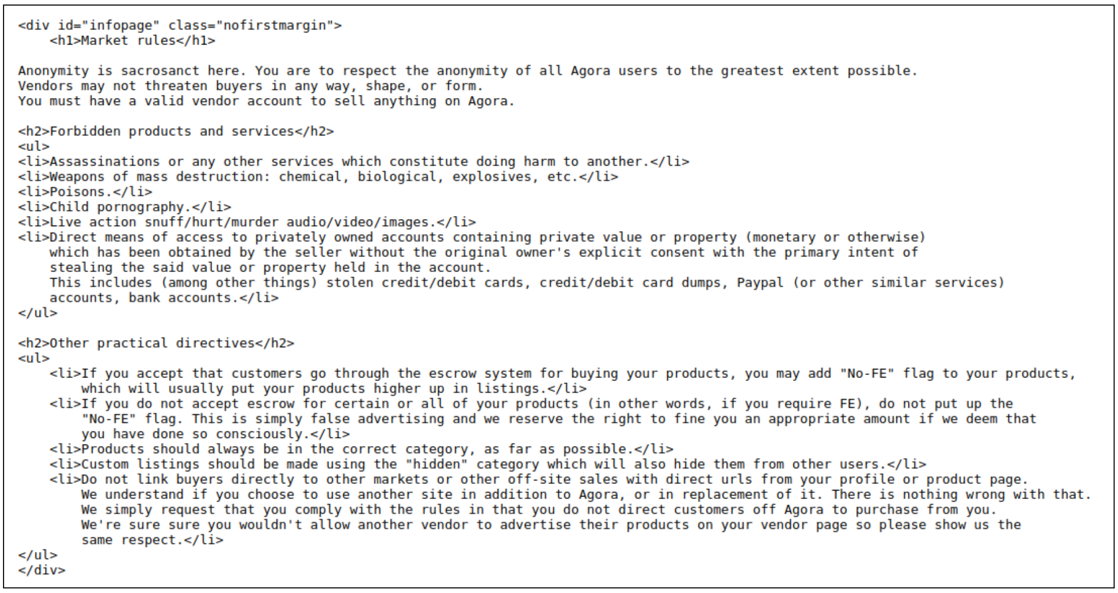
To refer another user, use your referral link which you can find on your Profile page, once registered. A referral link looks like this:

http://agorahooawayyfoe.onion/register/RFZ5gSM902

Fees

The base fee (from which the referral percentages are calculated) is currently 4%.

The fee is taken from the amount which is received by the vendor. The buyer always pays the actual amount that is displayed for every product. The vendor receives that amount minus the fee.



```html
<div id="infopage" class="nofirstmargin">
    <h1>Market rules</h1>

Anonymity is sacrosanct here. You are to respect the anonymity of all Agora users to the greatest extent possible.
Vendors may not threaten buyers in any way, shape, or form.
You must have a valid vendor account to sell anything on Agora.

<h2>Forbidden products and services</h2>
<ul>
<li>Assassinations or any other services which constitute doing harm to another.</li>
<li>Weapons of mass destruction: chemical, biological, explosives, etc.</li>
<li>Poisons.</li>
<li>Child pornography.</li>
<li>Live action snuff/hurt/murder audio/video/images.</li>
<li>Direct means of access to privately owned accounts containing private value or property (monetary or otherwise)
    which has been obtained by the seller without the original owner's explicit consent with the primary intent of
    stealing the said value or property held in the account.
    This includes (among other things) stolen credit/debit cards, credit/debit card dumps, Paypal (or other similar services)
    accounts, bank accounts.</li>
</ul>

<h2>Other practical directives</h2>
<ul>
    <li>If you accept that customers go through the escrow system for buying your products, you may add "No-FE" flag to your products,
        which will usually put your products higher up in listings.</li>
    <li>If you do not accept escrow for certain or all of your products (in other words, if you require FE), do not put up the
        "No-FE" flag. This is simply false advertising and we reserve the right to fine you an appropriate amount if we deem that
        you have done so consciously.</li>
    <li>Products should always be in the correct category, as far as possible.</li>
    <li>Custom listings should be made using the "hidden" category which will also hide them from other users.</li>
    <li>Do not link buyers directly to other markets or other off-site sales with direct urls from your profile or product page.
        We understand if you choose to use another site in addition to Agora, or in replacement of it. There is nothing wrong with that.
        We simply request that you comply with the rules in that you do not direct customers off Agora to purchase from you.
        We're sure sure you wouldn't allow another vendor to advertise their products on your vendor page so please show us the
        same respect.</li>
</ul>
</div>
```

Figure F.2: Agora's Market rules

Following text are from F.2:

Market rules

Anonymity is sacrosanct here. You are to respect the anonymity of all Agora users to the greatest extent possible.

Vendors may not threaten buyers in any way, shape, or form.

You must have a valid vendor account to sell anything on Agora.

Forbidden products and services

Assassinations or any other services which constitute doing harm to another.

Weapons of mass destruction: chemical, biological, explosives, etc.

Poisons.

Child pornography.

Live action snuff/hurt/murder audio/video/images.

Direct means of access to privately owned accounts containing private value or property (monetary or otherwise) which has been obtained by the seller without the original owner's explicit consent with the primary intent of stealing the said value or property held in the account. This includes (among other things) stolen credit/debit cards, credit/debit card dumps, Paypal (or other similar services) accounts, bank accounts.

Other practical directives

If you accept that customers go through the escrow system for buying your products, you may add "No-FE" flag to your products, which will usually put your products higher up in listings.

If you do not accept escrow for certain or all of your products (in other words, if you require FE), do not put up the "No-FE" flag. This is simply false advertising and we reserve the right to fine you an appropriate amount if we deem that you have done so consciously.

Products should always be in the correct category, as far as possible.

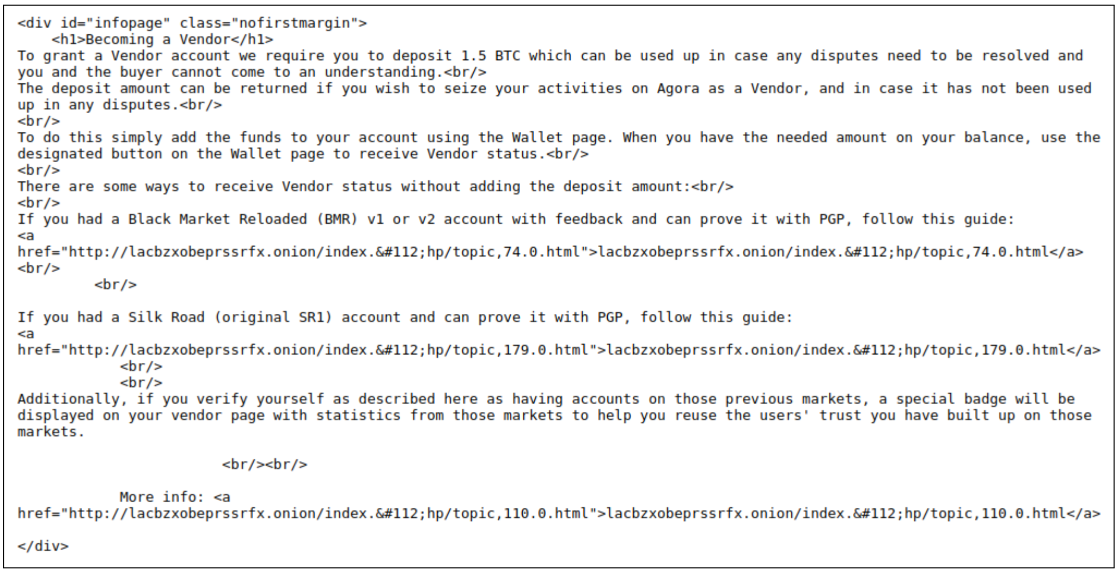
Custom listings should be made using the "hidden" category which will also hide them from other users.

Do not link buyers directly to other markets or other off-site sales with direct urls from your profile or product page.

We understand if you choose to use another site in addition to Agora, or in replacement of it. There is nothing wrong with that.

We simply request that you comply with the rules in that you do not direct customers off Agora to purchase from you.

We're sure sure you wouldn't allow another vendor to advertise their products on your vendor page so please show us the same respect.



```
<div id="infopage" class="nofirstmargin">
    <h1>Becoming a Vendor</h1>
To grant a Vendor account we require you to deposit 1.5 BTC which can be used up in case any disputes need to be resolved and
you and the buyer cannot come to an understanding.<br/>
The deposit amount can be returned if you wish to seize your activities on Agora as a Vendor, and in case it has not been used
up in any disputes.<br/>
<br/>
To do this simply add the funds to your account using the Wallet page. When you have the needed amount on your balance, use the
designated button on the Wallet page to receive Vendor status.<br/>
<br/>
There are some ways to receive Vendor status without adding the deposit amount:<br/>
<br/>
If you had a Black Market Reloaded (BMR) v1 or v2 account with feedback and can prove it with PGP, follow this guide:
<a
href="http://lacbzxobeprssrfx.onion/index.&#112;hp/topic,74.0.html">lacbzxobeprssrfx.onion/index.&#112;hp/topic,74.0.html</a>
<br/>
         <br/>

If you had a Silk Road (original SR1) account and can prove it with PGP, follow this guide:
<a
href="http://lacbzxobeprssrfx.onion/index.&#112;hp/topic,179.0.html">lacbzxobeprssrfx.onion/index.&#112;hp/topic,179.0.html</a>
            <br/>
            <br/>
Additionally, if you verify yourself as described here as having accounts on those previous markets, a special badge will be
displayed on your vendor page with statistics from those markets to help you reuse the users' trust you have built up on those
markets.

                        <br/><br/>

            More info: <a
href="http://lacbzxobeprssrfx.onion/index.&#112;hp/topic,110.0.html">lacbzxobeprssrfx.onion/index.&#112;hp/topic,110.0.html</a>

</div>
```

Figure F.3: Agora's requirements for new vendors

Following text are from F.3:

Becoming a Vendor

To grant a Vendor account we require you to deposit 1.5 BTC which can be used up in case any disputes need to be resolved and you and the buyer cannot come to an understanding.

The deposit amount can be returned if you wish to seize your activities on Agora as a Vendor, and in case it has not been used up in any disputes.

To do this simply add the funds to your account using the Wallet page. When you have the needed amount on your balance, use the designated button on the Wallet page to receive Vendor status.

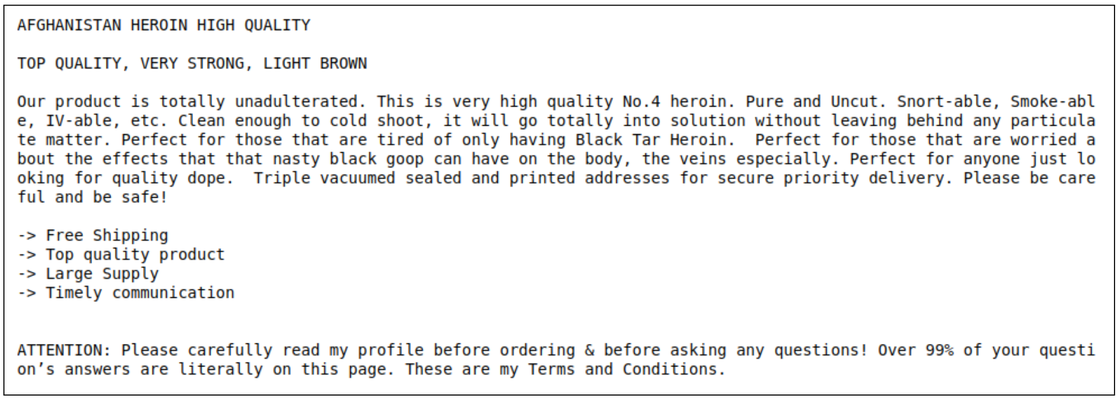
There are some ways to receive Vendor status without adding the deposit amount:

If you had a Black Market Reloaded (BMR) v1 or v2 account with feedback and can prove it with PGP, follow this guide: lacbzxobeprssrfx.onion/index.&#112;hp/topic,74.0.html

If you had a Silk Road (original SR1) account and can prove it with PGP, follow this guide: lacbzxobeprssrfx.onion/index.&#112;hp/topic,179.0.html

Additionally, if you verify yourself as described here as having accounts on those previous markets, a special badge will be displayed on your vendor page with statistics from those markets to help you reuse the users' trust you have built up on those markets.

More info: lacbzxobeprssrfx.onion/index.&#112;hp/topic,110.0.html



```
AFGHANISTAN HEROIN HIGH QUALITY

TOP QUALITY, VERY STRONG, LIGHT BROWN

Our product is totally unadulterated. This is very high quality No.4 heroin. Pure and Uncut. Snort-abl
e, IV-able, etc. Clean enough to cold shoot, it will go totally into solution without leaving behind any particula
te matter. Perfect for those that are tired of only having Black Tar Heroin.  Perfect for those that are worried a
bout the effects that that nasty black goop can have on the body, the veins especially. Perfect for anyone just lo
oking for quality dope.  Triple vacuumed sealed and printed addresses for secure priority delivery. Please be care
ful and be safe!

-> Free Shipping
-> Top quality product
-> Large Supply
-> Timely communication


ATTENTION: Please carefully read my profile before ordering & before asking any questions! Over 99% of your questi
on’s answers are literally on this page. These are my Terms and Conditions.
```

Figure F.4: Item description of an item from marketplace Evolution

Bibliography

Afilipoaie, A. and Shortis, P., 2015a. From dealer to doorstep âĂŞ how drugs are sold on the dark net. (06 2015). (cited on pages 3, 4, 5, 11, and 12)

Afilipoaie, A. and Shortis, P., 2015b. Operation onymous: International law enforcement agencies target the dark net in november 2014. (Jan 2015). (cited on page 16)

Ball, M.; Broadhurst, R.; Niven, A.; and Trivedi, H., 2019. Data capture and analysis of darknet markets. (03 2019). (cited on page 4)

Branwen, G.; Christin, N.; DÃl'cary-HÃl'tu, D.; Andersen, R. M.; StExo; Presidente, E.; Anonymous; Lau, D.; Sohhlz, D. K.; Cakic, V.; Buskirk, V.; Whom; McKenna, M.; and Goode, S., 2015. Dark net market archives, 2011-2015. https://www.gwern.net/DNM-archives. https://www.gwern.net/DNM-archives. Accessed: 2019-12-10. (cited on pages 4, 15, and 39)

Broadhurst, R., 2019. Child sex abuse images and exploitation materials. (10 2019), 310–336. doi:10.4324/9780429460593-14. (cited on page 4)

Buxton, J. and Bingham, T., 2015. The rise and challenge of dark net drug markets. (01 2015). (cited on pages 3, 4, 5, and 11)

Christen, P., 2006. A comparison of personal name matching: Techniques and practical issues. In *Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, ICDMW âĂŹ06, 290âĂŞ294. IEEE Computer Society, USA. doi:10.1109/ICDMW.2006.2. https://doi.org/10.1109/ICDMW.2006.2. (cited on page 12)

Christin, N., 2012. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. *Proceedings of the 22nd International Conference on World Wide Web*, (07 2012). (cited on pages 4 and 12)

Decary-Hetu, D.; Paquet-Clouston, M.; and Aldridge, J., 2016. Going international? risk taking by cryptomarket drug vendors. *International Journal of Drug Policy*, 35 (01 2016), 69–76. (cited on page 3)

Dolliver, D. S. and Kenney, J. L., 2016. Characteristics of drug vendors on the tor network: A cryptomarket comparison. *Victims & Offenders*, 11, 4 (2016), 600–620. doi:10.1080/15564886.2016.1173158. https://doi.org/10.1080/15564886.2016.1173158. (cited on page 12)

Europol, 2017. Drugs and the darknet. perspectives for enforcement, research and policy. Technical report. (cited on pages 3, 4, and 6)

Greenberg, A., 2014. 'silk road 2.0' launches, promising a resurrected black market for the dark web. (cited on page 16)

Greenberg, A., 2017. The dark web's top drug market, evolution, just vanished. https://www.wired.com/2015/03/evolution-disappeared-bitcoin-scam-dark-web/. (cited on page 16)

Manning, C. D.; Raghavan, P.; and Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA. ISBN 0521865719. (cited on page 25)

Meland, P. H.; Bayoumy, Y.; and Sindre, G., 2020. The ransomware-as-a-service economy within the darknet. *Computers and Security*, 92 (02 2020), 101762. doi: 10.1016/j.cose.2020.101762. (cited on page 4)

Mirea, M.; Wang, V.; and Jung, J., 2018. The not so dark side of the darknet - a qualitative study. *Security Journal*, (07 2018). doi:10.1057/s41284-018-0150-5. (cited on page 2)

Redman, J., 2016. Dark net market vendors reveal their day-to-day lives on reddit. https://news.bitcoin.com/dnm-vendors-reveal-lives/. (cited on page 5)

Wang, Y.; Liu, T.; Tan, Q.; Shi, J.; and Guo, L., 2016. Identifying users across different sites using usernames. *Procedia Computer Science*, 80 (12 2016), 376–385. doi:10.1016/j.procs.2016.05.336. (cited on pages 12 and 25)

Zhang, Y.; Fan, Y.; Song, W.; Hou, S.; Ye, Y.; Li, X.; Zhao, L.; Shi, C.; Wang, J.; and Xiong, Q., 2019. Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In *The World Wide Web Conference*, WWW âĂŹ19 (San Francisco, CA, USA, 2019), 3448âĂŞ3454. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3308558.3313537. https://doi.org/10.1145/3308558.3313537. (cited on pages 7 and 12)