Measuring Darknet Markets

by

B.P.J. Stinenbosch

to obtain the degree of Master of Science in Computer Science
at the Delft University of Technology,
to be defended publicly on Tuesday August 21, 2019 at 15:30.

Student number:	4370538	
Thesis committee:	Prof. dr. P. H. Hartel,	TU Delft, supervisor
	Prof. dr. M. J. G. van Eeten,	TU Delft, Second Reader
	Drs. R. S. van Wegberg,	TU Delft, Daily Supervisor
	Dr. G. J. van Hardeveld,	FIOD

TU Delft

Acknowledgements

I would firstly like to thank my daily supervisor Rolf van Wegberg. I could not have wished for a better supervisor. Rolf was always available to answer any of my questions about my research or writing. He gave me a lot of support while allowing this paper to be my own work.

I would like to thank Pieter Hartel and Michel van Eeten for their constructive criticism during the kickoff, midterm, and greenlight meeting. With their expertise in both the technical as social-economic aspects of this subject, they gave me the right feedback to get the most out of this thesis.

Special gratitude goes out to the Financial Advanced Cyber Team (FACT) of the Dutch Fiscal Information and Investigation Service (FIOD) for supplying me with the data that enabled me to do this research. Not only did they give me a place to work in their facilities, but they also gave me the opportunity to get to know other organizations like the Police, Europol, and the Public Prosecutor's Office. Everyone on this team was available to think along and share their expert opinion with me. From day one, you made me feel at home in your organization. Thank you!

A special mention goes out to Gert Jan van Hardeveld for sharing all his valuable knowledge and expert opinion. He helped me in removing any organizational obstacles that impeded my thesis. Also, he helped me by proofreading my thesis and giving me tips on how to improve it.

And finally, last but by no means least, a big thank you to all my family and friends that supported me in the past five years!

B.P.J. Stinenbosch
Delft, August 2019

Abstract

Background: A lot of scientists have tried to shed light on dark web markets. They did this by scraping these marketplaces over a period of time and describe what they were seeing. However, the methods used to measure these markets were never validated before. **Research goal:** This research will identify and validate the methods that are used in the literature to measure darknet marketplaces. **Methods:** To validate these methods, a novel dataset is used, namely the confiscated backend of a market. This dataset is cleaned and used to analyze the accuracy of these proxies on. **Results:** It is found that the number of transactions, revenue, and market share can be estimated with a high amount of explained variance. The predictions are precise enough to find prominent vendors on the market. The designed method of calculating the illegally obtained profits of a darknet vendor is easy to understand and could be used by law enforcement agencies. Finally, it is found that some vendors register to a market early as a strategy to ensure their business continuity.

Keywords: Cryptomarket, Darknet, Market, Measuring, Proxies, Validating

Contents

1 Introduction

1.1. background

Dark markets are booming! Since 2010 there has been an increasing number of online anonymous marketplaces (Christin, 2018). These markets enable criminals to trade in drugs, weapons, counterfeit goods, bank account details, and more alike. Anonymity and trust are two key principles that made these markets grown out to be an ideal place to facilitate criminal businesses (Mounteney et al., 2016, Tzanetakis et al., 2016).

Trust is an essential notion in dark markets. You have to trust your vendor that (s)he supplies you a good quality product and sends it discretely. Also, you have to trust the market and the vendor that they do not run away with your money. To give you some guarantees, markets implement systems like escrow and reviews (Verburgh et al., 2018). Escrow protects buyers and vendors by using a giving some certainty that the buyer pays for the product, and the vendor delivers the product. The review system lets the vendors build up a reputation. This way, new buyers can see if the vendor is more likely to provide a high-quality product and does not try to scam you.

The precautions that most markets take to let their users stay anonymous in these markets are using pseudonyms, Tor, PGP, and cryptocurrencies (van Hardeveld et al., 2018). Pseudonyms are used to hide one's real name; Tor routes the user's traffic through multiple servers making it difficult to reveal the users actual IP address (Calis, 2018); Cryptocurrencies are used to create anonymous transactions, and PGP enables secure communication.

On the one hand, the users of these markets enjoy these anonymity and trust systems with which the market facilitates them. However, from a law enforcement perspective, these measures make it hard to find out who is participating in criminal activity. To shine a light on the dark markets, multiple scientists tried new ways of measuring the size and scope of the operations on these markets and developing models to make predictions to generate insights on what is going on (Broséus et al., 2016, Calis, 2018, Christin, 2013, Kruithof et al., 2016, Soska and Christin, 2015, van Hardeveld et al., 2018, Van Wegberg et al., 2018).

Although the models and insights that came out of these studies look promising, all these studies have some biases. The biggest one being that these studies conducted their analysis on scrapes from dark markets. The administrators of the market could try to counteract the scraping by, e.g., only presenting a selection of their data. Additionally, the users of these markets can come up with adversarial techniques to mislead the currently used proxies. For example, a vendor could ask his buyers not to leave a review if he already has enough reviews to build up a good reputation. This way, a proxy based on reviews will underestimate his revenues. Another way to mislead law enforcement is by hiding the real sales price in the shipping costs. This way, you will stay under the radar as the shipping costs are not used in proxies. It is not unthinkable that criminals use these kinds of methods to deceive law enforcement. To validate the existing proxies and give insight into the way criminals mislead these proxies a study on a complete data source is needed.

In the summer of 2017, a multinational law enforcement operation took place named operation Bayonet (Verburgh et al., 2018). During this operation, two of the most prominent dark markets were shut down. First, the FBI shut down Alphabay. Next, Hansa was taken over and shut down after a month by the Dutch police. During a takeover, the database of the market can be seized. This data can lead to new operations, but it can also be used to verify the know proxies and refine them. By doing this, scientists will be able to measure the scale and scope of marketplaces more precisely, and law enforcement can get better insights on the suspects they are investigating. E.g., if the formula's used for predicting the revenue of a vendor are precise and easy to understand, then it could be used in the process of seizing the illegally obtained profit of the vendor like the formula for calculating the illegally obtained profit of a hemp nursery (OM, 2005).

1.2. Research goal

The main aim of this research is verifying the known proxies with backend data of a dark market. The main research question is: ”How valid are the peer-reviewed methods that are designed to measure dark markets with scrapes and what biases them?”. This question looks at the validity of the known proxies. The existing proxies for cryptomarkets are mostly built on scraped data. Because the scrapes are incomplete and the markets will try to withhold particular information, it is expected that these proxies will not always give reliable results. With the backend data from a seized market, these proxies will be verified. With this data, it is also analyzed how these proxies are biased.

1.3. Structure

The rest of this paper is structured as follows. Section two provides an overview of the know proxies and the biases they enclose. Next, chapter three will describe the data that is used and go over the limitations the dataset has. Section four will go over the methodology used to validate and refine these proxies. In Section five, the known proxies will be verified. Section six will discuss the results, and section seven will give the main conclusions of this study.

2

Related Work

This section will describe the proxies that are used in the literature to measure activity on cryptomarkets. The proxies will be discussed in groups based on the concept they try to measure. For every proxy, the biases will be reviewed. Before the proxies are described, it is important to understand the way cryptomarkets works to understand the biases the proxies enclose.

2.1. How a cryptomarket works

2.1.1. Placing an order

Anonymous online marketplaces look like regular online markets on the clear web. To access a cryptomarket, you first have to access the dark web using an application like Tor. Tor is a client to ensure the users to stay anonymous while browsing the dark web. After navigating to a cryptomarket, you first have to make an account. When logged in you can see a page with listings. The markets often have a menu with categories or a search bar to find the listings with items in which you would be interested.

After clicking on such an item, you redirect to the listing page. This page includes all the information of the listing. e.g., a title, description, price, place it ships to, and the places from which it can be shipped. Like a normal online market, you can read the reviews of other buyers who purchased the product.

If you like to proceed to purchase the item you enter the quantity you would like to buy and select a shipping option. Some vendors give out discount codes which can be used while placing your order. After placing the order, you will be asked to pay for the order. Paying on cryptomarkets is done with crypto valuta. The most common one being bitcoin. To pay for an order, you make a transaction with the correct amount of money to a given address.

This address is often an escrow address. Escrow required the market to sign off a transaction before the vendor gets his money. This way, the vendor will not run off with your money without delivering the product. When a product is not delivered, and the vendor is not reacting anymore, you can dispute your order. This way, the administrators of the market will look into your case.

When buying the item, you can contact the vendor. When having contact, it is often recommended to use some encryption like PGP. This way, in the scenario that other parties would see your conversations, they will not be able to see the content of the message — for example, the address where you like the items to be shipped.

When your order is delivered, you should complete your order. This way, the vendor will get his money. When an order is completed, you will be asked to leave feedback. Other potential buyers can see this feedback. The vendor would also be asked to leave feedback on the buyer.

2.1.2. Custom listings and holding prices

When a buyer wants to buy large quantities, the vendor can make a custom listing. A custom listing is a listing made especially for one buyer. This way, a vendor can offer (loyal) customers a discounted price. Christin (2013) uses the words “custom listing” to identify custom listings by their title. A custom listing is no different than a standard listing regarding the buying procedure. Everyone can try to buy a custom listing. However, a vendor would not accept the order when the custom listing is not meant for you.

The markets offer the option to hide a listing to reduce the chance that other buyers will buy a custom listing. The hidden listings (also know as stealth listings) would not appear while browsing the categories or using the search bar. A hidden listing can only be accessed if you know the correct URL. Some vendors also use the hidden option as a way to temporarily remove their listing without deleting it — this way the feedback on the listing is preserved. However, if someone still has the link you the listing they can try to order it. Therefore some markets also give the option to mark a listing as out of stock. This way, a buyer is not able to order the listing anymore. Hidden listings can be problematic while identifying the revenue of a vendor as a scraper will not automatically find hidden listings.

On sites where it is not possible to hide a listing or mark it as out of stock, some vendors are known to use another technique to prevent buyers from temporarily buying their listing. They do this by increasing the price so that no one will buy it anymore. These high prices are called holding prices. Holding prices can be problematic while trying to predict the correct revenue of a vendor.

2.2. Ways of measuring cryptomarkets

2.2.1. Dealing with holding prices

Because it is possible that holding prices exist on dark markets, methods were developed to deal with them, so you get a reliable listing price. A scraped data set could contain multiple (different) price observations for a listing of which some could be holding prices. Different scientists used different ways of handling these holding prices. A suitable method of dealing with holding prices should make it possible to get an indication of the real price the buyer paid. When holding prices are observed, their effect should be minimized. But if there exist no holding prices, the method of dealing with them should not influence the indication significantly while removing the prices that are not a holding price.

Kruithof et al. (2016) call holding prices problematic as they will distort the estimation of drug prices and revenues. To get a better indication, they recommend using the median of the costs of a listing collected over time. Their goal with this indicative listing price was to estimate the monthly revenue of a vendor.

Soska and Christin (2015) used the listing price to estimate the total revenue of a vendor. Their method of dealing with holding prices was as follows. They excluded a price x from their observations if x > 10.000 USD, x = 0 USD (It was a free listing), x > 5 times the median of the remaining observed listing prices, or x < 0.25 the median of the remaining listing price. Van Wegberg et al. (2018) Used this same method.

Aldridge and Décary-Hétu (2016) manually inspected all the listings above 10.000 USD. Then they removed the listings from vendors that specifically referred to using ‘holding prices’.

Décary-Hétu and Giommoni (2017) used a more rigorous approach for dealing with holding prices. While researching the effect of a police intervention on price changes, they removed an observed listing price P at observation t if: Pt > 2*Pt-1 or Pt > 2*Pt+1. This means that they remove an observed price when it is more than two times the previous or next observed price.

At first the method of Kruithof et al. (2016) sounds promising. However, it cannot be used if you would like to see how prices change over time. The method of Soska and Christin (2015) could take this change in price over time into account. However, with their method, they discard observations. Wholesale listings could be easily worth over 10.000 USD. But with their method, these observations would be discarded. The method of Aldridge and Décary-

Hétu (2016) would give a lot of false negatives because a vendor would not always state that they use holding prices. Décary-Hétu and Giommoni (2017) use a rigorous way of removing holding prices, which could lead to an even less accurate estimation.

2.2.2. Transactions

The next proxy indicates the number of transactions each listing has. In the past scientists have used the number of feedbacks as a proxy to estimate the number of transactions (Aldridge and Décary-Hétu, 2014, Christin, 2013, Décary-Hétu and Quessy-Doré, 2017, Dittus et al., 2018, Soska and Christin, 2015, Van Wegberg et al., 2018). After making a purchase, the buyers are asked to leave a review. By posting a review, you tell future buyers about the quality of the product and the delivery. The more good reviews a vendor has the more buyers may trust the vendor. The number of reviews on a listing will always underestimate the number of transactions (Aldridge and Decary-Hétu, 2016, Christin, 2013, Soska and Christin, 2015). This is because not all the buyers will leave a review. Therefore the number of feedbacks will give a lower bound of the number of transactions that are made.

Aldridge and Décary-Hétu (2014) Tried to indicate the yearly number of transactions and revenue of the silk road. However, they only scraped for three days. To get an indication of the annual number of sales of a vendor, they designed an extrapolation method by taking the number of feedbacks posted in the last 30 days before the day that the listing was scraped. This number of feedbacks was divided by days between the day of the first feedback (within the 30 days before the scrape) and the day of the scrape. This was done to get the daily feedback rate. This rate can then be multiplied by a number of days to extrapolate the estimated number of transactions. In the original paper Aldridge and Décary-Hétu (2014) extrapolated the number of transactions to a year (365 days).

Kruithof et al. (2016) use the same method. They tried to measure the number of transactions of a month based on a scraping period of five days. The month that was sampled was the thirty days previous to the five days of scraping. In this paper, the authors argued that the extrapolation compensates for the listings that went offline during the sampling period and therefore, can no longer be measured. They also argue that in the scrape, the active number of listings should be about the same as in the sample, and these listings should transact at the same rate for the extrapolation method to work.

As the feedback rate underestimates the number of transactions, Kruithof et al. (2016) looked for a way to compensate for this when using the estimated number of transactions in estimating the revenue of a vendor. To compensate for this, they used the overall feedback percentage (feedbacks/transactions). Aldridge and Décary-Hétu (2014) found that 88 percent of all the transactions had a review. They calculated this by looking at the number of feedbacks a vendor has and the number of transactions he had. This was conducted on scrapes of the Silk Road. A law enforcement representative suggested that 80.6 percent of all the transactions leave feedback Kruithof et al. (2016). Kruithof et al. (2016) found that on Dream Market this percentage was only 71 percent. Therefore, Kruithof et al. (2016) later used this feedback rate of 71 percent to calculate an upper bound of the revenue.

Nurmi et al. (2017) suggested a method for measuring transactions that was not built upon reviews. They used the change in stock as a proxy for sales. Some markets specify the number of items that are in stock. When monitoring this stock, it is seen when an item is bought and what the quantity is that was bought.

These proxies will give you a rough indication of the real amount of transactions. However, feedbacks will always underestimate the real number of transactions. Christin (2013) notes that you also cannot account for stealth (a.k.a hidden) listings. These are listings that do not show up while browsing the categories or using the search function on the market. This way, the hidden listings, and their feedback are not being scraped.

2.2.3. Time of a transaction

When looking at the reviews on a listing, it is hard to find out when the order was placed. However, the feedback gives a rough indication of when the feedback placed. How specific this time indication differs from market to market. Soska and Christin (2015) used the time indicated on the feedback as a proxy of the time an order was placed.

2.2.4. Paid price

When using the feedbacks as a proxy for the transactions, it is important to know for what price the transaction was made. Kruithof et al. (2016) used their median observed listing price as a proxy for the paid price. Soska and Christin (2015) were more specific. They tried to tie the real paid price to the feedback. For every feedback on one of the listings of the vendor, they searched for the closest (in time) observed listing prices that were not a holding price in their definition.

the downside of the method of Kruithof et al. (2016) is that it cannot be used when inspecting price changes over time. With the method of Soska and Christin (2015) this can be done. However, This method is biased by the time between the transaction and the review.

2.2.5. Revenue

Soska and Christin (2015) estimate the lower bound of the revenue of a vendor because their number of transactions was also a lower bound. To calculate the total revenue, they summed those observed listing prices connected to the feedback on the listings of the vendor. This method was adopted by Van Wegberg et al. (2018)

Aldridge and Décary-Hétu (2014) estimated the yearly revenue of vendors multiplying their extrapolated annual transaction estimate by the listing price that they found.

Kruithof et al. (2016) calculated the monthly revenue of a vendor in two ways. The first method was used to get a lower bound by multiplying their monthly number of transactions on a listing with the median observed price of that listing. Then they summed the revenues over all the listings of a vendor in that period to get the total monthly revenue of a vendor. To get an upper bound, they divided the lower bound by the review rate they found on Dream Market (71 percent). This was done to compensate for the orders that did not have a feedback. they also compensated for the listings that they did not scrape by dividing by their estimated scraping rate of 80 percent. It should be noticed that the revenue as it is calculated here is not the same as the profit of a vendor, because the vendor needs to hand-off a fee to the market.

2.2.6. Number of vendors

The number of vendors is often used as an indication of the size of a cryptomarket. To find out how many active vendors there are different scientists use different proxies. Christin (2013) used the number of vendors with an active listing at time T as a proxy for the number of active vendors at time T. like Christin (2013), Décary-Hétu and Giommoni (2017) used the number of vendors with one or more active listings that were online in a week to estimate the number of active vendors in that week. This is the same as the definition used in Christin (2013)

Soska and Christin (2015) found that this proxy did not take the vendors that were on vacation into account. To include those vendors, they changed their definition. To measure the active vendors at time T, they counted all the vendors that have a listing at t1 and the same or another listing at t2 where t1 >= T >= t2.

van Wegberg and Verburgh (2018) took a different approach. They did not use the listings but used the number of vendors that were active on the forum of Dream Market as a proxy for the number of users on the market. On Dream Market, there was a list available of users on the forum. For every user on the forum, it was possible to see if it was a vendor or not. Using the forum would lead to a lower bound as not all vendors from the market would be active on the forum.

2.2.7. Vendor matching

To get a clear indication of the number of vendors on the market and to see which vendors also operate on other markets, you need to be able to match multiple accounts that belong to the same vendor. Some vendors use the same name in different markets. Most of the time, vendors won't use the same username in an attempt not to get caught, thus smart ways to match vendors were developed. Most studies use the PGP key to match vendors Broséus et al. (2016), Kruithof et al. (2016), Soska and Christin (2015). Using the same PGP key can also be seen as a way to ensure business continuity when the market gets shut down. This way, your customers can find you in a new market. Kruithof et al. (2016) suggested another way to identify the same vendor by comparing profile descriptions. They were able to match 23 vendors based on identical profiles.

2.2.8. Market share

Décary-Hétu and Giommoni (2017) measured the market share per week of a vendor by dividing the number of sales of a vendor by the total number of sales in that week. In this case, the number of sales in a week was divided by the number of reviews in that week.

3 Data

To answer the research questions, a novel data source is used. By using the backend data of an anonymous online marketplace, unique and novel insights are generated. However, using this data has its implications. The market under study had a protection mechanism it purges orders purged. This section discusses the captured data in the database and the mechanism the market used to purge the data. It will also discuss the limitations this data has for this study and describe how the data was prepared for analysis. The code corresponding to this section can be found in appendix A.

3.1. Orders

Orders are an important part of the market. This market stored all the orders in one table. In this table, it kept the latest status of all the orders. This included the time at which an item was purchased, the number of goods that were ordered, the price that was paid for the items and the shipping (in BTC), and the percentage of discount the buyer got. The market had a safety mechanism that clean up completed orders. This safety mechanism helped to keep the database clean and provided some operational security in case law enforcement ever got hold of the database. The market had a script that dropped orders from the table if they were older than 30 days and completed. ”completed” includes orders that were finalized, declined or disputed, resolved and refunded. After deleting these orders, the market also purged the conversations and messages that were linked to this order.

The market, however, made backups of its database. Some of these backups were found during the law enforcement operation. When added to the set of orders that were still in the database during the law enforcement operation, we get the most complete data set we can get. Figure 3.1 shows the number of orders per day over time from the aggregated data set. In this figure, it can be seen that there are bulks of orders missing over time.

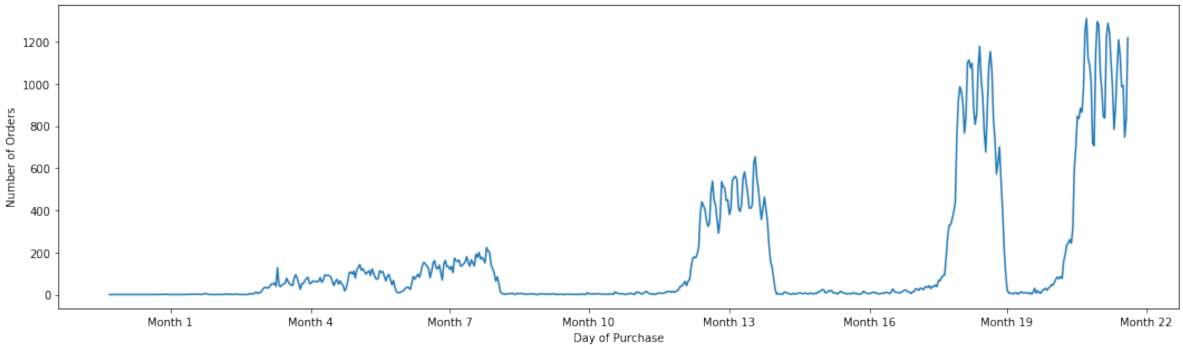

Figure 3.1: Orders per day

To get an indication of how much orders there are missing the count of the orders table was compared to the highest id in the order table. Assuming that the id in the order table is auto-incremented the ratio of count / highest id will give the percentage of orders that we have. By doing this, it is found that 53.0 percent of all the orders on this market are in the aggregated orders dataset.

Because the backups and the original order table are aggregated into one table, different time slices of the order table are added together. This means that the fields that change over time in the aggregated table are not reliable. Therefore it is tricky to give insights in the statuses of the orders as the order might be finalized just after the moment the backup was made. Because this is not known, I will only take into account the orders that were completed during one of the snapshots of the database. This way of dealing with missing values is called casewise deletion (Peugh and Enders, 2004). This method is chosen because there is a large sample and it is the best way to preserve the statistical relations. Also, one month of orders is removed from the dataset, as an external event influences this month.

3.2. Feedbacks

Feedbacks are an important component of the trust system on crypto markets. When an order was completed, buyers could leave feedback to inform other buyers about the quality of the product, the vendor, or the delivery. If the feedback was not true, the vendor could request the administrators of the market to remove the feedback. The administrators only removed feedback if the vendor had a good reason. Also, a vendor could set a number of days after which the feedback on him was removed. This number had to be a minimum of 180 days after the feedback was posted.

After doing the same count / highest id calculation, it is found that 99.0 percent of the feedbacks are still in the feedback table. It was also found that 0.03 percent of the vendors let the market remove their feedback. It is possible that the feedback on the orders of the vendors that remove feedback could have been removed. Therefore I will exclude these vendors with their listings and the corresponding orders and feedback from the dataset. This is done because leaving them in would give a false impression on the validity of the proxies.

The feedbacks not belonging to one of the finalized orders in the aggregated dataset will be removed. This is done because there could be feedbacks in the database corresponding to orders that were completed outside of one of the snapshots of the database. Because these orders are not included in the dataset, the corresponding feedbacks are also removed. This is done to preserve the real feedback rate.

3.3. Listings

The database has a table with listings. This table included the time the listing was created, the time the listing was updated, the corresponding vendor, whether the listing is hidden or deleted, the price of the listing (in the preferred currency of the vendor) and the stock of the listing.

3.4. Stealth listings

Stealth listings, also known as hidden listings, are listings are hidden when browsing the categories or using the search. The market advises using hidden listings to avoid cluttering up the market with custom listings. However, this does not imply that hidden listings are custom listings or vice versa. It is seen that some vendors also use this option to hide their listing temporarily without deleting it.

Hidden listings can be accessed by sharing the link of the listing. One may accidentally stumble upon a hidden listing while increasing the identifier in the URL of a listing. Dittus et al. (2018) found that this was also the case on other marketplaces. While being on the hidden listing page, there is nothing that indicates that it is a hidden listing. One can buy it like a normal listing and leave feedback as well.

In our dataset, we only know the latest status of a listing. So there is no way in telling whether a listing has always been hidden or not. At the last state of the market, 17.5 percent of all listings were hidden.

3.5. Custom listings

Custom listings are listings that are specially made for one buyer. Christin (2013) found that custom listings include the words ”Custom listing” in the title. By doing this, he missed custom listings that were not named ”Custom listing”. Therefore I expanded the search with the terms ”Special listing”, ”Listing for” or ”Order for”. After manually checking 200 samples, I found that this method gave 4 percent false positives. After marking the custom listings, it was found that 1.9 percent of all listings were custom listings. When checking the Pearson correlation between custom listings and hidden listings, it is found that the correlation coefficient is 0.006. This can be seen as a weak relation (Cohen, 1992, 2013), which means that normal listings and custom listings are about evenly likely to be hidden. This indicates that vendors do not always use the hiding option to hide their custom listings. Only 19 percent of all the custom listings are hidden. It is also found that vendors use the hidden option to (temporarily) delete their listing.

3.6. Price observations

When scraping a marketplace, a snapshot is taken. This snapshot contains a static view of the market in time. Making multiple scrapes of the market over time enables you to see the market dynamics. These dynamics include price changes over time. The proxies discussed in the previous section use the different price observations of a listing over time. So, to validate those proxies, some price observations are needed. The database, however, only kept the latest state of a listing. This means I only have the latest listing price. Which would not be enough to validate the proxies.

In order to validate the proxies, I came up with a new method to capture the dynamics of the price change of a listing over time. To capture the dynamics of the price observations of listings over time, the orders of that listing were used. Since the paid prices in BTC are known for the orders, and we know this BTC price is generated during the purchase of the order, it was simple to reverse engineer the price of the listing at the moment of purchase.

The problem with this is that the price of bitcoin fluctuates a lot. To compare prices over time, the prices were converted to the USD price. Since there is no information available about which service was used by the market to retrieve the USD/BTC rate, I decided to scrape the USD/BTC rate per minute from marketcap.io. This site calculates the bitcoin price based on a weighted sum of the bitcoin prices at exchanges where the weight is the percentage of trade volume of the exchange compared to the total trade volume.

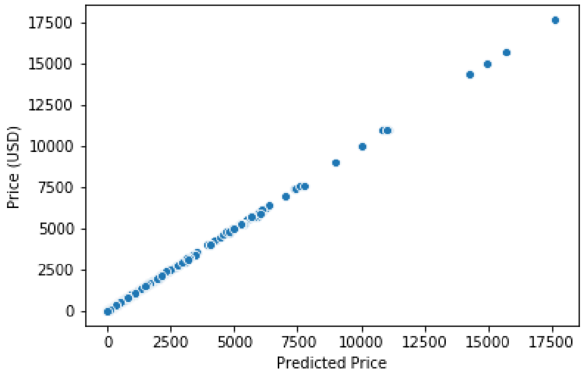

Figure 3.2: Predicted vs real USD listing price

To get an indication of how reliable these reverse-engineered price observations are, they were compared to the observed price of a listing. To do this, only the orders that were purchased after the moment the listing was last updated (i.e., a price change) would be taken into account. From this set, only the orders from listings where the vendor set their listing price in USD are kept to make a fair comparison. Because the listing prices are not normally distributed, namely a lot of low prices and little high prices, a Mann Whitney U-test was conducted to test whether there is a significant difference in the distributions. This test concluded that the distributions are not significantly different (U=175147889, P»0.05). When testing the quality of this estimator of the observed listing price, an R-square of 0.9999 is found with a standard error of the estimate of 4.86. In Figure 3.2 this strong linear relation can be seen.

4 Methodology

This section will go over how this paper is going to answer the research question, “How valid are the peer-reviewed methods that are designed to measure dark markets with scrapes and what biases them?”. This section will go over the methods used to answer the research question.

Shedding light on criminal behavior is not a novel thing to do. Multiple scientists have done this before (Broséus et al., 2016, Calis, 2018, Christin, 2013, Kruithof et al., 2016, Soska and Christin, 2015, van Hardeveld et al., 2018, van Wegberg and Verburgh, 2018). However, the proxies used up till now are not validated. In Section 5, the proxies that can be validated with the backend data of the market will be validated. The proxies that will be validated with the market data are:

- Dealing with holding prices
- Transactions
- Time of a transaction
- Paid price
- Revenue
- Number of vendors
- Market share

The vendor name matching and quantities cannot be validated as the dataset does not contain a ground truth of these subjects.

4.1. Dealing with holding prices

The “observed” listing prices in the used data set are reversed engineered from the paid prices in the order table. Also, it can be said that a buyer would not buy a product for the price of a holding price. Therefore the assumption is made that our observed listing prices don’t include holding prices. To check this section 5.1 will check for outliers in the list of observed prices. This check will be done by manually checking the observed listing prices that fall out of a range of 3 times the standard deviation from the mean observed listing price of a listing. Next, the methods of dealing with holding prices will be tested. The four methods that will be tested are

1. Removing a holding price x from the observations if x > 10.000 USD, x = 0 USD (It was a free listing), x > 5 times the median of the observed listing prices, or x < 0.25 the median listing price (Soska and Christin, 2015).

2. Removing an observed listing price P at observation t if: Pt > 2*Pt-1 or Pt > 2*Pt+1 (Décary-Hétu and Giommoni, 2017).
3. Removing observed prices over 10.000 USD where the listing also specified using a 'holding price' (Aldridge and Décary-Hétu, 2016).
4. Taking the median observed listing price (Kruithof et al., 2016).

To validate the methods that remove 'holding prices', the percentage of observed prices that were deleted will be given. Because the observed prices are reversed engineered from the paid price for the orders, we do not expect holding prices to be in the dataset. Therefore, validating these methods on this dataset will give an impression of the number of false positives these methods give us. Also, the marked holding prices will be manually checked to see why they are removed.

The method of taking the median observed price will be validated by comparing it to the mean observed price. This is done because the difference between the median and mean observed listing prices should be as small as possible when no holding prices exist.

4.2. Transactions

Often the number of feedbacks is used as a proxy for the number of transactions. Using the feedbacks will always give you a lower bound as not all buyers would leave feedback. To get a better insight in the relation between the number of feedbacks and the transactions first, the feedback rate will be explored. Next, section 5.2 will validate the method to extrapolate the number of feedbacks to estimate the number of transaction rate by Aldridge and Décary-Hétu (2014) on a sample like Kruithof et al. (2016) as there is only a month of data available for this.

To validate this method, a slice of the dataset is used that includes 30 days (sampling) and five days after (scraping). The chosen time slice falls at the beginning of the autumn and is not expected to be influenced by other external factors (Foster et al., 2015). Next, the method under study is used to estimate the number of transactions in the sampling time slice based on the scraping time slice. To determine the reliability of this estimation, the real number of transactions is compared to the estimated number of transactions. This is done by calculating the R-squared and the standard error of the estimate, as well as visually inspecting a scatterplot.

The R-squared is calculated by:

$$R^2 = 1 - \frac{\text{Sum of squared errors}}{\text{Sum of squares total}} \tag{4.1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y} - y)^2}{\sum_{i=1}^{n}(\bar{y} - y)^2} \tag{4.2}$$

This is the percentage of variance in the dependent variable that can be explained by a model. For a good prediction this variable should be high.

The standard error of the estimate is calculated by:

$$S = \frac{\sum_{i=1}^{n}(\hat{y} - y)^2}{n - 2} \tag{4.3}$$

95 percent time the real value falls within the prediction plus or minus two times the standard error of the estimate. For a good prediction this value should be low.

4.3. Time of a transaction

Soska and Christin (2015) use the time of a review as a proxy for the time of transaction. Section 5 will evaluate how precise this proxy is. To analyze how far the time of transaction and the time of review lie apart from each other, a cumulative plot will be made indicating how much time there is between the purchase and review. Also, it will be tested whether this time is influenced by the category of good or listing price (in USD). For testing the influence of the category of good, a Kruskal-Wallis H test will be used. For analyzing the influence of the price of the listing, the Pearson correlation coefficient will be given, and a scatter plot will be made.

4.4. Paid price

The paid price is used by Soska and Christin (2015) and Kruithof et al. (2016) as one of the inputs of their formulas to predict the revenue of a vendor. Therefore, to get a precise prediction of the revenue, the prediction of the listing price should also be as precise as possible. Therefore the two proxies for the paid price will be validated.

1. Using the median of the observed listing prices per listing as the paid price for that listing (Kruithof et al., 2016)
2. Using the closest (in time) observed listing price that is not a holding price to the time of the corresponding feedback. (Soska and Christin, 2015)

To validate these proxies, a Wilcoxon signed-rank test will be conducted to compare the predicted paid price to the actual paid price. Also, for both methods, the explained variance (R-squared) of the estimator and the standard error of the estimate (S) will be calculated to get a better indication of how precise these estimators are.

4.5. Revenue

Three methods to estimate the revenue will be tested:

1. Summing the observed listing prices connected to the feedback to get a lower bound of the revenue (Soska and Christin, 2015).
2. Multiplying the extrapolated number of feedbacks with the corresponding median listing price to get a lower bound of the revenue (Kruithof et al., 2016).
3. Calculating the upper bound by dividing the revenue of the previous proxy by the feedback ratio times the scraped percentage (Kruithof et al., 2016).

The method Aldridge and Décary-Hétu (2014) cannot be validated, as the real amount of revenue for the vendors of a whole year is not available. Also, they used the price at the moment of the scrape, which is not available in the data. The lower bound of Kruithof et al. (2016), however, will give a good impression of the accuracy of the method of Aldridge and Décary-Hétu (2014) as they used the same extrapolation method.

When validating and comparing the other proxies the context that they were used it should be taken into account. Therefore these proxies will first be evaluated in the context they were originally designed. Next, they will be put into another context to be able to compare the different methods.

First, the method of Soska and Christin (2015) will be validated. This is done by using all the finalized orders in the data set. The revenue is first calculated per vendor to get an indication of how good their lower bound is. This is evaluated by making a scatterplot and calculating the R-squared and the standard error of the estimate. Again the percentage of real vendor revenues that fall under the estimated lower bound will be given.

Secondly, the methods of Kruithof et al. (2016) will be validated by using their extrapolated monthly number of feedbacks and the calculated median listing price. For the calculation of the upper bound of the revenue, the feedback rate found on the market under study will be used. It is also assumed that we have a scraping percentage of 1 as all the transactions that are analyzed are available. The accuracy of the upper and lower bound will be analyzed by calculating R-squared and standard error of the estimate. Also, scatter plots are made to analyze the predictions. Finally, the accuracy of the upper and lower bound will be analyzed by calculating R-squared and standard error of the estimate. Also, scatter plots are made to analyze the predictions. Finally, the percentage of vendor revenues that fall under the lower bound, above the upper bound and between the upper and lower bound will be given.

To make a fair comparison between these methods, the method of Kruithof et al. (2016) will also be used on the sample on which the method of Soska and Christin (2015) is tested. This way, the bias of using their method of extrapolating the monthly feedbacks will be removed. Again a scatterplot will be given along with the R-squared and standard error of the estimate. Finally, the percentage of vendor revenues that fall under the lower bound, above the upper bound and between the upper and lower bound will be given.

As it is known that the revenue is biased by the discount, quantity and shipping cost it should be tested if the revenue prediction can be improved if these factors would be known. When designing a method to predict the revenue that can also be used by law enforcement when convicting a criminal, it should be tried to make this model as simple as possible. This should be done so that it is easy to understand how one came to an estimation.

This paper will not try to improve the revenue prediction by using the discount as it is not observable in a scrape. The shipping cost will also not be taken into account as the prices of the different shipping options over time are not known in the data. Therefore, this paper will try to improve the revenue prediction by using the quantities. In the current methods, a quantity of one per review is assumed. This will largely underestimate the as not every buyer leaves a review and some people order more than one product at a time. Nurmi et al. (2017) suggested that looking at the change of the stock of the listing would give a good indication of how many times a product is ordered. This would lead to a better prediction as not only the product can be the real number of purchases can be captured, but also the quantity that is bought at once leading to the real number of purchases. When taking this number instead of the number of transactions a big improvement in the estimation is expected. This will be tested by recalculating the revenue by multiplying the median observed listing price by the number of items purchased. In order to make a fair comparison, this is done using the same sample as the other two revenue models are tested on. A choice is made not to use a linear regression model or another kind of machine learning model that can be trained to compensate for the bias due to discounts and shipping costs. This is done because in order to seize the illegally obtained profit of a vendor every part of the calculation should be explainable.

4.6. Number of vendors

The methods of Christin (2013), Décary-Hétu and Giommoni (2017), Soska and Christin (2015) and van Wegberg and Verburgh (2018) can not be validated as the database does not include the a date until which a listing was able to be found. Also the market under study did not keep the dates on which vendors entered the forum like van Wegberg and Verburgh (2018) used with Dream Market.

To still get an indication of how good the number of vendors on the market can be observed by looking at the listings, it will be researched how much time there is between a vendor being approved and their first listing being posted. This will be analyzed by giving some descriptive statistics and inspecting a cumulative distribution plot.

4.7. Market Share

Décary-Hétu and Giommoni (2017) measured the market share per week of a vendor by dividing the number of sales of a vendor by the total number of sales in that week. We know that the number of feedbacks underestimates the number of transactions. But if the feedback rate were to about the same for all the vendors, it can be expected that the estimated market share based on reviews will lie close to the real market share based on transactions. To validate this proxy, eight weeks of orders and feedbacks are taken. As the database misses some data, we make sure that the chosen weeks include all the orders and feedbacks for that week. The proxy will be validated by calculating the real and estimated market share and then calculating the R-squared and the standard error of the estimate. Also, the Herfindahl–Hirschman Index will be calculated to get an indication of the amount of competition on the market of that week.

5 Validating proxies

In this section, the different proxies will be validated with backend data. The proxies are grouped per subject they measure. The code corresponding to this section can be found in appendix A.

5.1. Dealing with holding prices

Holding prices are prices that are sometimes used when a listing is out of stock. These are high prices (holding prices) that vendors use to ensure that buyers will not buy their listing. This is done to prevent deleting the listing together with the feedback. One could argue that if a marketplace supports features like hiding the listing or marking it as out of stock, it is no longer necessary to use holding prices. As the market under study supports these features and the estimated observed listing prices are based upon the reverse-engineered USD prices at the moment of the purchase, it is not expected that there will be holding prices present in our dataset. This is because it is unlikely that a buyer will buy a product for the price of a holding price. However, to check for holding prices in the dataset, there will be tested for outliers. Outliers will be defined as observed prices that are more than three standard deviations away from the mean observed listing price of the listing.

It is found that 2.7 percent of the observed listing prices are outliers. When inspecting these cases, it can be seen that these outliers are legit increases in the listing price over time. For example, a few USD price increase on a 99 cent listing price as can be seen in Figure 5.1. Therefore, it will be assumed that this dataset does not include holding prices.

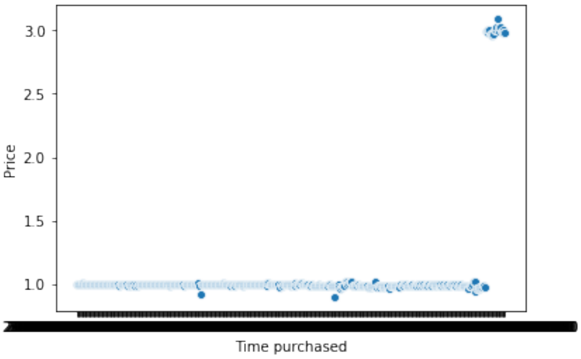

Figure 5.1: Price increase of an outlier over time

Now that we know that there are no holding prices in the dataset, the methods of dealing with holding prices can be tested. The three methods that will be tested are

1. Removing a holding price x from the observations if x > 10.000 USD, x = 0 USD (it was a free listing), x > 5 times the median of the observed listing prices, or x < 0.25 the median listing price (Soska and Christin, 2015).
2. Removing an observed listing price P at observation t if: Pt > 2*Pt-1 or Pt > 2*Pt+1 (Décary-Hétu and Giommoni, 2017).
3. Removing observed prices over 10.000 USD where the listing also specified using a 'holding price' (Aldridge and Décary-Hétu, 2016).
4. Taking the median observed listing price (Kruithof et al., 2016).

Method	N wrong	False positive rate
Soska and Christin (2015)	299	0.28 %
Décary-Hétu and Giommoni (2017)	573	0.53 %
Aldridge and Décary-Hétu (2016)	0	0.00 %

Table 5.1: Summary classification holding prices

Tabel 5.1 Gives an overview of how the methods compete. The holding price heuristic of Soska and Christin (2015) classifies 299 observed prices (0.28 percent) as holding price. These holding prices were observed over 151 different listings. It can be said that their heuristic marks free goods as holding price where it is no holding price (22 percent of the holding prices). Around 16 percent of the price observations are unfairly classified as a holding price because it was a listing of more than 10.000 USD. All other observations that were marked as a holding price can be explained by either being a cheap good getting higher in price (e.g., 1 USD to 5 USD) or by a vendor that has changed the base quantity of the listing.

The more rigorous approach of Décary-Hétu and Giommoni (2017) marked 573 observed prices (0.53 percent) as a holding price. Again, here, no real holding prices were found.

The method of Aldridge and Décary-Hétu (2016) did not lead to any prices being marked as holding prices. However, it can be expected that this method would miss some holding prices as not every listing with a holding prices puts in its description that it has a 'holding price'.

In contrary to the other methods, the method of Kruithof et al. (2016) does not remove any observations. It takes the median of the observed prices to discard the influence of the holding prices. When the dataset does not contain any holding prices, this method should give a price estimate that lies close to the mean. This is because the revenue will be calculated later by multiplying the number of transactions times this median price. A Wilcoxon test is used to test whether the distributions of the mean and median observed listing price are the same. The results of this test show that these distributions are the same (W = 14533268.0, p » 0.1). Also, the median explains 99.99 percent of the variance of the mean.

5.2. Transactions

In the past scientists have used the number of feedbacks as a proxy to estimate the number of transactions Aldridge and Décary-Hétu (2014), Christin (2013), Soska and Christin (2015), Van Wegberg et al. (2018). Not all buyers leave feedback. Therefore the number of feedbacks would be a lower bound of the number of transactions.

5.2.1. Exploring the feedback rate

To get a better insight in the relation between the number of feedbacks and the number of transactions, the feedback rate is explored. The overall feedback rate over all the orders in the dataset is 71.03 percent. This is interesting because it corresponds to the 71 percent found by Kruithof et al. (2016) on another market. In Figure 5.2, it can be seen that there is some variation in the feedback rate per listing. However, this feedback rate stabilizes when the listing has more orders.

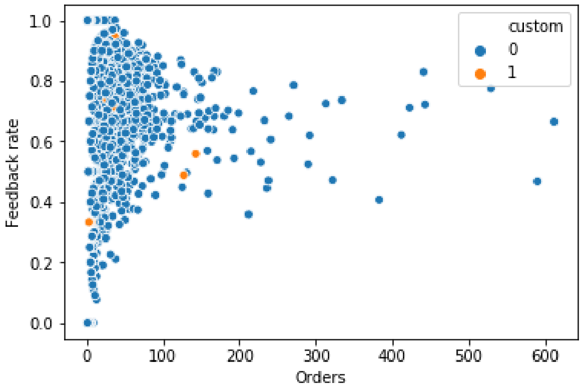

Figure 5.2: Scatterplot of the number of orders per listing against the feedback rate of that listing

A Kruskal-Wallis test was conducted to see whether the feedback rate differs per main category and drug category. It is found that there is a significant difference of feedback rate per listing between the different main categories (H = 1665, p « 0.001) and drug categories (H = 496, p « 0.001). This can be seen in the boxplots of the feedback rates in Figures 5.3 and 5.4.

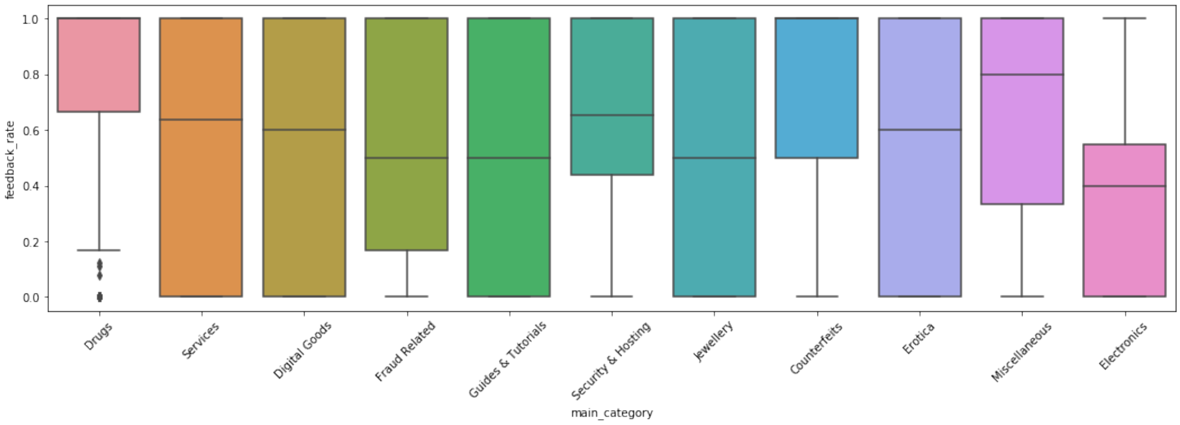

Figure 5.3: Boxplot of the feedback rates of the listings in the different main categories

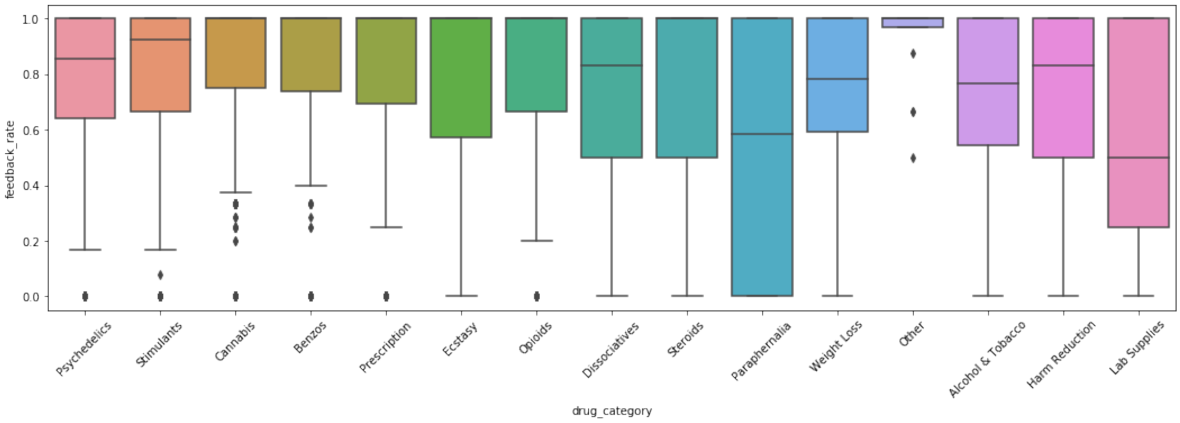

Figure 5.4: Boxplot of the feedback rates of the listings in the different drug categories

In Table 5.2, it can be seen that the counterfeit and drug categories have higher feedback rates than the other categories. In Table 5.3, the feedback rates for the different categories of drug listings can be found as well. It is unknown why different categories have different feedback rates.

Category	Mean feedback rate	SD	N
Counterfeits	0.723028	0.366023	198
Digital Goods	0.571558	0.399576	1573
Drugs	0.795641	0.307571	12825
Electronics	0.358984	0.340689	23
Erotica	0.564014	0.393550	508
Fraud Related	0.530281	0.379788	1239
Guides & Tutorials	0.514557	0.383386	1347
Jewellery	0.500000	0.500000	11
Miscellaneous	0.634644	0.411666	161
Security & Hosting	0.593216	0.345474	79
Services	0.572447	0.418330	391

Table 5.2: Feedback rate per main category

Category	Mean feedback rate	SD	N
Alcohol & Tobacco	0.711383	0.341372	17
Benzos	0.804460	0.300943	646
Cannabis	0.829508	0.284623	5659
Dissociatives	0.713179	0.352922	333
Ecstasy	0.744719	0.342163	1822
Harm Reduction	0.666667	0.471405	4
Lab Supplies	0.571429	0.449868	7
Opioids	0.812717	0.299697	318
Other	0.919271	0.159911	16
Paraphernalia	0.530769	0.456258	26
Prescription	0.810617	0.308774	953
Psychedelics	0.757990	0.311347	1261
Steroids	0.719879	0.373586	200
Stimulants	0.780934	0.301451	1531
Weight Loss	0.694258	0.371207	32

Table 5.3: Feedback rate per drug category

To test whether there is a significant difference in feedback rate between custom listings and normal listings, a Mann-Whitney U test was conducted. No significant difference was found (U = 3072873, p > 0.10). This is interesting because it shows that buyers also leave feedback on custom listings.

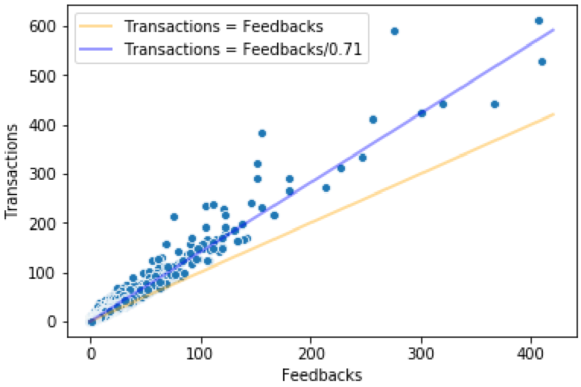

Figure 5.5: Scatterplot of the number of feedbacks against the transactions

To inspect to what extent the feedbacks predict the real number of transactions, a scatterplot was made (Figure 5.5). In this figure, it can be seen that the number of feedbacks underestimates the number of transactions. When estimating the transactions with the feedbacks, an R-squared of 0.85 is found with a standard deviation of the estimate of 6.5. However, the fit gets better when the estimate correct with the feedback rate (R-squared = 0.95, S = 3.7).

5.2.2. Validating the extrapolation method

To validate the extrapolation method of Aldridge and Décary-Hétu (2014), the monthly feedback rate of a sample was calculated by using their extrapolation method on a three-day ”scrape”. In the original paper, the feedbacks were extrapolated to a year. However, in our dataset, there is no full year of data available. Therefore, this method will be tested whilst extrapolating to a month.

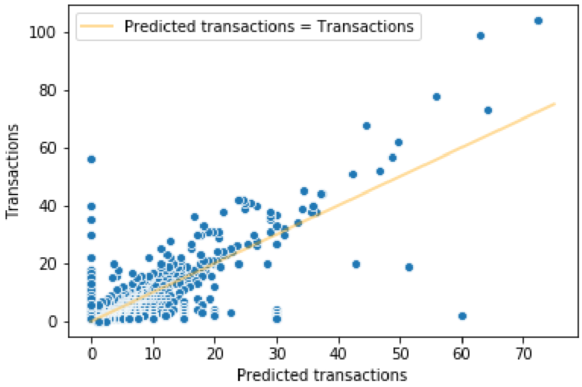

Figure 5.6: Predicted monthly feedback

In Figure 5.6, it can be seen that the extrapolation method is not accurate in estimating the actual number of transactions. The estimate has an R-squared of 0.24 and a standard error of the estimate of 5.24. The vertical line on the left at where the predicted feedbacks are zero exists because those listing were not observed in the ”scraping period”.

The reason why this extrapolation method works so severely can be found by looking at the daily orders (Figure 5.7). When looking at this graph, a clear weekly pattern of a higher number of sales at the beginning of the week and a dip at the end of the week emerges.

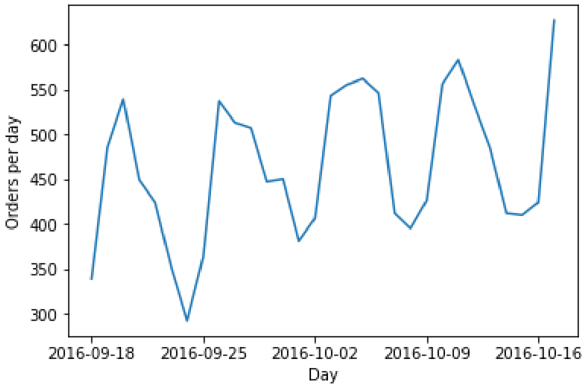

Figure 5.7: Daily number of orders in the month that is being estimated

When you do not capture a whole week (one period) of data, like happens in some cases where the listing is not yet online for a week, it can be expected that the results of extrapolation method will not be correct. This explains the horizontal line at the bottom of Figure 5.6 Also, the effect of the growth of the market (increasing number of orders over time) biases the extrapolation because the transaction rate is not the same over time.

5.3. Time of a transaction

Soska and Christin (2015) use the time of feedback as a proxy for the time of the sale. However, the time between making the transaction and leaving the feedback is not known. By using the backend data, insight can be given in the time it takes to leave feedback. Figure 5.8 shows a plot of the cumulative distribution function of how much time there is between making the purchase and leaving the feedback. On average, it takes 7.3 days before feedback is given.

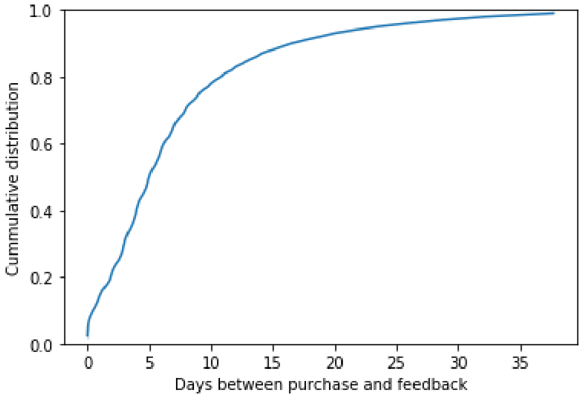

Figure 5.8: Cumulative distribution function of the days between purchase and feedback

Figure 5.9 shows the relation between the time between purchase and feedback and the price of a listing. Here, it can be seen that buyers often take longer to leave feedback on cheaper items. This can be due to a longer shipping time or the buyer waiting longer to leave a review after they received their order. The Pearson correlation coefficient is 0.03.

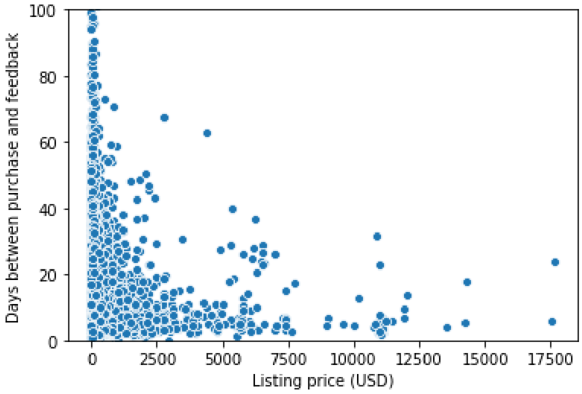

Figure 5.9: Scatterplot of the listing price against time between purchase and feedback

To analyze if there is a difference in the time between the transaction and feedback for the different categories of goods, a Kruskal-Wallis test was conducted. This test shows that there is a significant difference between the categories (K = 713121, p « 0.001). The boxplots in Figure 5.10 clearly show this difference. The categories of goods that physically need to be transferred have a longer time between purchase and feedback.

To get an indication of the time that the vendor receives the money, a cumulative distribution function plot is made. In Figure 5.11, it can be seen that 80 percent of the buyers leave their feedback directly when they finalize the order. For the other 20 percent, this can take up to a month. This is probably because these buyers want to test their product first. The

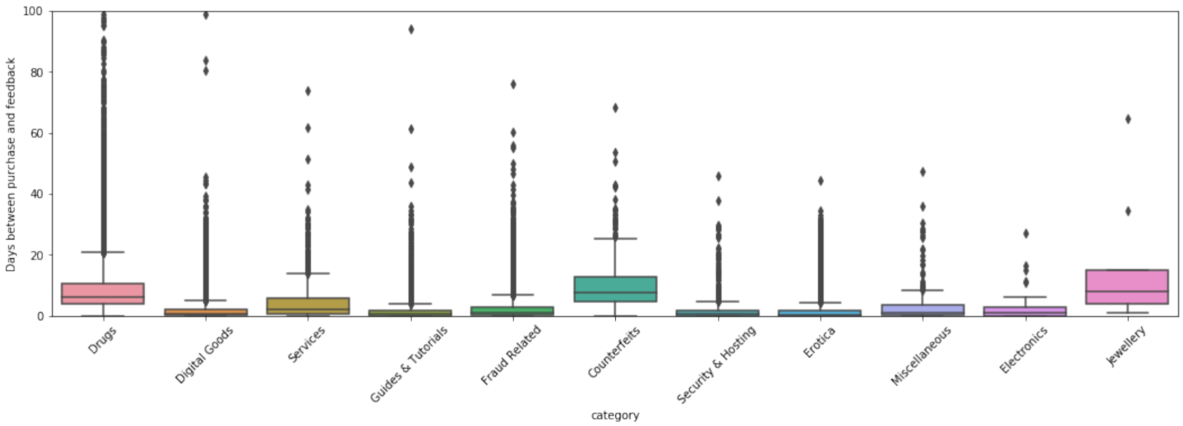

Figure 5.10: Time between purchase and feedback per category

80 percent of the buyers that leave feedback directly after the finalization show that the time of the feedback is a good proxy for the time that vendors receive their money. However, the time of the feedback does not give a precise indication of the time of the purchase, because this does not expect the product to be shipped.

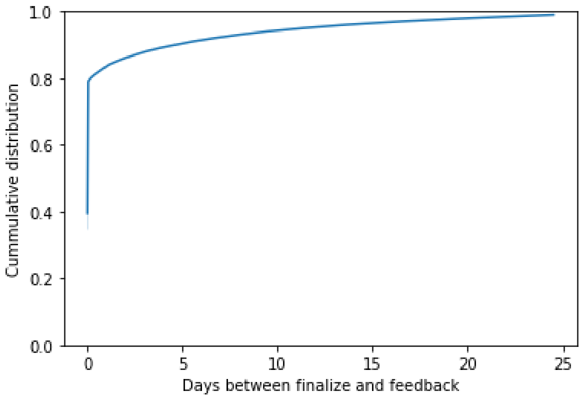

Figure 5.11: Cumulative distribution of the time between finalization and feedback

5.4. Paid price

In this subsection, the proxies for the paid price that will be validated are:

1. Using the median of the observed listing prices per listing as the paid price for that listing (Kruithof et al., 2016)

2. Using the closest (in time) observed listing price that is not a holding price to the time of the corresponding feedback. (Soska and Christin, 2015)

To understand why these proxies behave the way they do, it is essential to understand how the paid price is build up. In Figure 5.12, it can be seen how the revenue of a vendor is build up. In this figure, the red parts are observable. It can be seen that the price that a buyer paid is not the same a the listing price. The paid price is build up of two parts: the item price and the shipping cost. The item price is the listing price times the number of items minus the discount. The shipping price is the price paid for shipping. It is possible that there are multiple shipping options available for a listing (for different prices). To get a better insight into how the paid price is build up, we will take a closer look at the discounts, quantities, and shipping cost.

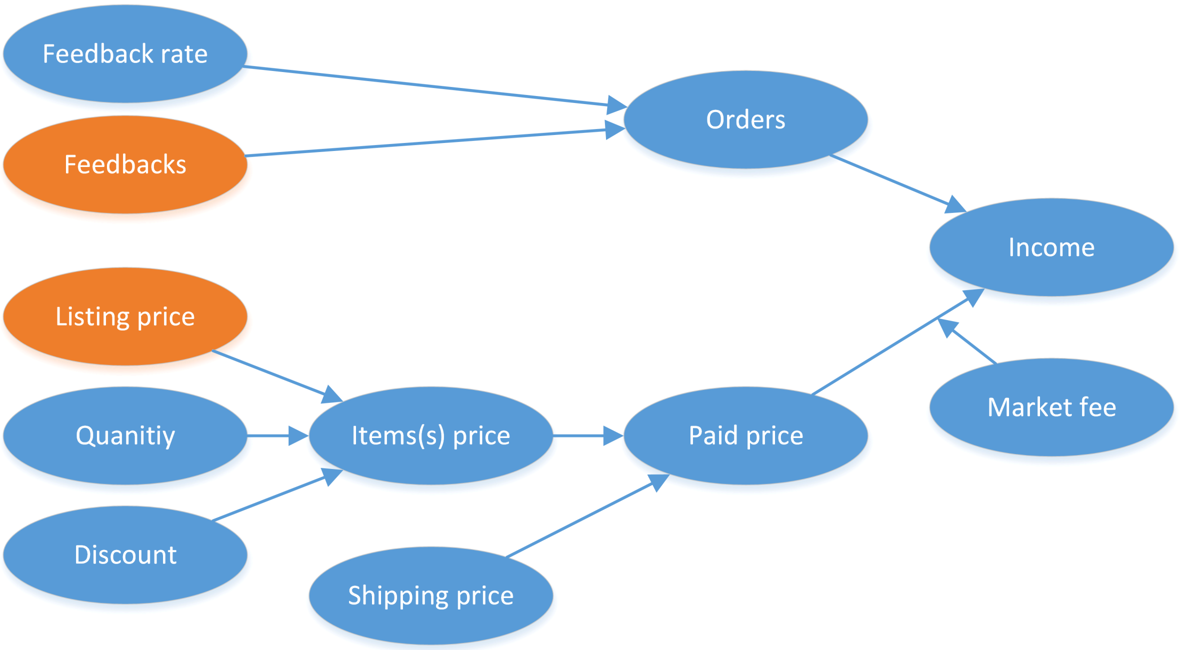

Figure 5.12: How income is calculated on a cryptomarket

5.4.1. Discounts

Vendors can give their buyers a discount on certain listings. By filling in a discount code, the buyer gets a discount on the listing price. 0.53 percent of all the orders have a discount. When looking at the distribution of discount one gets after removing the orders with no discount (Figure 5.13), it can be seen that lower discounts are more common than higher discounts. On average, who receives a discount receives an average discount of 22.7 percent.

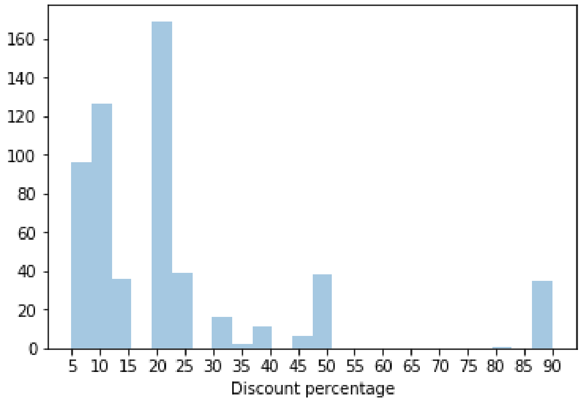

Figure 5.13: Histogram of the discounts on the orders with discounts

A Kruskal-Wallis test is conducted to test whether there is a significant difference in discount between the different categories of goods. There was no significant difference found (H = -216771654, p»0.25).

The influence of the listing price on the discount is visualized in Figure 5.14. It can be seen that higher discounts are given to orders with a lower listing price. This means that the bias of a discount of the estimate of the paid price will be lower when the listing price increases.

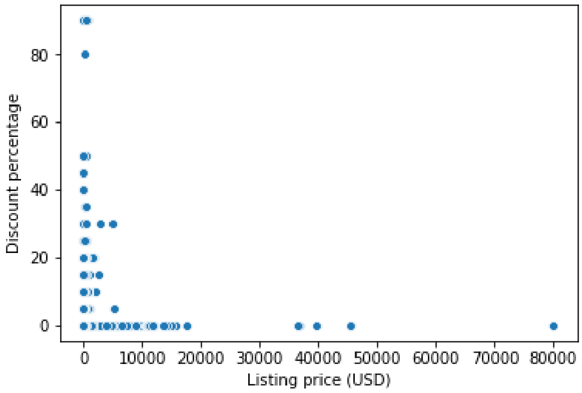

Figure 5.14: Scatterplot of the discount vs the listing price

To test if the discounts differ on orders from custom listings a Mann-Whitney U test is conducted. Here it is found that there is a significant difference between the discounts from regular listings and custom listings (U = 91260326, p « 0.001). On average orders from custom listings get a higher discount. This difference is 0.26 percent versus 0.12 percent on average.

5.4.2. Quantities

A buyer can order larger quantities at once. It is found that 90.9 percent of the orders have a quantity of one. On average, the quantity is 11.3. When looking a the box plots of the quantities in the different categories (Figure 5.15 and Figure 5.16) it can be seen that the drugs categories have more orders with a high quantity. Particularly cannabis and prescription drugs.

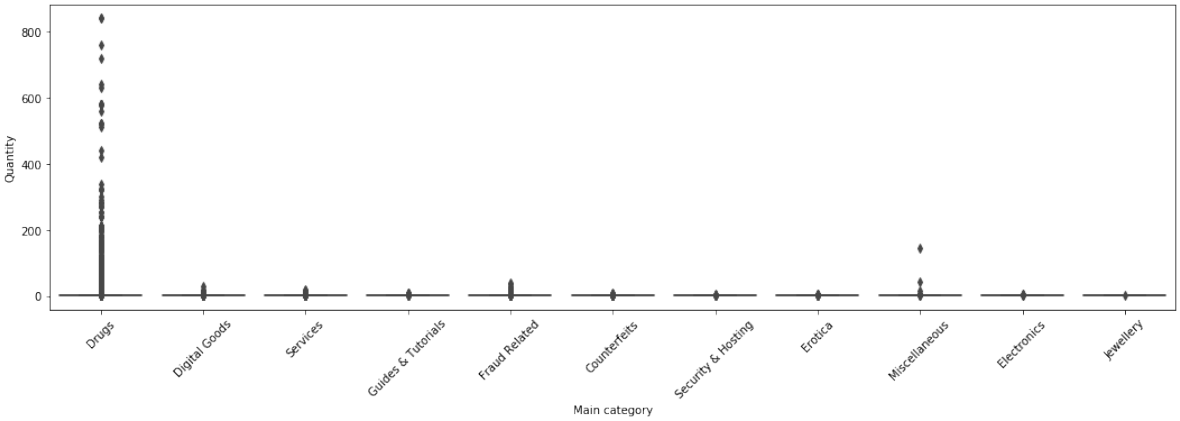

Figure 5.15: Boxplots of the quantities over the different categories

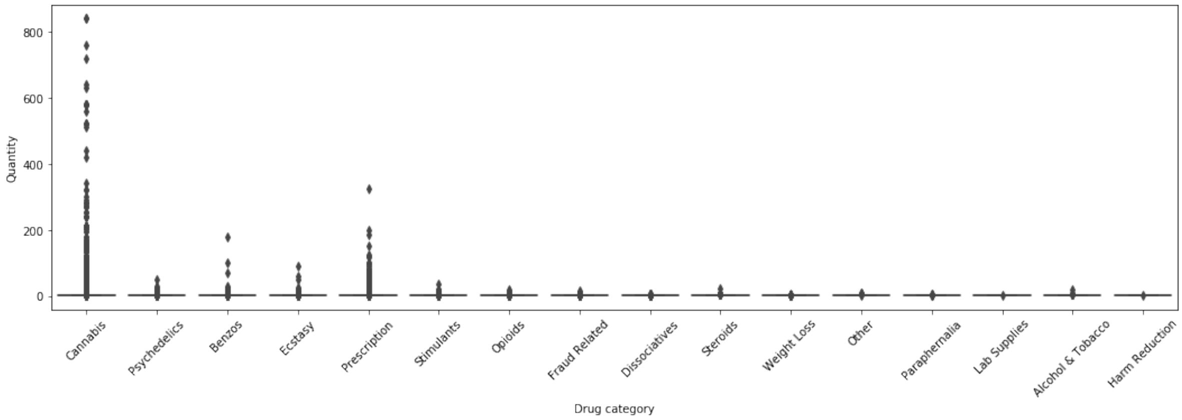

Figure 5.16: Boxplots of the quantities over the different drug categories

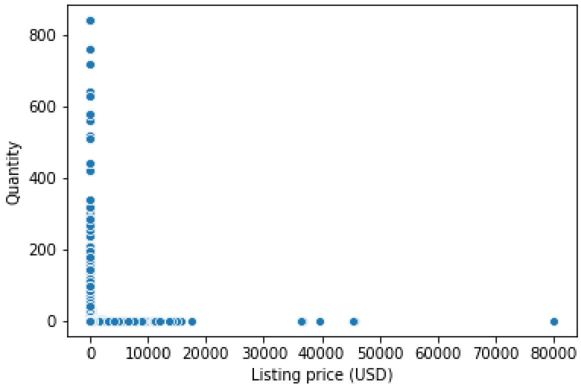

Figure 5.17: Scatterplot of the quantities vs the listing price

It is found that there exists a strong interaction between the quantity and the listing price. As can be seen in Figure 5.17 higher volumes are ordered on cheaper listings. However, higher than one could be problematic for the revenue estimation.

Figure 5.18 shows the interaction between quantities and discounts. It can be seen that higher quantities do not go together with discounts.

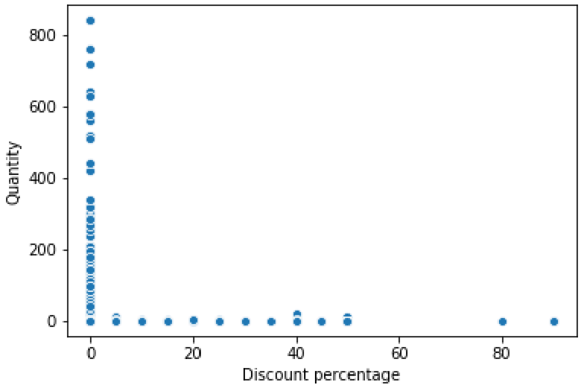

Figure 5.18: Scatterplot of the quantities vs discount

A Mann-Whitney U test is conducted to test if the quantity differs on orders from custom listings. It is found that there is a significant difference between the quantities from normal listings and custom listings (U = 90035647, p < 0.001). On average orders from normal listings (μ = 1.9, σ = 12.25) are ordered in higher quantities than orders from normal listings (μ = 1.17, σ = 0.95).

5.4.3. Shipping cost

The price paid for the item(s) is not the only income a vendor gets. A buyer also pays for the cost of shipping the items. On the market under study, the shipping options (and corresponding costs) are listed on the listing page. Normally a vendor gives the buyer multiple shipping options for different prices.

When analyzing the shipping costs, it is found that 52 percent of the orders are shipped for free. In Table 5.4 the percentage of free shipped orders, average shipping price, and the standard deviation of the shipping price are given for the different categories of goods.

Category	percentage shipped free	mean shipping cost	standard deviation
Counterfeits	35%	5.39	6.69
Digital Goods	100%	0.08	2.76
Drugs	31%	5.09	7.53
Electronics	97%	0.14	0.89
Erotica	100%	0.00	0.20
Fraud Related	96%	0.54	4.47
Guides & Tutorials	99%	0.14	1.83
Jewellery	43%	10.09	12.37
Miscellaneous	96%	0.48	3.13
Security & Hosting	99%	0.02	0.23
Services	89%	1.40	5.46

Table 5.4: Shipping cost per category

It can be seen that the products that need to be physically shipped have higher shipping costs. It stands out that jewelry listings have the highest average shipping costs.

Category	percentage shipped free	mean shipping cost	standard deviation
Weight Loss	4%	12.56	11.46
Prescription	27%	10.12	12.98
Lab Supplies	64%	9.52	20.06
Opioids	22%	9.16	10.55
Benzos	11%	7.11	7.39
Steroids	18%	5.74	4.21
Ecstasy	23%	5.63	11.43
Paraphernalia	4%	5.51	2.58
Dissociatives	33%	5.42	8.83
Stimulants	36%	4.78	7.17
Cannabis	37%	4.28	4.97
Psychedelics	24%	3.79	4.17
Alcohol & Tobacco	44%	3.71	5.91
Other	68%	1.74	2.87

Table 5.5: Shipping cost per drug category

When looking at the drug categories (Table 5.5) it can be seen that certain drugs have higher shipping costs. These higher shipping costs have a negative effect on the accuracy of the estimated paid price as they are not taken into account. It can, therefore, be expected that the revenue of vendors selling weight loss and prescription drugs are harder to estimate than, for example, a vendor selling alcohol with low shipping costs.

A Mann-Whitney U test is conducted to test whether there is a difference in shipping costs for orders from custom listings and regular listings. It is found that there is no significant difference between the shipping costs from the orders from regular listings and custom listings

(U = 90611138, p>0.05).

5.4.4. Validating median price

Now we know what influences the paid price of an order, the proxies can be validated. First, the proxy of Kruithof et al.(2016) will be validated. They used the median of the observed listing prices as a proxy for the paid listing price. In Figure 5.19, it can be seen that the most median prices fall together with the actual paid price (orange line). However, a strange pattern can be observed where some samples are off (red line). These samples all fall on the line paid price = 2 * median price. This can be explained by orders where the quantity of the orders is 2.

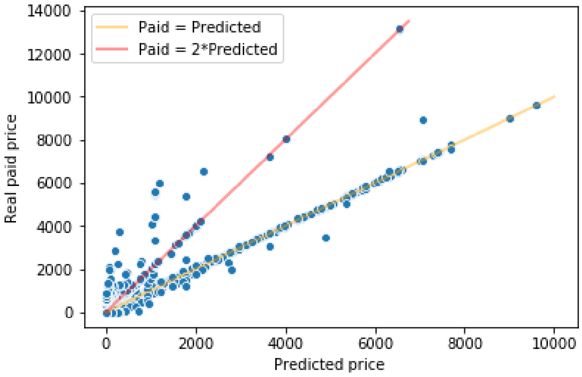

Figure 5.19: Median listing price vs the actual paid price

As can also be seen in the scatterplot, it looks like the median observed listing price underestimates the paid price. This is likely due to shipping costs and a quantity that is greater than one. However, in 22.1 percent of the time, the median listing price overestimates the paid price. To test whether the two distributions are the same, a paired Wilcoxon rank test is conducted. This test confirms that the two distributions indeed differ (W = 738018168, p«0.001). It is found that the R-squared of the estimate is 0.9024, and the standard error of the estimate is 96.98 USD.

5.4.5. Soska and Christin

Soska and Christin (2015) connected an observed listing price to each feedback they found. They did this by looking at the time the feedback was placed. They connected the observed listing price of that listing where the observation time was closest to the feedback time, and the observed price was not a holding price. This heuristic failed in 75 cases because the listing had no observations that were not a holding price (in their definition). Therefore I excluded these observations from the rest of the analysis.

In Figure 5.20, it can again be seen that there are different lines visible in the scatter plot. These lines can still be explained by orders with higher quantities. Mostly the method of Soska and Christin (2015) underestimates the real listing price, but in 14.7 percent of the time, it overestimates the paid price. A Wilcoxon paired rank test was conducted to see if the prices and predicted prices are the same. It can be concluded that there is a significant difference between the predicted price and the real price (W = 160927267, p « 0.001). It is found that their estimate has an R-squared of 90.16 percent and a standard error of the estimate of 97.38.

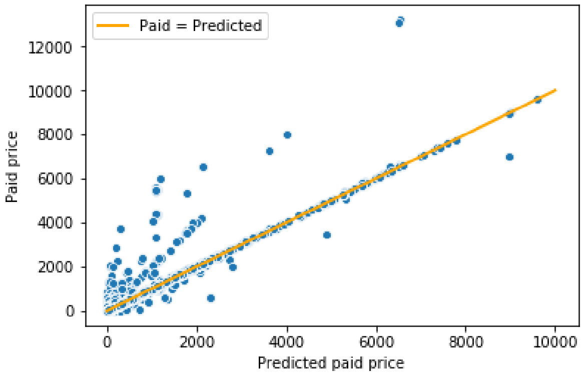

Figure 5.20: Soska and Christins price vs the actual paid price

In comparison, these methods work well. The method of Kruithof et al. (2016) performs a little better. This is due to the way that the higher prices are incorporated and not removed as in the method of Soska and Christin (2015). Also, this method is not biased by the difference in time between when the order and the feedback were placed.

5.5. Revenue

The three methods used to calculate the revenue of a vendor that will be tested are:

1. Summing the observed listing prices connected to the feedback to get a lower bound of the revenue (Soska and Christin, 2015).
2. Multiplying the extrapolated number of feedbacks with the corresponding median listing price to get a lower bound of the revenue (Kruithof et al., 2016).
3. Calculating the upper bound by dividing the revenue of the previous proxy by the feedback ratio times the scraped percentage (Kruithof et al., 2016).

5.5.1. Monthly revenue

Kruithof et al. (2016) used their method to calculate the monthly revenue based on a 5 day scrape. On this scrape, they used the extrapolation method Aldridge and Décary-Hétu (2014) to get the estimated number of transactions in the sampling month. Next, they used the number of transactions and their median price as inputs to calculate the monthly revenue. For their lower bound, they multiplied the number of monthly feedbacks with the corresponding median listing prices and summed this over all the listings of a vendor. For this research, the same method is used to predict the lower bound of the monthly revenue.

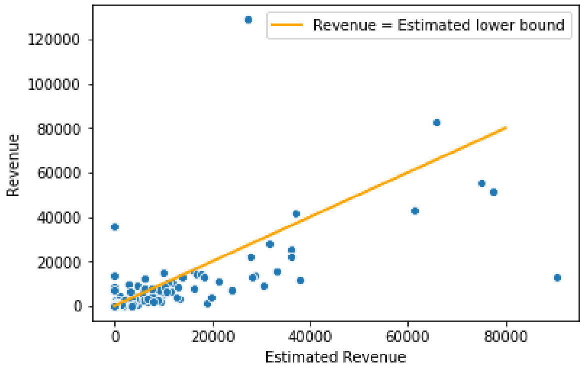

Figure 5.21: Predicted lower bound versus the real monthly revenue

In Figure 5.21 it can be seen that the predictions are far off. There is only one outlier found that is underestimated by about 90.000 USD, which can be explained by the vendor having custom listings of 30.000 USD without a review. Most of the other revenues are overestimated, which can be explained by the transaction rate increasing over time. The R-squared of this estimate is 0.36, and the standard error of the estimate is 7698 USD. It is found that the lower bound overestimates 60 percent of the vendors' revenues.

To predict the upper bound, they divided the lower bound by the feedback rate and the scraping rate. The feedback rate that was found in the dataset of this research (71 percent) is the same as the feedback rate that was found on Dream market by Kruithof et al. (2016). For the calculation of the upper bound in this analysis, a feedback rate of 71 percent is used as this is the average feedback percentage found in this study. The scraping percentage that was used is one as it was not possible to distinguish regular listings from hidden listings in the dataset. Therefore we use this method as if a full scrape was available.

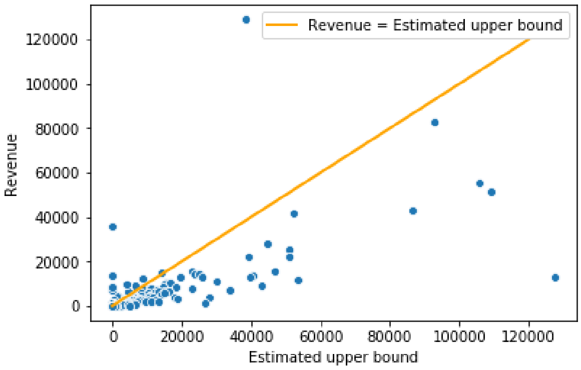

Figure 5.22: Predicted upper bound versus the real monthly revenue

In Figure 5.22, it can be seen that the prediction of the upper bound is even worse. There are still a lot of data points overestimated. The R-squared is -0.16, and the standard error of the estimate is 10.373 USD. It is found that the upper bound under estimates 30 percent of the vendors' revenues. In total, only 10 percent of all the revenues fall in between the predicted lower and upper bound.

5.5.2. Method Soska and Christin

When using the method of Soska and Christin (2015) a better prediction is found. When looking at Figure 5.23, it is found that their method also overestimates samples with the lower bound. This happens in 13.7 percent of the cases.

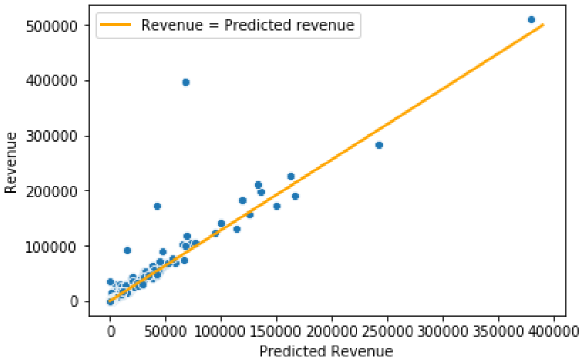

Figure 5.23: Predicted lower bound of Soska and Christin versus the total revenue

When calculating the statistics, an R-squared of 80.44 percent and a standard error of the estimate of 11954.32 USD are found. Whilst interpreting these results, one should take into account that 75 cases were removed because the listing price could not be determined with their method. If these ”holding prices” would be left in, a better fit may be achieved, leaving an R-squared of 82.5 percent.

5.5.3. Comparing the methods

To make a fair comparison the method of Kruithof et al. (2016) is also used on the sample used to validate the method of Soska and Christin (2015). This removes the bias of estimating the number of transactions. Instead, the real number of feedbacks per listing is given, which was also done whilst validating the method of Soska and Christin (2015).

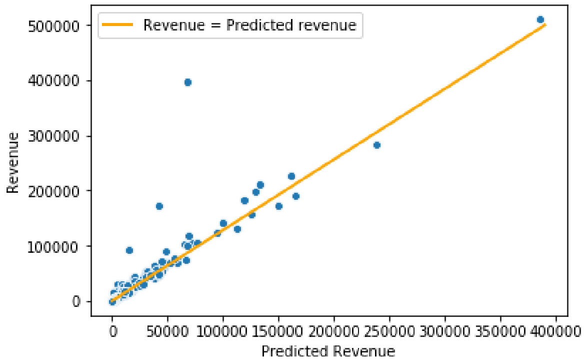

Figure 5.24: Predicted lower bound versus the real revenue

Figure 5.24 shows the scatterplot of the lower bound of the estimate calculated with the method of Kruithof et al. (2016). This estimate has an R-squared of 0.8049 and a standard error of the estimate of 12182.91 USD. This under bound overestimates 1.5 percent of the samples. This means that taking the median price instead of connecting the observed prices does not lead to a significant difference.

The custom listings, however, affect the estimated revenue. Because the prices of the custom listings are high, the revenue will be underestimated if the custom listing was ordered but does not have feedback. If the custom listing has feedback, it will primarily influence the revenue when this feedback is compensated for missing feedbacks.

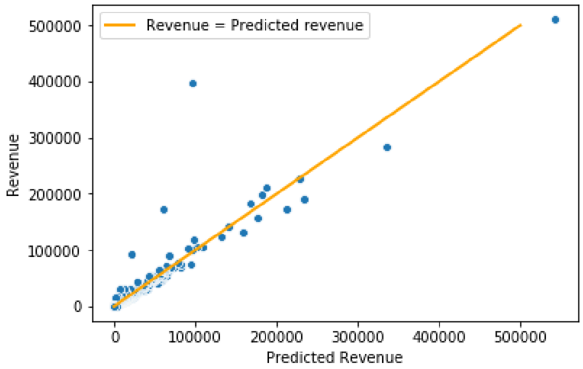

Figure 5.25: Predicted upper bound versus the real revenue

Figure 5.25 shows the scatterplot of the upper bound of the estimate calculated with the method of Kruithof et al. (2016). It has a good fit with an R-squared of 0.88 and a standard error of the estimate of 9542 USD. This is a significant improvement over the lower bound, which is due to the correction of the feedback rate. Again this upper bound should not be used as an upper bound because it underestimates 47.8 percent of the samples. This leaves 50.7 percent of the samples between the lower and upper bound.

5.5.4. Improving the prediction

In order to try to improve the estimation of the revenue the real purchased quantities are used as they can be scraped by looking at stock changes on some markets. When then multiplying the real purchased quantity times the median listing price a better prediction of the revenue is achieved.

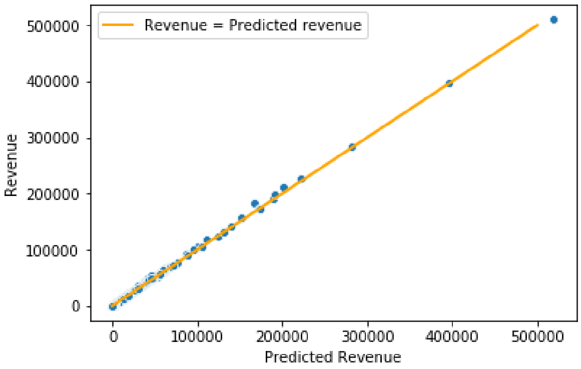

Figure 5.26: Predicted revenue when quantities are included.

Figure 5.26 the estimates of the revenue can be seen when the quantities are included in the prediction. This leads to amazing prediction with an R-squared of 99.9 percent and a standard error of the estimate of 960.23 USD. Although the estimation looks perfect, it still overestimates if a vendor gave the buyer a discount. The prediction underestimates when the shipping is not free.

5.5.5. Comparing the methods

When comparing the different methods (Tabel 5.6) we see that the extrapolation method performs the worst. The method of Kruithof et al. (2016) that corrects for the feedback rate performs the best if the stock cannot be scraped on a market. If the stock can be observed it is better to multiply the median price by the number of items purchased. Doing this will lead to an almost perfect estimation.

Method	R^2	S
Extrapolated transactions times median price (lower bound)	0.36	7698 USD
Extrapolated transactions times median price (upper bound)	-0.16	10373 USD
Sum of feedbacks times connected price	0.804	11954
Number of feedbacks times median price (lower bound)	0.805	12183
Number of feedbacks times median price (upper bound)	0.88	9542
Observed quantities times median price	0.999	960

Table 5.6: Summary of the accuracy of the different revenue prediction methods

5.6. Number of vendors

As explained in the methodology section, the methods of the number of vendors on the market cannot be validated with the data. However, insights can be given on how good of a proxy the active listings are in predicting the number of vendors on the market by looking at the time between a vendor getting approved and the creation of the first listing.

When inspecting this relation (Figure 5.27) it can be seen that about 90 percent of the vendors put up their first listing within a month after being approved. However, about 10 percent of the vendors wait longer to post their first listing. This seems strange, but it can be explained when looking at the time that these 10 percent of the vendors post their first listing.

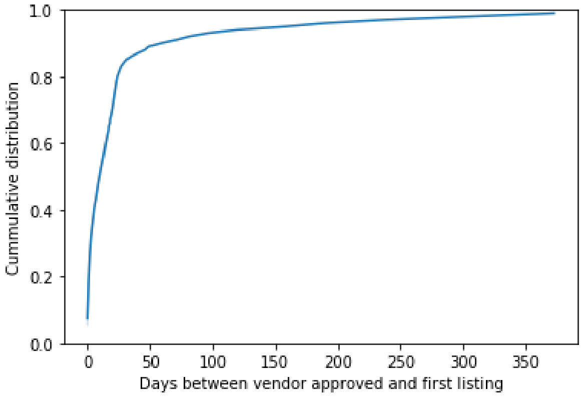

Figure 5.27: Cummulative distrubution function of the time between a vendor getting approved and the creation of the first listing.

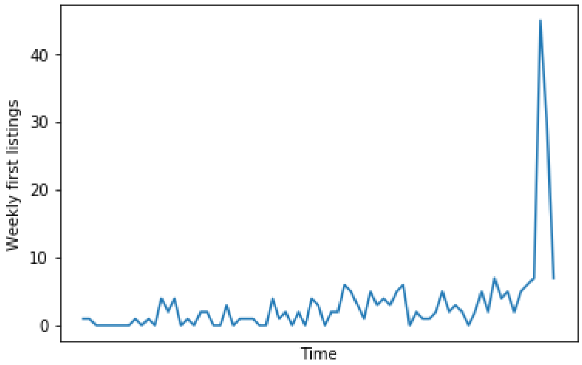

Figure 5.28: Number of first listings created per week over time

When looking at the time that this last 10 percent of the vendors post their first listing (Figure 5.28), it can be seen that there is one big spike. This spike corresponds to a period that another market was shut down. This could mean that some vendors already apply to other markets for the case that the market that they work on gets shut down.

5.7. Market Share

Décary-Hétu and Giommoni (2017) estimated the weekly market share of a vendor by dividing the feedbacks this vendor has in a week divided by the total number of feedback that week. To validate how good this estimation is, we compare eight weeks where we have all the transactions and feedbacks.

In Figure 5.29 it can be seen that the shape of the number of feedbacks over time follows the number of orders over time with a delay. Therefore, a bias is present in the prediction in the weeks that are analyzed (Figure 5.30).

To test how good this method estimates the weekly market share of a vendor, the estimates are calculated and compared to the real market shares. In Tabel 5.7, the R-squared and standard error of the estimate of the estimated market share per week are given. Also, the Herfindahl-Hirschman Index is included as a measure of competition on the market. It can be seen that the estimated market share has an average R-squared of 0.926 and an average standard error of the estimate of 0.0013. However, it can be seen that the standard error of the estimate decreases over time. This can be explained by the Herfindahl-Hirschman Indexes. When looking at the HHI's it can be seen that the competition on the market increases, making the average market share drop. This leads to the decrease of the standard error of the estimate over time.

In Figure 5.31, it can be seen that the market share is off in some cases. This is because

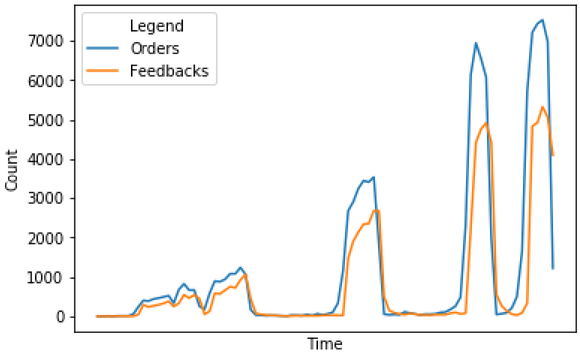

Figure 5.29: Weekly number of orders and feedbacks.

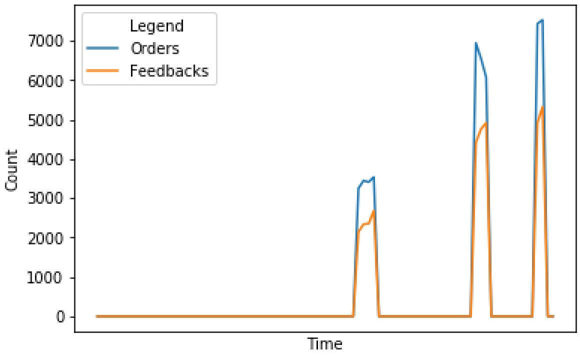

Figure 5.30: Weekly number of orders and feedbacks in the weeks that are taken into account.

feedback is not placed at the time of a transaction. Therefore a week could contain feedbacks of transactions that were done the week before, as well as transactions that don't have their feedback yet. This means that prominent vendors will come out bigger than they are. But thanks to the skew distribution of the market share, it is still possible to identify the ”big players”.

Week	R^2	S	HHI
1	0.951762	0.001537	0.017230
2	0.947444	0.001744	0.019789
3	0.934670	0.001556	0.014301
4	0.938146	0.001672	0.016728
5	0.914465	0.001125	0.009092
6	0.865432	0.001379	0.008707
7	0.907542	0.001237	0.009730
8	0.945873	0.000731	0.007300
9	0.931934	0.000849	0.007682

Table 5.7: R-squared, standard error of the estimate and Herfindahl-Hirschman Index per week

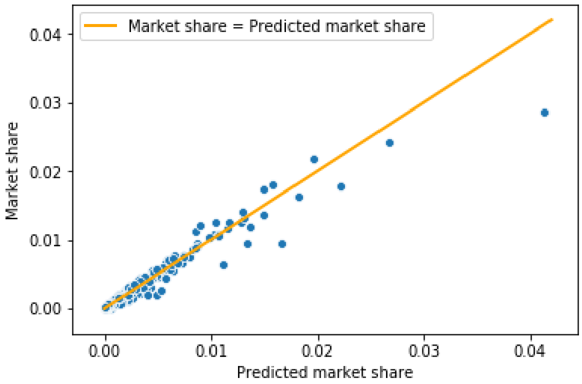

Figure 5.31: Predicted market share vs the real market share for the last week

6 Discussion

In this section, the results of the experiments will be discussed. It will extract all the relevant insights that can be used by scientists and law enforcement agencies to measure dark web markets.

6.1. Holding prices

Holding prices can have an impact on revenue predictions. However, it is not expected that all markets have holding prices. Some markets, like the market under study, give the vendor options to hide their listing or mark it as out of stock. This way, it cannot be bought. This makes the incentive to use holding prices lower compared to the early markets where these options were not implemented. However, it cannot be ruled out that holding prices are still used as it can be a habit of vendors that were also active in the earlier markets.

If it is unknown whether holding prices are used on a market, it is advised to use a method that deals with holding prices when calculating the revenue of a vendor or the total market. The known methods that remove ”holding prices” from the set of observed listings (Aldridge and Décary-Hétu, 2016, Décary-Hétu and Giommoni, 2017, Soska and Christin, 2015) are not good in identifying holding prices when:

1. There is a big increase or decrease in price over time. This happens when the base quantity of a listing is changed.
2. The set includes wholesale listings or custom listings with high prices.

Therefore it is advised to use the method of Kruithof et al. (2016). This method takes the median of the observed prices. The downside of this method is that it cannot be used if you want to observe price changes over time. In that case, the method of Soska and Christin (2015) should be used, as it has a low false positive rate and is expected to have a low false negative rate as well. In future research, it should be inspected the method of Soska and Christin (2015) can be changed to deal with price changes and wholesale listings.

6.2. Feedbacks

The number of feedbacks is a good proxy to use for the number of transactions. However, the number of feedbacks always underestimate the number of transactions since not every buyer leaves feedback. The feedback rate should be used to compensate for this. In the data, an average feedback rate of 0.71 feedbacks per transaction was found. This corresponds to the feedback rate Kruithof et al. (2016) found on Dream Market. This finding substantiates that this method could also be used in other markets. Nonetheless, it should be taken into account that other markets may have different feedback rates as they can be more or less aggressive in nudging the buyers into leaving feedback (Kruithof et al., 2016).

This study also looked at the method to extrapolate the number of feedbacks by using the daily feedback rate (Aldridge and Décary-Hétu, 2014). This extrapolation method does not seem to work. There are three reasons why this method did not work:

- The papers that used this method had a small scrape (< seven days) to extrapolate on, which did not capture the weekly pattern in sales.
- The markets grow over time as more users start to use them. This growth cannot be captured when multiplying the daily feedback rate to estimate a year or month.
- This method does not compensate for the listings that were not observable in the scraping period.

Therefore, it can be seen as bad practice to use the extrapolation method when measuring markets, as it biases the results by underestimating the real number of feedbacks.

When using the number of feedbacks to estimate the transactions, it should be taken into account that a vendor can fake feedbacks with different fake accounts to boost one's reputation. This is also known as a Sybil attack. If a vendor faked transactions, it could be the case that the number of ”real” transactions is overestimated.

6.3. Time of a transaction

The time at which feedback is placed is often (80 percent of the time) the same as the time at which the order is finalized, and thus the time that the vendor gets paid. However, the time of a order is not the same as the time that feedback is placed, which biases the estimation of the paid price corresponding to a listing as used by Soska and Christin (2015). Therefore, the best way to estimate a listing price when it is used to predict the revenue is by taking the median price of the observed listings. This method is also expected to be robust when holding prices would exist in the market.

6.4. Paid price

When estimating the paid price, both the method of taking the median (Kruithof et al., 2016) and the method of taking an observed price that is connected to a feedback (Soska and Christin, 2015) work well. However, in some cases, a feedback cannot be connected to a price observation when all the prices of a listing are marked as a holding price. The connecting method is also slightly biased by the time between the moment of order and moment of feedback. Therefore, it is advised to use the method of Kruithof et al. (2016).

The two methods of estimating the paid price are biased by the shipping cost, quantities, and discount a buyer gets. Shipping costs lead to an underestimation of the paid price as the extra price paid for shipping is not captured by the proxies. The same happens when higher quantities are bought at once, as the proxies only expect a quantity of one. This is problematic in estimating the correct price, as 9 percent of the orders have a quantity greater than one. When discounts are used by the buyers, the paid price is overestimated. This happens because it cannot be observed that a discount is used.

6.5. Revenue

As discussed, the extrapolation method is not precise when estimating the number of transactions. The same is the case when estimating the revenue, as the estimation errors are large. When estimating the revenue of a vendor, the median price (Kruithof et al., 2016) performs the same as connecting the correct price to the feedback (Soska and Christin, 2015). However, some revenues could not be calculated when using the ”connecting method” as some listings only had ”holding prices”. Therefore, the method of taking the median price (Kruithof et al., 2016) is more robust. The prediction of the revenue improves when the number of feedbacks is compensated for the transactions that do not have feedback by dividing by the feedback rate.

All the large outliers in the estimations are due to custom listings. Because the prices of the custom listings are high, the revenue will be underestimated if the custom listing was ordered but does not have feedback. The revenue of a vendor is biased by the shipping costs, quantities, and discounts. The discounts lead to an overestimation, and the shipping costs and quantities lead to an underestimation as they are not taken into account by the estimation methods. If the quantity can be measured, an almost perfect explained variance of 99.9 percent can be achieved.

It should be noted that the estimated revenue is not the same as the profit of a vendor as the market also takes a fee. On some markets, this is a fixed percentage, on other markets this percentage is determined by the number of experience a vendor has. The vendor also has costs for purchasing his products.

6.6. Waiting to post your first listing

The analysis of the time between a vendor getting approved and the first post being created showed an interesting result, for there were vendors that waited more than a month before posting their first listing. It was found that most of these vendors published their first listing at the moment that another big market was taken down. This could mean that these vendors already took precautions to have an approved account on this new market in case the other markets that they are active on are taken down. This can be seen as a strategy for criminals to ensure their business continuity.

6.7. Market share

The market share of a vendor can be estimated with an explained variance of 93 percent based on the number of feedbacks this vendor has. However, the estimation is biased in two ways. First, the result is biased if a vendor orders his listings and fakes some feedbacks to get a better reputation. Secondly, the time of a feedback is not the time at which a transaction is done. Therefore there is a slight delay in the curve of the market share over time. Nonetheless, this method can easily be used to identify the bigger players on the market since they are in the long right tail of the distribution. It should, however, be researched how this method works when a vendor has hidden listings, and the feedbacks cannot be observed.

6.8. Scraping bias

It is expected that the methods above will be influenced when using scraped data. This is because this paper uses market data as if it was a full scrape. However, in practice, it is hard to get a full scrape as crawlers would miss out on hidden listings. This would mean that the number of transactions, revenue, and market share of a vendor will be underestimated.

On some markets, however, it is possible to access hidden listings without getting the link from the vendor. This was possible for the market under study by increasing the identifier of the listing in the URL. Dittus et al. (2018) found that this was also possible on other markets.

If one was to have the scraped data and the corresponding backend data, it could be researched how large this bias is. It would be interesting to know what the scraping percentage is and how far the predictions are off.

6.9. Law enforcement perspectives

In this paper, it is already discussed how precise the proxies are and what biases them. However, when it comes to law enforcement and convicting criminals one should be precise. Therefore this paper would not be complete without writing a section about the usability of these methods in law enforcement investigations. This section will touch upon two use cases. The first being identifying ”big players” on a market. As big players could be defined in multiple ways, this paper talks about the players that earn the most. The second use case is using these methods when seizing the illegally obtained profits of a vendor.

6.9.1. Identifying big players

When trying to identify ”big players” the proxies for market share and revenue could be used. As both proxies measure a different thing - namely the size in revenue and the size in the number of transactions - it would be interesting to use them both. By applying these methods to a scrape of a darknet market, it would be easy to identify big players as they are in the long right tail of the distributions. The identification of ”big players” could be the start of new investigations. Although, it is still unknown how this method works with scrapes as they could contain hidden listings.

6.9.2. Calculating the illegally obtained profits

The estimations should be precise when seizing the illegally obtained profits of a vendor. It should also be possible to clearly explain how one calculated the illegally obtained profits of a vendor. This research found that calculating the revenue by using the median price works best. The formula to calculate the revenue is as follows:

$$\mathrm{R} = \sum_{i=1}^{n} \widetilde{P_i} * q_i \tag{6.1}$$

Where R is the revenue, n is the number of listings, $\widetilde{P_i}$ is the median observed price of listing i. The median is taken to compensate for holding prices. Lastly, q_i is the number of items that are bought of listing i.

The number of purchases should be estimated with as much accuracy as possible. This can be done with scrapes. The best option is counting the quantities that have been sold over time by monitoring the stock of a listing. If this is not possible one should take the number of reviews instead. This will underestimate the revenue in most cases. However, one should find out if the suspect had faked reviews. When feedbacks are faked, one should compensate for the feedbacks by subtracting the number of faked feedbacks from the number of observed feedbacks. To get a better indication of the real revenue, one could compensate for the orders that do not have a feedback by dividing the number of feedbacks by 0.71.

When a vendor has given discounts, the revenue can be overestimated. Therefore, one always needs more information on how much discounts were given and what the discount percentages (δ) were to compensate for this.

It should be noted that the revenue is not the same as the profit of a vendor. To calculate the profit of a vendor, his costs need to be subtracted. The method of calculating the revenue does not include the shipping costs. Therefore there is no need in subtracting the shipping costs. It is, however, needed to subtract the market fee (α). Another cost is the cost (c) that the vendor made to make or purchase the sold item. Lastly, the vendor could also have marketing costs (m) for promoting their items on the market. This leads to the formula for the profit of a vendor on a darknet market:

$$\pi = (\sum_{i=1}^{n} (\widetilde{P_i} * (1-\delta) * (1-\alpha) - c) * (q_i - f_i)) - m \tag{6.2}$$

Where π is the profit, n is the number of listings of a vendor, i is the i’th listing, $\widetilde{P_i}$ is the median observed listing price, c_i is the cost of the vendor to make or purchase the item, q_i is the number of items that are bought of listing i, f_i is the number of faked purchases of listing i, is the market fee percentage, δ is the average percentage of discount a vendor gave on the listing i, and m is the cost the vendor paid for extra promotion on the market.

7
Conclusion

This paper is the first paper that validated the methods of measuring darknet markets. To do so, a novel dataset was used, namely the confiscated database of a big darknet marketplace. This data made it possible to identify the weaknesses in these proxies and give new insights on how to improve those methods.

It is found that the number of transactions, revenue, and market share can be estimated with a high amount of explained variance. The predictions are precise enough to find the big vendors on the market. The designed method of calculating the illegally obtained profits of a darknet vendor is easy to understand and could be used by law enforcement agencies. Finally, it is found that some vendors register to a market early as a strategy to ensure their business continuity.

When the back end and a scrape of a market are available, it would be interesting to analyze what portion is missed when scraping a market. While doing this, it should be researched what the influence is on the proxies that build up to the revenue and market share. But for now, we are one step closer to measuring the criminal activities on dark web markets.

Bibliography

Judith Aldridge and David Décary-Hétu. Not an'ebay for drugs': the cryptomarket'silk road'as a paradigm shifting criminal innovation. *Available at SSRN 2436643*, 2014.

Judith Aldridge and David Decary-Hétu. Cryptomarkets and the future of illicit drug markets. *The Internet and drug markets*, pages 23–32, 2016.

Judith Aldridge and David Décary-Hétu. Hidden wholesale: The drug diffusing capacity of online drug cryptomarkets. *International Journal of Drug Policy*, 35:7–15, 2016.

Julian Broséus, Damien Rhumorbarbe, Caroline Mireault, Vincent Ouellette, Frank Crispino, and David Décary-Hétu. Studying illicit drug trafficking on darknet markets: Structure and organisation from a canadian perspective. *Forensic science international*, 264:7–14, 2016.

Thijmen Calis. Multi-homing sellers and loyal buyers on darknet markets. 2018.

Nicolas Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the 22nd international conference on World Wide Web*, pages 213–224. ACM, 2013.

Nicolas Christin. Measuring and analyzing online anonymous marketplaces. *Understanding the Dark Web and Its Implications for Policy*, 2018.

Jacob Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

David Décary-Hétu and Luca Giommoni. Do police crackdowns disrupt drug cryptomarkets? a longitudinal analysis of the effects of operation onymous. *Crime, Law and Social Change*, 67(1):55–75, 2017.

David Décary-Hétu and Olivier Quessy-Doré. Are repeat buyers in cryptomarkets loyal customers? repeat business between dyads of cryptomarket vendors and users. *American Behavioral Scientist*, 61(11):1341–1357, 2017.

Martin Dittus, Joss Wright, and Mark Graham. Platform criminalism: The'last-mile'geography of the darknet market supply chain. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 277–286. International World Wide Web Conferences Steering Committee, 2018.

Simon Foster, Gerhard Gmel, Natalia Estévez, Caroline Bähler, and Meichun Mohler-Kuo. Temporal patterns of alcohol consumption and alcohol-related road accidents in young swiss men: seasonal, weekday and public holiday effects. *Alcohol and alcoholism*, 50(5): 565–572, 2015.

Kristy Kruithof, Judith Aldridge, David Décary-Hétu, Megan Sim, Elma Dujso, and Stijn Hoorens. *Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands*. RAND, 2016.

Jane Mounteney, Alberto Oteo, and Paul Griffiths. The internet and drug markets: Shining a light on these complex and dynamic systems. *The Internet and drug markets*, pages 13–18, 2016.

Juha Nurmi, Teemu Kaskela, Jussi Perälä, and Atte Oksanen. Seller's reputation and capacity on the illicit drug markets: 11-month study on the finnish version of the silk road. *Drug and alcohol dependence*, 178:201–207, 2017.

Bureau Ontnemingswetgeving Openbaar Ministerie OM. Wederrechtelijk verkregen voordeel hennepkwekerij bij binnenteelt onder kunstlicht, Apr 2005. URL https://www.om.nl/publish/pages/17693/wederrechtelijk_verkregen_voordeel_hennepkwekerij.pdf.

James L Peugh and Craig K Enders. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research*, 74 (4):525–556, 2004.

Kyle Soska and Nicolas Christin. Measuring the longitudinal evolution of the online anonymous marketplace ecosystem. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*, pages 33–48, 2015.

Meropi Tzanetakis, Gerrit Kamphausen, Bernd Werse, and Roger von Laufenberg. The transparency paradox. building trust, resolving disputes and optimising logistics on conventional and online drugs markets. *International Journal of Drug Policy*, 35:58–68, 2016.

Gert Jan van Hardeveld, Craig Webber, and Kieron O'Hara. Expert perspectives on the evolution of carders, cryptomarkets and operational security. 2018.

Rolf van Wegberg and Thijmen Verburgh. Lost in the dream? measuring the effects of operation bayonet on vendors migrating to dream market. In *Proceedings of the Evolution of the Darknet Workshop*, pages 1–5, 2018.

Rolf Van Wegberg, Samaneh Tajalizadehkhoob, Kyle Soska, Ugur Akyazi, Carlos Hernandez Ganan, Bram Klievink, Nicolas Christin, and Michel Van Eeten. Plug and prey? measuring the commoditization of cybercrime via online anonymous markets. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 1009–1026, 2018.

Thijmen Verburgh, Eefje Smits, and Rolf van Wegberg. Perspectieven voor wetenschappelijk onderzoek naar dark markets. *Justitiele Verkenningen*, 44(5), 2018.

A

Code used in this thesis

In this appendix the code for my thesis can be found. The notebooks correspond to the chapters of my thesis. The following notebooks are added as appendices:

3. Cleaning data
   - 3.1 Orders.pdf
   - 3.2 Feedbacks.pdf
   - 3.3 Custom listings.pdf
   - 3.4 Scraping BTC-USD rate.pdf
   - 3.5 Building the data table.pdf
   - 3.6 Analyzing predicted listing prices.pdf
   - 3.7 Stealth and custom listings.pdf
5. Validating proxies
   - 5.1 Holding prices
     - 5.1.1 Testing for Holding prices.pdf
     - 5.1.2 Soska and Christing Holding Prices.pdf
     - 5.1.3 Decary Hetu and Giommoni Holding prices.pdf
     - 5.1.4 kruithof et al (Rand) Holding prices.pdf
   - 5.2 Number of transactions
     - 5.2.1 Feedback rate.pdf
     - 5.2.2 Validating Extrapolating.pdf
   - 5.3 Feedback time
     - 5.3.1 Time between finalize and feedback.pdf
     - 5.3.2 time between transaction and feedback.pdf
   - 5.4 Listing price
     - 5.4 Build up listing prices.pdf
     - 5.4.1 Discounts.pdf
     - 5.4.2 Quantities.pdf
     - 5.4.3 Shipping cost.pdf
     - 5.4.4 Kruithof median vs paid.pdf
     - 5.4.5 Soska and Christing Holding Prices.pdf
   - 5.5 Revenue
     - 5.5.2 Revenue Soska and Christin.pdf
     - 5.5.3 Median times observed feedbacks.pdf
     - 5.5.4 Median times quantities.pdf
   - 5.6 Number of users
     - 5.6.1 Number of vendors based on listings.pdf
   - 5.8 Market share
     - 5.8.1 Market share.pdf