Building a web- or forum crawler to extract (and analyze) scam/extortion emails

Johan Fisch Carl Haglund
Email: {johfi544, carha197}@student.liu.se
Supervisor: Niklas Carlsson, niklas.carlsson@liu.se
Project Report for Information Security Course
Linköpings universitetet, Sweden

Abstract

Scam email has become more and more common lately as more people move towards a more digitized lifestyle. With this comes problems which can be extortion of various kinds. This report gives an overview of the development and use of the web crawler tool and some analysis of the dataset gathered. A total of 495 scam emails were found and these emails showed a pattern of strange email addresses and an email content with poor grammar. Some words were found in a majority of these emails, for example "camera", "malware" and "adult", these can be used as a warning sign of scam emails. The conclusion to be drawn from the content of the report is that scam email content tends to lack proper grammar and often claims to have recorded footage of the targeted person and also that scam email addresses are in most cases unlike a normal email address and tend to look somewhat random.

1. Introduction

Every year more than 3 trillion scam emails are sent, where more than half are phishing scams [3]. This leading to more and more people getting blackmailed to pay a huge amount of money/bitcoin [4].

The purpose of this project is to build a web crawler or targeted forum-crawler to extract scam emails and scam email addresses that have something to do with bitcoin. This can be used in combination with project 9, which is supposed to find information about bitcoin addresses.

Addressing the problem of scam emails, and the use of bitcoin addresses in particular, is of greater importance than ever as people tend to get more and more digitized as well as with the rise of crypto currencies. This paper is aiming to help prevent people from falling for these types of scams. For example the recent outbreak of covid-19 has led to a surge of people receiving scam emails where the senders claim to be legitimate organizations with information about the coronavirus for payment through bitcoin [1-2].

2. Background

The purpose of this section is to give some background about tools and terms that are used throughout the report. This in order to understand the content of the report.

2.1. Reddit

Reddit is a forum where users themselves can create their own subforums about any topic to their liking, these are called subreddits [7]. These subreddits are then moderated by the user who created it, as well as any other moderators they choose to add. The posts in a subreddit can be posted by any reddit user and users can also vote up or down on each post. Posts that receive the most upvotes will be on the subreddits frontpage for a while (usually a day or two, depending on the size and activity of the subreddit). Posts that receive a lot of downvotes will end up disappearing quickly. In each post users can also send comments, vote on comments, as well as respond to other comments. In this report we have used the subreddit called "scams" where they have a post every 6 months (due to users no longer being able to comment once a post is more than 6 months old) that users fill with scam emails they have received. This has proven to be a resourceful place to find the scam emails that were needed.

2.2. Web crawler

A web crawler is a program/system whose task is to browse web pages on the internet for chosen content and to download this [6]. This allows users to retrieve a lot of data from different web pages. An important component of today's web search engines is the web crawler. This enables the search engines to match the issued queries from the users to the web pages which contains the queries.

2.3. API

An application programming interface is a specification of how different programs and systems communicate with each other [8]. This specification allows the parties to communicate without having to know how they are implemented. This can be seen as a contract between the parties since this serves as an agreement on how they talk to each other to send and retrieve data and information of various kinds.

2.4. Bitcoin

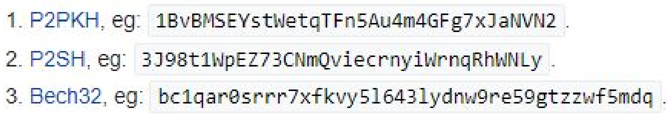
In the publishing article of bitcoin Satoshi Nakamoto starts by explaining how transactions are made with bitcoin [5]. Transactions are made by adding the transaction and the public key of the previous transaction to the blockchain of digital signatures and sending it to the next owner.

Nakamoto also talks about one problem which can occur when people are buying and selling stuff with bitcoin. He calls the problem “double-spend”, which means that a person is trying to spend the same money in multiple places at the same time, trying to trick the blockchain and the seller that he had transferred coins, while in the real case he has not. To avoid this a peer-to-peer network using proof-of-work is used.

Bitcoin helps increase the privacy of the buyer and seller. By using the blockchain and bitcoin people do not have to rely on a third party to handle the transactions (banks normally). Nakamoto also explains that transactions with bitcoin are like transactions on the stock market. People can see that something has been sold or bought but they cannot see who made the transaction.

2.5. Bitcoin addresses

There are a few different address formats for bitcoin addresses. They are rather easy to tell apart because they all start with a unique number or string of letters. The three formats currently used are P2PKH, P2SH, and Bech32. The start of the address is “1”, “3”, and “bc1” respectively, see Figure 1.



1. P2PKH, eg: `1BvBMSEYstWetqTFn5Au4m4GFg7xJaNVN2` .
2. P2SH, eg: `3J98t1WpEZ73CNmQviecrnyiWrnqRhWNLy` .
3. Bech32, eg: `bc1qar0srrr7xfkvy5l643lydnw9re59gtzzwf5mdq` .

Figure 1: Examples for each address format [11].

2.6. Email scams

There are many types of email frauds where a large part of them are about getting the recipient of the email to click on a shady link or to blackmail them into paying an amount of money or bitcoin to the sender [9]. A common type of email scam is the phishing scam. This scam is all about tricking the receiver to provide personal or financial information of some kind to the sender which then uses this to his advantage. Often this type of scam contains content that lacks grammar. You can also find strange URLs in these emails.

The type of email scam where the scammer tries to extort the receiver in some way to make him pay could for example be that the sender says that he hacked the webcam of the receivers computer and recorded

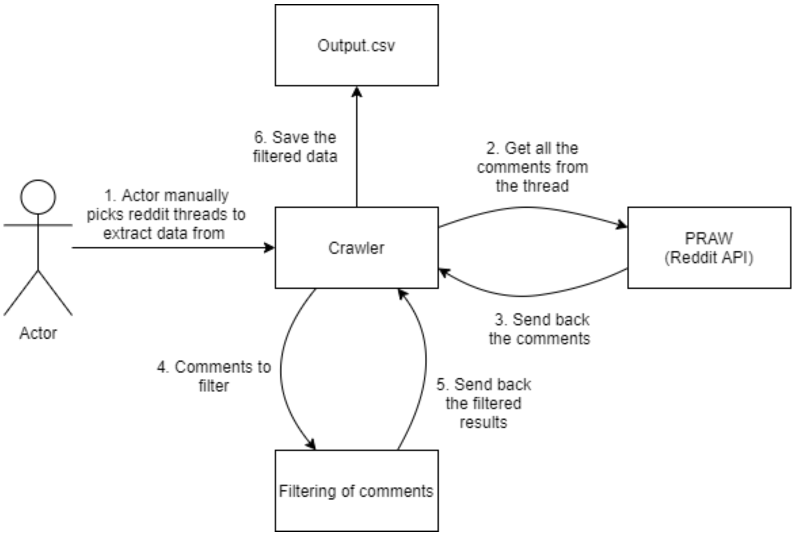

Output.csv
6. Save the filtered data
2. Get all the comments from the thread
1. Actor manually picks reddit threads to extract data from
Crawler
PRAW (Reddit API)
Actor
3. Send back the comments
4. Comments to filter
5. Send back the filtered results
Filtering of comments

Figure 2: Overview of the web crawler

inappropriate stuff and that he will leak it if he does not get paid [10].

3. Method

In this section we will talk about our method of developing the web crawler and the analyzing of the data collected by the web crawler.

3.1. Overview of the web crawler

The web crawler can be described by Figure 2. The first step to be able to run the crawler is for the user to manually collect reddit threads containing relevant information. Once the user has found some of these threads they add them to the list of threads in the crawlers code. When the user is satisfied with the number of threads they have found it is time to move to step two. Step two involves the crawler getting all of the comments from the thread using a Reddit API and in step three they are sent back to the crawler. When the crawler has all of the comments from a thread it will move on to filtering. In step four the comments are filtered for different things (explained further below). Once the filtering is done the filtered data is sent back to the crawler. These five steps are then repeated for every thread the user has added to the list. Once all of the threads have been filtered the crawler will move on to step number six, which is to save all of the data in a CSV (comma separated values) file.

3.2. PRAW - A Reddit API

For the web crawler we used an API called "PRAW" which stands for "Python Reddit API Wrapper". This API simply allows access to the data on Reddit in a quick and simple manner. The program takes a list of reddit threads to search. With help from PRAW, the program creates a reddit instance which is then used to access the threads. Using the reddit instance, a submission object from the reddit thread URL is now created and then used to access the comments in this submission.

3.3. Filtering of comments

When the comments have been collected the program will then search through all the comments that contain anything like "bitcoin", "btc" or "bitcoins" to find relevant scam emails. Once a relevant post has been found a search for a bitcoin address is done for the post. If there is no bitcoin address found, move on to the next post. The bitcoin address will be verified (explained further in 3.4) and if there is a bitcoin address found, a search for an email will begin. Only the bitcoin address and email content will be saved if no email address is found.

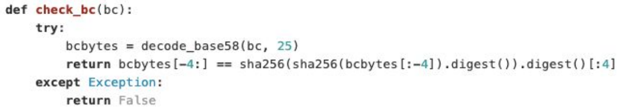
The program will also check if there are any of the bitcoin addresses that occur more than once. These will be displayed in the terminal with the number of times they appear, see Figure 4.

3.4. Verifying the bitcoin address

A bitcoin address is encoded with base58. This makes the addresses 25 bytes when encoded. To verify a bitcoin address you need to make sure that the checksum of the first twenty-one bytes compares to the last four bytes. This was made by the program which can be seen in Figure 3.

3.5. File output

Once all of the data has been filtered and searched the crawler will go on to save it in a CSV file. The format of the CSV file is: "Bitcoin address, Email address, Email content".



```
def check_bc(bc):
    try:
        bcbytes = decode_base58(bc, 25)
        return bcbytes[-4:] == sha256(sha256(bcbytes[:-4]).digest()).digest()[:4]
    except Exception:
        return False
```

Figure 3: Program to verify a bitcoin address [12].

Bitcoin address	#
1PL9ewB1y3iC7EyuePDoPxJjwC4CgAvWTo	4
15ZHnf1MPn6ybb8yUeAoCQ1AJtiKhg3NrP	4
3K3vVqkxeDeD8Qbex4MzXe2WdAcfw2WXzF	4
1NQrcoefW8Ky33oEMC57vqD6KuFY4h7crS	3
12DA8mpQCnTB1cEHLPFU7ckP44zN5Xmgu3	3
17zmnmqEUCesNz6UgXGbRk7fKnu8iq1q2J	3
1PzrJSAhZSiYK93qLZnKsRzQzS49j5Ugzc	3
15yF8WkUg8PRjJehYW4tGdqcyzc4z7dScM	3
17XHRucfd4kx3W5ty7ySLGiKHqmPUUdpus	2
1DEbZ7uqJYZpVcB3tBH2kGe9bucBjX3buS	2
194iizBy5K9AVDqTBvzDAWR6t9MrrqvseZ	2
1LisQiLaiwa1HZutVFKWhPqfKVUK96Yej8	2
16fMSCFEXULPbXNGEvemMy7dnz9Tu3qKWL	2

1D1DZAac5chXcvULdRAk8nbxB5HWWbffwc	2
1BzkoGfrLtL59ZGjhKfvBwy47DEb6oba5f	2
17iRfpgSwmJ6nLXR8evx6pUBo3R33S5LXB	2
1PhAzthZMqAaFHBAEDLinbNk6yZBVVfyrr	2
1PhAzthZMqAaFHBAEDLinbNk6yZBVVfyrr	2
1Cke5M3SejwCQZTANAHw4MnJA1zPCS8UFt	2

Figure 4: Some of the duplicate addresses found and how many times they appeared.

3.6. Analysis of extracted data

When the dataset of bitcoin addresses and scam emails are gathered an analysis of this will take place. This will partly be done by checking for duplicates of bitcoin and email addresses. By doing this different patterns can be found, for example if scammers tend to send email with the same content or if they change it from time to time.

3.7. Limitations

The project has been limited to only extract information about scam emails that contain a bitcoin address. This means that the web crawler developed only looks for posts containing bitcoin and bitcoin addresses.

4. Results

This section will cover the result of the web crawler tool and the data collected by it.

4.1. Bitcoin addresses

Looking at the dataset which was extracted using the web crawler, the number of bitcoin addresses found numbered to 448 unique addresses and 495 total found. Meaning that there are some duplicate addresses found through the search. We found a total of 432 addresses that were of the type P2PKH, 43 of the type P2SH and 20 of the type Bech32. This means that 87% of all addresses found were of the type P2PKH, 9% was P2SH and 4% was Bech32.

4.2. Email addresses

The email addresses which were gathered show a pattern. A majority of the addresses which were extracted by the web crawler were suspicious, like “yokoji-0629.0425@zeus.eonet.ne.jp”, “wwwbattlesnakes@ferrari2.serverchamber.com” and “David01@9973.com”. The addresses, however, could also look like any normal email address, for example “kentonvsomentenwa@outlook.com” which has a first and last name in it.

4.2. Email content

The content of the emails are often very similar, some kind of threat against the users privacy in an attempt to blackmail them. The bitcoin addresses that have occurred more than once usually have the same email content even if the scammer uses different email addresses. The emails are for the most part poorly worded and contain poor grammar. A recurring theme regarding the content of the emails is that the scammer claims that they have hacked into the recipients computer and recorded them through the webcam and that they now have some inappropriate video of the recipient. The scammer then wants the recipient to pay them some amount of bitcoin to not spread this video. The amount required is usually between 1000 and 3500 in US dollars.

The email in Figure 5 was sent four times by four different email addresses. The content and the bitcoin address remained the same.

> From Raymond Philips <Robertsn@yxzuche.com>
> "i know ******, is one of your password. I've recorded your cam while you were watching porn on XX sites, also I've installed a
> keylogger on ur pc & collected all your contacts on social networks, messenger & emails. If you want me to erase the recording, pay me 1128$ on bitcoin address: 3K3vVqkxeDeD8Qbex4MzXe2WdAcfw2WXzF (search in Google for "how to buy bitcoin"), [case SenSitiVe so copy & paste it]. If I don't get the bitcoins, I will definately send your video to all of your contacts, don't reply to this email it's hacked. LeZnqO"

Figure 5: Example of a scam email.

As seen in Figure 6 there are some words that are very common in these types of emails.

5. Analysis and discussion

This section will cover the discussion and analysis of the results presented in section 4.

5.5. Email and bitcoin addresses

More than 3 billion scam emails are sent each day [3], out of all of these we extracted 495 which included a bitcoin address. As seen in the result the webcrawler extracted 448 unique bitcoin addresses and 495 in total. It is interesting to see that there are some bitcoin addresses that appear more than one time. Especially since the source we used is rather small in contrast to how many scam emails that are sent. Statistically you could claim that there should not be people who have been targeted by the same scammer with the same bitcoin address.

From the email addresses obtained there is a noticeable pattern that many of them are somewhat random. This could be used as a warning signal, and if you receive an email from a random email address the risk that it is a scam is rather high. The randomness could also be thought of if you were to develop a scam email filter of some kind. By looking for randomness in the email address you will blacklist many of the scam addresses. Although, this would not be near enough to be a good spam/scam filter and it could also blacklist legit people who for some reason use a random email address.

By looking at the graph in Figure 4 it is easy to see that the scammers tend to use the act of visiting adult websites as a main tactic, judging by the high number of related terms such as “porn” and “adult”. The scammer will usually also claim to have the targets contact list or their social network contacts and that they mean to spread the footage to these contacts.

The poor grammar of the emails could mean that the blackmailer is from a country where the english education is lacking or the blackmailer might not even have studied english for very long. The poor and weird grammar might also be a way for the scammers to avoid email filters as the words are no longer as easy to detect.

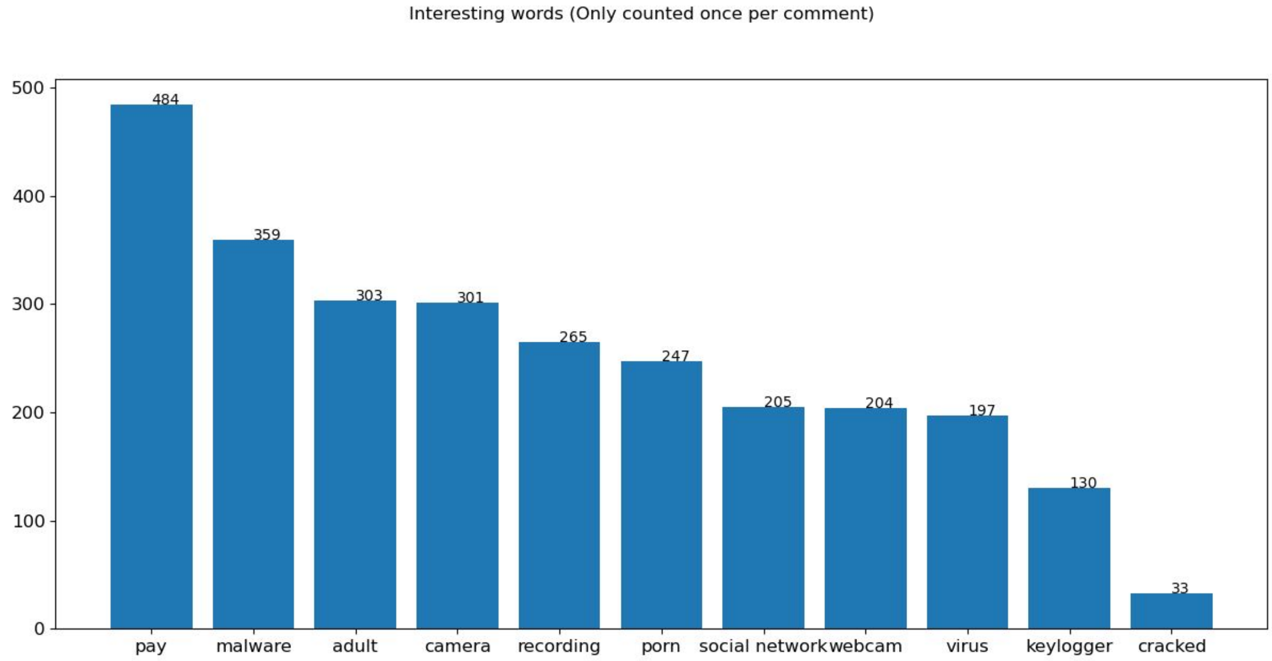


Figure 6: Plot of recurring words

5.5 Email content

As mentioned before the emails are often related to the blackmailer claiming to have some kind of footage of the target which is confirmed by the high number of appearances of words such as “webcam”, “recording”, and “camera”. The high volume of words like this might originate in some kind of fear humans have for being exposed naked on the internet which in turn leads to a higher chance for the scammers to get paid. This since you can not really be 100 percent sure that someone has not recorded you if you did not cover your camera.

6. Conclusion

From this project and the results and analysis of it the following conclusions can be drawn:

- Scam email content tends to lack proper grammar and often claims to have recorded footage of the targeted person.
- Scam email addresses are in most cases unlike a normal email address and tend to look somewhat random.
- A majority of the emails included words like “camera”, “malware” and “adult”. This can be seen as a red light if a person is not sure if an email is a scam email.

7. Future work

For future development of the crawler one addition is to add validation support for Bech32 bitcoin addresses. The crawler in its current state is only able to validate P2PKH and P2SH formatted bitcoin addresses (it is still possible to collect Bech32 addresses, but they’re not validated as real addresses).

References

[1] NortonLifeLock Inc. [Online; accessed 15-April-2020] https://us.norton.com/internetsecurity-online-scams-coronavirus-phishing-scams.html

[2] Sinclair, Sebastian. Mar 27 2020. [Online; accessed 16-April-2020] https://www.coindesk.com/uk-counties-warn-of-bitcoin-scams-using-coronavirus-as-a-hook

[3] Spadafora, Anthony. June 12 2019. [Online; accessed 16-April-2020] https://www.techradar.com/news/one-trillion-phishing-emails-sent-every-year

[4] Fazzini, Kate. Email sextortion scams are on the rise and they’re scary — here’s what to do if you get one. June 17 2019. [Online; accessed 16-April-2020] https://www.cnbc.com/2019/06/17/email-sextortion-scams-on-the-rise-says-fbi.html

[5] Nakamoto, Satoshi. Bitcoin: A peer-to-peer electronic cash system. Manubot, 2019.

[6] Christopher Olston and Marc Najork (2010), "Web Crawling", Foundations and Trends® in Information Retrieval: Vol. 4: No. 3, pp 175-246. http://dx.doi.org/10.1561/1500000017

[7] Widman, Jake. What is Reddit?. 2020. [Online; accessed 18-April-2020] https://www.digitaltrends.com/web/what-is-reddit/

[8] Red-Hat. 2020. [Online; accessed 18-April-2020] https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces

[9] Rafter, Dan. Norton. 2020. [Online; accessed 19-April-2020] https://us.norton.com/internetsecurity-online-scams-phishing-email-examples.html

[10] Gralla, Preston. Broadcom. 2019. [Online; accessed 18-April-2020] https://symantec-blogs.broadcom.com/blogs/feature-stories/why-email-extortion-schemes-are-skyrocketing-and-how-protect-yourself-against-them

[11] Address. Bitcoin Wiki. 2019. https://en.bitcoin.it/wiki/Address [Online; accessed 18-April-2020]

[12] Bitcoin/address validation. Rosetta Code. 2020. https://rosettacode.org/wiki/Bitcoin/address_validation#Python, [Online; accessed 20-April-2020]