Crawling and Analysis of Dark Network Data

Ying yang
Qilu University of Technology (Shandong Academy of Science) Shandong Computer Science center (National supercomputer Center in Jinan), Shandong Provincial Key Laboratory of Computer Networks Jinan, China
yangy@sdas.org

Guichun Zhu
Qilu University of Technology (Shandong Academy of Science),Shandong Computer Science center(National supercomputer Center in Jinan), Shandong Provincial Key Laboratory of Computer Networks Jinan, China
sdcenter@163.com

Lina Yang
Qilu University of Technology (Shandong Academy of Science), Shandong Computer Science center (National supercomputer Center in Jinan), Shandong Provincial Key Laboratory of Computer Networks Jinan, China

Huanhuan yu
Qilu University of Technology (Shandong Academy of Science) Shandong Computer Science center (National supercomputer Center in Jinan),Shandong Provincial Key Laboratory of Computer Networks Jinan, China

ABSTRACT

Due to its anonymity and non-traceability,it is very difficult to research websites on the dark network. The research of the dark network is very important for our network security. Now there is very little data for studying the dark network,so we independently developed dark web crawler that runs automatically. This article will detail the implementation process of our dark web crawler and the data analysis process of crawled data. Currently,we can use crawled data to detect if multiple urls belong to the same site. We can use data to extract features of similar websites and we have generated an ever-increaing data set that can be used for simple website classification.We use the crawled data as a categorical dataset to categorize newly discovered urls.When we get the a certain number of new urls,we crawl again and the crawled data will be added to the previous data set. After multiple rounds of crawling,our data sets will be more and more abundant. through our approach,we can solve the problem that the dark network data is small,researchers can use our method to get enough data to study all aspects of the dark network.

CCS Concepts

• **Security and privacy→Security services→Pseudonymity, anonymity and untraceability.**

Keywords

Dark net ; crawl ; data analysis ; darknet data set

1. INTRODUCTION

1.1 Dark Net

The dark net is only a small part of the deep net, but because of its anonymity,it has become a paradise for criminals,a growing number of illegal behavior take place in the dark networks [9], because it is anonymous,we effectively regulatory control is difficult. So the dark network is full of drugs, pornography,and violent content. A large number of illegal transactions occur every day.Dark net also has a good side,such as protecting privacy and communicating anonymously.Facebook in launch its own online website in dark net, after about a year,the number of Facebook pages that visited the dark has increased to 52.5 million. As of april 2016, this number has further increased,eventually breaking the record of 100 million records per month.[1],as we can see, the dark network is not only the dark side.Common dark net includes P2P networks, Freenet and Tor net,if you want to access the dark net,it requires specialized tools such as tor (onion router), I2P and Freenet, Tor is the most popular tool.

1.2 tor network

In the dark net,the most people used tor.Tor uesd all based on TCP application protocol,when we pass tor access to relevant network resources,such as visiting the site,it will downloads all tor node information from the directory server, and then based on each node announcement bandwidth, online duration, the set exit access policy and other factors select three nodes as the entry section [2]. In the tor network, the communication between the user and the server needs to go through many routers. In addition, communication between any two routers uses a different encryption key [6]. Therefore, no one can track where the real users and servers are, so anonymous services can be guaranteed.

1.3 Data collection

Dark Web site data with the particularity and sensitivity characteristics,publicly available data set is not much, at present, Gwern Branwen collect published data sets are popular, the data set has been published for several years, because the dark network website generally has a short life cycle, many websites in the data sets have been invalidated . If this data set is used again, it will not reflect the characteristics of the current dark network. The more popular dark data collection tool is onionscan, which can scan out the privacy leakage risks in the dark web site, as the dark web develops, less things can be scanned. The network has a number of open-source darknet collection system, but they can’t meet our needs, they lack versatility, and some of the collection system are complex and difficult to install successfully, so the best way to collect data is to write our own data collection tools, this article has independently developed a crawler system that is easy to install and use.

1.4 Introduce the structure of the article

In the first section, we introduces the basic knowledge of the dark net, which can have a basic understanding of the dark net. The second part provides a brief overview of the dark web crawler

ICCDE '20, January 4 – 6, 2020, Sanya, China

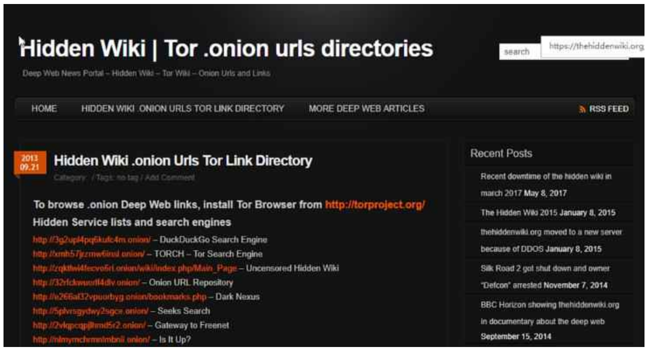
ACM ISBN 978-1-4503-7673-0/20/01…$15.00
https://doi.org/10.1145/3379247.3379272

system and the purpose of the analysis, which helps to understand the general function of our crawler system. The third part introduces the specific implementation steps of the crawler and the analysis of the data set. We'll explain the key steps in running a crawler system to help more people develop their own crawler systems. The fourth part introduces the usefulness of the dataset. we verify the effect of the dataset implementation classification and use the dataset to find sites with multiple urls.The five sections summarize our papers and give a brief introduction of our future work.

2. OVERVIEW OF THE CRAWLER

2.1 The difficulty of crawler

It is difficult to obtain the dark web address from the normal web, the dark web URL is not named regularly and different from the web address www.google.com that is simple and easy to remember. The dark web urls is a random combination of a series of letters and numbers. The dark URL is similar to hfdjsaj3jndsjcnjschffhvjdwwkn34n.onion, such a URL is impossible to remember by our brain, because of the special nature of the dark network and its short life cycle, it is difficult to obtain a dark web site, If you want to study a specific category of dark network, one URL is not enough, there must be enough urls ,to get a lot of web site also is difficult. There are a lot of posts on the Reddit website [3],dark web url also can be found on the dark web site. we found some dark web site urls in the normal web, but 70% website has been shut down, the figure1 is a normal network website that published out of the dark network website.



Hidden Wiki | Tor .onion urls directories

Deep Web News Portal – Hidden Wiki – Tor Wiki – Onion Urls and Links

search
https://thehiddenwiki.org
HOME
HIDDEN WIKI .ONION URLS TOR LINK DIRECTORY
MORE DEEP WEB ARTICLES
RSS FEED
2013
09.21

Hidden Wiki .onion Urls Tor Link Directory

Category: / Tags: no tag / Add Comment

To browse .onion Deep Web links, install Tor Browser from http://torproject.org/

Hidden Service lists and search engines

http://3g2upl4pq6kufc4m.onion/ – DuckDuckGo Search Engine
http://xmh57jrzrnw6insl.onion/ – TORCH – Tor Search Engine
http://zqktlwi4fecvo6ri.onion/wiki/index.php/Main_Page – Uncensored Hidden Wiki
http://32rfckwuorlf4dlv.onion/ – Onion URL Repository
http://e266al32vpuorbyg.onion/bookmarks.php – Dark Nexus
http://5plvrsgydwy2sgce.onion/ – Seeks Search
http://2vlqpcqpjlhmd5r2.onion/ – Gateway to Freenet
http://nlmymchrmnlmbnii.onion/ – Is It Up?

Recent Posts

Recent downtime of the hidden wiki in march 2017 May 8, 2017

The Hidden Wiki 2015 January 8, 2015

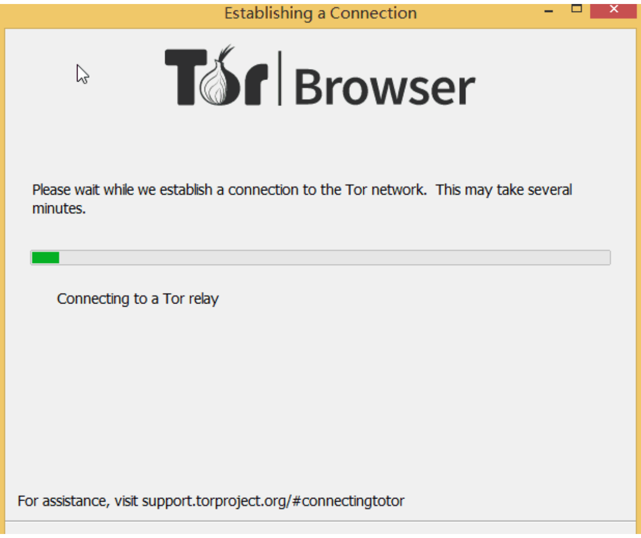
thehiddenwiki.org moved to a new server because of DDOS January 8, 2015

Silk Road 2 got shut down and owner "Defcon" arrested November 7, 2014

BBC Horizon showing thehiddenwiki.org in documentary about the deep web September 15, 2014

Figure 1 darknet url

The dark network login requires a tor agent, so it requires a specific browser to log in.The dark network is different from the ordinary network, it is anonymous, so access to it requires special tools, if the terminal logged into the dark network, use the pip install tor command to install it, and then set up a proxy, in general, the window of the login method is more convenient, we used tor browser, first go to the official website to download the tor browser, set up the proxy and then you can access the Internet like a normal browser. the page of the tor browser is as shown figure 2.



Please wait while we establish a connection to the Tor network. This may take several minutes.

Connecting to a Tor relay

For assistance, visit support.torproject.org/#connectingtotor

Figure 2 Tor browser startup

The dark web site has a short life cycle, because many websites are engaged in illegal activities, they are either shut down by law enforcement agencies or due to poor management. So getting the crawled darknet site may expire in a few days, so our crawler needs to crawl for a long time.

2.2 Achieve the function

Automatic crawling: we give the crawler a set of urls that are the same category of darknet, the crawler will crawl automatically. The system crawl the website by category. Later, the crawler will use the crawled data to generate a data set and directly classify the newly discovered URLs, so the data sets will become more and more abundant.

File save: the crawler saves the required content,the current content includes the text content of the website,the title and the onion address that we have not crawled.

Data processing: We need data that can be used for classification. In general, crawling data cannot be used directly, so data needs special processing to meet our requirements. When crawling dark web, some websites are not accessible,so you first need to remove empty files,after a simple process, we got the data that we wanted.

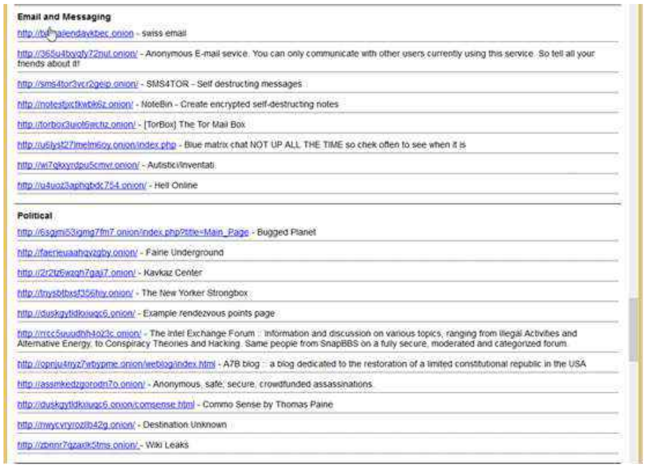
2.3 Data Analysis

Our first task is to generate a data set for classification. the data set obtained by our crawling. For the data, we will further analyze it. we can find a website with multiple url addresses, and we can get some specific information for a specific category of websites.

3. CRAWLER IMPLEMENTATION

3.1 Get the dark web URL

There are some sites or blog will provide some dark web site, such as: https://thehiddenwiki.org/, Through the open source project Onionscan, you can not only discover the threat of privacy breaches in dark networks, it also can add new URLs to the dark web list and it can crawl new urls continuously, through it we can get urls [10] continuously, but the list is messy. We get url through the dark web, some dark webs will provide dark web URLs, and they are classed, which is what we need, through the dark web dirnxxdraygbifgc Onion, we can get to a variety of specific categories [8] of urls, which greatly facilitated our work , Figure 3 illustrates the URL we obtained.



Email and Messaging

http://tt[illegible]alendaykbec.onion - swiss email

http://365u4bxqfy72nut.onion/ - Anonymous E-mail sevice. You can only communicate with other users currently using this service. So tell all your friends about it!

http://sms4tor3vcr2geip.onion/ - SMS4TOR - Self destructing messages

http://notestjxctkwbk6z.onion/ - NoteBin - Create encrypted self-destructing notes

http://torbox3uiot6wchz.onion/ - [TorBox] The Tor Mail Box

http://u6lyst27lmelm6oy.onion/index.php - Blue matrix chat NOT UP ALL THE TIME so chek often to see when it is

http://wi7qkxyrdpu5cmvr.onion/ - Autistici/Inventati

http://u4uoz3aphqbdc754.onion/ - Hell Online

Political

http://6sgjmi53igmg7fm7.onion/index.php?title=Main_Page - Bugged Planet

http://faerieuaahqvzgby.onion/ - Fairie Underground

http://2r2tz6wzqh7gaji7.onion/ - Kavkaz Center

http://tnysbtbxsf356hiy.onion/ - The New Yorker Strongbox

http://duskgytldkxiuqc6.onion/ - Example rendezvous points page

http://rrcc5uuudhh4oz3c.onion/ - The Intel Exchange Forum :: Information and discussion on various topics, ranging from Illegal Activities and Alternative Energy, to Conspiracy Theories and Hacking. Same people from SnapBBS on a fully secure, moderated and categorized forum.

http://opnju4nyz7wbypme.onion/weblog/index.html - A7B blog :: a blog dedicated to the restoration of a limited constitutional republic in the USA

http://assmkedzgorodn7o.onion/ - Anonymous, safe, secure, crowdfunded assassinations.

http://duskgytldkxiuqc6.onion/comsense.html - Commo Sense by Thomas Paine

http://nwycvryrozllb42g.onion/ - Destination Unknown

http://zbnnr7qzaxlk5tms.onion/ - Wiki Leaks

Figure 3 website list

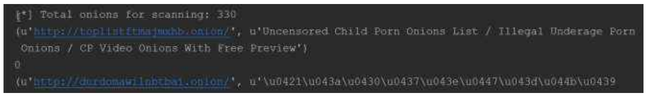
3.2 Running dark web crawlers

Our crawler system use python 's selenium module to simulate login, using beautifulsoup module to get data, crawler can run in any python environment, if we want to run it, we mainly need to install selenium,re and beautiful soup module, after the installation is complete, system call selenium module to crawl the dark web site. The specific steps are as follows:

Tor is essential for logging in to the dark network[7]. We use the tor browser. The browser downloads from the official website. After the installation is completed, if the network cannot directly connect to the tor, it will need the local proxy settings. The proxy type has socks4, socks5 and http//https. If you use the gaogent proxy, just set the proxy type to " http/https ", set the address to " 127.0.0.1 ", and set the port to " 8087 ". Similarly, we can set other SCOKS class agents and so on.

The main modules used are the selenium module and the beautifulsoup module. The selenium module is used to simulate and control the browser. To use this module, you also need to download the kernel driver of the browser[5]. The tor browser is based on firefox, so we can download the firefox driver, after the download is complete, this module will be available. The beautifulsoup module is a module that is used to match specific content. Then system crawl the dark web by category, we crawl the URL of each website and extract the title, new onion url and text content in the web page. For text extraction we use the beautifulsoup.text method. For url extraction we use the re module, if there is a new dark web addresses,we will saved it, eventually we will get a list of darknet urls. Later, for new urls, we can use the constructed categorical dataset to categorize it.

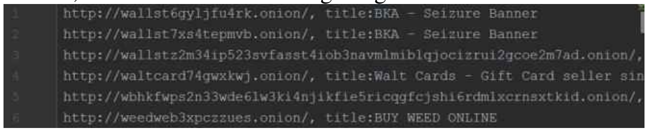
When you start running the crawler, the page will display the number of darknet urls. When you crawl a website, the page displays the URL of the crawling website and the website title. If the title shows a messy character, it indicates that this url failed to crawl. In figure4, the first URL is successfully crawled, the second url crawl failed:



```
[*] Total onions for scanning: 330
(u'http://toplistftmajmxhb.onion/', u'Uncensored Child Porn Onions List / Illegal Underage Porn
 Onions / CP Video Onions With Free Preview')
0
(u'http://durdomawilnbtbai.onion/', u'\u0421\u043a\u0430\u0437\u043e\u0447\u043d\u044b\u0439
```

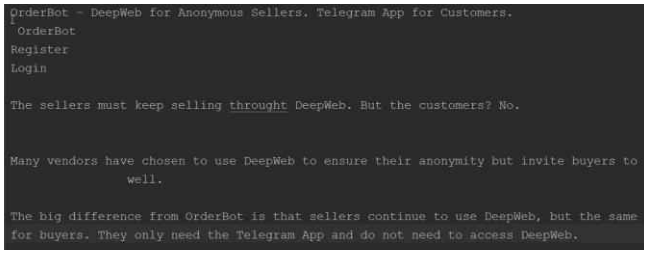
Figure 4. crawl page

We constructed two datasets, one containing only the website URL and the title, and the other containing all the text content of the site, as shown in the following image:



```
http://wallst6gyljfu4rk.onion/, title:BKA - Seizure Banner
http://wallst7xs4tepmvb.onion/, title:BKA - Seizure Banner
http://wallstz2m34ip523svfasst4iob3navmlmiblqjocizrui2gcoe2m7ad.onion/,
http://waltcard74gwxkwj.onion/, title:Walt Cards - Gift Card seller sin
http://wbhkfwps2n33wde6lw3ki4njikfie5ricqgfcjshi6rdmlxcrnsxtkid.onion/,
http://weedweb3xpczzues.onion/, title:BUY WEED ONLINE
```

Figure 5. data set one



```
OrderBot - DeepWeb for Anonymous Sellers. Telegram App for Customers.
 OrderBot
Register
Login

The sellers must keep selling throught DeepWeb. But the customers? No.


Many vendors have chosen to use DeepWeb to ensure their anonymity but invite buyers to
            well.

The big difference from OrderBot is that sellers continue to use DeepWeb, but the same
for buyers. They only need the Telegram App and do not need to access DeepWeb.
```

Figure6 data set two

3.3 data processing

We use the crawled text content to construct our dataset, we need to remove the url address for the first data set, if we fail to access the site, we will generate an empty file.It belong to the abnormal data ,we need to remove the abnormal data. Using the data set, we can classify the new URLs during our crawling process, so our data sets will become more and more abundant.

3.4 Data Analysis

For the title, if two sites have the same site title, the two websites are likely to be the same website. Therefore, the keywords of the two websites are extracted,if the keywords are the same, then the two sites can be considered a website.The method can reduce the repetition rate of the website, which is very helpful for evaluating the size of the dark network [11]. For one class site we have a comprehensive analysis, the analysis results will generate a word cloud,by figure7 and figure8, we can have a general understanding of the characteristics of such websites, for the two categories of websites, we can see that there are significant differences.

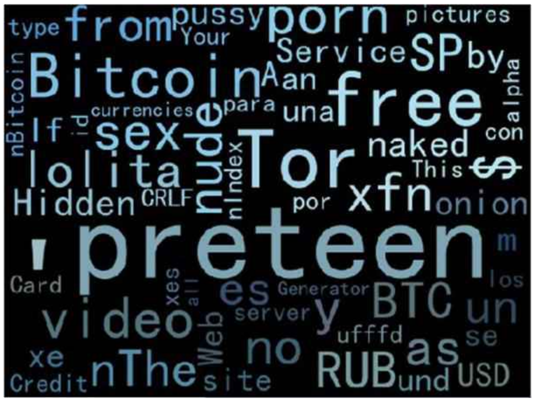


Figure 7 . all darknet word cloud

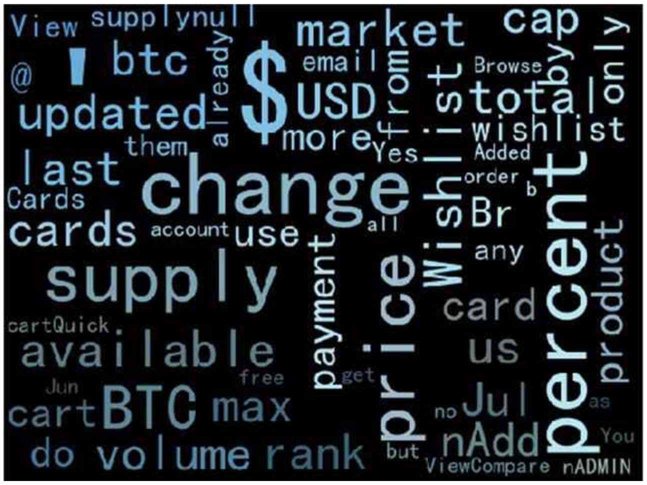


Figure 8. Encrypted market word cloud image

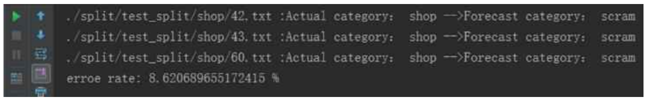
By analyzing the crawling results, we briefly summarize some of the most frequently occurring words on various websites. The frequency of occurrence of the words is shown in the table1:

Table 1 dark net high frequency word list

Word rank	All website	Market website	Financial website
1	Preteen	$	BTC
2	Tor	Change	Number
3	Free	Percent	Card
4	Bitcoin	Supply	Credit
5	Free	Price	Information
6	Sex	BTC	Account

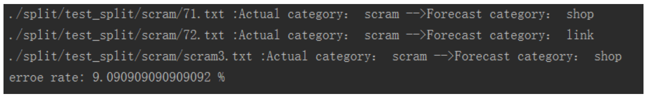
4. DATA SET VERIFICATION

We use the support vector machine to run in the generated data set, and the running classification result reaches 90% correct rate, which is enough for our research needs. In actual research work, we do not need a particularly high accuracy rate, as long as we can get enough websites of the same category.We use our crawled data to generate two data sets, a data set that contains the site title, the data named data_set1,another data set that contains the title and text of the site, the data named data_set2, the results show they all can achieve a good classification effection, the running results are as follows :



```
./split/test_split/shop/42.txt :Actual category:  shop -->Forecast category:  scram
./split/test_split/shop/43.txt :Actual category:  shop -->Forecast category:  scram
./split/test_split/shop/60.txt :Actual category:  shop -->Forecast category:  scram
erroe rate: 8.620689655172415 %
```

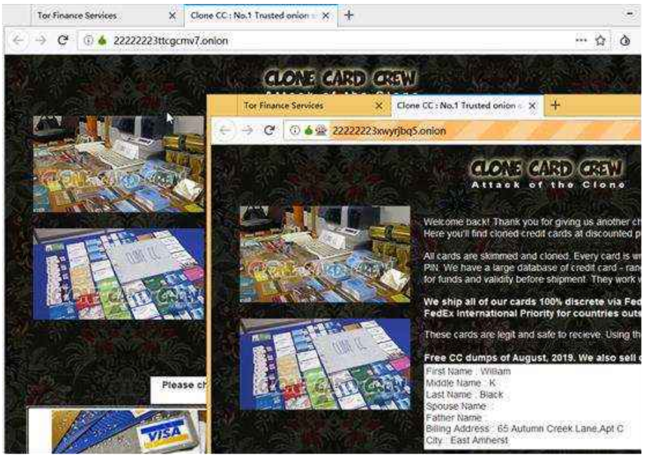
Figure 9   data_set1 running result



```
./split/test_split/scram/71.txt :Actual category:  scram -->Forecast category:  shop
./split/test_split/scram/72.txt :Actual category:  scram -->Forecast category:  link
./split/test_split/scram/scram3.txt :Actual category:  scram -->Forecast category:  shop
erroe rate: 9.090909090909092 %
```

Figure 10 data_set2 running result

Through data analysis, we found that many websites have different urls. Previous studies have estimated the size of dark nets by the number of dark web sites, this is inaccurate. We need to further explore whether different urls are the same site. As figure11 shown, although the URLs are different, they are the same website.



Welcome back! Thank you for giving us another ch
Here you'll find cloned credit cards at discounted p

All cards are skimmed and cloned. Every card is w
PIN. We have a large database of credit card - ran
for funds and validity before shipment. They work w

We ship all of our cards 100% discrete via Fed
FedEx International Priority for countries outs

These cards are legit and safe to recieve. Using th

Free CC dumps of August, 2019. We also sell c
First Name : William
Middle Name : K
Last Name : Black
Spouse Name :
Father Name :
Billing Address : 65 Autumn Creek Lane,Apt C
City : East Amherst

Figure11 same site with different urls

5. CONCLUSION

We developed the dark web crawler and generated a data set. when we get a new URL by our crawlers, we will further crawl and add it to our dataset,through this method, we can get enough data. As the data set continues to increase, it will satisfy our understanding of the dark world. In the future work, we will continue to enrich the data set and develop a more convenient dark network analysis platform.

6.REFERENCES

[1] Jardine, E. (2018). Privacy, censorship, data breaches and Internet freedom: The drivers of support and opposition to Dark Web technologies. *new media & society*, *20*(8), 2824-2843.

[2] He Gaofeng, Yang Ming, Luo Junzhou, & Zhang Wei. (2013). Tor Anonymous Communication Traffic Online Identification Method. Journal of Software, 24(3), 540-556.

[3] Reddit. Available online: https://www.reddit.com/r/darknetmarkets (accessed on 6 August 2019)

[4] Ghosh, S., Das, A., Porras, P., Yegneswaran, V., & Gehani, A. (2017, August). Automated categorization of onion sites for analyzing the darkweb ecosystem. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1793-1802). ACM.

[5] Wu, Y., Zhao, F., Chen, X., Skums, P., Sevigny, E. L., Maimon, D., ... & Zhang, Y. (2019, July). Python Scrapers for Scraping Cryptomarkets on Tor. In *International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage* (pp. 244-260). Springer, Cham.

[6] Alex Biryukov , Ivan Pustogarov , Fabrice Thill , Ralf-Philipp Weinmann, Content and Popularity Analysis of Tor Hidden Services, Proceedings of the 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops, p.188-193, June 30-July 03, 2014 [doi>10.1109/ICDCSW.2014.20]

[7] A. Biryukov, I. Pustogarov, and R.-P. Weinmann. 2013. Trawling for Tor Hidden Services: Detection, Measurement, Deanonymization. In IEEE-SP.

[8] Nicolas Christin, Traveling the silk road: a measurement analysis of a large anonymous online marketplace, Proceedings of the 22nd international conference on World Wide Web, May 13-17, 2013, Rio de Janeiro, Brazil [doi>10.1145/2488388.2488408]

[9] Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., ... & Shakarian, P. (2016, September). Darknet and deepnet mining for proactive cybersecurity threat intelligence. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)* (pp. 7-12). IEEE.

[10] G. Owen and N. Savage. 2016. Empirical analysis of Tor Hidden Services. IET Info. Sec. 10 (2016). Issue 3.

[11] Tai, X. H., Soska, K., & Christin, N. (2019). Adversarial Matching of Dark Net Market Vendor Accounts.