ISSN 2007-9737

Methodology for Identification and Classifying of Cybercrime on Tor Network through the Use of Cryptocurrencies Based on Web Textual Contents

Julio Jesús Salas Conde[1], Manuel Martín Ortíz[2], Víctor Manuel Carneiro Díaz[3]

[1] Benemérita Universidad de Puebla,
Facultad de Ciencias de la Computación,
Mexico

[2] Benemérita Universidad de Puebla,
National Super-Computing Laboratory,
Mexico

[3] Universidade da Coruña,
Computer Science and Information Technology,
Spain

julio.salas@sspc.gob.mx, manuel.martin@correo.buap.mx, victor.carneiro@udc.es

Abstract. The Fourth Industrial Revolution has propelled global society into a new era of information and knowledge, transforming the economy and society of many countries. The global digitization process impacts more than half of the world population with internet access and the increase in the incidence of crimes in cyberspace, affecting the population mainly online fraud, crimes that attack vulnerable groups such as girls, boys and adolescents, as well as the diversity of cyberattacks with an impact on the availability, integrity and confidentiality of essential data and information systems of public, private and academic institutions. Most of these antisocial behaviors are published on the Deep Internet due to its anonymity, one of these being the TOR browser project (The Onion Router), in order to address this problem, a methodology was developed that allows the authorities in Mexico have a database that allows correlating data published on this network with the investigations they carry out derived from reports of cybercrimes to obtain lines of investigation based on the identification and classification of cybercrime, and using language engineering techniques and of knowledge as the methods of creation of ontologies of Ding, Y; Foo, S; recovery tool for large information files on websites such as wget, security measures for browsing the Deep Internet such as "Whonix Gateway", "Text Cleaning" techniques, extraction and classification features such as "Jaccard and Cosine Similarity Calculation", among other.

Keywords. Cybercrime, cryptocurrencies, fraud identification, web textual contents, detection tools.

1 Introduction

The web can be can be roughly divided into Surface Web and Deep Web. Surface Web is the part of the web that standard search engines like Google or Bing can crawl and index. However, despite its existence, there is still a huge part of the web without indexing due to its large size and the lack of hyperlinks, that is, not referenced by the rest of the web pages. This part, which cannot be found using a search engine, is known as the Deep Web [1].

The root of many, if not most, cybersecurity threats are not at the edge of the Internet, but within it, on the Dark Web [2]. However, the Dark Web is increasingly difficult to crack as privacy and encryption techniques become more sophisticated. The sites will be much less visible and can only be accessed by invitation. Furthermore, Bitcoin, the cryptocurrency of choice on the Dark Web, is rapidly being replaced by Monero, which offers stealth mechanisms that

prevent direct and indirect tracing of those who conduct transactions, a vulnerability that has affected bitcoin. The same open source tools promoted by privacy advocates to protect personal data and evade government censorship and surveillance are also fueling wide-spread criminal activity.

Given the sophisticated infrastructure of the Dark Web and the superior technical capabilities of many of its inhabitants, traditional forensic techniques are unlikely to have a substantial or lasting effect. However, new machine learning, data mining and analytics tools are about to become formidable offensive weapons in the fight against cyber crime.

According to [3], the Internet is built around web pages that refer to other pages, if you have a destination web page that does not have inbound links, that page has been hidden and cannot be found by users or engines search (not published with a link). An example of this would be a blog post that has not yet been indexed. The blog post may exist on the public internet, but unless the exact URL is known, it will never be found.

The Surface Web is a part of the Internet that can be found through link tracking techniques, known as Link-crawling, which means that the linked data can be found through a hyperlink from the home page of a domain and the search engine can extract this data.

The Deep Web is a part of the Internet not accessible to link-crawling search engines such as Google, Yahoo, and Bing, for instance (direct browsers).

The Dark Web is a part of the World Wide Web that needs a special type of software to be accessed and specifically refers to a collection of Web sites that exists on an encrypted network that cannot be accessed by traditional search engines or even visiting traditional web browsers. Once inside the Dark Web, Web sites and other services can be accessed through a browser in the same way as on a traditional Web. However, there are some sites that are effectively hidden, which means that traditionally they have not been indexed by a search engine and therefore such sites can only be accessed if you know the address of the site [4].

Virtual private networks are another aspect of the Deep Web, existing within the public Internet, and often require additional software to access them. TOR (The Onion Router) is a great example. Hidden within the public network is this private network of different content and which can only be accessed through the TOR network.

The National Crime Agency (NCA) is leading the UK's fight to end organized crime, noting that the market for illicit use of cryptocurrencies is estimated at US $ 603 million in bitcoin transactions on the Darknet in 2018 (National Crime Authority, 2019) [5]. In its 2020 publication, it notes that high-impact criminals exploit the vulnerable through modern slavery, human trafficking and child sexual abuse; dominate communities through wholesale drug supply networks and the illegal firearms trade, and undermine the UK economy and infrastructure through criminality, causing money to be diverted through illicit transactions and cybercrimes, some of its chronological success events in 2019 are:

- January: Three men who ran a business on the Dark Web that sold lethal drugs fentanyl and cafentanil to customers across the UK and around the world were imprisoned for a total of 43½ years. The criminals mixed fentanyl, which is up to 100 times stronger than morphine, and carfentanil, which is 10,000 times stronger, with freight forwarders at an industrial unit in Leeds. They then sold them through the dark web. There have been over 125 deaths in the UK related to fentanyl or carfentanil since December 2016.
- March: A 26-year-old man was jailed for 22 years for raping a five-year-old boy and sexually abusing a three-year-old girl. Videos of his abuse had been posted on TOR's "Welcome to Video" website, which was run from South Korea. The videos were discovered by the NCA and, following an investigation, the suspect was traced to his home in the UK. Although his face was not visible in the abuse videos, investigators used specialized techniques to identify him. After his arrest, officers found a large number of abuse images and videos on his laptop. The "Welcome to Video" site, which contains more than 250,000 child sexual abuse videos, was removed by an international task force supported by the NCA.

- April: A London-based cybercriminal was imprisoned at Kingston Crown Court for six years and five months for blackmail, fraud, money laundering and computer misuse. Criminal Targeted Hundreds of Millions of Computers with Lockdown Ransomware. The blackmailed victims would be forced to pay the ransom demands using virtual currency, which was then laundered through an international network of financial service providers. The offender received more than £ 270,000 through this online blackmail campaign.
- May: Two men were sentenced to a combined 29 and a half years in prison for possession and conspiracy to sell handguns made at an illegal East Sussex gun factory. The factory was discovered following an NCA investigation in conjunction with Sussex Police in August 2018, the first time UK law enforcement has found a weapons factory for criminals of this nature. Officers discovered three firearms and their components, indicating that another 121 firearms were being manufactured. Since then, 11 additional firearms manufactured at the unit have been recovered, with one known to have been used in two assassination attempts in London.
- November: Working with international law enforcement partners and the North West Regional Organized Crime Unit, the NCA coordinated the UK effort against an online site selling a popular hacking tool. The Imminent Monitor Remote Access Trojan, once covertly installed on the victim's computer, allowed the hacker to have full access to the infected device, allowing him to disable antivirus software, steal data or passwords, record keystrokes and observe the victims through their webcams. More than 14,500 people around the world had purchased the tool for as little as $25 USD. In November, coordinated action across the UK resulted in nine arrests and more than 100 items recovered. The website selling the tool was taken down by the Australian police.
- December: Russian national Maksim Yakubets was indicted in the United States in connection with two separate international hacking and bank fraud schemes, following an unprecedented collaboration between the NCA, the FBI, and the UK's National Center for Cyber Security. Yakubets ran Evil Corp, the world's most damaging cybercrime group that created and deployed malware that caused financial losses totaling hundreds of millions of pounds in the UK alone. In 2014, a dedicated team at the NCA began working with various partners to investigate one of the group's leading malware strains, Dridex. These officers developed intelligence and identified evidentiary material over several years to support the US allegations, as well as the sanctions against Evil Corp.

The NCA and the Metropolitan Police Service have also targeted the Yakubets money launderer network. in the UK who have funneled the profits to Evil Corp. So far, in the UK, eight people have been sentenced to a total of more than 40 years in prison.

This article is divided into 5 sections, the conclusion and the Bibliography, section 2 presents previous works related to classification on the Internet and Deep Internet, section 3 deals with the Construction of the Database resulting from the proposed methodology, section 4 presents the methodology carried out, and section 5 shows preliminary results regarding the experimentation of the methodology.

2 Related Work

In recent years [6], many researchers have investigated the Surface Web classification [7-17]. However, the Darknet classification literature is still in its early stages and specifically the classification of illegal activities [18, 19].

Kaur [10] presented an interesting survey covering various algorithms to classify web content, paying attention to its importance in the field of data mining. Additionally, the survey included preprocessing techniques that could aid feature selection, such as removing HTML tags, punctuation marks, and the root. Kan and collaborators [8, 9] explored the use of Uniform Resource Locators (URLs) in web classification by extracting characteristics using parsing and segmentation.

These techniques cannot be applied to Tor Hidden Services (HS) since, onion addresses are

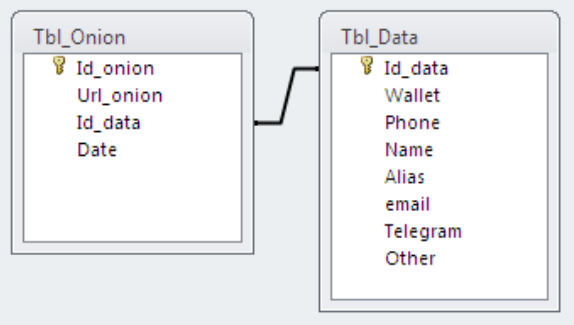

Fig. 1. Database with main data to be correlated in an investigation by the authorities

constructed using 16 random characters. However, tools like Scallion[1] and Shallot[2] allow Tor users to create custom .onion addresses based on brute force technique. For example, Shallot takes 2.5 years to build 9 custom characters out of 16. The use of Support Vector Machine (SVM) has been helpful to classify web content by leveraging context features, for example, HTML tags and hyperlinks in addition to textual features to create the feature set.

Regarding the Deep Web classification, Noor et al. (2011) [1] discussed common techniques used for content extraction from DeepWeb data sources called "Query Probing", which are commonly used for supervised learning algorithms, and "Visible Form Features" (Xian et al., 2009 [16]) and Su et al. (2006) [11] have proposed a combination of SVM with Query Probing to hierarchically classify the structured Deep Web. Barbosa et al. (2007) [13] proposed an unsupervised machine learning clustering pipeline, in which they used Term Frequency Inverse Document Frequency (TFIDF) for the representation of the text and the cosine similarity for the distance measurement for the k-means.

Regarding Darknet, Moore et. al. (2016) [19] presented a new study based on Tor hidden services to analyze and classify the Darknet. Initially, they collected 5K Tor onion page samples and classified them into 12 classes using the SVM classifier. Graczyk et al. (2015) [18] proposed a pipeline to classify the products of a famous black market on the Darknet, called Agora, into 12 classes with 79% accuracy. Its pipeline architecture uses TFIDF for text feature extraction, PCA for feature selection, and SVM for feature classification.

Various attempts have been proposed in the literature to detect illegal activities, either on the World Wide Web (WWW) (Biryukov et al., 2014 [20]; Graczyk y Kinningham, 2015 [15]; Moore and Rid, 2016 [19]), peer to peer networks (P2P) (Latapy et al., 2013 [21]; Peersman et al., 2014 [22]) and in chat messaging systems (Morris y Hirst, 2012 [23]). Latapy el at. (2013) [21] investigated P2P systems, for example, eDonkey, to quantify pedophilia activity by building a tool to detect child pornography queries by performing a series of lexical word processing. They found that 0.25% of the queries entered are related to the pedophilia context, which means that 0.2% of the eDonkey network users are entering such queries. However, this method relies on a predefined list of keywords that cannot detect new or previously unknown words.

3 Database Construction

In order to help the authorities to correlate crypto-currency accounts of criminal behavior, it is necessary to create databases that contain the information to be consulted. For this, a methodology that can be replicated was carried out to identify onion networks, download the content of these identified networks, classify them with respect to criminal behavior based on an ontology of crimes, through their similarity cosine and through the distance of Jaccard, to finally create one databases that contains key fields such as cryptocurrency accounts, names, telephones, emails, among others, that allow the information to be correlated in an investigation by the authorities (Figure 1).

4 Methodology

In general, the methodology has the next steps:

1. Preparing tool box,
2. Identification of onion networks,
3. Security measures to navigate,

[1] www.github.com/lachesis/scallion

[2] github.com/katmagic/Shallot

4. Conformation of an ontology of terms,
5. Download crude information from Dark Web from onion networks identified,
6. Cleaning text,
7. Characteristics extraction,
8. Jaccard and cosine similarity calculation,
9. Field tests,
10. Data results,
11. Data output analysis,
12. Dataset and continuous growing.

4.1 Identification of Onion Networks and Tools Used

In the first place, there is a service called "Hunchly"[3], which is a tool designed for online research. When subscribing, it sends via email a daily Excel file classified in; active, inactive and discovered onion networks as of the day in question. As of 2019 in December 1, 3,302 active onion networks were counted.

Another option to identify onion networks and even scan their vulnerabilities to obtain information about your hosting is the OnionScan tool (https://github.com/s-rah/onionscan), this free tool has two main objectives:

- Help troubleshoot hidden onion network service operators as well as misconfigurations.
- Help crime investigators monitor and track Dark Web sites.

In a scan carried out, a total of 8,386 networks were obtained, of which 1,143 onion networks were active.

4.2 Security Measures for Navigation in TOR

To carry out these actions, certain security precautions must be taken, among others we have the following:

- Use of a bootable operating system from a USB (such as Tails [24]), in such a way that the hard disk of our equipment is not used and privacy and anonymity is preserved.
- Use of virtual machines for connection to TOR and for browsing. For this, two Linux virtual machines are used, the first will have the virtual machine "*Whonix Gateway*" [25] that will pass all the traffic to the TOR network and the second virtual machine will be used to navigate through a "browser" that will connect to the network of the first. virtual machine, this will help if the Tor network is compromised, it is not working on it.
- When navigating within the Tor browser, do not allow the use of Scripts on the pages that are visited, in addition to not running programs downloaded from this network, if it is necessary to download and run any of these programs, check it first with an antivirus or in a free service like Virus Total [26] on the Surface Web.
- It is recommended to remove the drivers for the microphone and the webcam when they are available and if possible cover the computer camera with a tape.
- It is recommended to implement your own Virtual Private Network (VPN) even though Tor handles its own encryption and VPN.
- Do not enter personal data, emails from Surface Web services, usernames previously used, or the same password for the sites where you register.
- Avoid the use of personal credit cards, if necessary use single use prepared cards and verify that the website is secure by verifying the web address starting with "https://" and not with "http://". Here the "s" means that the sent and received data travels encrypted (Rafiuddin, 2017 [4]).

4.3 Conformation of an Ontology of Terms

From the methods of work of Ding and Foo [29] and in order to understand the different classes of crimes that are tried to be discovered within the information of the identified onion networks, we worked on creating an ontology of terms, the concepts were determined from the definitions of crimes of the Federal Law Against Organized Crime of Mexico and the distinction between “is-a” and “assoc-with” relations based in the linguistic property of noun compounds of this crimes, of course that this ontology can be adapted to the

[3] www.hunch.ly

Fig. 2. Ontology of crimes against Organized Crime in Mexico and crimes against information technologies

```
nginx;;;;;;;;46082;;keep-alive;text/html;;keep-alive
nginx/1.6.2;;;;;;;;2859;Cookie;keep-alive;text/html;;keep-alive
nginx;;SAMEORIGIN;;;;;;39982;;keep-alive;text/html;;keep-alive
nginx;nosniff;SAMEORIGIN;;;;;;16489;Accept-Encoding;close;text/html;;keep-alive
nginx;;;;;;;;;1240480;Accept-Encoding;keep-alive;text/html;;keep-alive
nginx;;;;;;;;;1240480;Accept-Encoding;keep-alive;text/html;;keep-alive
Apache/2.2.22;;;;;;;;;61999;Accept-Encoding;Keep-Alive;text/html;;keep-alive
Apache;;;;;;;;26818;;Keep-Alive;text/html;keep-alive
Apache;;;;;;;;17462;Accept-Encoding;Keep-Alive;text/html;;keep-alive
;;;;;;;;;12699;;keep-alive;text/html;keep-alive
Apache/2.4.18;;;;;;;;;21649;Accept-Encoding;Keep-Alive;text/html;;keep-alive
nginx;;;;;;;;;681911;Accept-Encoding;keep-alive;text/html;;keep-alive
nginx;;;;;;;;;52170;Accept-Encoding,;keep-alive;text/html;;keep-alive
```

Fig. 3. File separated by semicolons with characteristics of the onion network headers

current laws of each country or adapt to a specific type of crime, such as cybercrime.

The purpose of having an ontology of crime terms is that in order to classify the identified networks, there must be an understanding of a specific domain of knowledge and a conceptual framework that represents the knowledge so that there is a reference to the type of crimes on which one works, its definition and legal support. In figure 2, we can see the result of this Ontology created with the "*Protége*" tool [27].

4.4 Download from Identified Onion Networks

Our choice and proposal are use open source tools that do not require licensing and so that this process can be replicated anywhere, the wget tool (Linux) is used to extract data and channel it through the anonymous TOR network and avoid blocking of Servers IP, TOR is a SOCKS proxy that works sending the information (data) over a network quite anonymously.

The problem with TOR is that it does not offer an HTTP proxy, which is what wget requires. So, to solve this the *Privoxy* package [28] is installed and used, which will allow to connect to TOR through a simple HTTP proxy.

For it to be possible to download information from onion networks with the wget tool in a Linux environment, by example, it is necessary to carry out an installation and configuration process of programs such as TOR, privoxy, wgetc and selector. In the following link, there is a discussion how to assemble and coordinate the cited tools[4].

It is possible to recover up to five levels the information of the onion networks defined (we use a text file to make the list of them, by instance: "def_file.txt"), with an instruction like the following one we can do the information retrieval, online and in real time:

```
wget -r -i def_file.txt --save-headers -o gnulog.
```

We experiment for this job with information corresponding to "crimes against the people in their patrimony", in addition to obtaining the headers of said files.

[4] www.dejonck.be/2013/05/data-mining-using-wget-with-tor-for.html?m=1

4.5 Cleaning Text

A series of index.html files are obtained, each one contains their header and their html structure, it will be necessary to convert them to text format, remove the html tags and remove control characters, converting them to UNIX format to be able to treat them later, for this the following procedure was fo lowed:

1. Using the html2text package, that converts html files to text and removes their html tags, the files are converted from the recovered html files into text, for this job was executed the following instructions in Linux:

   # apt i<Anstall html2text (1)
   # html2text index.html.* > index.txt.* (2)

2. As there are control characters that each operating system handles with its files, it's necessary to convert the recovered files to a Unix format, for this was used the dos2unix command as shown below:

   # dos2unix index.html.* (3)

This process allows us to have information to be treated ready and clean.

4.6 Extraction of Characteristics

In order to categorize the similarities and differences that exist between the dissimilar types of criminal behaviors, we create text files separated by commas that include the characteristics of their headers, as well as their Jaccard similarity value, cosine similarity value, mean, minimum, maximum, median and variance. To do this, we use the following procedure:

1. To retrieve the characteristics of the headers of the index.html files for the class of crime selected that are already available as can be seen in Figure 3, we use the *awk* tool and generate a text file separated by semicolons with these characteristics.
2. To achieve this, we process files with the following command:

   # awk -f cmd.txt index.html.1 (4)

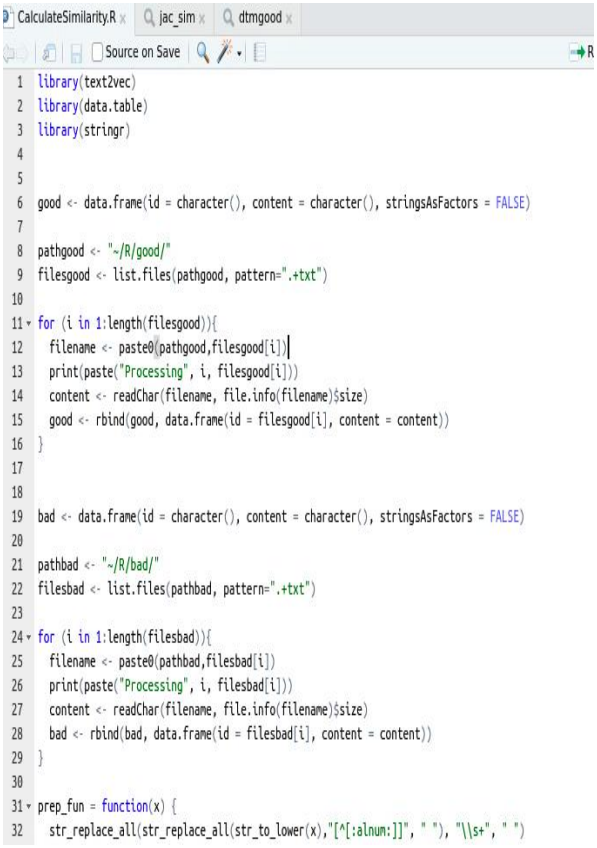

Fig. 4a. Program to calculate the similarity of Jaccard and Cosine

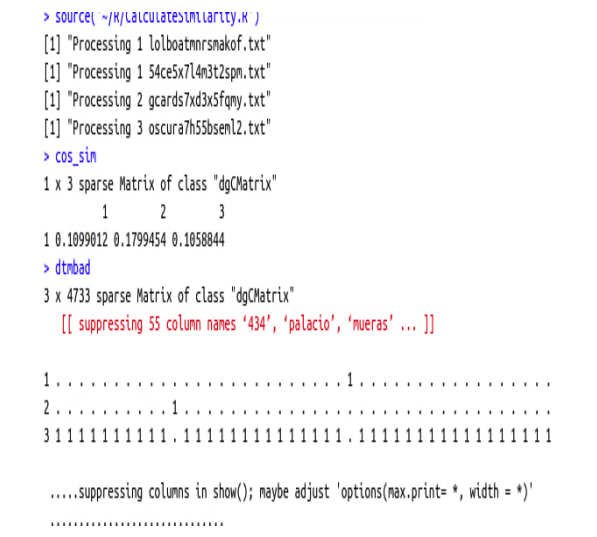

Fig. 4b. Console output of the process to calculate the similarity of Jaccard and Cosine

Table 1. Number of onion networks for criminal conduct

Criminal Conduct	Onion Network Number
Crimes against the free development of personality	74; Child Pornography, Corruption of Minors, Lenocide
Crimes against Liberty and Normal Psychosexual Development	22; Pornography, Sex Tourism, Scorts
Crimes Against Health	61; Sale of drugs, steroids, narcotics
Crimes Against the Security of the Nation	11; Cyber Guerrilla, Anarchist and Terrorist Groups
Crimes Against Public Safety	17; Sale of weapons and ammunition
Crimes Against Life and Bodily Integrity	11; Murders and violence
Crimes Against People in their Estate	208; Bitcoin transactions, card cloning, scams
Copyright Crimes	3; Books, Video Streaming, Broadcasts, Movies, Video Games
Falsification	11; Counterfeit currency, passports, visas, works of art, apocryphal documents
Disclosure of secrets and illegal access to computer systems and equipment	96; DDOS, Hacked Accounts, Hacking Services, Spam Services, Malware, system access, surveillance

where cmd.txt is a text file containing the characteristics extraction from the header of the given html file.

3. A script was generated that automatically converts all the index files to UNIX format and later extracts the headers and generates a file separated by semicolons that collects all the characteristics of the headers of all the indexes. Subsequently, the script was run for all the folders classified by crime to obtain the headers with the characteristics in files separated with commas for analysis:

```
# ./proc1.sh                    (5)
```

Resulting in this way one file with name csv1.txt for each folder

4.7 Jaccard and Cosine Similarity Calculation

Once we have the text files separated by commas with the characteristics of their headers of the onion networks with criminal behaviors, it is required to attach to each record the calculation of the Jaccard index and the Cosine similarity, this in order to know how much are they look like two files, one with criminal behaviors and the other without criminal behaviors, and even between different criminal behaviors and between themselves. To do this, we create a program using Rstudio.

Figure 4a shows the program CalculateSimilarity.R, two folders bad and good are obtained, in the bad folder we have 3 html files and their corresponding txt (without html tags) with criminal behaviors; and in the good folder we have a domain without criminal conduct html and txt. Figure 4b shows terminal output.

The program begins by storing the content of the text file in the good folder in good, and the content of the three files contained in the bad folder in bad, then it creates two vocabularies, one good and one bad, containing their words, and finally calculates the Jaccard index and the Cosine Similarity.

In this way, we obtain a matrix of results that will be the basis for classifying new onion networks by criminal conduct or without criminal condition.

5 Experimentation

A sample of 930 onion networks was analyzed within the set identified with the process of discovery inside the Dark Web, Table 1, presents the number of onion networks studied according to the ontology created, making a total of 514 pages of onion networks with possible criminal behavior and 416 pages without criminal behavior.

Total of 930 onion networks, 514 onion networks with criminal conduct, and 416 onion networks without criminal conduct.

From this sample the methodology was followed and later a database was created that allows us to be consulted by blockchain checkbook number, one of our goal is to aid to authorities to respond to the investigations of criminal conduct through crypto assets.

6 Conclusions

In this article, we propose and describe a methodology to solve a problem raised by the authorities, in particular by the cybercrime task forces, to enface the increase in criminal behavior through the TOR network, the methodology proposed can help to identify cryptocurrency checkbooks and correlate them with information from onion networks within TOR.

The methodology has been proved using an incremental dataset, this dataset has been extracted from the Dark Web applying the discovery step and further steps each month, and identifying which onion networks survive, which one of them disappears and obtaining new onion networks.

The database has been updated, and the cyber-crime sites has been tracked if they are operative using full fake accounts, the accounts are changed periodically.

The full process is in use, and one cybercrime office has as daily task, where a task force is dedicated to this. Several cases have been tagged for judicial actions.

As future work we have, the improvement of each of the stages of the methodology, and an evaluation of criminal behavior to verify the effectiveness of the proposed method, including metrics such as precision, recovery and the f-measure of the classification.

References

1. **Noor, U., Rashid, Z., Rauf, A. (2011).** A survey of automatic deep web classification techniques. International Journal of Computer Applications, Vol. 19, No. 6, pp. 43–50. DOI: 10.5120/2362-3099.
2. **Hurlburt, G. (2017).** Shining light on the dark web. Computer, Vol. 50, No. 4, pp. 100–105. DOI: 10.1109/MC.2017.110.
3. **BrightPlanet. (2013).** Understanding the deep web in 10 minutes. https://brightplanet.com/2013/03/12/whitepaper-understanding-the-deep-web-in-10-minutes/.
4. **Rafiuddin, M., Minhas, H., Singh, P. (2017).** A dark web story in-depth research and study conducted on the dark web based on forensic computing and security in Malaysia. IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), pp. 3049–3055.
5. **National Crime Agency (2019).** National Strategic Assessment of Serious and Organized Crime.
6. **Nabik, M., Fidalgo, E., Alegre, E., de Paz, I. (2017).** Classifying illegal activities on TOR network based on web textual contents. 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 1, pp. 35–43.
7. **Dumais, S., Chen, H. (2000).** Hierarchical classification of web content. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 256–263. DOI: 10.1145/345508.345593.
8. **Kan, M. (2004).** Web page classification without the web page. Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, pp 262–263. DOI: 10.1145/1013367.1013426.
9. **Kan, M., Nguyen, H. (2005).** Fast webpage classification using URL features. Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 325–326.
10. **Kaur, P. (2014).** Web content classification: a survey. International Journal of Computer Trends and Technology (IJCTT), Vol. 10, No. 2, pp. 97–101. arXiv.1405.0580. DOI: 10.14445/22312803/IJCTT-V10P117.

11. **Su, W., Wang, J., Lochovsky, F. (2006).** Automatic hierarchical classification of structured deep web databases. Proceedings 7th International Conference on Web Information Systems, pp. 210–221. DOI: 10.1007/119128 73_23.
12. **Xu, H., Hao, Wang, S., Hu, Y. (2007).** A method of deep web classification. International Conference on Machine Learning and Cybernetics, Vol. 7, pp. 4009–4014. DOI: 10.1109/ICMLC.2007.4370847.
13. **Barbosa, L., Freire, J., Silva, A. (2007).** Organizing hidden web databases by clustering visible web documents. IEEE 23rd International Conference on Data Engineering, pp. 326–335. DOI: 10.1109/ICDE.2007.367878.
14. **Lin, P., Du, Y., Tan, X., Lv, C. (2008).** Research on automatic classification for deep web query interfaces. International Symposiums on Information Processing, pp 313–317. DOI: 10.1109/ISIP.2008.140.
15. **Zhao, P., Huang, L., Fang, W., Cui, Z. (2008).** Organizing structured deep web by clustering query interfaces link graph. International Conference on Advanced Data Mining and Applications, pp. 683–690. DOI: 10.1007/978-3-540-88192-6_72.
16. **Xian, X., Zhao, P., Fang, W., Xin, J., Cui, Z. (2009).** Automatic classification of deep web databases with simple query interface. International Conference on Industrial Mechatronics and Automation, pp. 85–88. DOI: 10.1109/ICIMA.2009.5156566.
17. **Khelghati, M. (2009).** Deep web content monitoring. DOI: 10.3990/1.9789036541237.
18. **Graczyk, M., Kinningham, K. (2015).** Automatic product categorization for anonymous marketplaces.
19. **Moore, M., Rid, T. (2016).** Cryptopolitik and the darknet. Survival, Vol. 58, No. 1. pp 7–38. DOI: 10.1080/00396338.2016.1142085.
20. **Biryukov, I., Pustogarov, I., Thill, F., Weinmann, R. (2014).** Content and popularity analysis of Tor hidden services. IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW), pp 188– 193. DOI: 10.1109/ICDCSW.2014.20.
21. **Latapy, M., Magnien, C., Fournier, R. (2013).** Quantifying paedophile activity in a large p2p system. Information Processing & Management, Vol. 49, No. 1, pp. 248–263. DOI: 10.1016/j.ipm.2012.02.008.
22. **Peersman, C., Schulze, C., Rashid, A., Brennan, M., Fischer, C. (2014).** iCop: Automatically identifying new child abuse media in p2p networks. 2014 IEEE Security and Privacy Workshops, pp. 124–131. DOI: 10.1109/ SPW.2014.27.
23. **Morris, C. Hirst. Graeme. (2012).** Identifying sexual predators by SVM classification with lexical and behavioral features. CLEF (Online Working Notes/Labs/Workshop), Vol. 12, pp. 29.
24. **Tails (2021).** Tails, Secure computer portable operating system on USB. https://tails.boum.org
25. **Whonix (2021).** Whonix Gateway. https://www.whonix.org/wiki/VirtualBox
26. **Virustotal (2021).** Virus Total, Software to find Virus, Malware over URLs. https://www.virustotal.com
27. **Protégé (2021).** Protégé, A free, open-source ontology editor and framework for building intelligent systems. https://protege.stanford.edu/
28. **Privoxy (2021).** Privoxy, non-caching web proxy. https://www.privoxy.org/
29. **Ding, Y., Foo, S. (2002**). Ontology research and development. Part 1-a review of ontology generation. Journal of Information Science, Vol. 28, No. 2, pp. 123–136.

Article received on 29/07/2021; accepted on 30/09/2021.
Corresponding author is Manuel Martín Ortíz.