Scraping The Deep Web : A 3-Dimensional Framework For Cyber-Threat Intelligence

VICTOR ADEWOPO, University of Cincinnati, United States
BILAL GONEN, University of Cincinnati, United States

ABSTRACT The cyberspace is one of the most complex systems ever built by humans. The utilization of cybertechnology resources are used ubiquitously by many, but sparsely understood by the majority of the users. In the past, cyberattacks were usually orchestrated in a random pattern of attack to lure unsuspecting targets. However, the cyber virtual environment is an ecosystem that provided a platform for an organized and sophisticated approach to launch an attack against a specific target group or organization by nefarious actors. In 2019, the average cost of cyber-attack in the US was about $1.6 million. This paper propose a 3D framework to signal new threat alert before the actual occurrence of the threat on surface web to alert cybersecurity experts and law enforcement agencies in preventive measures or means of mitigating the severity of damage caused by cyberattacks. The proposed methodology combines information extracted from the deep web through a smart web crawler with socio-personal and technical indicators from twitter which is mapped with OTX (Open Threat Exchange). The OTX is an open-source cyber threat platform managed by security experts. The OTX endpoint security tool(OTX python SDK) will be used to identify a new type of cyber threats. The effectiveness of the framework will be tested using the machine learning algorithm precision-recall rate.

Additional Key Words and Phrases: Cyberattack, Deepweb, Cybersecurity, Machine learning

ACM Reference Format:
Victor Adewopo and Bilal Gonen. 2020. Scraping The Deep Web : A 3-Dimensional Framework For Cyber-Threat Intelligence. In *IT Research Symposium '20: School of Information Technology IT Research Symposium, April 14, 2020, Cincinnati, OH*.https://scholar.uc.edu/

1 INTRODUCTION

Cyberspace holds an enormous quantity of information that is astute in gathering threat intelligence useful for cybersecurity experts in preventing cyberattacks and protecting an organization's network system. Cyber attacks are now performed in an organized and sophisticated manner launched against a specific target group in contrast with the old random pattern of attack leaving a majority of netizens who posses minute knowledge of the powerful resources embedded in cyberspace vulnerable. The high level of restriction on accessible information in the deep-web prompted many organizations to put their database on the deep web. The quality information embedded in the deep-web is essential for cybersecurity experts in gaining insights on preventive and security measures. In the quest for a high quality of information, developing methods to crawl the deep web and excavate the rich information becomes a paramount concern.

Some contents of the deep-web database can be accessed through the surface web interface by filling the necessary query forms to generate the requested information[24]. There are often misconceptions with the terms “Surface web”, “Deep web”, and “Dark web”, which are relatively interconnected but do not mean exactly the same thing. The surface web is web pages that are unencrypted and can be accessed using traditional search engines (e.g. Google, Bing, Yahoo). The

IT Research Symposium '20, April 14, 2020, Cincinnati, OH

surface web consists of billions of static webpages and it occupies only 10% of the internet space [25]. The deep web contains about 90% of contents available on the internet [1]. The deep web contents are available in the cyberspace but cannot be indexed or accessed by using search engines. Several techniques and tools have been designed to understand and crawl the deep web. A recent report indicated the low harvest rate of the deep web, about 647,000 distinct web forms, were found by sampling 25 million pages using Google index [26][6]. The dark web is a layer within the deep web and is not accessible using a standard search engines. They are intentionally hidden part of the web but can be accessed using a specific URL address. Darknet is a network technology that gained recognition in 1971, through illegal drug trading using ARPANET, and it’s mainly characterized by illegal activities[2]. A Cybersecurity firm “UpGuard” reported 419 million Facebook user’s data breach in April 2019. Data exposed publicly on the Amazon server includes; passwords, user IDs, and check-ins [18]. Similarly, in September 2018, over 50 million Facebook users’ breached data was auctioned at a bitcoin value of 3$ per each user’s data on the darknet marketplace [3]. Nefarious actors have been able to build a digital ecosystem that provides platforms famously acclaimed to promote activities such as cybercrime, terrorism, hacking, procurement of illegal drugs, arms deals, and anonymous markets[10]. The hidden nature of darknet websites aided fraudsters in utilizing the cyberspace for criminal activities. Distributed Denial Of service is reported to be the largest amount of cyberthreat with over 48% of total cyber-attacks [7]. In the past, illegal drug transactions are carried out at a high level of secrecy and high level of restrictions that serves the dual purpose of evading law enforcement interference and also reducing recondite participants. In modern practices, illegal drug deals are carried out in the cyberspace through hidden services such as TOR without restrictions. Deep-web platforms serve as a consummative ecosystem shielded by boundaries of trust, information sharing, trade-off, and review system that allows vendors with a common interest to carry out transactions across the global community. Some bright spots on the darknet include, freedom of digital media, academic research, secure drop, anonymous emails, and Ad-free search engines.

Deep-web is a dynamic and agile environment for darknet market activities, vendor activities declined after operation onymous which lead to the termination of 410 hidden services and arrest of site vendors. The closure of a new darknet service leads to the evolution of another in a new form [22]. Over the years, criminal activities progressed geometrically in a new dimension through the use of cryptocurrency[4]. Crypto markets are online marketplace on darknet websites that provides an anonymous platform for trading of illicit drugs and services. The use of bitcoins and escrow services in darknet markets increased the reliability of customers in making transactions on darknet markets [17]. This paper seeks to develop a 3D-framework that analyzes and extracts information available in deep web forums to provide cybersecurity experts with insights that can rapidly be used in responding and prevention of cyber threats before the actual occurrence. An NSA cybersecurity exploit tool “ExternalBlue" was leaked by a hacker group surfaced in the dark web before being disclosed on surface web[5]. Using data mining tools, correlation analysis will be done to unravel the relationship between dark-web and surface-web virtual environments in orchestrating criminal activities.

2 RELATED WORK

In this section, we present related work in crawling the deep web, extraction of information from the deep web using smart crawlers and how the use of TOR services have aided anonymous criminal activities in cyberspace.

2.1 Crawling the Deep-Web

In the era of big data, crawling the deep web is very important to get insights into information embedded in the deep web, a large amount of data exists on the web today and can be accessed through interfaces. The inability of search engines to crawl deep web limits the information provided to users. Oftentimes, users have to spend extra time and effort querying a website manually to access the contents of a website. The study of Manvi et al.,[15] developed an Ontology-based adaptive crawler for the hidden web by mapping relationships between webpages using four components (Ontology Builder, Hidden Web Miner, Result Processor, and Domain Specific - Flow Diagram). The research of Dong et al.,[5] utilized a lightweight framework to predict cyber threat from darknet data. The method applied a scrapy crawler with proxy VPN to parse the web-page for relevant information which is processed through verification techniques to generate cyber threat warnings. Research work in [13][16][20] adopted a framework in predicting the future occurrence of a cyber attack by mapping different datasets from the dark web, socio-personal and technical indicators with online cybersecurity report forums. A combination of human and automated techniques was used to extract live data from the dark web, insights gathered from online forums and community can be used to predict malicious acts before they occur. The use of search engines and spider services on the TOR network was used to collate a list of more than 100 malicious hacker websites. One of the major challenges encountered in crawling the websites was CAPTCHA authentication, username and password, invitation code and filling of query forms for automatic crawlers. The prototype system built in [11], tackled the challenge of query generation, empty page filtering, and URL duplication.

2.2 Extracting Information using Smart-Crawlers

Little literature exists on deep web crawlers. The unavailability of public datasets makes it difficult to measure the effectiveness of deep web crawlers developed by researchers against a standard benchmark. The survey of deep web crawlers carried out in [12] reported that a typical web crawling process was broken down to contain at least 5 basic steps spanning from entry points to crawling path learning. The deep web is estimated to be relatively 5,000 times bigger than the surface web. Accessing the deep web requires the identification of entry point and analyzing features of the deep web. Most researchers identified automated form filling as a major setback in crawling the deep web. Crawling the deep web using automatic crawlers entails that the forms are filled appropriately to elicit meaningful results. A focused crawler for locating deep web entry, "FFC" crawler, was used initially based of the limitation of features training, "ACHE" an evolution of FFC was used in crawling the deep web before the advent of smart crawlers used in [26][12][14]. Wang et al., describe the DELA system (Data extraction and label Assignments) as a method for extracting data from crawled web pages and fitting them into a table with an appropriate label based on the structure of the data without human intervention. The DELA system adopts an existing hidden webcrawler "HiWe" to perform the automated task of filling the labels with relevant values that send automatic queries to HTML form to generate regular expression wrappers to extract information from the web pages[23].

2.3 Use of TOR for Cyber Attack

TOR relay servers are managed by freelancers and organizations across the globe to secure anonymous means of communication among netizens. Cyberspace is a virtual hub to create new innovative ideas to enhance the global community. The continuous use of cyberspace is faced with susceptibilities by fraudsters to spawn nefarious activities such as spam, identity theft, DDoS, and DRDoS. DDoS is the largest amount of cyber threat with over 48% of cyber-attacks[7]. An alarming DDoS attack hit a peak of 400Gbps using a Network Time Protocol (NTP), CompTIA research revealed that a greater percentage of security breaches comes through inadvertent oversight of non-technical staffs. A novel approach is proposed to train and raise awareness of end-users on the overlooked basic security measures[9]. Tor services are considered one of the most popular anonymizing services. The survey of Saleh et al.,[21] classified anonymity and security as the highest recurring keywords associated with TOR services in 26 years of research on TOR services. “FireEye” a major cybersecurity firm, identified that the Dyre Banking Trojan, designed to steal credit card information exploited the windows vulnerability detected in 2015[20]. Darknet can be used in monitoring cyber threats through the deployment of technology visualization and image processing techniques. Darknet based monitoring system is designed to trap unused IP addresses with non-interactive hosts as a source of investigating and gathering intelligence [8].

3 METHODOLOGY

In this section, we describe our data collection method and the features of the proposed framework. The proposed methodology combines a 3-dimensional approach in providing information that can be used to alert cybersecurity experts of potential threats and also with information that can be used to prevent cyber attacks before actual occurrence.

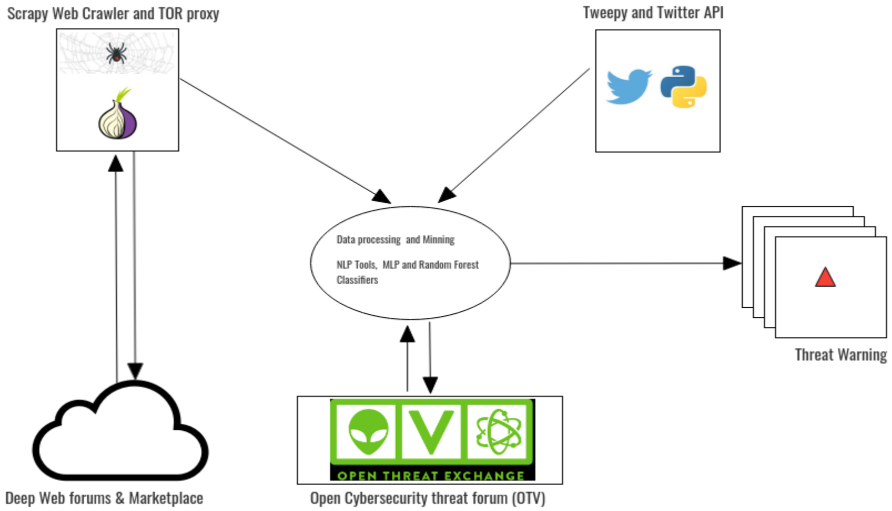
Scrapy Web Crawler and TOR proxy
Tweepy and Twitter API
Data processing and Minning
NLP Tools, MLP and Random Forest Classifiers
Threat Warning
OPEN THREAT EXCHANGE
Deep Web forums & Marketplace
Open Cybersecurity threat forum (OTV)

Fig. 1. Proposed 3D-Framework

(1) Crawling Deep-Web: An open source web crawler “Scrapy web-crawler” will be used to crawl selected darkweb forums to extract information of listings and communications between vendors and consumer in the forums. Scrapy is a fast high-level web crawling and web scraping framework, used to crawl websites and extract structured data from their pages. Information will be extracted at intervals for three months and stored in a table. Algorithms will be configured with the link of the specified forums to deploy the spyder crawlers in the python shell.

(2) Extracting Twitter Data: Information from twitter relating to cyber threats and cyber attacks will also be collected through the Twitter Api using TWeepy, a python library, to collect and store tweets with selected keywords. The selected keywords include; XSS, spam, malware,

data, attacker, DNS, DDOS, code, ciphertext, cryptography, hacked, breach, sniffer, buffer, firewall,hijacking, checksum, virus, and vulnerability from cybersecurity domain expert in correlation with the research in [19].

(3) Cybersecurity forum OTX: An open-source cybersecurity threat forum (OTX) will be used to map data generated from Twitter API and Deepweb to identify if a particular threat corresponds to an already existing threat on the OTX platform before generating a threat alert. If a new potential threat is scanned on the OTX endpoint security pulse and yielded results greater than 2, this denotes that it is not a new type of threat. Otherwise, a new threat alert will be generated to alert LEAs and Cybersecurity analysts. OTX provides open access to a global community of cyberthreat researchers and security professionals across the globe with over 100,000 participants in 140 countries, who contribute over 19 million threat indicators daily to deliver community-generated threat data.

(4) Data Preprocessing and Text Cleaning: Data extracted from deep web forum typically consists of titles, descriptions and special characters which serve as noise to the classifier such as (%, !, ; *, &). To mitigate these challenges in data processing, we will remove all nonalphanumeric characters from the data we will use stop-words remover an NLP toolkit. Tokenizers will be used to break up a sequence of strings into pieces such as words and phrases. Misspellings and Word Variations will be corrected by using the standard library bag-of-words approach to correct misspellings. Variations of words will also be considered in data processing (e.g. running, run, runner, etc.). Word stemming and lemmatization are commonly used to solve word variations, but for efficiency and speed performance, portstemmer will be used to solve word variations.

4 DISCUSSION & CONCLUSION

In this study, we have been able to provide a detailed overview on the sophisticated pattern of cyberattacks and how the use of anonymizing services such as TOR has shielded nefarious actors in utilizing the cyberspace for criminal activities. Cyberattacks and data breaches have a significant high-cost effect and dire consequences on organization activities and people's privacy. Hence, our proposed 3-dimensional framework aims at providing rapid support to law enforcement agencies in monitoring related cybercrime threats, and also provide cybersecurity analysts with a tool that can be easily deployed in preventing cyberattacks and also responding spontaneously to a newly identified threat before emerging on the surface web.

The framework utilizes the high value of information embedded in dark web forums to generate new cyber threat insights. The unstructured data extracted from deep-web and social media (Twitter) will be processed and filtered to capture only unique information that is relevant to new cyber threats. Our algorithm is efficient in automatic querying of the deep-web forum to extract information of interest and also eliminate de-duplication of web pages. The data will be mapped with the OTX - Endpoint an open-source cybersecurity threat platform to trigger alert for a new type of cyber threats that are not identified in OTX-Endpoint. Finally, we will also use data analysis tools to establish the correlation and relationship between dark-web and surface-web virtual ecosystem in orchestrating criminal activities by nefarious netizens.

REFERENCES

[1] 2020. The Deep Web: Cyber Dangers Lurking on the Dark Web. https://usa.kaspersky.com/resource-center/threats/deep-web

[2] Julia Buxton and Tim Bingham. 2015. The rise and challenge of dark net drug markets. *Policy brief* 7 (2015), 1–24.

[3] Anthony Cuthbertson. [n.d.]. *Facebook hack: People's accounts appear for sale on dark web | The Independent.* Technical Report. https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-hack-data-dark-web-login-details-cost-dream-market-a8564671.html

[4] David Décary-Hétu and Luca Giommoni. 2017. Do police crackdowns disrupt drug cryptomarkets? A longitudinal analysis of the effects of Operation Onymous. *Crime, Law and Social Change* 67, 1 (2017), 55–75.

[5] F Dong, S Yuan, H Ou, L Liu 2018 IEEE Conference on Big, and undefined 2018. [n.d.]. New Cyber Threat Discovery from Darknet Marketplaces. *ieeexplore.ieee.org* ([n. d.]). https://ieeexplore.ieee.org/abstract/document/8629658/

[6] Eduard C Dragut, Weiyi Meng, and Clement T Yu. 2012. Deep web query interface understanding and integration. *Synthesis Lectures on Data Management* 7, 1 (2012), 1–168.

[7] Claude Fachkha, Elias Bou-Harb, and Mourad Debbabi. 2015. Inferring distributed reflection denial of service attacks from darknet. *Computer Communications* 62 (may 2015), 59–71. https://doi.org/10.1016/j.comcom.2015.01.016

[8] Claude Fachkha and Mourad Debbabi. 2016. Darknet as a Source of Cyber Intelligence: Survey, Taxonomy, and Characterization. *IEEE Communications Surveys and Tutorials* 18, 2 (apr 2016), 1197–1227. https://doi.org/10.1109/COMST.2015.2497690

[9] Steven Furnell, Maria Papadaki, and Kerry Lynn Thomson. 2009. Scare tactics - A viable weapon in the security war? *Computer Fraud and Security* 2009, 12 (dec 2009), 6–10. https://doi.org/10.1016/S1361-3723(09)70151-4

[10] Clement Guitton. 2013. A review of the available content on Tor hidden services: The case against further development. *Computers in Human Behavior* 29, 6 (2013), 2805–2815. https://doi.org/10.1016/j.chb.2013.07.031

[11] Yeye He, Dong Xin, Venkatesh Ganti, Sriram Rajaraman, and Nirav Shah. 2013. Crawling deep web entity pages. In *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. 355–364. https://doi.org/10.1145/2433396.2433442

[12] Inma Hernández, Carlos R. Rivero, and David Ruiz. 2019. *Deep Web crawling: a survey*. Vol. 22. 1577–1610 pages. https://doi.org/10.1007/s11280-018-0602-1

[13] Masashi Kadoguchi, Shota Hayashi, Masaki Hashimoto, and Akira Otsuka. 2019. Exploring the dark web for cyber threat intelligence using machine leaning. In *2019 IEEE International Conference on Intelligence and Security Informatics, ISI 2019*. Institute of Electrical and Electronics Engineers Inc., 200–202. https://doi.org/10.1109/ISI.2019.8823360

[14] Milly Kc, Markus Hagenbuchner, and Ah Chung Tsoi. 2008. A scalable lightweight distributed crawler for crawling with limited resources. In *Proceedings - 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT Workshops 2008*. 663–666. https://doi.org/10.1109/WIIAT.2008.234

[15] M. Manvi, Ashutosh Dixit, and Komal Kumar Bhatia. 2013. Design of an ontology based adaptive crawler for hidden Web. *Proceedings - 2013 International Conference on Communication Systems and Network Technologies, CSNT 2013* (2013), 659–663. https://doi.org/10.1109/CSNT.2013.140

[16] Ericsson Marin, Mohammed Almukaynizi, and Paulo Shakarian. 2020. Reasoning About Future Cyber-Attacks Through Socio-Technical Hacking Information. Institute of Electrical and Electronics Engineers (IEEE), 157–164. https://doi.org/10.1109/ictai.2019.00030

[17] James Martin. 2014. *Drugs on the dark net: How cryptomarkets are transforming the global trade in illicit drugs*. Springer.

[18] Janet Perez. [n.d.]. Unprotected server held 419 million Facebook user records including phone numbers. | Komando.com. https://www.komando.com/happening-now/593814/another-huge-facebook-data-breach-exposes-419-million-user-records

[19] Priyanka Ranade, Sudip Mittal, Anupam Joshi, and Karuna Joshi. 2018. Using deep neural networks to translate multi-lingual threat intelligence. In *2018 IEEE International Conference on Intelligence and Security Informatics, ISI 2018*. Institute of Electrical and Electronics Engineers Inc., 238–243. https://doi.org/10.1109/ISI.2018.8587374 arXiv:1807.07517

[20] John Robertson, Ahmad Diab, Ericsson Marin, Eric Nunes, Vivin Paliath, Jana Shakarian, and Paulo Shakarian. 2017. *Darkweb cyber threat intelligence mining*. 1–137 pages. https://doi.org/10.1017/9781316888513

[21] Saad Saleh, Junaid Qadir, and Muhammad U. Ilyas. 2018. Shedding Light on the Dark Corners of the Internet: A Survey of Tor Research. *Journal of Network and Computer Applications* 114 (jul 2018), 1–28. https://doi.org/10.1016/j.jnca.2018.04.002

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

[23] Jiying Wang and Fred H. Lochovsky. 2003. Data extraction and label assignment for web databases. *Proceedings of the 12th International Conference on World Wide Web, WWW 2003* (2003), 187–196. https://doi.org/10.1145/775152.775179

[24] Shaohua Wang. 2010. *Crawling Deep Web Using a GA-based Set Covering Algorithm*.

[25] Wikipedia. [n.d.]. Surface web - Wikipedia. https://en.wikipedia.org/wiki/Surface{_}web

[26] Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, and Hai Jin. 2016. SmartCrawler: A two-stage crawler for efficiently harvesting deep-web interfaces. *IEEE Transactions on Services Computing* 9, 4 (jul 2016), 608–620. https://doi.org/10.1109/TSC.2015.2414931