2023 2nd International Conference for Innovation in Technology (INOCON)
Bangalore, India. Mar 3-5, 2023

Dark Web Crawling for Cybersecurity: Insights into Vulnerabilities and Ransomware Discussions

Ashwini Dalvi
Veermata Jijabai Technological Institute,
Mumbai, India
aadalvi_p19@ce.vjti.ac.in

Parth Kulkarni
K. J. Somaiya College of Engineering
Mumbai, India
kulkarni.pp@somaiya.edu

Aditya Kore
K. J. Somaiya College of Engineering
Mumbai, India
aditya.kore@somaiya.edu

S G Bhirud
Veermata Jijabai Technological Institute,
Mumbai, India
sgbhirud@ce.vjti.ac.in

2023 2nd International Conference for Innovation in Technology (INOCON) | 979-8-3503-2092-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/INOCON57975.2023.10101162

***Abstract*— Security professionals recognize the importance dark web to proactive and reactive security measures. The dark web data is considered a potential source for cyber-attack investigation. In the present work, the authors elaborate on the distinct tone of discussion about vulnerabilities and attacks on the dark and surface web. The authors crawl the surface and dark web for vulnerabilities and ransomware discussions. Authors considered log4j vulnerability, REvil ransomware, and WannaCry ransomware for discussion. The scope of the result is confined to offering evidence from dark web crawling. A dark web crawling mechanism is designed to crawl with entered keywords and harvest the hidden services to create a word-cloud visualization. The future scope of the proposed work is set to develop an analytic engine to investigate each hidden service from the pool of commonly harvested links to gather attack-specific insights.**

Keywords*— *Log4j Vulnerabilities, REvil Ransomware, WannaCry Ransomware, Dark web

I. INTRODUCTION

The surface web refers to the part of the internet that is easily accessible to the general public and can be found using search engines like Google. In contrast, the dark web is a part of the internet that is not indexed by search engines and can only be accessed using specialized software such as the TOR browser. The dark web is often associated with illicit activities and is known for hosting illegal markets and facilitating anonymous communication.

The dark web is a unique and largely unexplored part of the internet, and there is a great deal of research to be done on its content and activities. The dark web is where cybercriminals buy and sell illegal goods and services, including malware and hacking tools. Monitoring the dark web can help security researchers stay up-to-date on the latest threats and understand how they are being developed and disseminated. In addition, security researchers may use the dark web as a data source for academic or other research projects.

Researchers discussed different types of the web (surface, deep, and dark) and focused on the dark web and its structure and technologies [1]. Researchers also discuss law enforcement agencies' challenges in preventing and combating crime and terrorism on the dark web [2].

The dark web is a complex and often anonymous space, which makes it challenging for researchers to determine the motivations and intentions of its users. To address this, researchers may use data mining techniques to search through the vast amount of data available on the dark web and identify specific discussions or activities that may be relevant to their research [3]. By sifting through the dark web and identifying relevant discussions, researchers can better understand the potential threats to external enterprises and take appropriate action to mitigate them. For example, it can involve tracking the sale of illegal goods or services, identifying cyber threats, or investigating the activities of specific individuals or groups [4].

When studying vulnerabilities and attacks on the surface and dark web, the authors of the proposed work may be interested in understanding the different types of threats on these two parts of the internet. A vulnerability is a weakness in a system or network that attackers can exploit to gain unauthorized access or perform other malicious actions. An attack is an attempt by an individual or group to exploit vulnerabilities in a system or network to achieve a specific goal, such as stealing sensitive data or disrupting services.

On the surface web, vulnerabilities and attacks may include phishing scams, malware infections, and unsecured websites that allow attackers to access sensitive information. On the dark web, vulnerabilities and attacks may include illegal marketplaces selling stolen data or hacking tools and anonymous communication platforms that can be used to plan and coordinate cyber attacks.

Vulnerability monitoring and prioritizing could be challenging for enterprise security stakeholders. With several vulnerabilities published regularly, few of the vulnerabilities result in real-life attacks. However, no matter how slight the possibility of vulnerabilities could result in an exploit, proactive attention is given to vulnerability-related information by the cybersecurity community.

Researchers aim to consolidate the information regarding vulnerabilities, including but not limited to CVSS, vulnerability description, and vulnerability mentioned on different platforms, including social media platforms and the dark web.

The authors proposed to identify discussion on vulnerabilities and attacks on both the surface and the dark web. This information could improve the security of systems and networks on both the surface and the dark web. In the proposed work, the authors intended to comprehend the discussion about vulnerabilities and attacks on the surface and dark web.

The following paper includes related background work from the related work, methodology, result, and conclusion.

II. RELATED WORK

Researchers have studied vulnerabilities discussed on the dark web and surface web.

The researchers suggest that the number of software vulnerabilities discovered and publicly disclosed increases yearly [5]. However, only a small fraction of these vulnerabilities are exploited in real-world attacks. Organizations may have limited time and resources to patch all vulnerabilities and need a way to identify the most likely to be exploited.

To develop their exploit prediction model, the authors collected data on vulnerability mentions from online sources, including the white-hat community, vulnerability researchers community, and dark web/deep web sites.

Researchers compared the discussion of security vulnerabilities on three digital platforms: Reddit, Twitter, and GitHub [6]. It found that more vulnerabilities are discussed on Twitter, but conversations about them become more widespread on Reddit more quickly. Additionally, the study found that activity on Reddit and Twitter can be used to predict activity related to security vulnerabilities on GitHub accurately. The authors suggest that social media can be helpful information for understanding the public's perception and discussion of security vulnerabilities.

The researchers have developed a new method for predicting when attackers will exploit cyber vulnerability based on data from Twitter discussions [7]. Researchers did not use CVSS scores (a standard measure of the severity of vulnerabilities) in their approach. Instead, they proposed a framework based on a concept called "CVE-Author-Tweet (CAT) graphs" and a set of novel features derived from them. However, researchers did not provide further details on their framework's specific algorithms or techniques.

Researchers have developed a hybrid machine learning and knowledge representation model to predict whether a software vulnerability will be exploited in the real world based on data from online sources, including white hat communities, vulnerability research communities, and dark web websites [8]. The model outperformed standard scoring systems and a benchmark model using Twitter data, achieving a high true positive rate and low false positive rate, and was also able to withstand adversarial examples. The model was most effective as an early predictor of exploits that could appear in the wild.

The researcher proposed a system for identifying systems vulnerable to cyber-attacks based on discussions among hackers on the dark web [9]. The system combines DeLP (Defeasible Logic Programming) and machine learning classifiers in a hybrid approach. The researchers evaluated their system on hacker discussions collected from nearly 300 dark web forums and marketplaces

Researchers propose a deep learning model called the exploit-vulnerability attention deep structured semantic model (EVA-DSSM) that can proactively predict vulnerabilities likely to be exploited on the dark web [10]. The model is intended to aid in proactive cyber threat intelligence (CTI) by helping to prioritize vulnerabilities.

The researchers in this paper have developed a search tool to link published vulnerabilities to known exploits and vice versa [11]. This tool is meant to help enterprise security stakeholders prioritize their vulnerabilities by providing information about which vulnerabilities are most likely to be exploited. The tool uses data from various sources, including social media platforms and the dark web, to identify relevant vulnerabilities for a given exploit or vice versa.

Based on related work, the authors concluded that various research had been conducted to examine the effect of vulnerabilities and attacks referring to platforms ranging from surface and dark web to social media platforms. Therefore, the proposed work aims to observe differences in communication when attacks /vulnerabilities are discussed on the surface and dark web.

Due to the study's limited scope, the proposed work covers a discussion on Log4j vulnerabilities, REvil ransomware, and WannaCry ransomware.

A. Log4j vulnerabilities

Every security researcher knows about the log4j vulnerability that occurred on December 9, 2021. An open-source logging framework built on Java called Apache Log4j gathers and organizes data about system activities. Log4j is well-liked by Java developers, who have integrated it into countless other software products since it is easy to use, free to use, and successful in its intended usage. Log4j.

Log4j is a Java-based logging utility used to output log statements from applications to various output targets. For example, it can output log statements to a file, the console, a database, or other output targets.

In the past, several vulnerabilities have been discovered in Log4j that could allow an attacker to execute arbitrary code or gain unauthorized access to a system.

Researchers investigated the attitudes of Twitter users towards the Log4j library after it was discovered vulnerable to exploitation by hackers through remote code execution in December 2021 [12]. The study collected Twitter data using the VADER sentiment analysis tool and analyzed it using the CRISP-DM methodology. The results showed that tweets about Log4j were mainly positive before but primarily negative after the incident. However, there was a shift towards more positive sentiment in the five months following the incident, with the first month exhibiting a predominance of negative sentiment and the following month being predominantly positive. This study provides insights into the way discussions on social media circulate in response to significant security threats.

B. REvil Ransomware

Ransomware is malware that encrypts a victim's files and demands a ransom from the victim to restore access to the files. REvil, also known as Sodinokibi, is ransomware first discovered in April 2019. REvil is known for targeting large enterprises and has been used in several high-profile attacks.

Once REvil infects a system, it will encrypt the victim's files and display a ransom demand on the victim's screen. The ransom demand typically includes a deadline for payment and a specific amount of money the victim must pay to restore access to their files.

Researchers use fictitious characters and real-world hacking processes and outfits, which likely allows the case to present a realistic and engaging scenario demonstrating the

importance of comprehending ransomware exchange on the dark web [13].

A. WannaCry Ransomware

WannaCry is a type of ransomware that was first discovered in May 2017. It quickly spread to infect millions of computers in over 150 countries, causing widespread disruption and damage.

WannaCry was particularly disruptive because it was able to spread rapidly through a network, infecting all the computers on the network and encrypting their files. As a result, the attack significantly impacted organizations, including hospitals, banks, and government agencies.

During the WannaCry attack, an infected system attempts to connect to an onion server on the dark web [14]. It sends certain information about itself, including the user name and hostname. The onion server, also known as the command and control (C&C) server, is used by attackers to control the infected systems and receive ransom payments.

The response from the onion server may include an updated bitcoin address in the form of "c.wnry.". The bitcoin address "c.wnry" is likely a reference to a specific bitcoin address that the attackers have set up to receive ransom payments.

The following section covers surface and dark web communication about Log4j vulnerabilities, REvil ransomware, and WannaCry ransomware.

II. METHODOLOGY

The authors executed the following steps to conduct the study:

i. Identify the research question: Before beginning their research, security researchers must determine the specific questions they are trying to answer to guide the rest of the research process and ensure that the data collected is relevant and valuable.

The research question RQ formulated here is as follows:
RQ: Investigating the difference between discussion and communication o the surface and dark web regarding vulnerabilities and attacks.

ii. Determine the appropriate research methods: Researchers must then decide on the best methods to gather data on the dark web. For example, it may involve using specialized software to crawl the dark web and collect data and manually searching for and collecting data from specific websites or forums.

For further investigation, the authors decided to execute a customized crawler with input keywords—a dark web search engine to refer to pick up a keyword.
Katana -*Katana is a tool for gathering links from the surface, deep, and dark web. This tool needs a new word and finds all the Links related to it.*
Ahmia - *It is a search engine that is used to search for onion websites. It is a suitable means to find open community forums*

iii. Obtain the necessary tools and resources: To access the dark web, researchers will need to use specific tools and resources, such as a TOR browser or a virtual private network (VPN). They may also need specialized software or hardware to collect and analyze data.

Authors used self-developed customized crawlers to crawl surface and the dark web [15].

iv. Conduct the research: Once researchers have all the necessary tools and resources, researchers can begin their research. It may involve searching for and collecting data from specific websites or forums and analyzing it to look for trends or patterns.

The authors presented results in the result section.

v. Analyze and interpret the data: After collecting the data, researchers must analyze it to extract meaningful insights.

The authors offered word cloud visualization to quickly and easily identify trends or patterns in the data and highlight the important or relevant information.

III. RESULTS

A. Log4j Vulnerabilities

Several links are collected after crawling the surface web with the keyword 'Log4j' vulnerabilities. Table I shows a few sample links with their categories.

TABLE I. THE SURFACE WEB WITH THE KEYWORD 'LOG4J' VULNERABILITIES

Sr No	Category	Link
1	Github	https://0xsapra.github.io/website/CVE-2019-17571
2	Media	https://youtu.be/IB6YTr184Tw
3	Blog	Log4Shell: RCE 0-day exploit found in log4j, a popular Java logging package | LunaTrace (lunasec.io)
4	Article	https://www.zdnet.com/article/log4j-rce-activity-began-on-december-1-as-botnets-start-using-vulnerability/

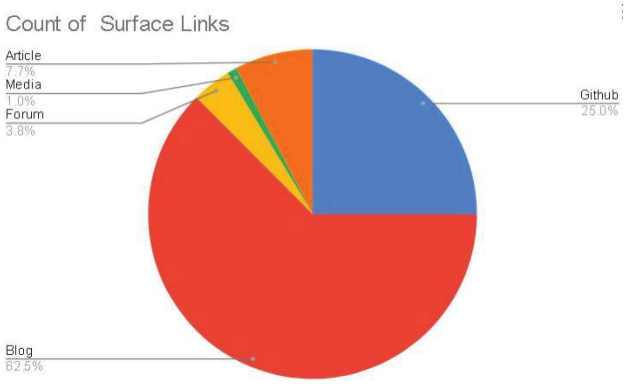

Fig. 1. Surface web link distribution in different categories with Log4j vulnerability keyword

Figure 1 represents surface web link distribution in different categories with the Log4j vulnerability keyword.

Further, dark web crawler crawled onion services; Table II presents a sample of collected services.

TABLE II. THE DARK WEB WITH THE KEYWORD 'LOG4J' VULNERABILITIES

Sr No	Category	Link
1	Bitcoin Cash Index	cashkxxqpzbgggg7.onion
2	cryptostorm	stormwayszuh4juycoy4kwoww5gvcu2c4t dtpkup667pdwe4qenzwayd.onion
3	RSS Bridge	mo2s6juoepmoob6d43mic7nctlp4gg66kk h7bdii3vwiwp626h6b2bqd.onion
4	nanochan	nanochanqzaytwlydykbg5nxkgyjxk3zsrct xuoxdmbx5jbh2ydyprid.onion

Table II depicts that dark web links need further processing for categorization. Also, most of the onion links become unavailable because of dark web hidden services.

The dark web hidden services can be difficult to locate and access. As a result, it can be challenging to categorize and organize the data collected from the dark web. Therefore, authors need to process the collected links further to identify and categorize them according to specific criteria, such as the type of content they contain or the language in which they are written.

It is also worth noting that many dark web links may become unavailable due to the ephemeral nature of hidden services on the dark web. Hidden services can be taken down or moved to new locations at any time, making it difficult to maintain a comprehensive list of active links. As a result, researchers may need to continually update and refresh their list of collected links to ensure they can access and analyze the most current data.

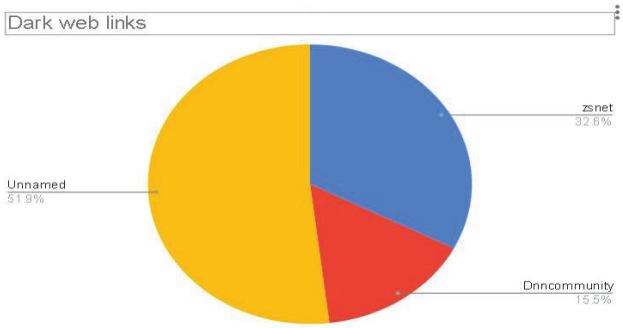

Fig. 2. Dark web link distribution in different categories with Log4j vulnerability keyword

Figure 2 represents dark web link distribution in different categories with the Log4j vulnerability keyword.

Authors proposed in the future scope to process the collected links further to identify and categorize them according to specific criteria by using specialized software or algorithms to extract relevant information from the data and organize it in a meaningful and helpful way.

For example, the authors proposed to use natural language processing techniques to analyze the content of the collected links and categorize them based on the content they contain (e.g., information about vulnerabilities, discussions about ransomware attacks, etc.).

The interpretation from crawled results includes that various blogs and newsletters were found discussing the log4j vulnerability and how hackers could exploit it. In addition, remote code execution codes for the vulnerability were published on GitHub, and tools for large-scale scanning for the vulnerability were developed.

There was widespread discussion of the vulnerability on social media platforms such as Instagram, Twitter, and Telegram, as well as on dedicated communities focused on exploiting vulnerabilities. The log4j vulnerability was found to be primarily targeted in the United States, with a particular focus on Alaska. U.S. agencies played a crucial role in responding to and mitigating the vulnerability.

B. REvil Ransomware

The 'revil' Keyword was used to search in the Google search engine, and 19 links were crawled using depth 4.

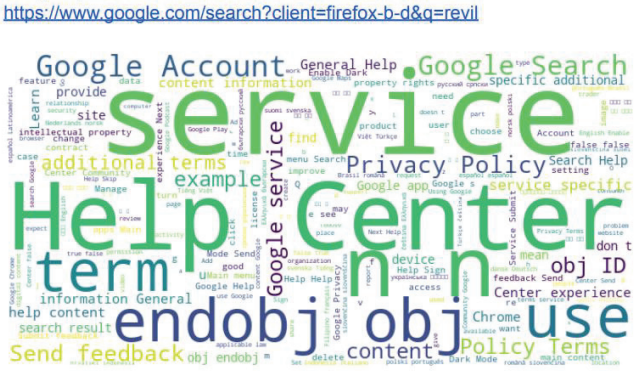

Fig. 3. Word Cloud of Revil keyword on the surface web

On the dark web, the 'REvil' Keyword was used to search in Torch(onion link), and a total of 1200 links were crawled using BFS and Depth 3 as arguments.

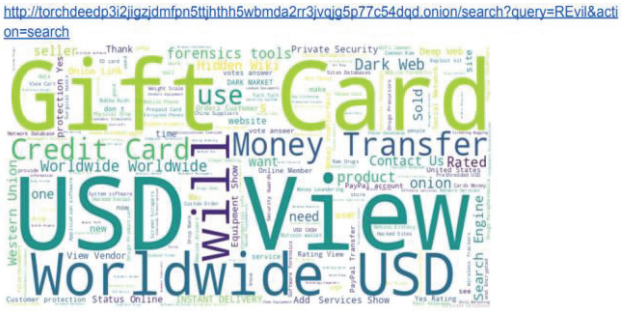

Fig. 4. Word Cloud of Revil keyword on the dark web

On the dark web, the 'Sodinokibi' Keyword was used to search Malwiki, a wiki for malware, and a total of twelve links were parsed using BFS and depth two as arguments. Sodinokibi is an alternative name to REvil, as this wiki did not give any results for the REvil keyword; Sodinokibi was used, and so far, it provided the most accurate word cloud.

Fig. 5. Word Cloud of 'Sodinokibi' keyword on the dark web

On interpretation, the authors concluded that much more info about REvil ransomware could be found on the surface web compared to the dark web. Even the analysis of the ransomware is available on the surface web.

C. *WannaCry Ransomware*

On the dark web, the 'WannaCry' keyword was used to search in Ahmia(onion link), a deep web search engine, and 40 links were parsed using a multithreading crawler with depth two as arguments.

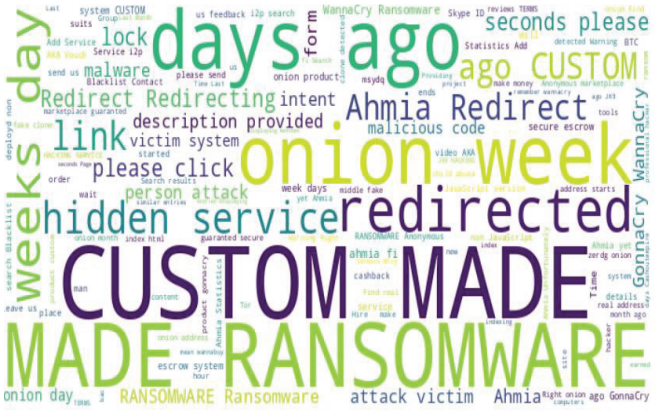

Fig. 6. Word Cloud of WannaCry keyword on the dark web

On the dark web, the site shown in figure 6 was found where the gonnacry/wannacry ransomware is being sold

http://kw4zlnfhxje7top26u57iosg55i7dzuljjcyswo2clgc3mdlivis wwyd.onion/product/gonnacry-wannacry/

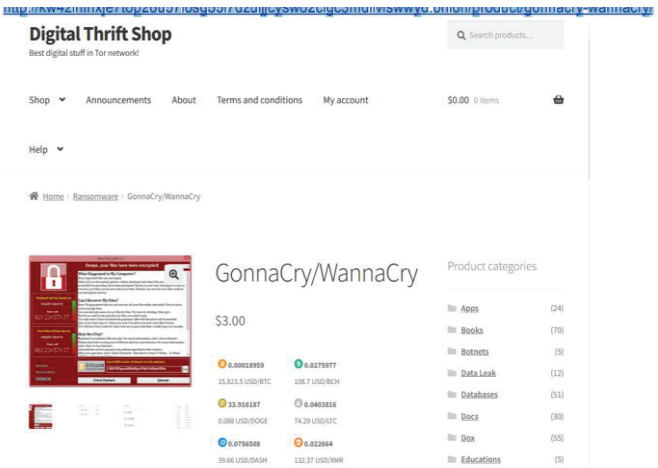

Fig. 7. WannaCry ransomware selling link on the dark web

The authors also found another website selling these three ransomware Ranion, Phobos, and Jigsaw. Figure 8 depicts the hidden service page of the same.

http://loomoostvvtblhkxtxxgtauxw7veooggkjl4cc53tezutmwwn oncpsad.onion/index.php/product/custom-made-ransomware

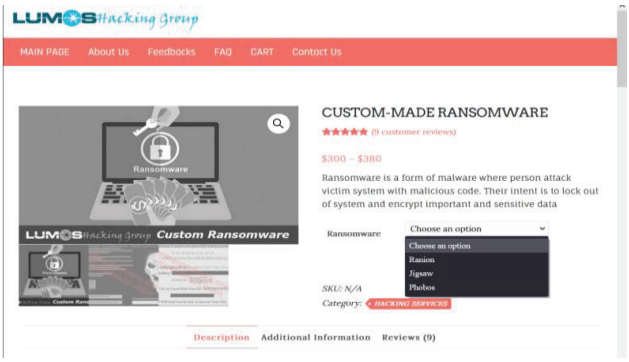

Fig. 8. Ranion, Phobos, and Jigsaw ransomware selling link on the dark web

III. CONCLUSION AND FUTURE SCOPE

Security researchers and professionals need to monitor the surface and dark web alike. By studying vulnerabilities and attacks on both the surface and the dark web, the authors attempt to understand the different types of threats on these two parts of the internet and how they can be addressed.

It is observed that communication on the dark web is sometimes not much reveling unless and until researchers enter invite-only forums or groups. At the same time, vigilant monitoring on the surface web will lead to potential security information.

The offered word cloud visualization can be used to quickly and easily identify trends or patterns and highlight important or relevant information. It may also be a starting point for further analysis or investigation.

The future scope of the proposed work includes the development of an analytic engine to investigate each hidden service from the pool of commonly harvested links. The authors plan to use the collected data to identify and analyze specific hidden services in more detail. The goal of this analysis is to gather "attack-specific insights," which likely refers to a deeper understanding of the types of attacks or vulnerabilities that are present on

REFERENCES

[1] Kavallieros, D., Myttas, D., Kermitsis, E., Lissaris, E., Giataganas, G., & Darra, E. (2021). Understanding the dark web. In Dark Web Investigation (pp. 3-26). Springer, Cham.

[2] Hurlburt, G. (2017). Shining light on the dark web. Computer, 50(04), 100-105.

[3] Dalvi, A., Salve, S., Zape, G., Kazi, F., & Bhirud, S. G. (2022). Security of Cyber-Physical Systems Through the Lenses of the Dark Web. In Proceedings of International Conference on Intelligent Cyber-Physical Systems (pp. 39-50). Springer, Singapore.

[4] Saini, J. K., & Bansal, D. (2019). A comparative study and automated detection of illegal weapon procurement over dark web. Cybernetics and Systems, 50(5), 405-416.

[5] Almukaynizi, M., Nunes, E., Dharaiya, K., Senguttuvan, M., Shakarian, J., & Shakarian, P. (2017, November). Proactive identification of exploits in the wild through vulnerability mentions online. In 2017 International Conference on Cyber Conflict (CyCon US) (pp. 82-88). IEEE.

[6] Horawalavithana, S., Bhattacharjee, A., Liu, R., Choudhury, N., O. Hall, L., & Iamnitchi, A. (2019, October). Mentions of security vulnerabilities on reddit, twitter and github. In IEEE/WIC/ACM International Conference on Web Intelligence (pp. 200-207).

[7] Chen, H., Liu, R., Park, N., & Subrahmanian, V. S. (2019, July). Using twitter to predict when vulnerabilities will be exploited. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3143-3152).

[8] Almukaynizi, M., Nunes, E., Dharaiya, K., Senguttuvan, M., Shakarian, J., & Shakarian, P. (2019). Patch before exploited: An approach to identify targeted software vulnerabilities. In AI in Cybersecurity (pp. 81-113). Springer, Cham.

[9] Nunes, E., Shakarian, P., & Simari, G. I. (2018, May). At-risk system identification via analysis of discussions on the darkweb. In 2018 APWG symposium on electronic crime research (eCrime) (pp. 1-12). IEEE.

[10] Samtani, S., Chai, Y., & Chen, H. (2022). Linking exploits from the dark web to known vulnerabilities for proactive cyber threat intelligence: An attention-based deep structured semantic model. MIS Quarterly, 46(2), 911-946.

[11] Dalvi, A., Ambekar, A., Kazi, F., & Bhirud, S. G. (2021, October). BM25 Algorithm Driven Search Tool for Linking Exploits to Vulnerabilities. In 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) (pp. 1-5). IEEE.

[12] Froissart, I., & Ring, J. (2022). Attitudes Towards Log4j: A Sentiment Analysis Study on Twitter Data.

[13] Datta, P. M., & Acton, T. (2022). From disruption to ransomware: Lessons From hackers. Journal of Information Technology Teaching Cases, 20438869221110246.

[14] Da-Yu, K. A. O., Hsiao, S. C., & Raylin, T. S. O. (2019, February). Analyzing WannaCry ransomware considering the weapons and exploits. In 2019 21st International Conference on Advanced Communication Technology (ICACT) (pp. 1098-1107). IEEE.

[15] Dalvi, A., Paranjpe, S., Amale, R., Kurumkar, S., Kazi, F., & Bhirud, S. G. (2021, May). SpyDark: Surface and Dark Web Crawler. In 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC) (pp. 45-49). IEEE.