Monitoring dark web for organization-specific information leakage

Sonam Chophel[1], Pema Wangmo[2], Tashi Wangchuk[3]
Department of Information and Technology, Jigme Namgyel Engineering College
jnec05180241@jnec.edu.bt[1], jnec05190273@jnec.edu.bt[2], tashiwangchuk@jnec.edu.bt[3]

ABSTRACT
The use internet has become a part of every organization's daily activity, and the information of the organization is becoming difficult to be protected from malicious actors even though the security mechanisms are put in place by the organizations. When a data breach occurs, organizations' sensitive information is being made available out on the dark web by cybercriminals. On the other hand, the organization finds it difficult to detect the presence of their data on the dark web. The dark web provides a platform for users who seek to maintain their privacy, but it is also a platform for hosting and carrying out prohibited activities. This paper presents the development of a dark web monitoring tool to detect the presence of leaked information of the organization in the dark web and alert the information security officer of the concerned organization. The tool was developed using Python and successfully experimented with the functionality of crawling, scrapping, and alerting components of the tool. The tool can be customized and used by any organization to monitor the presence of their organization-specific data in the dark web.

I. INTRODUCTION

With the rapid growth and development of the internet and with the online platform, data privacy is also becoming a major concern. And everyone knows the internet is one of the greatest inventions of humans, where many researchers of the various field will add various services to be used by any kind of the users and also maintain security at the same time.

There is a part of the internet known as the dark web, where an individual can do their work away from being attacked and monitored. This part of the internet provides a secure place for many individuals, who are worried about the privacy of their connection via the internet and still want to access the resource on the web while maintaining the privacy of their work. This can be achieved with the help of technologies that encrypt the connection and such users are mostly universities and education centers, business, and commercial industries.

However, many cybercriminals take advantage of the flip side of the technologies to conduct unethical activities within the dark web, like trading weapons, selling sensitive personal or corporate information, illegal drug businesses, providing access to child pornographic materials, buying and selling of credit card numbers, and etc. Cybercriminals aren't likely to stop anytime soon, so it becomes extremely important to do everything individuals can to safeguard personal information.

This paper explores the concepts for accessing the dark web, layers of the internet, types of activities carried out in the dark web, and the reasons and the technicalities behind the operation of the tor network. Finally, this study experimented with the monitoring of the dark web for detecting the presence of organization-specific information in the dark web.

II. AIMS AND OBJECTIVES

The main aim of this study was to monitor the dark web for organization-specific information leakage with the tool developed using Python, which will crawl the dark web pages and perform the pattern search to detect the presence of organization-specific data and alert the information security personnel of the organization.

III. LITERATURE REVIEW

The Surface Web, Deep web, and the Dark web, Surface web are part of the Internet. The surface web according to [1] [2], refers to the unencrypted part of the internet that the search engine like chrome can index, and the surface web contains the publicly available documents, contents, and information. Visible web, Indexed web or clear web are some of the terms used for surface web. Most individuals view the internet as the glacier at the sea and the only viewable part above the water is the surface web that makes about 4% of the internet.

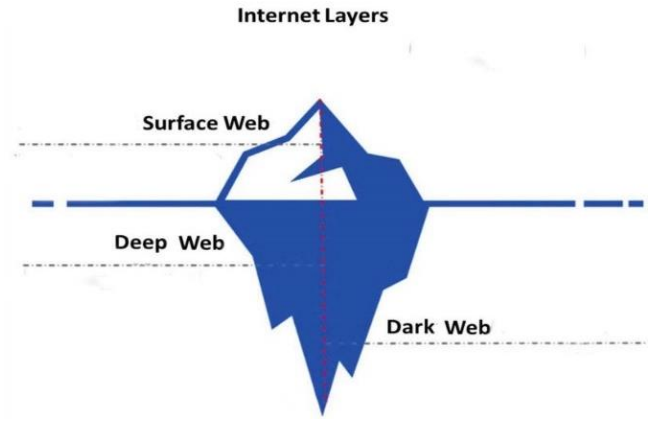

Figure 1: Parts of the Internet

3.1 The Deep Web

The deep web is the part of the internet that contains the webpages that cannot be indexed, which means the search engines cannot reach the particular webpage but the Deep web is accessible by using any standard browser stated by [2] and the reason mentioned for not able to index a webpage was because the web pages may be protected with a password and also the webpage can be unlined from other web pages which makes it unreachable unless one knows the correct URL.

3.2 The Dark web

Most of the activities per [2] are illegal activities such as drug trading, child abuse, weapons trading, and many more; Guccione [3] also refers to the dark web as the sub-set of the deep web.

The dark web is only accessible with special browsers or software. A few examples of such software or special browser are TOR, the Invisible Internet Project (I2P), or free net. So, Darknet is an infrastructure underneath the Dark web, which is the content and website that can be accessed by only a special browser.

The darkweb sites can be found pretty much like any other site, but Guccione [3] states that there are some important differences such as naming structure. Instead of ending in [.]Com, dark websites will end with [.]onion. Dark web sites also use a jumbled naming structure that forms URLs that are not easily identified and can hardly remember the correct URL.

Dark web monitoring is the method of looking for and keeping updates or track of personal information found over the section of the internet that is not accessible through normal standard browsers. According to Liu et al. [], more than 845,000 stolen credit/debit cards, over 1.2 billion stolen account credentials, and 1.3 million personally identifiable information were being detected while monitoring the activities over the dark web.

To access the dark web requires the use of an anonymous browser called TOR. TOR stands for "The Onion Router". The TOR according to Guccione [3] uses a series of proxy servers operated by thousands of volunteers around the planet, rendering the individual IP address untraceable and unidentifiable.

3.3 Activities of the dark web

According to [4] stated that due to nature its anonymity and privacy, many individuals that act as criminals use the dark web. The law enforcement agencies try to shut down unethical marketplaces and also, they use the dark web to reduce the exposure of government IP addresses and will ensure their anonymity over the dark web.

There is various marketplace within the dark web that deals with the vast number of drugs and illegal substances. And Silk Road is one of the most popular marketplaces and was like eBay over the dark web stated by many sources. The cryptocurrency is not similar to banking online even though the transitions occur online and [5] stated that cryptocurrency is the main standard method used for the operation of the dark web.

3.4 Web Crawling and scraping

Web crawling refers to the process that makes use of bots or an automated script to go through the contents on websites for archiving and indexing purposes. These automated scripts or bots are very well known by many multiple names, spider, crawler, and often spider bot. Crawlers have a lot of uses in various applications and research areas, aim to get updated data.

Web scraping according to [6], is the process that is mainly used to collect and parse raw data from the Web Pages, but some websites prohibit individuals from scrapping the data with automated tools like the ones we have created for this project because according to [6] the sites has a legit explanation to protect their data and also by making many repetitive requests to the website's server may utilize the bandwidths that will lead to slow down the website for other legit user and will potentially overload the server and will take the websites down completely.

IV. Methodology

4.1 Hidden Services URLs

The first obstacle in scraping the dark web is looking for and collecting the hidden services links called seed URLs to scrap. Some of the places which provide hidden URLs are Hidden Wiki Directories and the Hunchly Service. The URLs provided by the Hidden Wiki and the Hunchly service were saved to a file so that the Crawler can use them as the seed URLs.

4.2 To Build Crawler

A web crawler was developed that will fetch all the URLs and will go to the next URL and also fetch until all the URLs are collected and will be stored in a file. The crawler is designed such that it will not collect the same URL more than once.

4.3 Monitoring Process

The process of crawling the dark web, scraping, and alerting the user is shown in Figure 2. Firstly, the crawler makes anonymous connections to the TOR websites, then crawls the sites looking for the patterns to detect the presence of the organization-specific information. If specific information such as the email address is found in those TOR sites, the user will be alerted to make the timely intervention if in case the specific information has been leaked.

4.4 Alert System for the User

This system is designed to send the emails to the particular individual if any information is found after monitoring the dark web and with this system, it will be easy to alert the particular individual too with the message containing the URL from where scraper got the specific data and will let the individual take necessary action.

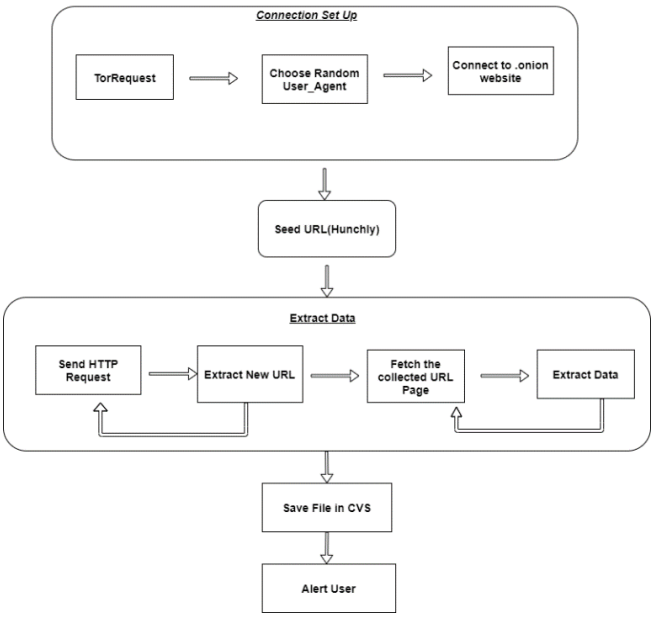

Figure 1: The monitoring process

V. RESULT

The developed tool was able to connect anonymously to the TOR site, crawl, and look for patterns specified in the code to find out whether the information of a specific pattern is present or not. If the information matching the pattern is found present, then the tool sends the alert to the specified email of the user. The following figure shows the tool successfully making the anonymous connection to the TOR site masking the actual IP address of the system.

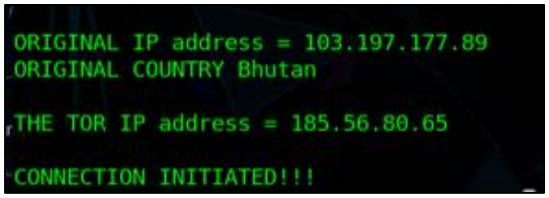

Figure 2:Making an anonymous connection

Figure 4 is the collection of [.]onion URLs saved to a file after visiting the seed URLs which can be used for scraping and pattern matching in the next.

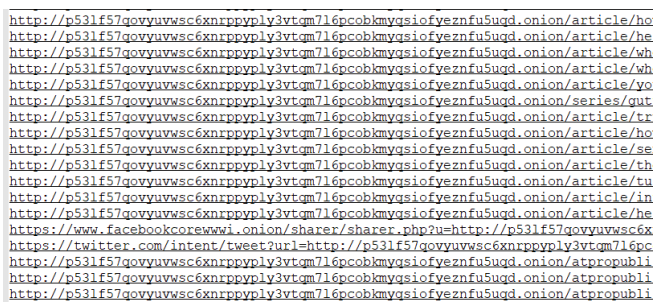

Figure 3:URLs collected by the crawler

The following figure shows a redacted URL link sent as a part of the alert to the user's email id whenever the particular pattern has been matched.

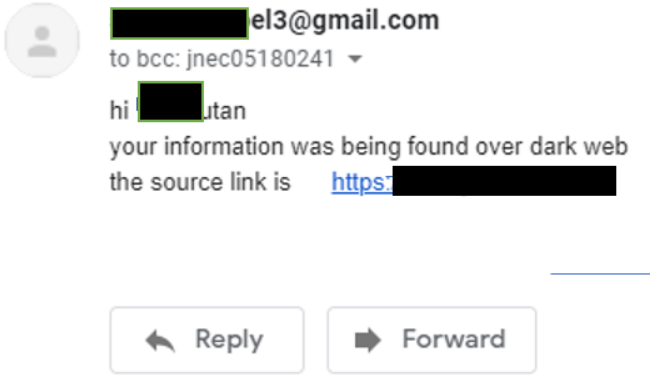

Figure 5: Alert sent to the user

VI. CONCLUSION

This paper experimented the monitoring and detecting the presence of organization-specific sensitive information in the dark web using the tool developed with Python. The experimental result showed that the tool was able to make the anonymous connection to the TOR sites, crawl, scrap, and perform the pattern matching. After finding the information matching the specified pattern, the tool can send the alert to the user successfully, thereby giving time to prevent further damage caused by the data breach or the information leak. This tool can be customized by any organization to suit their needs; change the patterns to focus on different keywords or patterns to suit the time, nature, and organization.

VII. REFERENCE

[1] L. Grustniy, S. Pike, A. Starikova, and H. Aver, "Darknet, dark web, deep web, and surface web - what's the difference?," *Daily English Global blogkasperskycom*. [Online]. Available: https://www.kaspersky.com/blog/deep-web-dark-web-darknet-surface-web-difference/38623/. [Accessed: 10-Dec-2021].

[2] B. AlKhatib and R. Basheer, "Crawling the dark web: A conceptual perspective, challenges and implementation," *Journal of Digital Information Management*, vol. 17, no. 2, p. 51, 2019.

[3] D. Guccione, "What is the dark web? how to access it and what you'll find," CSO Online, 01-Jul-2021. [Online]. Available: https://www.csoonline.com/article/3249765/what-is-the-dark-web-how-to-access-it-and-what-youll-find.html. [Accessed: 10-Dec-2021].

[4] S. Retzkin, *Hands-on dark web analysis: Learn what goes on in the dark web, and how to work with it*. Birmingham: Packt Publishing Ltd., 2018.

[5] N. Denic, "Government activities to detect, deter and disrupt threats enumerating from the dark web," 2017.

[6] D. Amos, "A practical introduction to web scraping in Python," Real Python, 06 march 2021. [Online]. Available: https://realpython.com/python-web-scraping-practical-introduction/. [Accessed 20 August 2021].