Cyber threat intelligence tool for leaks and cyber-criminal groups investigation in the Dark Web

Carolina Amado Fernández
carolamfdz@gmail.com
Computer Science and Engineering Department
Master in Cybersecurity
Universidad Carlos III de Madrid
Leganés, Madrid, SPAIN

ABSTRACT

A tool for monitoring and analysis of different cybercriminal groups dedicated to RaaS (Ransomware as a Service) has been developed for use mainly by the Spanish Law Enforcement Agencies, since its objective is to detect leaks of entities published by these groups on the Dark Web. The tool collects information about the different entities whose data is on sale and identifies the country of this attacked entity.

KEYWORDS

Cybersecurity, crawler, leaks, Dark Web, law enforcement

ACM Reference Format:
Carolina Amado Fernández. 2024. Cyber threat intelligence tool for leaks and cyber-criminal groups investigation in the Dark Web. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

According to studies, cybercrime is increasing exponentially in recent years, estimating that in 2024 the costs of the crimes committed will amount to $9.5 trillion USD [6]. One of the most common weapons used by these criminal groups are the data they obtain through the use of different attack techniques such as the following presented by Europol: trojan, worm, RAT, spyware, ransomware...

By means of these attacks they obtain an unimaginable amount of data, both information about the victim entities and personal data of the workers/clients/users of these companies... All this is published on the page of the criminal group located in the Dark Web, once here it can be divided into two phases: a first one in which with a countdown, they give a time to the attacked company to pay for not publishing the data or ask for a lower amount to extend the time of this payment, if the time established by the criminals passes, the data are put on sale for everyone, showing a summary of what was stolen or evidence to prove that the data have been stolen and give reliability to potential buyers. Figure 1 shows an screenshot of some LockBit3.0 leaks on sale.

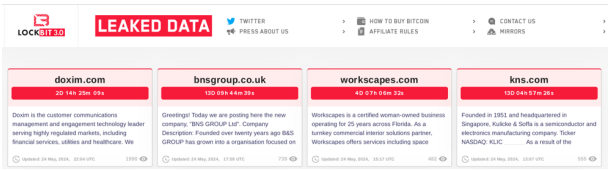

Figure 1: Screenshot of LockBit3.0 leaks on sale

The number of criminal groups[1] is increasing, being "Ransomware" the main attack directed towards companies with two main purposes: to steal information and use it for industrial espionage and/or get money. These purposes can be carried out at different stages of the attack. We are interested in the stage in which the stolen data is taken to the Dark Web to continue with an extortion or monetize them to a third party buyer.

One of the many problems that this type of attack can cause is that companies do not know that their information and that of their users is published on the Dark Web and can be bought and viewed by any user in the world, from a regular Internet user to a rival company using this information to its advantage. Although there are commercial tools that search for leaks in the Dark Web, the cost of these tools is very high and they are not within the reach of all organizations. In addition to the aforementioned problems, the affected entity can be affected at a legal level by the "Law of Protection of Personal Data and guarantee of digital rights [3]".

The proposed solution is the creation of a tool for monitoring and analysis of different criminal groups, using techniques such as web scraping and crawling to obtain the name of the entities whose data are exposed on the pages of these groups, and then analyze them and detect the country of the entities (e.g., Spanish Law Enforcement Agencies will be interested in leaks affecting Spanish entities to warn them and act accordingly).

2 METHODOLOGY

The methodology includes the following phases:

(1) **Requirements definition** After identifying a website publishing leaked data, the tool goes through the site and collects information about each leak. It is absolutely necessary that such collection be done anonymously, to prevent users of the tool from being identified and becoming the target of an attack, so the tool is provided to end users in a container.
(2) **Identification of sources**. CTI (Cyber Threat Intelligence)[5] and OSINT (Open Source Intelligence) techniques are used: through GitHub repositories, web pages, Deep Web and the

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Dark Web itself using Tor Browser, different pages of cyber criminal groups are detected.

(3) **Tool development**. Two techniques have been used to collect the relevant information from the web pages:
   - **Web scraping:** extracts information from a web page in an automated way by using a program that analyzes the HTML structure and collects the data.
   - **Crawling:** known as web crawling, automated process for the program (crawler or spider) to navigate the web systematically from one link to another collecting information, i.e., performing web scraping.

(4) **Acquisition and analysis of leaked data**. The data are collected and analyzed in detail as explained next.

(5) **Tool validation**. This tool will be shown to different profiles of cybersecurity professionals for validation.

3 TOOL DEVELOPMENT

For the development of the tool has been implemented a user interface in which you can choose between different criminal groups, once selected, you can access another page where you can see the companies attacked by date and filtering by country.

3.1 Inputs:

The urls of cybercriminal groups are obtained through different channels, these web addresses end in ".onion", this means that the websites are hosted on the Tor network, a decentralized network that provides anonymity and privacy to users. These addresses cannot be accessed from a conventional browser but must use Tor Browser to mask the user's identity and location.

3.2 Data processing:

Once accessed the website and with the scraping done the desired HTML code is obtained, all pages have a section within them in which the attacked entities are listed, so we have sought to obtain the HTML having to simulate the navigation through the page dynamically.

Once the HTML has been obtained, a program has been implemented for each criminal group in order to obtain a CSV with the urls of the attacked companies, there has to be a program underneath for each group since the programming of each page and the declaration of the information of the entities is different.

Then, with the url you get as much information as possible using the command line tool "whois"[8], which allows you to obtain information from the domain owner, in this case the entity, knowing the location of the same (this information is used to identify the entity's country), the associated name servers, the date of registration…The most significant data is stored in a CSV. Additionally the cases where the victim is identified as a Spanish entity are also recorded in a separate CSV.

It may happen that there are companies for which no useful information is obtained with the execution of the command "whois". Then, the tool can be asked to access the victim's website and by using NPL (Natural Language Processing) techniques[4] detect the probability of being written in certain language.

To address the continuous updating of the data leaks in the cyber criminal web sites[1], after each crawling the tool compares the freshly collected leaks with the ones collected in the previous crawling, and identifies if new entities are affected.

3.3 Outputs:

The tool user will be able to see, as explained in previous sections, the extracted information by accessing the different CSV files, but also can explore the collected data through the web interface. For example, in Figure 2

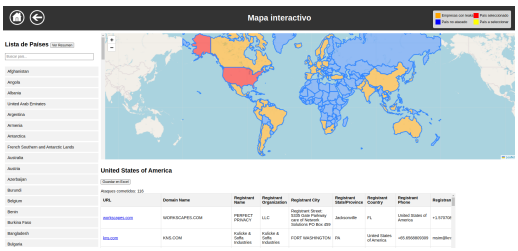

Figure 2: Screenshot of one of the tool's outputs: Map highlighting the countries with leaks and table.

Finally, to make the API we have used different Python and Golang libraries. The tool can be deployed as a Docker to avoid compatibility problems and to make its deployment easy, favoring the portability and usability of the tool.

4 CONCLUSIONS AND LIMITATIONS

The tool is a very useful cyberintelligence tool and will be of great use to the Spanish Law Enforcement Agencies as it will significantly speed up the process of searching and processing leaks.

The main limitation of the tool is that, due to the constant change and evolution of the criminal groups, the tool must be updated with the most active groups at all times. Also, as each group follows a different web site design (including anti-crawling and anti-DoS features) and they change their webs from time to time, it will be necessary to modify the crawler code.

REFERENCES

[1] MITRE ATT&CK. 2023. CTI/Groups. https://attack.mitre.org/groups/.
[2] Jesper Bergman and Oliver B.Popov. 2023. Exploring Dark Web Crawlers: A Systematic Literature Review of Dark Web Crawlers and Their Implementation. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10064292.
[3] BOE. 2018. Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. https://www.boe.es/buscar/pdf/2018/BOE-A-2018-16673-consolidado.pdf.
[4] DeepLearning.AI. 2023. A complete guide to Natural Language Processing. https://www.deeplearning.ai/resources/natural-language-processing/.
[5] CIS: Center for Internet Security. 2023. What is Cyber Threat Intelligence? https://www.cisecurity.org/insights/blog/what-is-cyber-threat-intelligence.
[6] Cybercime magazine. 2023. Cybersecurity Facts, Figures, Predictions and Statistics Sponsored by eSentire. https://cybersecurityventures.com/cybercrime-to-cost-the-world-9-trillion-annually-in-2024/.
[7] Yaman Roumani. 2021. Detection time of data breaches. https://doi.org/10.1016/j.cose.2021.102508.
[8] WHOIS. [n. d.]. WHOIS Search, Domain Name, Website, and IP Tools. https://who.is/.
[9] Cheng Huang Yong Fang, Yusong Guo and Liang Liu. 2019. Analyzing and Identifying Data Breaches in Underground Forums. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8686093.

[1] Usually, the leaks information does not include its publishing date.