A Proposed Method of Information Mining for Cyber Black Market Oriented to Tor Anonymous Network

Guangxuan Chen[1], Guangxiao Chen[2*], Qiang Liu[1], Bo Hu[1], Lei Zhang[3]
1. Zhejiang Police College, Hangzhou, China
2. Wenzhou Public Security Bureau, Wenzhou, China
3. Joint Logistics College, National Defence University, Beijing, China
Corresponding Author: Guangxiao Chen Email: ericcgx@qq.com

***Abstract*—With the development of information technology, the network black industry continues to expand, and cybercrime is becoming increasingly rampant. As a hotbed of crime, the darknet, with its anonymity, has also become a new front for online black market, bringing huge challenges to investigation and digital forensics. This article designs an automated darknet data crawling program that can achieve protocol conversion between darknet and surface networks, as well as data crawling of target sites on darknet. At the same time, the developed visualization program and data analysis program can achieve the display of target data in multiple dimensions. Combined with the periphery investigation technology to the website and social engineering, it can realize the mining of the criminal intelligence of the hidden network and provide reference for the investigation department.**

Keywords—cyber black market; tor; information mining; cyber crime; crawler

I. INTRODUCTION

With the continuous development of massive data, efficient algorithms, and super computing power, the scale of the digital economy continues to expand, ushering in one development opportunity after another. But at the same time, various types of illegal and criminal black markets are increasingly spreading. According to a research report, the black market that rely on the Internet for survival have formed a huge and intertwined benefit chain, and a platform-based, automated, specialized, refined, organized, and large-scale industrial network has been formed through multiple channels and strict division of labor, even reaching a scale of hundreds of billions.

From false account registration to online video traffic fraud, from illegal data theft to online "pornography and gambling", the cyber black market has been continuously upgraded and expanded, and even presented a new phenomenon of cross-platform implementation, making the existing challenges of combating, evidence collection, governance, and safeguarding rights more severe. For example, in the United States, many cyber criminals have begun to use Twitter as the main platform to cross-platform phishing attacks on unsuspecting PayPal users by registering false accounts. By imitating the official Twitter page of PayPal, fraudsters impersonate PayPal's customer service personnel to post malicious links to users on Twitter, thus tricking users to disclose bank account information. In China, the industrial chain of online black market also shows the trend of multi-platform expansion and migration. From post bars to WeChat, from QQ to Weibo, the cyber black market has formed a complete and diversified industrial chain, capturing profits based on upstream, midstream, and downstream links, posing enormous challenges and threats to the industrial operation ecology.

With the continuous strengthening of the crackdown on cyber black market by law enforcement departments, many criminals have gradually transferred the cyber black market to a more secret darknet. Because of its inherent hidden characteristics, the darknet is now widely used by criminals in cyber crimes. From personal acts of cyber hackers stealing secrets, digital currency transactions, private sales of illegal banned goods to the spying actions of the state will, all rely on the hidden services of the darknet. Therefore, the existence of the darknet poses a serious challenge to the prevention of new types of network crime, economy crime, endangering national security and other criminal acts, especially the public security and national security departments in tracing the source of the crime. At present, there are more and more illegal transactions of personal information, hacker software, hacker services, illegal payment platforms, bank cards, mobile phone cards, etc. on the darknet trading platform, which used to trade drugs, weapons, controlled drugs and other prohibited substances. Therefore, the darknet is also listed as one of the new network threats.

As the first link of combating the crime of cyber black market, collecting and mining information is an important means to carry out crime prediction, crime prevention, crime early warning and crime combating. Criminal intelligence mining refers to the activity of extracting, discovering, or identifying potential and valuable criminal intelligence from a large amount of collected data. This process often involves data crawling technology, data

mining technology, data cleaning technology, data analysis technology, etc. As more and more cyber black markets are gradually transferred from the open web to the darknet, it also brings greater challenges to information mining.

II. Related Work

At present, the rise of cyber black market has aroused the research interest of academia, industry and law enforcement departments. In 2020, Baidu and a research institute jointly released the "2020 Research Report on Cyber Black Market Industry Crime", which defined the cyber black market as an organized, purposeful, divisive and large-scale cyber crime with the help of Internet technology and network platform. The Report points out that the cyber black market and related industrial chain can be divided into upper, middle and lower reaches. Among them, the upstream cyber black market is an important source of resources in the entire industry chain, mainly including illegal collection and trafficking of personal information of citizens, theft of business secrets, illegal handling of telephone cards and other industries; Midstream cyber black market is responsible for developing and customizing a large number of illegal products that can be used for network crimes, such as building phishing websites, developing fake APP, illegal network promotion and other industries to use various illegal resources to carry out various network crimes in an automated way; Downstream cyber black market are responsible for realizing the results of illegal products and activities through transactions, involving many black network transactions and illegal payment channels, and tend to transfer the illegal and criminal gains after the successful transfer of illegal and criminal acts. The main ways of committing crimes are money laundering and cash withdrawal.

The cyber black market is a global problem, and scholars and research institutions in various countries have conducted extensive research and analysis on them. The following is a summary of scholars' research on the cyber black market in different countries. The United States is an important source and combatant of the cyber black market. American scholars mainly focus on issues such as hacking attacks, malicious software, and online scams, and propose various countermeasures and suggestions, such as strengthening network security, increasing vigilance, and enhancing cooperation. The United Kingdom is also an important source and combatant of the cyber black market. British scholars mainly focus on issues such as online gambling, false advertising, and online pornography, and propose suggestions such as strengthening legal regulation and establishing cross-departmental cooperation mechanisms. German scholars mainly focus on the economics and business models of the cyber black market and propose issues such as the connection between the cyber black market and traditional industries, and the impact of the cyber black market on society and the economy. Japanese scholars mainly focus on issues such as online scams, malicious software, and propose suggestions such as establishing cross-border cooperation mechanisms, strengthening technical defenses, and cultivating network security talents.

China is an important consumption market for the global black market. Chinese scholars mainly focus on the industrial chain, profit model, vulnerability mining, and other issues related to the black market, and propose suggestions such as strengthening technical research, enhancing legal regulation, and increasing international cooperation.

At the same time, some scholars have begun to study the strike strategies and governance methods of the cyber black market. Du studied the development trend of the cyber black market in China in recent years and used the LPA algorithm for empirical analysis; Lan Zhihan has established an online fishing fraud warning intelligence model based on Bayesian analysis; Gong Liu Hua discussed the importance and necessity of data mining technology in public security intelligence; Shao Dan proposed how to combine criminal hidden words with network tools in the process of handling illegal trading cases in network to collect information; Huang Xiaogen investigated it how the "fake base station" works in the process of telecommunications network fraud crimes; Xu Yongsheng and Liu Fangjie analyzed the fund transfer model in new fraud crimes involving networks and proposed investigations and control measures; Xie Ling launched a investigation and research on telecommunications network fraud crimes; Tan Yuxuan and Huang Juanjuan launched an in-depth research on the exploration of "network flow" data in the new type of fraud in telecommunications networks.

In summary, scholars in different countries have conducted research on the cyber black market in multiple fields and proposed different suggestions and measures for different problems. It requires strengthening cooperation and communication between countries to jointly respond to the challenges of the cyber black market.

III. Design and Implementation of Darknet Black Market Information Mining Framework

To conduct intelligence mining on target websites or target trading platforms suspected of committing crimes, it is first necessary to conduct peripheral investigations on the websites or platforms.

On the open web, the primary purpose of conducting peripheral detection on a website is to obtain basic information such as the website's registration information and owner information, which can be obtained in different ways. For example, we can query website owner information through website filing information, or use some commonly used network commands, network tools, and online query platforms.

However, for anonymous networks, such as darknet, the above methods cannot work. To this end, we have designed a darknet black market information mining

framework, aimed at crawling the black market data on the darknet. By classifying various types of black markets, we analyze the main forms, characteristics, and trends of black market activities and related network crimes, so that law enforcement personnel can monitor the darknet more targeted.

The framework consists of modules such as darknet domain name address collection module, darknet black market data crawling module, and black market data analysis module. The data analysis module also includes darknet data classification and popularity analysis. The specific darknet black market information mining framework is shown in Figure 1.

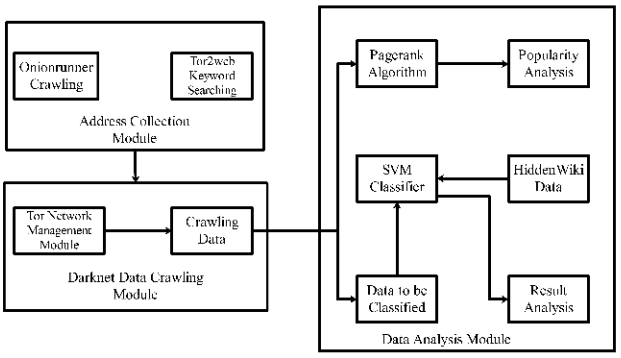

Fig. 1. Darknet black market information mining framework.

A. Design of Domain Name and Address Collection Module for Darknet

In order to obtain more darknet data content, it is first necessary to collect a certain amount of darknet domain names and addresses. The difference between a darknet domain name address and a surface web domain name address is that the hidden service address cannot be obtained through ordinary search engines. At the same time, in order to avoid tracking by law enforcement officials, addresses are frequently changed, and some have a very short life cycle, even only a few days to weeks. Therefore, only by discovering more domain names and addresses can we further crawl darknet data, classify and analyze the popularity of darknet data content.

Currently, there are several main ways to obtain domain names from darknet: UGC (User Generated Content) website data extraction, Tor anonymous network acquisition, onionscan crawling, and keyword queries in search engines.

B. Design of Data Crawling Module for Darknet Black market

Based on the active darknet domain name and address obtained, a darknet data crawling module is designed to obtain black market data. The darknet data crawling module includes two sub modules: Tor network protocol conversion module and data crawling module. First, perform network protocol conversion for Tor. Here, we achieve the conversion of SOCKS and HTTP protocols during the darknet crawler process through a program written using Python. Secondly, the method of combining the Scrapy framework with the depth first algorithm is used for data crawling. At the same time, different crawler frameworks have been designed for different types of darknet data, such as darknet markets and darknet forums.

The core idea of a crawler in surface web is to parse a given seed URL, and then access it through a simulator to automatically download web content. The crawling of content on Tor's dark web pages is similar to that on the surface web, except that Tor's darknet links do not have regularity and require active disclosure to obtain them. More data sources can be obtained through the comprehensive link extraction and active link acquisition methods. In addition, as access to Tor's darknet requires multiple nodes, and the response speed is low, in order to improve crawling efficiency, the following optimizations have been made during the crawling experiment in this paper:

(1) Direct protocol conversion between HTTP and SOCKS5. During the experiment process, the Tor browser client was opened, and then the protocol conversion was directly performed using the web driver under the selenium framework while calling the simulator for access.

(2) Separate crawlers from storage and caching, deploy a multi-node distributed system, and improve data processing efficiency. Here, we use Redis's memory cache technology to process crawl queue information, combined with multi-process crawling, it achieves parallel processing of task queues by multiple nodes, thereby improving crawl efficiency.

C. Crawling workflow of parent process

During the experiment, distributed crawling of the task queue was performed through multiple nodes, and multi-process crawling was deployed on each node. The workflow of the parent process of the crawl task is as follows:

- Step 1: Initialize a Tor dark network link and add the collected link list to the queue;
- Step 2: Parsing Tor darknet links in the queue;
- Step 3: Judge whether the link ends with .onion. If yes, proceed to Step 4, otherwise proceed to Step 2;
- Step 4: Perform a page count through the Redis cache to determine whether the number of URL pages crawled under the current link is less than a given number. If yes, proceed to Step 5, otherwise proceed to Step 2;
- Step 5: Count the secondary domain names through the Redis cache to determine whether the number of secondary domain names is less than a given number. If yes, proceed to Step 6, otherwise proceed to Step 2;
- Step 6: Enable child process to conduct crawling tasks

- Step 7: Determine whether the queue information can be obtained. If yes, proceed to Step 2. Otherwise, the queue is empty and end the task.

The crawling workflow of parent process is shown in Fig. 2.

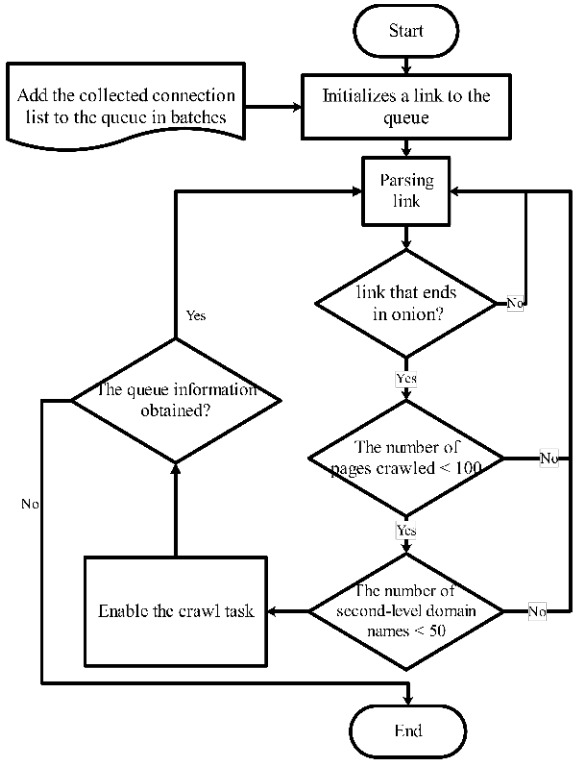

Fig. 2. Crawling workflow of parent process

D. Crawling workflow of child process

After obtaining the link to be crawled from the queue, due to the large amount of page content to be crawled during the crawling process, it is necessary to minimize repeated crawling of the URL. Thus, the Bloom_filter algorithm is used to avoid crawling the same link address. The detailed workflow of the child process is as follows:

- Step 1: Determine whether the link has not been crawled. If yes, proceed to Step 2, otherwise proceed to Step 8;
- Step 2: Add a link to the crawled list;
- Step 3: Insert the new link into the tor_url database through regular matching;
- Step 4: Configuring Socks5 agents through the selenium framework;
- Step 5: Use the non-interface mode to call the Chrome driver and obtain the page content through the get method;
- Step 6: Save the page text in html form to the tor_url_detail database;
- Step 7: Use the scraper framework to extract links that exist within a page;
- Step 8: De-duplicate links;
- Step 9: Add a new link to the queue.

IV. Experiment and Analysis

A. Data acquisition

After conducting peripheral investigations on the target website, then crawling site data for later analysis. The experimental environment for crawling is shown in Table 1.

TABLE I. EXPERIMENTAL ENVIRONMENT

CPU	Intel i5-8300H 4 cores, 2.3GHz
Memory	16G
OS	CentOS 8.2
Programming language	Python 3.11
Development software	PyCharm 4
Python Library	Scrapy, beautifulsoup, stylecloud, etc.

Scrapy is a powerful web crawling framework that can quickly extracts the required web page content by writing corresponding code in each project automatically generated by Scrapy. As shown in Figure 3, the Scrapy framework used in this article contains a total of 8 components, namely: the Scrapy Engine responsible for transmitting data signals, the Scheduler responsible for storing requests in the queue, the Downloader responsible for downloading and pushing requests, the Spiders responsible for obtaining specified entities (items) in the webpage, the Item Pipeline responsible for processing entities, the Downloader Middleware and Schedule Middleware responsible for processing requests and responses Spider Middleware and others responsible for data input and output.

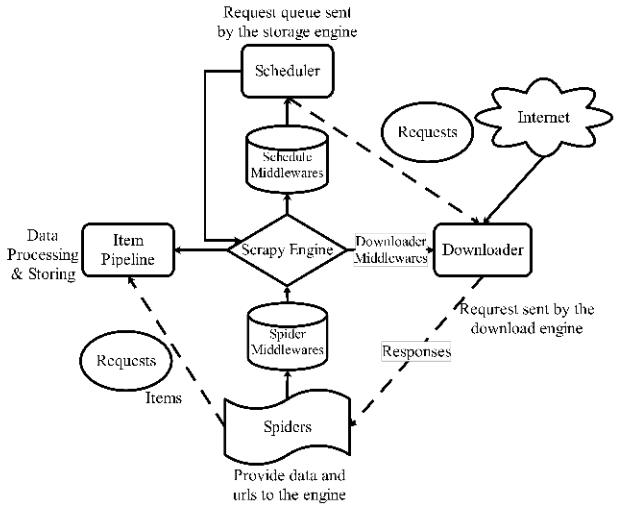

Fig. 3. Scrapy framework

B. *Data processing and analysis*

The collected raw data needs to be cleaned and organized, otherwise errors may occur during the later application and visualization of the data model. If there is garbled code in the CSV file, the text can be encoded and converted. If the encoding method is converted to ANSI, it can be displayed normally.

Then, it is necessary to eliminate duplicate data from the dataset. Here, we call the Pandas library in Python, use the duplicated function to find duplicate data, and use the sum function to query the number of duplicate data items. If there is duplicate data, use the drop function combined with 'inplace=true' to delete duplicate data from the source data based on the duplicated function query.

Similarly, using the isnull function to query missing data, if there is missing data, the loc function is used to complete it. After completing text encoding conversion, data deduplication, and missing data completion, the required data is basically cleaned.

To make the data more intuitive, it can be presented through visualization. A word cloud map can is created to present high-frequency words in the text and filter out a large amount of useless text information, highlighting high-frequency words in the data. Making word cloud charts requires the use of the jieba library and the call to the cut function in the library to achieve word segmentation. After completing word segmentation, a stop word library needs to be created, adding meaningless words that appear in the text to the stop word library to avoid presenting in the final word cloud charts. When creating word cloud images, the stylecloud library was mainly the used and the gen_ Draw in it. Figure 4 is the word cloud chart of the case data.

Fig. 4. Word cloud of personal information trading

It can be seen from the figure that the target website data obtained is filled with word such as data, ID card, US dollars, information, and business, indicating that the website may be a platform for trading of personal information, which is a key link in the cyber crime industry.

In addition, the evolution of cyber black market and their relationship with cybercrime can be analyzed from both temporal and spatial dimensions, in order to obtain some models on the overall situation and provide data support for combating and governing cybercrime.

CONCLUSIONS

This article designed an automated dark network data crawling program oriented to the tor dark network trading platform, which can achieve protocol conversion between dark networks and surface networks, and data crawling of target sites in dark networks. At the same time, the developed visualization program and data analysis program can achieve the display of dark network target data in multiple dimensions.

The future research will focus on the evolution of dark network black market transactions in time, space, and other dimensions, as well as how to combine peripheral investigation techniques to provide support for combating dark network crimes.

ACKNOWLEDGMENT

In this paper, the research was sponsored by the Science and Technology Plan of the Ministry of Public Security of the People's Republic of China under Grant No.2021JC36.

REFERENCES

[1] GX Chen, GX Chen, L Zhang, Q Liu, "An Incremental Acquisition Method for Web Forensics," International Journal of Digital Crime and Forensics, vol. 13, no. 6, pp. 1-13, 2021.

[2] GX Chen, GX Chen, D Wu, Q Liu, L Zhang and XS Fan, "An improved Simhash algorithm based malicious mirror website detection method," Journal of Physics: Conference Series, vol. 1971, pp. 1-7, 2021.

[3] Guangxuan Chen, Guangxiao Chen, Di Wu, Qiang Liu, Lei Zhang, Xiaoshi Fan, "A Selenium-based Web Application Automation Test Framework," Proceedings of 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence, pp.257-261, 2021.

[4] Gopal L S, Prabha R, Pullarkatt D, "Developing Efficient Web Crawler for Effective Disaster Management," EGU General Assembly Conference Abstracts. EGUGA, 2021.

[5] Jin, Yuping, "Development of Word Cloud Generator Software Based on Python," Procedia Engineering, 2017, pp.788-792.

[6] Sport T, "Fighting Cyber Crime:Stopping ransomware, and what happens when it can't be stopped," Grain journal, 2022(1):50.

[7] Ahdi I, Abdulajid S, Suwarti S, "The Effectiveness of Cyber Crime Handling by the Special Criminal Directorate North Maluku Polda," International Journal of Multicultural and Multireligious Understanding (IJMMU), 2021(9).