Cyber Security and Automation in Darkweb Ecosystem

Muhammad Daniyal Zakir
Department of Computer Science
Federal Urdu University, Karachi
daniyalzakir03@gmail.com

Abstract*— The Tor browser is used to access the Tor network; it automatically anonymizes a person identity and location. Cyber-criminals, hackers and weapons dealers use Tor browser to stay anonymous while conducting their illegal activities and business, because it avoids getting caught. Tor browser helps to access content that are geo-restricted, to bypass censorship and visit specific websites such as those on the dark web. Dark web is the dark part of the internet that contains multiple illegal networks, people buy and sold all sorts of illegal items, such as illegal drugs, stolen vehicles and weapons. The currency of the dark web is bitcoins an online digital cryptocurrency with no ties to a central bank or government, purchases made with bitcoins have no taxes and no chance of an account being frozen. Machine learning, data mining, automation tools and analytics tools are poised and offensive weapons in the fight against dark web illegal activities and cybercrime. For this purpose, cyber-security intelligence is use for information gathering and analysis of large amounts of unstructured data, we present BlackWidow, a highly automated Web Crawler Bot system that monitors Dark Web services automatically and collects data from Dark web webpages. For enhancement, the ATOL framework is used (an automatic keyword discovery and analysis algorithm) by using the available small volume of training data which came from the BlackWidow (Crawler). ATOLKeyword algorithm finds those keywords that are difficult to detect by humans, e.g. "phpcredlocker" for the hacker category "neurogroove" for the drugs category and "flash-ball" for the weapons category, all these keywords are relevant and valid keywords, but very difficult for a human to identify. ATOLKeyword can categorize millions of keywords in the Dark Web. By implementing the ATOL framework with BlackWidow, system accuracy increases by 12%.

Keywords—Dark Web, Automation in Dark Web, Dark Web analysis, Cyber intelligence in Dark Web, BlackWidow Crawler, ATOL Framework.

I. INTRODUCTION

Tor (The Onion Router) helps to browse the internet anonymously, it's a free open-source software developed by the US navy. Tor browser is used to access the Tor network, Tor-browser will automatically anonymize a person identity and location, the traffic coming to it is heavily encrypted and slowly decoded one layer at a time at the different nodes, it's a worldwide network of servers specifically made for private communication.

Tor browser is used to access content that are geo-restricted, to bypass censorship and visit specific websites such as those on the dark web, websites on the dark web won't be visible to a person, when the person is using normal web browsers such as Chrome, Safari, Opera, UC Browser or Firefox. Tor is completely legal in most jurisdictions, especially in the western world. Cyber-criminals and hackers use Tor to stay anonymous while conducting their illegal business and it avoids getting caught when they're going about their illegal activities because it provides access to the dark web [1], [2], [3]. Dark web is the dark part of the internet that contains multiple illegal networks, people buy and sold all sorts of illegal items, such as illegal drugs, stolen vehicles and weapons.

The currency of the dark web websites are Bitcoins, an online digital cryptocurrency with no ties to a central bank or government, purchases made with bitcoins. There are no taxes, no chance of an account being frozen, and rampant opportunity for the currency to be used for things like prostitution, gambling, buying stolen art, purchasing weapons and more. Machine learning, data mining, automation tools and analytics tools are poised and offensive weapons in the fight against dark web illegal activities and cybercrimes [3]. For this purpose, cyber-security intelligence is use for information gathering and analysis of large amounts of unstructured data, we present BlackWidow (Crawler). A Bot system that monitors dark web services automatically and collects data from dark web webpages [1]. For enhancement, a framework is used with BlackWidow known as ATOL (an automatic keyword discovery algorithm) by using the available small volume of training data which came from the BlackWidow. ATOLKeyword algorithm discover those keywords that are difficult to detect by humans [2].

II. AUTOMATION IN DARK WEB

Emerging machine learning, data mining, automation tools and analytics tools are poised to become formidable offensive weapons in the fight against dark web illegal activities and cybercrime.

A. BlackWidow Crawler

In this section, we analyse and discuss the challenges related to information gathering in the Dark Web for cyber-security intelligence purposes. To facilitate information collection and the analysis of large amounts of unstructured data, we present BlackWidow, a highly automated Web Crawler (Bot) system that monitors dark web services automatically and fuses the collected data in a single analytics framework [1]. BlackWidow architecture relies on a Docker-based microservice that uses both pre-existing and customized machine learning (ML) tools [1]. Less than two days of monitoring time, BlackWidow collects years of information in the areas of fraud monitoring and cyber-security and it can infer relationships between forums and authors. And can detect trends for illegal activities and cybersecurity-related topics [1], [2]. BlackWidow architecture has many capabilities that we will discussed in our research paper.

1) Identifying Dark Web Forums

To overcome the issues and challenges, target forums are recognized by hand in order to bootstrap the process [1]. BlackWidow seeks to analyze and investigate the content of Dark Web forums in order to gain additional links and addresses to other targets in a more automated way.

2) Establishing Anonymous Access to Forums

By using Docker containers, we establish anonymous gateways to the identified forums [1]. Tor will use to access the (VPN) Virtual Private Networks and other hidden services for regular Dark Web sites. It will become necessary to add a custom function to the BlackWidow, that emulates typing and clicking actions to automatically log in and determine whether or not gateway has successfully logged into the target.

3) Parsing Raw HTML Data

The data to be collected as BlackWidow in the (HTML) Hypertext Markup Language is collected over a headless window. Based on the web forum layout, it can be complex to extract organized and unorganized data from HTML web pages, we adapt to each forum's layout [1]. At first, this approach may seem expensive; several forums have a consistent layout [1], so that different forums can reuse the same parsers. HTML parser output for each page is a structured file with only the HTML web page text information.

4) Translation of Raw Data in Foreign Languages

Google's translation API is required to convert all non-English content into English content, because the collected raw data contains content in several languages [1], obtaining state-of-the-art translations which enable complex data modelling and relationship analysis over forums in different languages.

5) Identify Topics

Messages in forums are generally structured into the categories and threads, seeing which threads cover the same topics, is not always obvious. To facilitate the anal ysis of the patterns across different threads and forums. Through predictive topic analysis [1], BlackWidow correctly defines topics. BlackWidow uses unsupervised text clustering techniques to recognize or identify messages into categories based on the Latent Dirichlet Allocation (LDA). So, these groups are assigned to higher interest categories like exploits, botnets, leaks, databases and (DDOS) (Distributed Denial Of Service) attack.

6) Identify Cyber Security Trends

For identifying trends in cyber-security, BlackWidow fuses the messages, topics, and categories from the different forums and computes aggregated time series. These time-series form the basis to identify trending topics, e.g. when the time series experiences a high growth or decline over short periods [1]. Long-term trends are also detectable given that all collected messages are time-stamped and thus provide information over the whole lifetime of the forum.

7) *Gaining Access*

BlackWidow needs personal accounts to authenticate on each site for crawling data. Since most of the forums require some sort of login to access the site. The way to acquire such logins differs on each site. While certain sites on the dark web require a valid email address, while other sites have higher entry barriers with reputation systems, dealings with active participation [1], or even requiring users to first buy some credits (digital currencies).

8) Real-Time Intelligence

The collected data from the Dark Web is concentrated on a static environment. In contrast of collecting one or multiple snapshots of the targeted environment, BlackWidow aims to provide much faster intelligence and insights. Real-time capability is a core requirement due to the limited lifetime of the target forums. A high grade of automation is required to enable these functionalities, from the collection to the live analysis of the data.

By implementing BlackWidow as a collection and analysis tool, show that monitoring of the Dark Web can be done with relatively little resources and time.

B. ATOL Framework

For automatic categorization and for complex analysis of the Dark Web, a very helpful framework is discussed (ATOL) [2]. ATOL uses an algorithm (ATOLKeyword) for automatic keyword discovery and analysis, by using the available small volume of training data which came from the BlackWidow (Crawler). ATOLKeyword find those keywords that are problematic to detect by humans, e.g. ATOLKeyword detect "phpcredlocker" for the hacker category "neurogroove" for the drugs category and "flash-ball" belongs to the weapons category all these keywords are relevant and valid keywords [2], but very difficult for a human to identify. ATOLKeyword can categorize millions of keywords in the Dark Web. ATOLKeyword algorithm uses three core components, keyword discovery mechanism, classification framework and a clustering framework [2].

1) Keyword Discovery Mechanism

For each onion keyword category, an analyst will manually provide a set of keywords, e.g., the "Weapons" category has keywords like gun, calibre, silencer and Glock etc [2]. The goal of ATOL, is to automatically discover relevant keywords using data from multiple sources, e.g., title/content words from onion text as well as existing manually tuned keywords. This can be achieved by using the method ATOLKeyword.

2) Classification Framework

The approach is to trained Naive Bayes, Logistic Regression classifiers and SVM, using different kinds of weighting schemes e.g. TFICF, BOW and TFIDF [2]. to represent the training/test data points. By using the

classification framework (ATOLClassify), system accuracy increases to 12%.

3) Clustering Framework

The approach of unsupervised and semi-supervised clustering used in ATOL. ATOL use unsupervised clustering in the absence of labelled training data and use semi-supervised clustering in the presence of sufficient labelled training data [2] to train a classifier. By using the clustering framework (ATOLCluster), system accuracy increases to 7%.

By implementing the ATOL framework, system accuracy increases to 12% by using a classification model. It discovers categories on unlabelled onions-sites data and discusses applications of ATOLKeyword algorithm. ATOL can also be used for multi-classifier analysis, and thematic census mining, theme learning, graph analysis and supporting various analyses and investigations on the Dark Web.

III. FUTURE WORK

In the future, we would like to use ATOL framework on both, Surface Web and Dark Web, by extending the categorization framework of ATOL and using it, for automatic generation of the portal, where websites like blogs, wikis and forums can be added automatically to different categories of the portals. We would also like to use ATOL with DarkOwl database because it allows us to access the world's largest database of Dark Web contents and monitor the presence of your data on the darknet. DarkOwl Vision anonymously, automatically and continuously collects, indexes and ranks actionable data coming from the darknet, DarkOwl can enhance our system performance and give more accuracy.

IV. CONCLUSION

Dark Web is the dark part of the internet, it can be accessed by using Tor browser which anonymizes a person identity and location. We use a combination of BlackWidow (Crawler) for extracting Dark Web pages and ATOL framework for analysis, monitoring and auto keyword generation, this combination is highly beneficial, for detecting cyber-crimes and other illegal activities, happening on the Dark Web. By using the ATOL framework with BlackWidow (Crawler), our system accuracy increases by 12%.

ACKNOWLEDGMENT

The author of this research paper would like to thank. Dr. Steven Cheung for developing the BlackWidow (Crawler) and (DARPA) Defence Advanced Research Projects Agency for developing ATOL framework.

REFERENCES

[1] Matthias Schäfer and Markus Fuchs, “BlackWidow Monitoring the Dark Web for Cyber Security Information,” Jun. 2019.

[2] Shalini Ghosh, Phil Porras, Vinod Yegneswaran and Ashish Gehani, “Automated Categorization of Onion Sites for Analyzing the Darkweb Ecosystem,” Aug. 2017.

[3] Hsinchun Chen, “Uncovering the Dark Web: A Case Study of Jihad on the Web,” Jan. 2008.