Crawling the Dark Web: A Conceptual Perspective, Challenges and Implementation

Bassel AlKhatib, Randa Basheer
Syrian Virtual University, Syria
t_balkhatib@svuonline.org
randa.s.basheer@gmail.com

Journal of Digital Information Management

ABSTRACT: *Internet and network technologies have evolved dramatically in the last two decades, with rising users' demands to preserve their identities and privacy. Researchers have developed approaches to achieve users' demands, where the biggest part of the internet has formed, the Deep Web. However, as the Deep Web provides the resort for many benign users who desire to preserve their privacy, it also became the perfect floor for hosting illicit activities, which generated the Dark Web. This leads to the necessity of finding automated solutions to support law and security agencies in collecting information from the Dark Web to disclose such activities. In this paper, we illustrate the concepts needed for the development of a crawler that collects information from a dark website. We start from discussing the three layers of the Internet, the characteristics of the hidden and private networks, and the technical features of Tor network. We also addressed the challenges facing the dark web crawler. Finally, we presented our experimental system that fetches data from a dark market. This approach helps in putting a single dark website under investigation, and can be a seed for future research and development.*

Subject Categories and Descriptors
[**H.3.5 Online Information Services**]; Web-based services: [**C.2.1 Network Architecture and Design**]

General Terms: Web Processing, Deep Web, Dark Web, Web Networks

Keywords: Dark Web, Web Crawler, Tor Network, Dark E-Markets, Information Collection

Received: 11 October 2018, Revised 15 December 2018, Accepted 24 December 2018

Review Metrics: Review Score: 4.5/6, Revise Scale: 0-6, Inter-reviewer Consistency: 76%

DOI: 10.6025/jdim/2019/17/2/51-60

1. Introduction

Internet is one of the most extensive human achievements that witness fast-pacing evolvement, attracting researchers of different fields to add more services and facilities to it to make it available for all different users, from individuals to societies and institutions, but insuring privacy and security at the same time. There lies the special part of the internet where users can perform their activities away from tracking and monitoring, and even the geographical location of the hosting service cannot be identifiable. That part forms a secure place for many users, who are concerned about preserving the privacy of their connections through the internet, and desire to use information resources on the web while preserving the secrecy of their activities. They can achieve that with the help of special technologies that encrypt the connection and redirect traffics throughout several nodes on the network [1]. Such users are like universities, educational and scientific centers, business companies and commercial institutions.

However, many of the malicious activities take advantage of the technologies used in the hidden web, like trading drugs, weapons, pornography, child abuse, malware and hacking software, fraud, forgery, identity theft, and many others.

The term Dark Web originated since the beginning of 2000s [2], and many researches have studied it. However, most of these researches specialized in the field of terrorism and extremism, like determining the identity of the terrorist groups and defining their ideologies.

As for interests in other aspects on the dark web, like the drugs trading traffic, showed up just recently, and there are few researches about the subject.

In this paper, we summarize the concepts needed for researching on the Dark Web. We start from the internet layers, identifying the characteristics and the type of activities practiced in each layer. We describe the dark web in a more extensive explanation, and the technology behind operating Tor network. We illustrate the crawlers mission in general (as in surface web), and their specific mission in the dark web, and the challenges crawlers face during these missions. Finally, we demonstrate how it is possible to crawl a dark website by our experimental system.

2. Related Work

2.1. Web Invisibility

Sherman and Price [3] held the first discussion on what they stated as “the four types of invisibility” in the hidden web. These types generated from two reasons: it is either because search engines crawlers exclude some web contents, or due to the technical characteristics of the website or parts of it.

Furthermore, other studies held several discussions about the most hidden part of the invisible web, hosted by the private networks, or the “Hidden Services”. [4] [5] [1]

2.2. Accessing Hidden Web

Raghavan and Garcia-Molina [6] introduced the “Human-assisted approach” while developing a Hidden Crawler, a crawler that can crawl and fetch data from hidden databases, by a semi-automatic operation to process search forms and get data from hidden databases.

Other studies in the form filling aspect discussed methods to make crawlers able to send requests containing form entries to local search interfaces of hidden databases, and then continue crawling according to these entries. [7] [8]

2.3. Data retrieval from Dark Web

From the first 2000s, researchers introduced approaches for studying activities that started to take place increasingly and fast on the Dark Web. Most of these researches focused on terrorists groups around the world. They suggested the basic practices needed to collect data from dark websites, to analyze them later, or to be a part of a knowledge management system. These practices start from identifying terrorists groups from reliable resources, identifying a list of websites to start from, expanding the list by link analysis that adds more links to the list, and finally collecting data from the visited pages.[9]

Baravalle et al. [2] focused on studying e-markets on the dark web, and specifically “Agora”, an e-market for trading drugs and fake IDs and documents. The crawler makes a simulation of the authentication process for a user login to the market, and then it collects data using the classic web development platform LAMP Stack.

Kalpakis et al. [10] designed a focused crawler for the purpose of searching topics about Homemade Explosives (HME) tutorials and the places where vendors trade materials and tools used in HME, it works according to previously defined classifications to find required web resources starting from related “seed pages”.

Zulkarnine et al. [1] discussed in a more extensive way the technical characteristics of websites hosted on Tor hidden services presenting an approach of a dark crawler to collect data from terrorists and extremists’ websites, depending on social network analysis and content selection.

Pannu et al. [11] presented a crawler approach that serves a search engine specialized for detecting information from suspicious and malicious websites. Its main process starts from downloading the HTML files of the pages from a previously instantiated list of seed URLs, following the links to other Tor websites for scrapping as well. After archiving the pages, the system scans them to fetch information.

3. Internet Layers

To reach the depth, first we have to distinguish the parts of the web (or the internet in general) and determine the characteristics of each part.

Researchers divide the internet into three layers:

3.1. Surface Web

It represents part of the internet that search engines (like Google) can index. It has other terms such as Visible Web, Lightnet, Indexed Web, Clear Web, and others. To complete the indexing process, search engines use partial software pieces called Web Crawlers. The basic role of the crawlers is to discover webpages on the internet (in which researchers estimate the size of the indexed pages on the Surface web by more than 4 billion pages).When the crawler visits a page, it looks for any outgoing links from that page to other pages, and then visits those pages and so on, while sending the extracted data to the search engine that stores the data in indexes in the form of keywords describing the webpage and its location. Thus, when a user makes a search, the search engine matches the words of the search query with those stored in the index, extracts the matching pages, and displays them to the user. [12] [13]

3.2. Deep Web

It represents the part of the internet that contains un-indexable webpages, i.e. the search engines cannot reach, and they are unlinked to other pages on the Surface Web. It also has other terms like Deepnet, Hidden Web and Invisible Web, and researchers estimate that Deep Web forms 96% of the internet. [12] [14]

The inability to index a webpage may be due to many reasons, such like: [3] [12]

- The webpage owner protects it by a password so it prevents the crawlers from accessing it.
- A specified number of accessing times may restrict reaching the page, and after that number, the page might become unavailable before the crawler reaches it.
- The *robots.txt* file of the website is set to tell the crawler not to crawl on that site or parts of it, and this file is located on the root of the website.
- The page is hidden or unlinked to any other page on the website or other sites, and it is unreachable unless the whole URL is well known.

We can divide Deep Web into two areas according to the activities users perform on it:

1) Legal Activities: Including databases and virtual academic libraries, research papers libraries of periodical journals, or just browsing the web anonymously or the user prefers to be untracked. Many parties practice their activities on Deep Web to preserve their privacy, such as police, security and military forces, press and media, and others.

2) Illegal Activities: Including every action categorized as illicit or criminal, and this part is what forms the Dark Web.

3.3. Dark Web

It is where most of the illicit activities take place. Such activities are drugs trading, weapons trading, child abuse, trading sensitive information, malwares and spywares, sharing Software Exploits information that hacktivists discover in computer systems, or renting a Botnet, which is a full-equipped network connected to the internet that hackers can operate to perform a wide range security breach. In addition to trading documents, fake IDs, stolen credit cards, patients' medical records, and any other Personally Identifiable Information (PII). It also includes financial fraudulence, publicizing criminal ideologies, even employing hitmen, and a lot more. *Dark Web Hidden Service* stake place in the Dark Web as well, they are special services that host the security breach activities, and work as the hosting environment for malwares. [12] [2]

We noticed the difference in definitions of Deep Web and Dark Web among researchers. Some researches define Deep Web as the part that consists of everything the search engines cannot discover or index, while Dark Web consists of hidden networks that use specialized protocols and software [1] [15]. Other researches define Deep Web as the web that consists of all of the aforementioned, in addition to the hidden and private networks that may contain legal and benign activities, while Dark Web is that part of the hidden and private networks where illicit and criminal activities take place. [9] [12]

Figure (1) illustrates what we consider it as the layers of the internet:

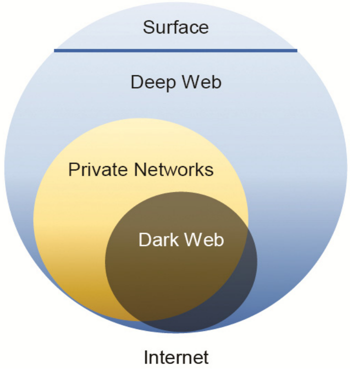

Figure 1. Layers of the Internet

4. Dark Web

In 2011, Ross William Ulbrichtý launched the most famous electronic market on the dark web with the name "Silk Road", while Ulbricht used the username "Dread Pirate Roberts". In September 2013, FBI had detected and shut down the website, and arrested Ulbricht in October of the same year. He had a sentence to life imprisonment in 2015, after he gained more than 13 million dollars from his trades and commissions on Silk Road, which was specialized in trading drugs, electronic products like malware and hacking services, hacked multimedia, in addition to fraud, passports and social cards forgery [14].

Researchers aim to form a vision about activities on darknets, the motives and interests of their performers, and the environments they activate in, the importance of this vision lies in introducing a clearer and deeper understanding of those activities through data and statistics that can benefit many specialists in different fields. For example, this information can provide security forces with evidences that help in investigations or prosecutions, and help cyber security agencies in detecting security vulnerabilities and anticipating the cyber-attack before it happens. It also provides financial institutions with important information about money laundering, currencies forging, accounts faking, stolen social and credit cards, and others.

To fulfill this vision, we see that a study must start from creating a clear understanding of the network that hosts these activities, in both its technical structure and social constituents. In the technical side, researchers determine the basic characteristics of private networks as follows: [1]

1. Decentralization: They mostly use peer-to-peer techniques where the data stays stored in a group of personal computers distributed on the network around the world instead of using a central server.

2. They take advantage of the infrastructure of the public internet.

3. They use non-standard computer protocols and ports that make it hard for the users outside the network to reach.

One of the most famous software that insures these features is Tor (The Onion Router), in addition to I2P, Freenet, Hyperboria Network, M-Web, Shadow Web and others.

From a social aspect, activities on the dark web depend on the strong community structure that the members of dark websites take notice on. As websites on the dark web, and especially electronic markets, need someone who administrates them and preserves their security and privacy to allow vendors and traders just to concentrate on their trades. The administrators are responsible of operating the websites, controlling traffics, promoting products, and often working as third parties during commercial transactions, where trust plays a basic role. [15]

They can achieve this trust by retaining consistent usernames through the different websites on dark web. In other words, the individual chooses the same username on all platforms he/she uses to execute his/her trading, thus they can conserve their reputation and build themselves consistent virtual identities, despite the fact that this can put them at risk of disclosing their identities or tracing them by the security agencies.[15]

5. Tor at a Glance

The Onion Router (Tor)[1] is the most famous software used for the purposes of hidden networking and anonymity. In its operation, Tor relies on a group of volunteering servers that transfer data through the network. Their number might reach over 3000 servers, and it takes its name- The Onion Router -because it encrypts the data with several layers of encryption before sending it through the network, and each transporting server on the randomly chosen path removes these layers one by one. [4] [5]

The U.S Navy initially developed Tor in the 1990s with the purpose of protecting U.S intelligence communications online. In 2002, they officially launched into the public to allow open source information gathering as the main objective. [1]

Despite of the main goal of developing such network, Tor network became the perfect place for many websites to deliberate illicit and malicious activities while preserving anonymity of their individuals. One of the most famous electronic markets was Silk Road (which was shut down by the FBI on October 2013, but was relaunched after only one month), in additions to Black Market Reloaded (which was shut down by its administrator on December 2013), Agora, and Pandora [4]. One can easily find some of these sites by surfing special pages on the network. Those pages work as dictionaries that contain lists of links to these sites (like Hidden Wiki), or by using specialized but slightly available search engines on Tor network (like DuckDuckGo, TorSearch and Grams), but all these resources cover a very limited number of hidden services on the network.

Tor works differently on TCP transport layer, and uses communication technique through socket connections. When a new user - called “source” - joins the network through Tor browser, Tor builds a virtual circuit with a number of randomly chosen nodes on the network. It uses this virtual circuit for approximately ten minutes, moving afterwards to creating a new circuit, and so on. [1] [16]

The circuit consists of three types of nodes:

1. Entry Node: The first node in the circuit and it receives the incoming traffic.

2. Intermediate Nodes: Pass the data from one node to the next one.

3. Exit Node: The last node in the circuit and it delivers the traffic to the open internet.

When a “source” request access to a website, Tor encrypts the request with several layers, sends it to the entry node, and transports it through a number of intermediate nodes distributed around the world and randomly chosen. With every jump to a node, the current node removes one layer of encryption from the encrypted request before passing it to the next intermediate node. When the request arrives to the exit node, the latter removes all remaining layers of encryption and sends the original unencrypted request to the web server on the public internet. With this mechanism, all information about the source is lost and the identity of the user remains ambiguous, and it is possible to return only to the last node in the circuit. [1] [16]

Tor also allows deploying websites without revealing the location of their hosting servers, and those sites use the extension “.onion” that cannot be processed and rendered outside Tor network. These characteristics combined allow anonymous browsing, and make it a safe means of communication among users.

[1] Tor Project, https://www.torproject.org

6. What is a Crawler?

Cheong defines Crawler as “software programs that traverse the World Wide Web information space by following hypertext links and retrieving web documents by standard HTTP protocol”. [17]

Crawlers have many uses in different applications and research areas, especially in search engines, which aim to gain up-to-date data, and where crawlers create a copy of all pages they visit for later processing. In other words, search engines index webpages so they can retrieve them easily and quickly when a user searches for some topic. Web administrators also use crawlers for automatically maintaining a website, like examining hyperlinks and validating HTML tags, or for collecting specific types of information like email addresses and especially harmful or spam emails. [18]

Another common use of Crawlers is Web Archiving, where crawlers collect and archive huge groups of pages periodically for the future benefits. In addition to web-monitoring services that allow users to insert queries about specific topics, and these queries form triggers for the crawler to crawl the web continuously and send alerts about new pages that match those queries to the users. [19]

The necessity of developing these crawling softwares started from two main factors: the massive size of information on the World Wide Web and the decentralization of control over this network, because the network allows any computer user to participate in sharing information globally in the open space of the internet. Therefore, this forms a big challenge to any computing or statistical process to work on this information, especially that it is stored in distributed databases. The main challenge here, which is Scalability, can be worked on by creating central repository developed specially to store webpages for wide range calculations, it starts from creating a database structure of URLs, then fetching the content from the chosen links, and updating the repository with new links, and so on. Researchers call this process “Crawling” or “Spidering”. [20].

Figure (2) illustrates how a crawler works:

In the last two decades, crawling software development noticed a great interest in dark web, but with the technical particularity of that part (which we have previously discussed), developing such software needs extra techniques integrated with it, so crawlers would be able to find malicious websites, accessing them, and fetching their pages for later analysis.

When designing a crawler, we must be aware enough of the characteristics of the crawled network. For the crawler to be able to access Tor network anonymously, proxy software (like Privoxy[1]) should be used to provide a proxy connection on HTTP protocol without saving any data cache about the currently occurring connection, and this proxy connects the crawler with Tor network.

7. Challenges

Crawler mission can be theoretically simple: starting from seed URLs, downloading all pages under the chosen addresses, extracting hyperlinks included in the pages and adding them to the list of addresses, and iteratively crawling on the extracted links, and so on.

Though it is not as easy as it looks, as web crawling faces several challenges in general, most importantly: [19] [1]

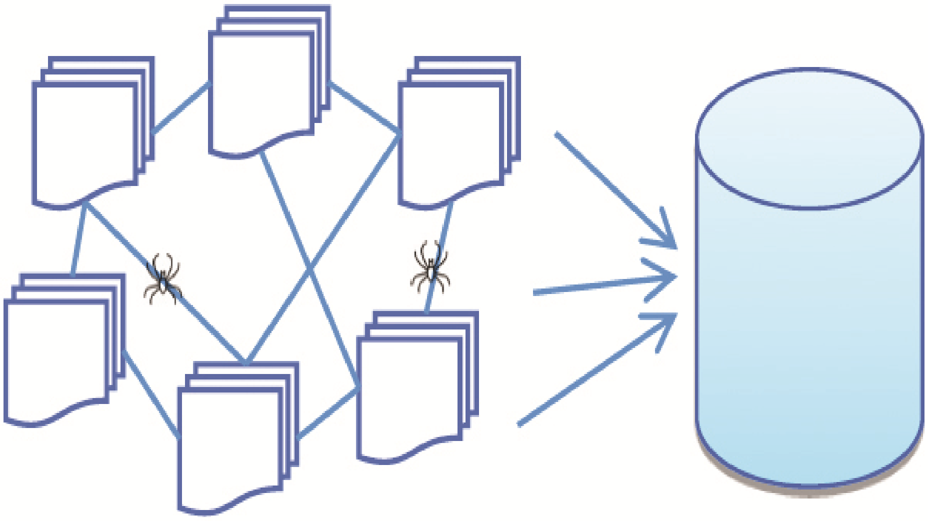

Figure 2. Illustration of how a crawler crawls linked pages and stores extracted data in a database

[1] Privoxy, https://www.privoxy.org

1. Scalability: The crawler must achieve a very high productivity against the massive size of the web and its fast pacing non-stopping evolvement, a distributed crawling can solve this issue, i.e. operating the crawler on several devices, and dividing the URL space where each device is responsible of a partial group of the URLs.

2. Social Obligations: One of the most important features the crawler should have is to not be an over-loader on the website it crawls, as it might unintentionally lead to a Denial of Service (DoS) attack, thus the crawler should be set where it can establish a very few connection requests to each site (e.g. two requests).

We can add to those general challenges some special ones facing the crawler when it's operated to specifically crawl the dark web, these challenges result from Tor network features, especially the uncorrelated websites, i.e. the links among sites are very scarce and therefore making them hard for the crawler to follow.

Of the most important challenges we mention: [21] [15]

1. The short lifecycle of websites hosted on a private encrypted network, compared to those on the surface web, as they immigrate frequently through several addresses, making their reliability and operability time untrusted. In additions, web administrators rely on shifting the websites among many web addresses, especially dark web electronic markets, to prevent monitoring. It is worth mentioning that the platforms working on encrypted networks suffer from technical difficulties like bandwidth limitation, therefore the availability of such websites is much less reliable than those hosted on the surface web, and the tunnel-like transportation through several nodes makes loading the websites hosted on Tor take longer time than those with direct connections.

2. Accessibility: Most of these sites require user registration and approvement on their community rules to access them. The registration and login processes often include completing CAPTCHA, graphical puzzles or quizzes to prevent automated logins or Denial of Service (DoS) attacks, which all requires manual handling.

3. Web administrators take notice of professionalism and the effectiveness of the electronic community they operate. This might include creating a social layering system that works according to the activeness of their members, their skills, and their professional level. They also employ a procedure that terminates accounts of inactive members to prevent attempts of hidden surfing, which they consider a suspicious behavior.

8. Suggested System Components

We can summarize the basic components of any dark web crawling system as follows:

1. Crawling Space: The crawler starts from a list of websites of illicit activities. We can depend on a number of sources like security resources - if available - (i.e. resources published by governments, or official non-governmental organizations) [9], or electronic resources like indexes of Tor network. Links to dark websites can be attainable even on the surface web or by using generic search engines (like Google). The crawling space can expand by adding new links that the crawler finds on the retrieved webpages, which leads to other pages and so on.

2. Website Preprocessing: This includes processing access obstacles in case it needs membership registration, login validation, fetching and saving session cookies.

3. Storage and Analysis: Storing and analyzing the retrieved webpages.

Crawler needs to access a permanent storage space to save extracted webpages before analyzing and processing them, and that can be achieved by two methods, according to the capabilities of the used equipment: connecting a database, or using a simple file storage system where pages are saved like separated files. [22]

Crawler is set by means of particular parameters - according to the servers and networks capabilities, and the blocking mechanisms used on some sites - to insure ease of access to the targeted sites, in addition to human-assisted approach [6], and employing dynamic proxies. Such parameters are number of crawls for each site, number of connections to the site, allowed period for downloading a content, speed and timeout of a connection, and others.

9. Our Experimental System Used for Crawling and Data Extraction

We have developed a system, *Darky*, using Scrapy1 (a programming library written in Python), as we provided it with a connection to dark websites on Tor network through Tor software integrated with Privoxy (a software for Virtual Private Network (VPN)) to insure the most possible security and anonymity of the crawler against those sites, by relocating the IP address.

After establishing the Tor-Privoxy connection, we operate the crawler starting from the website URL, and it processes the login interface with credentials that we have created earlier on the website for the purpose.

We designed the crawler especially for the website under study, according to its structure and the hyperlinks structure among its pages, i.e. we must customize a different crawler design for each website for different interfaces handling methods and different HTML structures.

Figure (3) Illustrates system architecture:

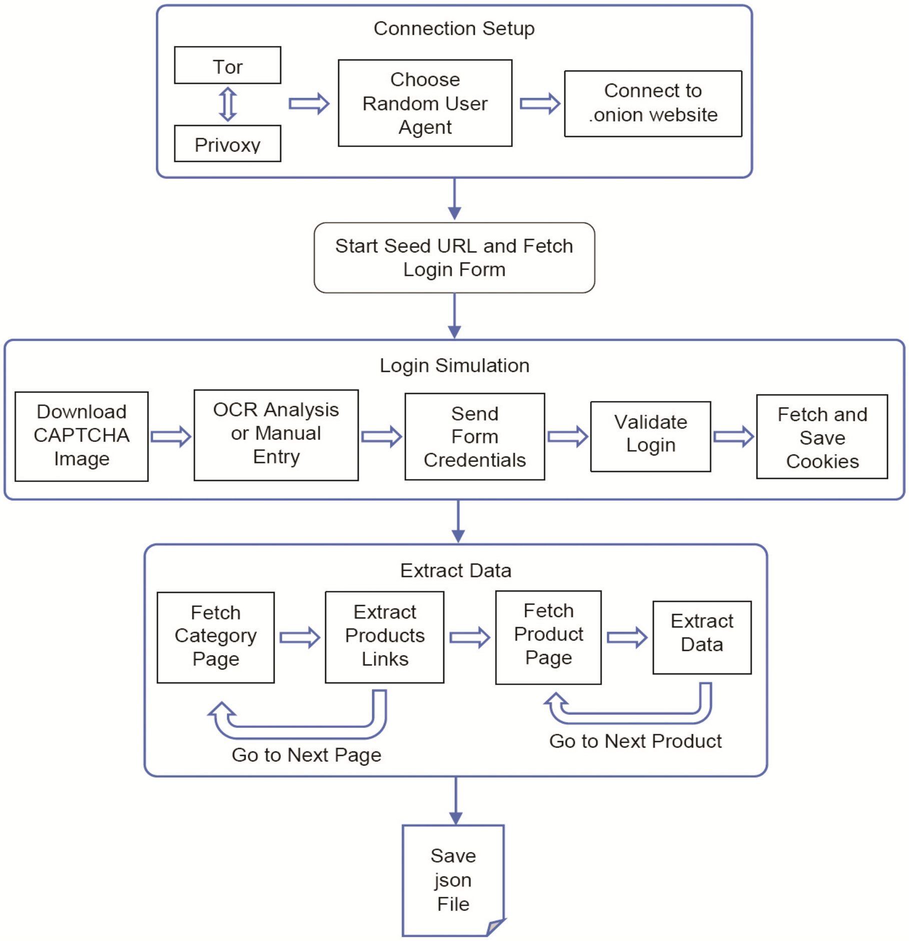

Figure 3. Proposed System Architecture

9.1. Using DOM, CSS and XPath for Data Extraction

To fill data tables, we take advantage of the HTML tags structure of each page, especially the part represented by the <body> tag, which includes the requested content. This content also consists of many other tags with different types and levels of the HTML structure. Therefore, we define the Document Object Model (DOM) nodes using CSS and XPath (response.css and response.xpath) whichever suitable for the job.

The goal behind this process is to reduce the extracted elements into rows of corresponding data, in other words transforming the unstructured data into structured data initiating it for analysis, it also preserve disk space by pulling out only the required data from the fetched pages instead of downloading the whole pages.

9.2. Steps to walk through

We have experimented our system on a dark web market (which we consider not mentioning for security reasons), and the developed algorithm for that market is as follows: (Figure (4) illustrates the structure of the market under study)

1. The crawler starts from the market URL where a login interface shows first.

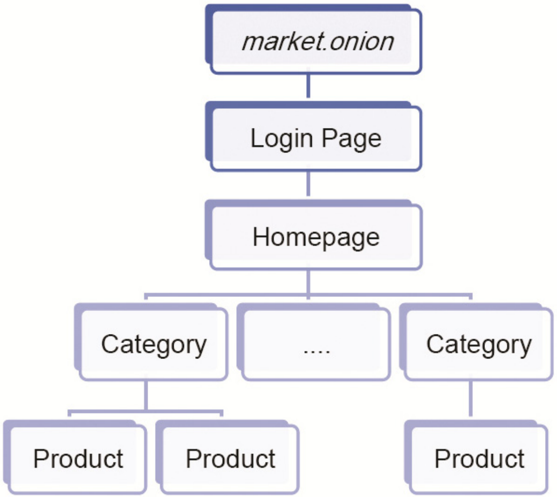

Figure 4. The dark market structure

2. Extracts the path of the CAPTCHA image using XPath.

3. Opens the path of the image and saves it to local disk.

4. Processing the downloaded image with *pytesseract* library for OCR analyzing.

[1] Scrapy, https://scrapy.org

5. In case OCR couldn't be done properly, it gives the option for manual entry.

6. Sends the credentials of the login form consisting of the triple: username - password - CAPTCHA value.

7. Calls the procedure of login validation.

8. In case login failed, crawler stops, otherwise, it extracts Cookies value for the opened session and sends it in the header of every request.

9. Redirecting the request to the main page and calling the function of processing the categories of products listings.

10. For each category, the crawler fetches the category page and extracts hyperlinks to products pages from the listing presented in the page, then calling the function of product processing for each extracted link. After finishing the current page, the crawler moves forward to the next page in the category by following its link in the pagination section at the bottom of the category page, also using XPath, and so on until the crawler follows all pages in the category.

11. In the product processing function, the crawler fetches the product page and extracts the product data using CSS, the available information about the product includes:

- Product title
- Vendor name
- Origin
- Shipment destination
- Price (in USD)
- Number of sales (since a specific date)
- Number of views (since the same date)
- Date of when mentioned sales and views started
- Moreover, we added the category name to the fields.

In another version of the crawler we also extracted the vendors' data, following the same algorithm but calling the function of vendor processing instead of product, the available information about the vendor includes:

- Vendor Name
- Vendor Level
- Trust Level
- Positive Feedback (percentage)
- Date of Membership
- Some mysterious number without any prefix, presented only between two parentheses, and we couldn't find its meaning in any section in the website, but looks like a number with a valuable significance against the Vendor Level and Trust Level

The crawler stores the data in *json* files for later processing.

The feature of dividing the market into categories helped us to split the crawler mission into parts, as the crawler was operated separately for each category, to reduce the crawling time, which if it was too long, it may lead the session to end before the crawler finishes its work, or the crawler to get detected.

9.3. Customized Settings

We also customized the general settings of the crawler. We:

1. Direct the IP address to the local host, which we have already redirected to Tor network, i.e. using HTTP protocol through Tor.

2. Hide the crawler name, or Bot Name.

3. Inform the crawler to disobey the non-tracking feature specified in *robots.txt* if such file exits in the website server.

4. Determine the *Download Delay* between a request and the next one by five seconds, with the ability to choose a random delay with the *Randomize Download Delay* setting.

5. Activate the feature of saving cookies.

6. Use *RotateUserAgentMiddleware* library, which helps to choose a random User Agent from a previously defined list of agents, to simulate the operation of a browser, this feature helps in reducing the probability of detecting the crawler.

Vendors' Total = 179		
Products Total = 6387		
	Category	**Number of Products**
1	Carded Items	27
2	Counterfeit Items	92
3	Digital Products	2179
4	Drugs & Chemicals	1569
5	Fraud	887
6	Guides & Tutorials	987
7	Jewels & Gold	15
8	Other Listings	243
9	Security & Hosting	50
10	Services	139
11	Software & Malware	185
12	Weapons	14

Table 1. Number of vendors of the site, and the products extracted from each category

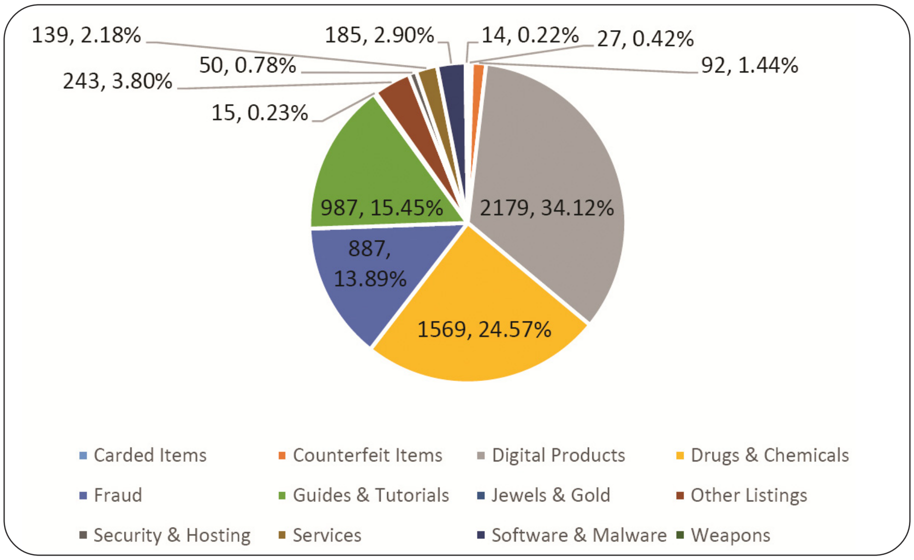

Figure 5. Percentages of products amounts in each category

9.4. Results

The table 1 shows the number of vendors and the quantities of products extracted from the market in 15 September 2018, in a total time equals to 15.63 hours, with an average of seven elements per minute, covering the whole content available on the website on that date:

Figure (5) shows a percentages pie chart representing the amounts of products in each category. As noticed, “Digital Products” takes the first place as the most traded products with 34.12%, “Drugs & Chemicals” comes secondly with 24.57%, “Guides & Tutorials” and “Fraud” follows with 15.45% and 13.89% respectively, while the other categories come with minimum percentages:

Example of the extracted products data:

Title: Bolivian Crack Cocaine 1Gr.
Vendor: *Vendor*
Origin: Europe
Ship to: Europe
Price: USD 90.40
Views: 251
Sales: 2
Since: July 06, 2018
Category: Drugs & Chemicals

Example of the extracted vendors’ data:

Vendor Name: *Vendor*
Vendor Level: 4
Trust Level: 4
Positive Feedback: 99.41%
Member since: April 11, 2018
Number: 211

10. Conclusion and Future Work

In this paper, we presented the basic concepts behind the three parts of the Web: Surface, Deep and Dark Webs, and the differences among them. We found that the deep web includes the biggest part of the web, which is unindexed by search engines, in addition to services of the hidden browsing and privacy preserving, which is not necessarily aimed for illicit activities, but mostly consists of benign and legal activities. However, in the deepest and darkest point of the deep web, a different kind of internet users exploits these services to deploy and deliberate malicious and criminal activities as well.

Thus, we discussed how it is possible to reach that point of the web, employ crawling methods on dark websites, and extract useful information to help security and law agencies in investigating activities that take place on the dark web. In our experimental environment, we designed a crawler that is able of simulating a user login to a dark market, crawling the whole website and fetching the required data from its pages. This data can then be prepared for processing with many different data analysis applications, such as data mining, which is our next goal. For the future, we aim to feed the gained results to a data mining system, thus the integrated crawling and mining systems together would form an approach for *Mining the Dark Web*.

References

[1] R. F. B. M. J. M. G. D., Ahmed, Zulkarnine, T. (2016). Surfacing Collaborated Networks in Dark Web to Find Illicit and Criminal Content, *Intelligence and Security Informatics (ISI), 2016 IEEE Conference,* p. 109-114, 28 (September).

[2] M. S. L. S. W. L. Andres Baravalle. (2016). Mining the Dark Web: Drugs and Fake Ids, in *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference*,

Barcelona, Spain, 2016.

[3] Sherman, Chris G. P. (2003). The Invisible Web: Uncovering Sources Search Engines Can't See, *Library Trends,* 52, (2) 282-298, 2003.

[4] S. V. M. v. S. Martijn Spitters. (2014). Towards a Comprehensive Insight into the Thematic Organization of the Tor Hidden Services, *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint,* p. 220-223, 24 September 2014.

[5] Gareth Owen. N. S. (2015). The Tor Dark Net, 30 September 2015. [Online]. Available: https://www.cigionline.org/publications/tor-dark-net.

[6] Sriram Raghavan, H. G.-M. (2001). Crawling the hidden web, *In*: Proceedings of the 27th VLDB Conference, Roma, Italy.

[7] Satinder Bal, R. N. (2008). Crawling the Hidden Web using Mobile Crawlers, *In:* 3rd International Conference on Advanced Computing & Communication Technologies (ICACCT2008).

[8] Mundluru, Dheerendranath., X. X. (2008). Experiences in Crawling Deep Web in the Context of Local Search, *In:* 5th Workshop on Geographic Information Retrieval.

[9] J. Q. G. L. H. C. E. R. Yilu Zhou, (2005). Building knowledge management system for researching terrorist groups on the Web, *In:* AMCIS 2005 Proceedings.

[10] T. T. C. I. T. M. S. V. J. M. U. W. I. K. George Kalpakis. (2016). Interactive discovery and retrieval of web resources containing home made explosive recipes, *InInternational Conference on Human Aspects of Information Security, Privacy, and Trust,* p. 221-233, 17 (July).

[11] Pannu, Mandeep., I. K. H. (2018). Using Dark Web Crawler to Uncover Suspicious and Malicious Websites, in *In:*International Conference on Applied Human Factors and Ergonomics.

[12] Hawkins, B. (2016). Under The Ocean of the Internet - The Deep Web, 15 May 2016. [Online]. Available: https://www.sans.org/reading-room/whitepapers/covert/ocean-internet-deep-web-37012.

[13] M. B. D. B. Onur Catakoglu. (2017). Attacks landscape in the dark side of the web, *In:* Proceedings of the Symposium on Applied Computing, ACM, 2017.

[14] Finklea, K. Dark Web, 10 March 2017. [Online]. Available: https://fas.org/sgp/crs/misc/R44101.pdf.

[15] A. D., E. M., E. N., V. P., J. S., P. S. (2017). John Robertson, Darkweb Cyber Threat Intelligence Mining, Cambridge: Cambridge University Press, 2017.

[16] Ahmad, I. (2018). A New Look at the TOR Anonymous Communication System, *Journal of Digital Information Management,* 16 (5) 223-229, (October).

[17] Cheong, F. C. (1996). Internet agents: spiders, wanderers, brokers, and bots, New Riders Publishing, 1996.

[18] Li Yang, F. L. J. M. K. R. K. E. (2009). Discovering topics from dark websites, *In:* Computational Intelligence in Cyber Security, 2009. CICS'09. IEEE Symposium, p. 175-179, 30 (3).

[19] Olston, Christopher., M. N. (2010). Web crawling, *Foundations and Trends in Information Retrieval,* 4 (3) p. 175-246, 1 (3).

[20] Ling Liu, M. T. O. z. (2018). Encyclopedia of database systems, Springer.

[21] Tianjun Fu, A. A. H. C. (2010). A focused crawler for Dark Web forums, *Journal of the Association for Information Science and Technology,* p. 1213-1231.

[22] Thelwall, A. A. H. C. (2001). A web crawler design for data mining, *Journal of Information Science,* p. 319-325, 27 (October).