PLAGIARISM REPORT OF RESEARCH PAPER

Kartikeya Srivastava

Submission date: 15-May-2024 07:44PM (UTC+0530)
Submission ID: 2264818995
File name: Dark_web_Crawler_Research_Paper.pdf (471.09K)
Word count: 6731
Character count: 37968

A Comprehensive Survey of Dark Web Crawlers

Mentor : Prof Gaurav Parashar
Computer Science and Engineering
KIET Group of Institutions
Kartikeya Srivastava, Rishi Srivastava, Really Singh

***Abstract*— Due to the widespread use of powerful encryption algorithms and advanced anonymity routing, the field of cybercrime investigation has greatly changed, posing difficult obstacles for law enforcement organizations (LEAs). Consequently, law enforcement agencies (LEAs) are increasingly relying on unencrypted web information or anonymous communication networks (ACNs) as potential sources of leads and evidence for their investigations. LEAs have access to a significant tool for gathering and storing potentially important data for investigative purposes: automat web content harvesting from servers. Although web crawling has been studied since the early ys of the internet, relatively little research has been done on web crawling on the "dark web" or ACNs like IPFS, Freenet, Tor, I2P, and others.**

This work offers a thorough systematic literature review (SLR) with the goal of investigating the characteristics and prevalence of dark web crawlers. After removing pointless tries, a refined set of 34 peer-reviewed publications about crawling and the dark web remained om an original pool of 58 articles. According to the review, most ark web crawlers are written in Python and frequently use Selenium or Scrapy as their main web scraping libraries.

The lessons learned from the SLR were applied to the creation of an advan d Tor-based web crawling model that was easily incorporated into an already-existing software toolbox designed for ACN-focused research. Following that, a series of thorough experiments were conducted to thoroughly analyze the model performance and show that it was effective at extracting web content from both conventional and dark This work provide more than just a review; it advances our knowledge of ACN-based web crawlers and provides a eliable model for digital forensics applications including the crawling and scraping of both clear and dark web domains. The study also emphasizes the important ramifications of retrieving and archiving content from the dark web, emphasizing how crucial it is for generating leads for investigations and offering crucial supporting evidence.

To sum up, this study highlights how important it is to keep researching dark web crawling techniques and how they might be used to improve cybercrime investigations. It also highlights promising directions for future study in this quickly developing sector, highlighting how crucial it is to use cutting-edge technologies to effectively fight cybercrime in a digital environment that is becoming more complicated.

I. INTRODUCTION

Due to the high level of secrecy and restricted traceability provided by sophisticated encryption and anonymity protocols, cybercrime investigations on the Internet are becoming more and more difficult, especially inside elusive dark networks. The extensive security of data traveling across these networks presents a considerable challenge to law enforcement organizations (LEAs) in their efforts to obtain evidence, requiring a substantial investment of time, labor, experience, and technology. The associate editor, Tiago Cruz, oversaw the review and approval of this paper for publication.

There are several software programs that allow access to the about six dark n works that are currently operational. Modern encryption and network traffic routing algorithms that leave little traces are among the aspects that these products have in common, notwithstanding their variances. Because there aren't many traffic traces and it's not feasible to decrypt data in these networks, LEAs have to look for proof in oth ways. Tor, the biggest and most well-known anonymous communication network (ACN), is made up of a network of computers, some f which are web servers that are a part of the so-called "dark web." Other ACNs, such as I2P, Freenet, IPFS, and Lokinet, also have servers that make up black webs unique to their networks. For individuals residing in non-democratic nations, journalists wanting complete anonymity, and whistleblowers, Tor has become an indispensable resource. Tor's anonymity, however, is indiscriminate, helping both criminals and whistleblowers, creating a challenging environment for digital policing. Using the most recent encryption methods, Tor's network encrypts data in many layers between servers, using a different key for each tier. With more than 8,000 servers, or relays, around the world, Tor distributes transmission pathways among numerous relays to create encrypted layers that resemble an onion and protect privacy. Considering the shortcomings of network traffic analysis and decryption inside ACNs, online content collecting shows itself to be a useful workaround. One effective method of extracting unencrypted data from Tor websites without requiring a lot of manual labor is web crawling, often know as automated online content collecting. Web crawling is widely utilized on the open web for commercial and archive purposes, but it has also been important in criminal investigations, with previous screenshots of illegal websites frequently serving as proof. The fleeting nature of Tor servers emphasizes how crucial it is to consistently record and store online information in order to preserve data integrity and ensure legal admissibility. Even though there are many different web crawler programs available, ACNs have not received as much attention in the study as the clear web.

Because there is a dearth of research on dark web crawlers, it is critical to investigate this area in order to comprehend the particulars of dark web crawling in ACNs. This knowledge can help with the creation of useful tools for practitioners

and researchers who crawl webpages on ACNs.

This work not only adds to the body of information but also describes the development and assessment of a Tor-based crawler, utilizing the knowledge gained from the literature review to improve an already-existing dark web research toolbox.

Within the scholarly community, the investigation of dark web crawlers is still a largely unknown field. Through further exploration of this topic and defining the distinct features of dark web crawlers—particularly in light of the peculiarities of anonymous communication networks (ACNs) in contrast to the surface web—a thorough synopsis of the state of dark web crawler technology can be formulated. The purpose of this analysis is to improve knowledge about the design and workings of dark web crawlers. For academics and professionals working on the creation and implementation of efficient instruments for locating and examining content on the dark web, these insights may prove to be quite beneficial.

In addition, this work presents and evaluates a Tor-based crawler, demonstrating its usefulness. The architecture of this crawler incorporates knowledge from the previous literature review.

II. LITERATURE SURVEY

A. DIPOSTION

The work which consists of two main research contributions—a systematic literature analysis and an experiment-based web crawler implementation—is organized into seven thorough chapters and a bibliography.

The foundational overview is given in the first chapter, which also explores the nuances of online crawling, website content, and anonymous communication networks, highlighting their importance in digital investigations. Chapter two expands on the scientific foundations that give rise to the research challenge and the development of research questions, building on the introduction.

B. THE WORLD WIDE WEB

Before HTTP became the standard protocol for transferring information, other internet protocols, such as Gopher, fought for supremacy in the early 1990s when the World Wide Web was beginning to take shape. The World Wide Web was widely adopted by 1994–1995 thanks to the graphical features and open nature of HTML, which finally overtook Gopher, a previous text-based protocol centered on network file exchange. Using the Internet Protocol (IP) and the Transport Control Protocol (TCP), HTTP evolved into the industry standard for data transmission, including photos, videos, and HTML pages. The fundamental process of fetching an HTML page via HTTP has remained unchanged since its inception, with a client (typically a web browser) sending a GET request to a web server and receiving a corresponding response. A successful retrieval returns an HTTP code of 200, while a failed attempt yields a 404 code, in line with the HTTP standard that includes various other response codes. Despite revisions from HTTP versions 0.9 to 1.1, 2.0, and the latest HTTP 3, backward compatibility ensures uniformity across all requests and responses, maintaining consistency with HTTP version 1.0. Essentially, web crawling automates the procedure of sending GET requests to websites, tracking embedded URLs, and saving the results that are obtained. Web crawlers can be standalone software entities or browser-based programs designed to communicate over HTTP. These days, a wide range of HTTP communication libraries for various programming languages—such as Lisp, Go, and Haskell—make it easier to create HTTP-based clients and servers, which speeds up the process of developing web crawling programs.

C. An Adaptive Crawler

An Adaptive Crawler for Locating Hidden Web Entry Points Luciano Barbosa and Juliana Freire

Adaptive crawling strategies have been demonstrated to be exceptionally efficient at locating the entry locations of concealed web sources. These techniques focus the content of the retrieved pages by giving priority to links that are most relevant to the subject. This method maximises the application of learned information, allowing for the identification of connections displaying hitherto unidentified patterns. As a result, the approach shows resilience and the capacity to correct for biases introduced throughout the learning process.

Mangesh Manke, Kamlesh Kumar Singh, Vinay Tak, and Amit Kharade's research article presents an advanced integrated crawling system designed specifically for exploring the deep web. The researchers of this extensive investigation introduce a novel adaptive crawler that utilizes offline and online learning mechanisms to train link classifiers. As a result, the crawler effectively gathers concealed web entries.

D. WEBSITE ACQUISTION TOOLS

Both open-source and closed-source technologies are available for forensic website acquisition that are designed to archive and maintain web material according to forensic science guidelines. Although some of these tools can be used for basic web crawling, and some can even be used in dark web contexts, their main purpose is not as powerful web crawlers.

One notable feature of OSIRT, an intuitive web browser designed specifically for investigators, is its ability to support both ordinary and dark websites (such as Tor). OSIRT is a widely used tool by law enforcement agencies in the United Kingdom. It helps to preserve the integrity of evidence in investigative processes by facilitating functions such as video recording, screenshot generation, and audit log generation.

Police departments throughout the world use FAW, a proprietary internet forensic collection tool that looks like a web browser, to collect web content from popular websites, social media platforms, and the dark web (Tor). Notably, FAW's collection of forensic tools includes the ability to crawl websites.

Another proprietary tool, called Hunchly, is available as an add-on for web browsers and is intended to suit the demanding needs of law enforcement personnel conducting

investigations. It supports the acquisition of content from both the clear and dark web (Tor).

E. WEB CRAWLERS

A number of privacy risks, such as the possibility of request leaks and other serious privacy breaches, might result from the mishandling or incorrect configuration of the Tor network. Users are exposed to grave privacy threats when the Tor network is not configured appropriately, creating opportunities for possible leaks.

Request leaks pose a serious risk since they can unintentionally reveal private information about a user's online activities outside of the Tor network. Misconfigured settings or bugs in the Tor client or connected apps may be the cause of this leakage. The main goal of Tor is to anonymize users and shield their identities and activities from monitoring or interception. Request leaks weaken this goal. A number of privacy risks, such as the possibility of request leaks and other serious privacy breaches, might result from the mishandling or incorrect configuration of the Tor network. Users are exposed to grave privacy threats when the Tor network is not configured appropriately, creating opportunities for possible leaks.

Request leaks pose a serious risk since they can unintentionally reveal private information about a user's online activities outside of the Tor network. Misconfigured settings or bugs in the Tor client or connected apps may be the cause of this leakage. The main goal of Tor is to anonymize users and shield their identities and activities from monitoring or interception. Request leaks weaken this goal.

Inadequate Tor network settings can also put users at danger of other privacy issues, like: IP Address Exposure: Users' anonymity may be jeopardized if Tor is not configured correctly and their actual IP address is made public. This exposure may be the result of Tor browser leaks or incorrectly setup proxies. Insufficient setup can expose users to traffic analysis, which is the process by which attackers track and examine encrypted data flows over the network. The identification of users or their online activities may result from this analysis.

DNS Leaks: When DNS requests are made outside of the Tor network, they betray the websites that are being browsed. This might happen as a result of improper configuration. Inadequate usage of Tor-specific DNS resolution techniques or incorrectly configured settings can result in DNS leaks.

Risks Associated with Exit Nodes: Incorrect setups may affect the security and choice of Tor exit nodes. Users' traffic may be intercepted or monitored by malevolent actors if insecure or hacked exit nodes are used. Following recommended methods when configuring Tor is crucial to reducing these dangers and guaranteeing strong privacy protection within the network. These best practices include:

using the most recent version of the official Tor software. setting up programs to effectively use Tor proxies.

enabling the Tor browser's recommended privacy settings. avoiding third-party plugins and custom changes that can jeopardize anonymity. analyzing and adjusting configuration parameters on a regular basis in response to security alerts and Tor network updates.

Users can reduce the risk of request leaks and other privacy vulnerabilities by adhering to these instructions and keeping a close eye on Tor network parameters, protecting the integrity and efficacy of Tor's anonymization capabilities.

F. Google's Deep Web Crawler

Jayant Madhavan, David Ko, Ju-wei Chiu, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy started working together on a project that would transform how deep-web material is found and used. They encountered and overcame the numerous obstacles that come with uncovering and using the deep web's enormous reservoirs, and as a result, they created a clever solution. This technology is a ground-breaking advancement made possible by an incredibly sophisticated and flexible algorithm.

Fundamentally, this algorithm is the driving force behind the effective navigation of the complex network of possible input combinations. It navigates the complex terrain of the deep web with systematic accuracy and forethought. The system finds and isolates those difficult combinations that are the key to unlocking valuable URLs through methodical analysis and deliberate selection. These URLs have been thoroughly examined and are ready to be included in our web search index.

The expedition made by Madhavan, Ko, Chiu, Ganapathy, Rasmussen, and Halevy is evidence of the inventiveness and spirit of cooperation of people. Through their combined endeavors, they have not only surmounted technical obstacles but also shed light on novel avenues for investigating and capitalizing on the concealed riches of the deep web. Their ground-breaking system is expected to have a significant and wide-ranging impact on the digital world as it develops and matures.

G. THE TOR WEBSITE AND NETWORK

Anonymous communication networks (ACNs) or dark networks use the same transport protocols (TCP/IP) as the clear web, but they use different anonymous protocols. The clear web uses TCP/IP to send HTTP requests. For example, Tor uses its onion routing (OR) protocol, and I2P uses garlic routing (GR), both of which are TCP/IP wrapped. Within ACNs, these protocols make it easier for HTTP to be transmitted and used for web page serving.

The idea of Onion Services, formerly known as Hidden Services, is essential to the Tor network. These services host websites that are only accessible through known URLs and are a component of the "dark web," like the following example: In contrast to the ordinary internet, it is not possible to search through a variety of IP addresses to find Onion Services on the Tor network. Furthermore, while theoretically feasible, trying to guess Onion Service URLs pseudorandomly is not a viable strategy. Tor websites, also called "onionsites," are identical to regular web pages in appearance and structure. They are made up of text, graphics, HTML, CSS, JavaScript, and other elements that are delivered over

HTTP. Onionsites' material however, usually captures the special qualities of existing on an anonymous communication network. Onionsites put anonymity, privacy, and secrecy above usefulness and speed. They also frequently refrain from using JavaScript because of the possibility that doing so could reveal a Tor user's true identity. Onionsites stand out from their obvious online competitors thanks to their attention on security and anonymity, which influences their design and content decisions within the Tor network.

III. RESEARCH MOTIVATION

Although there is a wealth of literature on clear web crawlers, there isn't a single thorough review that concentrates on dark web crawlers. Unlike the ordinary Internet and the conventional clear web, anonymous communication networks function under different protocols and setups, requiring programming and configuration specific to their own features.

As was noted in the sections before this one, both public and commercial organizations have created a variety of dark web crawlers, but a thorough scientific analysis of these instruments is noticeably absent. Doing a thorough assessment of the dark web crawlers that are now in use and have been reported in scholarly publications is one of the main goals of this research project. This evaluation aims to improve our knowledge of the changing field of dark web crawling technologies by offering insightful information about the strengths and weaknesses of different crawlers. Implementing the most popular dark web crawler found by academic research and customizing it to fit neatly into an already-existing toolset that is currently devoid of a reliable and all-inclusive crawler solution is another important goal. The performance evaluation that follows will provide important information about this integrated crawler's effectiveness and suitability for use in investigative and analytical settings.

By tackling these goals, the study intends to close important information gaps about dark web crawling techniques and make a significant contribution to the creation and improvement of instruments for examining and navigating the complex world of anonymous communication networks. This thorough method emphasizes how important it is to have reliable and specialized crawling strategies that are suited to the particular opportunities and problems that the dark web ecosystem presents.

IV. CRAWLING PATHS LEARNING

Investigating the Deep Web necessitates a multimodal strategy that goes beyond the traditional techniques of surface-level web crawling. Deep Web crawlers are frequently tasked with traversing through layers of content to uncover subsets of information pertinent to particular users or processes, as opposed to the straightforward tasks of traditional crawlers, which consist of completing out forms and retrieving result pages.

Central to deep web crawling methodologies are crawling paths, which comprise sequences of pages that are crucial for accessing the intended content. These pathways cover not just page navigation but all of the complex interactions that are needed at every stage, like form submissions, user event simulations, and link traversals. Although certain approaches integrate form interactions directly into crawling paths, others trigger the procedure from result pages acquired subsequent to form submissions. Because Deep Web information is so diverse, different pages have different levels of relevance, which has led to the creation of different crawling strategies.

The most basic method is represented by blind crawlers, which gather as many pages as they can from a website. They commence their expedition from a seed page and methodically adhere to each link it furnishes until each page that is accessible has been downloaded. Thus, all of the URLs that are reachable within the website's domain are included in the crawling pathways that they take. Conversely, targeted crawlers take a more discriminating approach, focusing on links that are likely to direct users to pages with relevant content related to a given topic. Crawlers utilise advanced classification methods to evaluate the pertinence of downloaded pages and proceed to follow links that are considered pertinent.

Conversely, ad-hoc crawlers ignore topical alignment in favour of the unique requirements and preferences of each user. They adjust the crawling experience to each user's specific preferences and needs by carefully selecting links that connect to pages that are judged relevant.

Concentrated and ad-hoc crawlers require more complex path-generating techniques than blind crawlers, which usually use simple algorithms to queue URLs for fast traversal. Crawlers may be classified as recorders, supervised learners, or unsupervised learners, according to the employed methodology and the necessary level of oversight. These classifications reflect the unique strategies utilized for path generation and content discovery by the crawlers.

Essentially, a deep comprehension of crawling patterns and the nuances of content retrieval is necessary to navigate the complexity of the Deep Web. Crawlers have the ability to discover concealed treasures of information that are beyond the reach of conventional web crawlers by employing sophisticated crawling techniques that are customized to achieve particular goals.

V. SYSTEMATIC LITERATURE REVIEW

Planning the Systematic Literature Review (SLR) is the first phase, or (1) planning. This includes developing strategies for data capture and dissemination, defining the research topic, setting criteria for study selection and quality assessment, and summarising the research background.

More in-depth work is done in the second phase, which is (2) doing the literature review. It consists of four main tasks: (1) choosing studies, (2) evaluating research quality, (3) extracting data, and (4) synthesising data. The document's later sections go into great depth about each of these tasks.

The definition of the dissemination mechanisms, report formatting, and report evaluation are all part of the

third phase, (3) reporting. Given the nature of the document—which is intended to be peer-reviewed—this phase is essential. This guarantees that the research will be subjected to a thorough assessment and made available for public use.

A. RESEARCH QUESTIONS

After following the instructions for each activity, the following tangible results were obtained: research questions unique to the SLR (i.e., this section of the article), which differ from the research questions for the full article:

1) What types of crawlers and/or scrapers have been utilised to gather data from the Tor network in scientific publications?

2) How are traffic routes made by crawlers and/or scrapers that gather information from the Tor network?

3) Which frameworks and programming languages are most frequently used to create crawlers and/or scrapers on the Tor network?

B. STUDY SELECTION STRATEGY

The search parameters TITLE-ABS-KEY ((dark AND web AND crawler) OR (dark AND web AND scraper) OR (tor AND crawler) OR (tor AND scraper)) AND LANGUAGE (english) yielded 59 items in total that were retrieved from the database. In this case, "TITLE-ABS-KEY" refers to searches that concentrate on the metadata elements included in the articles' titles, abstracts, and keywords. The prefix "LANG" signifies that only English-language items were found; results for searches in other languages were routinely filtered out.

Following identification, the articles were downloaded and locally saved with all of their metadata (authors, DOI, title, abstract, and keywords) intact, as shown in the example below. Keeping these items locally instead of depending on web services made data processing simpler. Publication the source code of the script used to find and choose these articles also contributes to the transparency of the study approach. This procedure guarantees a better comprehension of the search and selection procedure in addition to helping to replicate the study.

C. INCLUSION AND EXCLUSION CRITERIA

It is imperative to define precise criteria for the inclusion and exclusion of studies in order to discover pertinent research for the original questions that have been addressed, as suggested by Kitchenham [45]. Only English-language articles that were relevant to the predetermined search criteria were included in the initial database search phase. Further inclusion and exclusion criteria for the papers in this systematic literature review, which mainly focuses on content crawling and scraping on the Tor network, are described in this section.

It was agreed, therefore, that articles that did not specifically address the Tor network would be added if they alluded to or explored possible effects on the network. This strategy was used to reduce the possibility of leaving out research that were either slightly or somewhat relevant.

Qualifications for Inclusion:

articles that concentrate on information collection, monitoring, crawling, and scraping on the Tor network. studies in which data was gathered from the Tor network using a crawler or scraper. Criteria for Exclusion:

articles that discussed crawling and scraping without mentioning the Tor network. research that don't involve downloading content from distant Tor servers. publications that have not undergone peer review, such as articles published outside of journals, conference proceedings, or workshop proceedings.

A manual evaluation was possible because the search query produced a tolerable amount of articles. Initially, depending on the aforementioned criteria, the abstracts of all 59 papers were reviewed and either included or excluded. After comparing the titles, it was discovered that papers [20] and [18] had the same title; the former being a fourteen-page journal article, while the latter was a shorter conference proceeding of nine pages. Since the conference piece was thought to be a truncated version of the fuller journal publication, it was disregarded.

Similar to this, the journal article was chosen over the conference version for articles [39] and [40], which were conference and journal papers, respectively, because of its greater length and level of detail. This was for journal articles also applies to the following pair: the conference article [72] was excluded in favour of the more thorough journal publication because it had the same title and DOI as the journal article [72].

There were only 56 papers left in the systematic literature review for additional analysis after these duplicate entries between conference and journal articles were eliminated. The rigorous selection procedure makes sure that more in-depth, peer-reviewed sources are prioritised, which improves the calibre and dependability of the review's conclusions.

D. DATA EXTRACTION STRATEGY

Forty-one documents were found to be appropriate for additional analysis following the quality evaluation. A META data link pointing to the complete text was supplied with every article that was downloaded from the Scopus database. Each of these papers was downloaded separately in order to carefully extract the data.

Table 2 provides a complete list of the selected relevant articles. The matching ACN-based web crawler or scraper for each article is displayed in this table along with the research instrument that was used. Furthermore, a link to the crawler/scraper's open-source code is supplied where it is accessible, which improves transparency and permits the study to be replicated.

However, seven publications were deemed irrelevant during the data extraction stage and were thus removed from the review. Among the exclusions were articles about crawling that weren't expressly on the dark web, like those that were described.

VI. CHALLENGES

Web scanning, although seemingly uncomplicated in concept, presents an array of intricacies and difficulties that transcend its fundamental principle. The overall task appears straightforward: it begins with pre-specified seed URLs, then iteratively crawls through the network of links, extracting embedded hyperlinks, downloading every page under the specified addresses, and so on. However, there are many challenges in the way of this operational execution. In addition to being intrinsic to the vastness and dynamism of the internet as a whole, the dark web environment possesses particular qualities and subtleties that further compound these difficulties.

A. SCALABILITY

One of the most significant problems with web crawling efforts is scalability. Attaining optimal productivity in crawling operations is significantly challenging due to the web's immense scale and nonstop evolution. The web is growing at an exponential rate, and new ways of crawling are needed to keep up with this growth. One potential solution to this problem is distributed crawling, which involves running the crawler over several devices. Distributed crawling expands coverage of the web landscape by increasing efficiency and scalability through the partitioning of the URL space and allocation of specific subsets of URLs to individual devices.

B. FUTURE TRENDS AND IMPLICATIONS

Technological Innovations and Evolving Threat Landscape Technological innovations and the evolving threat landscape pose challenges and opportunities for the future of the dark web. Advances in encryption, blockchain technology, and decentralized networking may enhance user privacy and security, while also enabling new forms of criminal activity and regulatory evasion.

Regulatory Trends and Policy Directions Regulatory trends and policy directions shape the legal and regulatory environment surrounding the dark web. Initiatives such as the EU Cybersecurity Act and the US Cybersecurity Enhancement Act aim to enhance cybersecurity and combat online crime through legislative measures and regulatory frameworks.

Social and Cultural Shifts in Dark Web Usage Social and cultural shifts in dark web usage reflect broader trends in technology adoption and online behaviour. Changes in user demographics, platform preferences, and content consumption patterns may influence the future trajectory of the dark web and its societal impact.

Web crawlers have unique obstacles in addition to these general difficulties while attempting to navigate the dark web, an obscure and secretive area of the internet distinguished by anonymity, encryption, and covert activity. In this underground world, the peculiar characteristics and workings of the Tor network—the main entry point to the dark web—make crawlers' struggles much more difficult. Notable difficulties unique to crawling the dark web include:

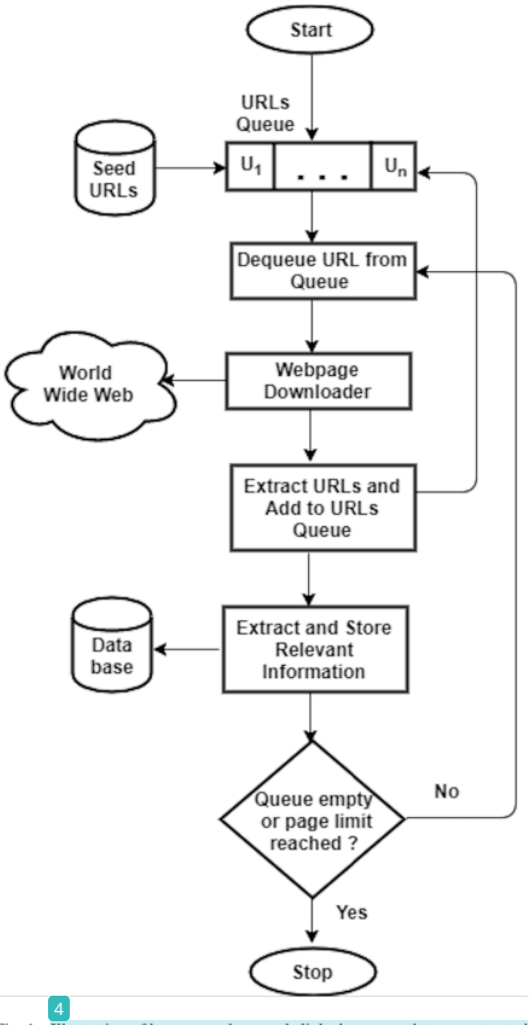

Fig. 1. Illustration of how a crawler crawls linked pages and stores extracted data in a database

C. WEBSITE LIFE CYCLE

Websites hosted on private encrypted networks, like the dark web, have a substantially shorter lifespan than their counterparts on the public internet. These websites are temporary because they move around a lot using different IP addresses; this is a strategy used by web administrators to avoid being tracked down and discovered. Particularly on the dark web, electronic markets are infamous for being transient, with site owners frequently changing their IP addresses to avoid detection by law enforcement and preserve operational security. The technical limitation of bandwidth further exacerbates the problem by jeopardizing the availability and dependability of dark websites. Furthermore, the intricate process of traffic routing across numerous nodes

inside the Tor network lengthens the loading duration of black web pages, making them more difficult for crawlers to access.

The accessibility of dark web sites is hindered by an extensive array of obstacles, including the need to perform complex security measures to prevent automated crawling and rigorous authentication requirements. Before allowing access, many dark web sites demand that users register and follow community guidelines. This means that completing registration procedures and getting past security measures sometimes requires direct intervention. Moreover, to discourage automated login attempts and reduce the vulnerability to DoS attacks, crawling is further complicated by the implementation of authentication mechanisms such as CAPTCHA, graphical riddles, and quizzes.

Community Dynamics and Management: Communities on the dark web function within a unique socio-technical environment that is defined by social dynamics, hierarchies, and strict regulations. Webmasters have a great deal of control over how these communities are run and governed; they put policies in place to keep their online forums professional and efficient. Potential measures to address this issue include the adoption of social stratification systems that consider the professional backgrounds, skill sets, and activity levels of members. Additionally, protocols may be established to identify and exclude inactive members to discourage suspicious conduct. Effectively navigating dark web communities as a crawler operator requires a nuanced comprehension of community management strategies and social dynamics due to their dynamic nature.

VII. APPLICATIONS

A. Use Cases

Crawling the dark web has uses in academic research to study behavioral trends, cybersecurity for threat intelligence, and assist law enforcement in countering unlawful activity. It helps reveal patterns and insights that are concealed behind the shadowy world of the Dark Web.

B. Applications of web crawler

In a variety of businesses, web crawlers are essential tools for competition analysis and market research. These crawlers gather a multitude of useful information from competitor websites, including pricing tactics, product details, customer reviews, and marketing efforts. They accomplish this by methodically browsing the websites of their rivals. Businesses can learn a great deal about the product offers, consumer sentiment, and market positioning of their rivals thanks to this data. Equipped with this data, businesses may decide on product development, pricing policies, and marketing campaigns with knowledge. Web crawlers also help firms remain flexible and responsive in ever-changing market circumstances by enabling real-time rival activity monitoring. All things considered, web crawlers are essential to helping companies obtain a competitive advantage, spot new industry trends, and seize expansion prospects.

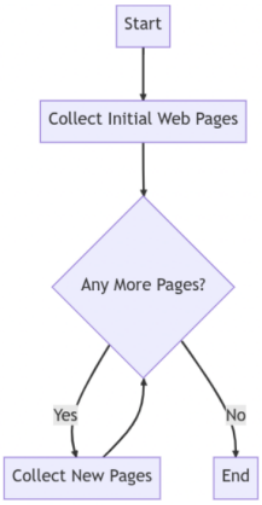

Fig. 2. Illustration of how a crawler crawls linked pages and stores extracted data in a database

VIII. RESULTS

The results from the first segment of this research study, the systematic literature review, were presented in the previous section. In this section, the results from the implementation and evaluation of the developed clear web and dark web crawler are presented. The crawler was implemented for scraping both clear web and dark websites, and the data collected from each web type was compared using a couple of different techniques and measures. First, the semi-manual inspection of the website pairs was done using GNU Diffutils. Diffutils identified discrepancies between the scraped web content files. The crawler fetched the same number of pages from both the clear web- and the dark websites for Debian, QubeS, and CIA. In the case of the CIA's website, however, the index.html was downloaded twice from the clear web crawler. This was due to a programmatic error related to internal URLs in the clear web in the crawler where both the index referrers "/" and "http://cia.gov/" were downloaded. On the Guardian's website, 12 files were not retrieved from its onion site. The random wait was set to 0-4 seconds to avoid blocking and therefore the complete scraping of the 201 web pages took circa 26 minutes. The scraping of the clear web version of The Guardian took circa six minutes with the same random delay of 0-4 seconds between each HTTP request. The files that were missing from the scraping of The Guardian's Onion were URLs that were not available over their Onion site B. DARK MARKETPLACE CRAWLING RESULTS The SIDE CT2S dark web crawler was used to crawl a dark marketplace to demonstrate and validate that fits its purpose as a digital investigation tool.

IX. CONCLUSIONS

The difficulties in gathering evidence for cybercrime investigations—especially those utilising the dark web—are addressed in this study paper. It draws attention to the need for specialised tools to support dark web investigations and presents a prototype built to satisfy particular specifications that are critical for software used for dark web investigations. This development was based on a rigorous study of the literature, which examined 58 papers on dark web crawling.

The principal aim of this investigation was to construct a thorough comprehension of dark web crawlers in the academic domain. A dark web crawler was created as an adjunct to the D3 cybercrime toolkit based on this insight.

With the use of databases of previously annotated online pages, the recently created crawler—which is coupled with an annotation-based machine learning classifier within the D3 toolset—aims to automate the collecting and classification of web material. The goal of this automation is to lessen the amount of manual labour that cybercrime investigators must expend to sort through enormous volumes of online content without sacrificing the crawling process's control or the investigation's forensic integrity. The interactive features, which include user login, crawling parameters, and URL selection, are intended to sustain the requisite level of investigator involvement.

Subsequent studies ought to concentrate on improving and assessing this collection of tools using input from experts or users. It's also critical to comprehend and combat the Tor network's crawler blocking algorithms.

The main purpose of this research study was to establish knowledge regarding dark web crawlers in academic research. From this knowledge, a dark web crawler was developed to fit an already existing dark web cybercrime toolset called D3. In combination with machine learning-based annotation and categorisation tools in D3, the crawler developed and presented in this article, will capacitate the toolset to automatically collect and classify web content based on previously annotated web pages. Ultimately, this will save manual labour for cybercrime investigators, without losing control over the crawling process. Neither will it compromise the forensic soundness of the overall process, since a certain amount of operator presence and interaction is necessary for URL selection, crawling scope specification, and user authentication for example. A logical continuation of this research would be to further elaborate on and test the toolset, and also make an expert or user evaluation of it.

REFERENCES

[1] Darkweb research: Past, present, and future trends and mapping to sustainable development goals by Raghu Raman

[2] Deep Web, Dark Web, Dark Net: A Taxonomy of "Hidden" Internet by Masayuki HATTA

[3] THE DARKNET: AN ENORMOUS BLACK BOX OF CYBERSPACE by Ms. Paridhi Saxena

[4] The Dark Web Phenomenon: A Review and Research Agenda by Abhineet Gupta

[5] "Frederick Barr-Smith and Joss Wright. "Phishing With A Darknet: Imitation of Onion Services". In: 2020 APWG Symposium on Electronic Crime Research (eCrime). IEEE, Nov. 2020.

[6] "Van Buskirk, J., Roxburgh", A., Farrell, M., and Burns, L. 2014. "The Closure of the S Ilk R Oad: What Has This Meant for Online Drug Trading?," Addiction (109:4), pp. 517-518.

[7] Dark Web Illegal Activities Crawling and Classifying Using Data Mining Techniques by Abdul Hadi M. Alaidi

[8] "Winkler, I., and Gomes, A.T. 2016. Advanced Persistent Security: A Cyberwarfare Approach to Implementing Adaptive Enterprise Protection, Detection, and Reaction Strategies. Syngress.

[9] "Weimann, G. 2016b. "Terrorist Migration to the Dark Web," Perspectives on Terrorism (10:3), pp. 40- 44.

[10] "Smart Crawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin.

[11] "Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. Focused crawling: a new approach to topic-septic web resource discovery. Computer Networks, 31(11):16231640, 1999.

[12] "Luciano Barbosa and Juliana Freire. Combining classier to identify online databases. In Proceedings of the 16th international conference on World Wide Web, pages 431440. ACM, 2007.

[13] "Jayant Madhavan, David Ko, ucja Kot, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy. Google's deep web crawl. Proceedings of the VLDB Endowment, 1(2):12411252, 2008.

[14] "Andre Bergholz and Boris Childlovskii. Crawling for domain specific hidden web resources. In Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on, pages 125133. IEEE, 2003.

[15] "Kevin Chen-Chuan Chang, Bin He, Chengkai Li, Mitesh Patel, and Zhen Zhang. Structured databases on the web: Observations and implications. ACM SIGMOD Record, 33(3):6170, 2004.

[16] Samtani, S., Chinn, R., Chen, H., and Nunamaker Jr, J.F. 2017. "Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence," Journal of Management Information Systems

[17] Sapienza, A., Bessi, A., Damodaran, S., Shakarian, P., Lerman, K., and Ferrara, E. 2017. "Early Warnings of Cyber Threats in Online Discussions," IEEE International Conference on Data Mining Workshops

[18] Sun, Y., Edmundson, A., Vanbever, L., Li, O., Rexford, J., Chiang, M., and Mittal, P. 2015. "Raptor: Routing Attacks on Privacy in Tor," USENIX Security Symposium, pp. 271-286.

[19] Tanenbaum, A.S., and Van Steen, M. 2007. Distributed Systems: Principles and Paradigms. PrenticeHall.

[20] Jansen, R., Juarez, M., Gálvez, R., Elahi, T., and Diaz, C. 2017. "Inside Job: Applying Traffic Analysis to Measure Tor from Within," Network and Distributed System Security Symposium: IEEE Internet Society.

[21] Jansen, R., Tschorsch, F., Johnson, A., and Scheuermann, B. 2014. "The Sniper Attack: Anonymously Deanonymizing and Disabling the Tor Network," Office of Naval Research, Arlington.

[22] Ahmad, A. 2010. "Tactics of Attack and Defense in Physical and Digital Environments: An Asymmetric Warfare Approach," Journal of Information Warfare.

[23] The Anonymity of the Dark Web: A Survey by Javeriah Saleem

[24] Dark Web 101 by Major Jeremy Cole

[25] Dark Web: A Web of Crimes by Shubhdeep Kaur

[26] Beneath the Surface: Exploring the Dark Web and its Societal Impacts by Hasan Saleh

[27] The Dark Web: An Overview by Kristin Finklea

[28] The Dark Web Dilemma: Tor, Anonymity and Online Policing by Eric Jardine

[29] Dark Web by Kristin Finklea

[30] The Dark Web: A Dive into the Darkest Side of the Internet by Divya Yadav

KS

ORIGINALITY REPORT

15%	2%	14%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Jesper Bergman, Oliver B. Popov. "Exploring Dark Web Crawlers: A systematic literature review of dark web crawlers and their implementation", IEEE Access, 2023 Publication	10%
2	Jesper Bergman, Oliver B. Popov. "Exploring Dark Web Crawlers: A Systematic Literature Review of Dark Web Crawlers and Their Implementation", IEEE Access, 2023 Publication	3%
3	www.irjet.net Internet Source	1%
4	pdfcoffee.com Internet Source	<1%
5	John M. Carroll, Mary Beth Rosson. "A Trajectory for Community Networks Special Issue: ICTs and Community Networking", The Information Society, 2003 Publication	<1%
6	rest.neptune-prod.its.unimelb.edu.au Internet Source	<1%

7	"Applied Cryptography and Network Security", Springer Science and Business Media LLC, 2017 Publication	<1%
8	mafiadoc.com Internet Source	<1%
9	www.hindawi.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		