Efficient Crawler to Enhance Dark Web Investigation for Cybercrime Detection

BY

Richard Jayson Sarkar

ID: 241-31-003

Presented in Partial Fulfilment of the Requirements for the Degree of

Master of Science in Electronics and Telecommunication Engineering (M.Sc. in ETE)

Supervised By

Engr. Md. Zahirul Islam

Assistant Professor, Department of ICE

Daffodil International University

Dhaka, Bangladesh

November 2025

Approval

The thesis, which is titled **"Efficient Crawler to Enhance Dark Web Investigation for Cybercrime Detection"** and was written by Richard Jayson Sarkar (ID:241-31-003) while he was a student in the Department of Electronics and Telecommunication Engineering at Daffodil International University, has been accepted as sufficient for partial fulfillment of the requirements for the Master of Science degree in Electronics and Telecommunication Engineering. Furthermore, the thesis has been approved for both its content and its style. In November 2025, this presentation was given.

Board of Examiners:

1. **Engr. Md. Zahirul Islam**

   Assistant Professor

   ICE, DIU

   Signature: _____

2. **Md. Taslim Arefin**

   Associate Professor and Head

   ICE, DIU

   Signature: _____

3. **Dr. M. Quamruzzaman**

   Professor and Head

   EEE, Dean, FSE, WUB

   Signature: _____

Declaration

Under my direct supervision, the work in this study was completed as a dissertation and project in the Department of Information and Communication Engineering (ICE) at Daffodil International University (DIU), Faculty of Engineering. I declare that this thesis is my own original work, accomplished without any input from any unauthorized sources, with due acknowledgment to all references that were used.

Supervised By:

Signature: _____

Engr. Md. Zahirul Islam

Assistant Professor, Department of ICE, DIU

Submitted by

Signature: _____

Richard Jayson Sarkar

ID: 241-31-003

Department of ETE, DIU

Acknowledgement

I wholeheartedly give thanks to the Lord from every bit of my heart for his mercy on me to complete the thesis. The thesis would have been impossible for me to successfully complete without the Lord's mercy.

I would like to acknowledge the opportunity given to me by Daffodil International University, Bangladesh, to join the Masters' program. I, therefore, want to specially thank the whole Daffodil family, my teachers, and the university's management for counseling, motivating, and advising me in every step of my educational journey.

I am thankful to my families and colleague for their support and direction in my educational endeavors. I would like to pass a special acknowledgment to my beloved mother, Jyotsna Sarkar, for the motivation and guidance that made me overcome every difficulty that came along my way toward Master graduation. I would also wish to thank my lovely wife and my sister, whose encouragement, support, and understanding have kept pace with my educational journey. I want to give my deep heartfelt thanks to all my colleagues for their valuable ideas, suggestions, and help that enabled lightened my heavy workload so that I could concentrate on my thesis improvement. I feel that I am blessed to have a number of caring and understanding individuals around me.

I also wish to give a heartfelt thanks to my research supervisor, Engr. Md. Zahirul Islam, for his useful direction, motivation, strong ideas, and perceptions that guided all the time in my research. His guidance and direction always provided me with the strength's foundation that I wholly relied upon and this was also greatly potential for my study to be completed.

Finally, I would like to thank my committee members and external examiners whose valuable comments and suggestions which encouraged me to improve this thesis. I also want to acknowledge each and everyone in the university who guided me all through my study. It was your generosity that made me achieve this degree.

Abstract

Historically, the Dark Web has been an online area where cybercrime has grown. This area serves mainly as a secret or hidden marketplace for reportedly goods and services exchange secretly to avoid identification. Among other operations want to sell harmful items, information such as stolen, secret information. Due to the absence of programs paired with CAPTCHA security, data encryption and the complexity of the method, technology like Tor, I2P, or Freenet, and dynamic URLs make the WebCrawler completely useless for identification and investigation. This work will suggest designing and implementing an effective dark web crawler for adaptation for use in cybercriminal activities and for qualified professional forensic analysis. During this work, many varieties of elementary data processing methods were realized from seed detection, information base clustering, machine learning categorization and structural storage to protect from crawling to facilitate forensic analysis. The feature of adaptive algorithms was acknowledged and fundamentally inflexible to access and cross the barriers. Besides, it is significantly cross-network compatible which shows the potential for sufficient coverage of services for maintain privacy. From the point of view of this experimental verification, the crawler can be significantly more accurate for the detection method and certainly meaningless regarding the amount in relation to the standard method and approaches. This research breakthrough brings in real-time investigative tools that can be both precise and cost-effective, while keeping the data undoubted forensically. In a country like Bangladesh, for instance, these devices will prove an especially strong boon. Because of these results, it's clear that crawlers do not have to operate alone. It also has ample capacity to aid law enforcement, academicians, and big business in alleviating the swiftly growing threat of cyber criminality by boosting digital resilience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The dark web is a smaller, yet intriguing subset of the Internet when compared to the vastness of the online world. There are three layers of the Internet: the surface web, the deep web, and the dark web. The surface web is a small part of the Internet that is publicly accessible and indexed by search engines [1, 2]. It is home to databases, private resources, and any other content that search engines will not crawl and index. The dark web is one of the smallest parts of the Internet, no more than 6%, and yet so much of what happens there seems outsized when compared to its actual size. [3–5]. Despite making up only 6% of the Internet in total, the dark web does seem to punch above its weight in many respects. Only 4% of the Internet is indexed by search engines and available on the surface web. The deep web accounts for more than 90% of existing web information and is responsible for databases and other resources that cannot be indexed [1,2]. But the dark web is still a space with its ground-level illicit marketplaces and extremist propaganda, just more localised and less so. The dark web can help for good, like protecting the privacy of journalists and activists under oppressive regimes [6]. This kinds of services frequently change their domains for identification and depend heavily on encryption, applying CAPTCHA or other similar comparable technics that common web crawlers cannot bypass easily [7]. Many new technologies face inactivity, limitation and a lack of complexity to adapt to the constant shifting environment of the dark net. Although we built a established methods, we also faced additional unresolved difficulties. Overall, we developed an adaptable and durable crawler for the Dark web investigations and future research of improvement. [8].

Below Three facts inspired this work:

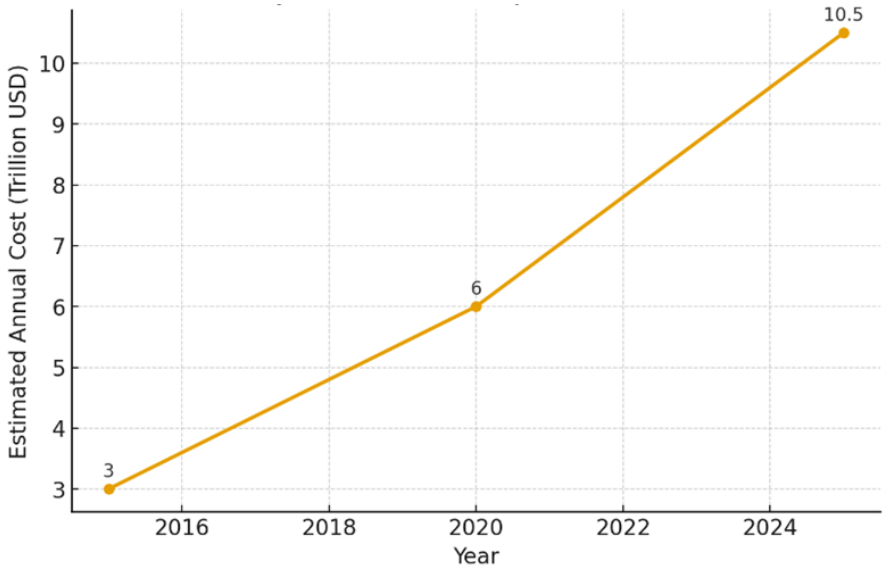

Figure 1.1: Cost Projection of Global Cybercrime

by Morgan et. al.

Operational Requirement: Additionally, by 2025 cybercrime is expected to bear cost over $10.5 trillion a year and consequently the demands for adaptable tools for investigation [9–11]. Dark web platforms are frequently associated with ransomware, common fraud, stolen and confidential identities, data compromises, which presenting a significant hazard to governments, business companies and common people as well.

Technology gap: The existing crawling solutions focus mainly on Tor-based services, such as OnionScan and TorBot, and have little to no support for I2P and Freenet [12, 13]. Most of these tools fail to avoid new countermeasures such as CAPTCHAs, secret directories, and adaptive authentication, which can keep some areas hidden (blind spots) and lead to inefficient data retrieval (which is far worse than having actual gaps in the collected data).

Academic Interest: There has been little research on the use of AI and machine learning for Dark Web investigations. Research on AI attacks for hidden services, anti-crawling bypassing, and especially adaptive volatile content monitoring is still limited. Building on this thread of research opens space for more efficient, scalable, and forensically verifiable cybercrime detection.

1.2 Understanding the Dark Web and Its Challenges

The dark web is not simply another hidden location on the Internet; it is something more than that. Deep web is a place where material is hidden behind paywalls or login restrictions, while the dark web is made purposefully difficult to locate.

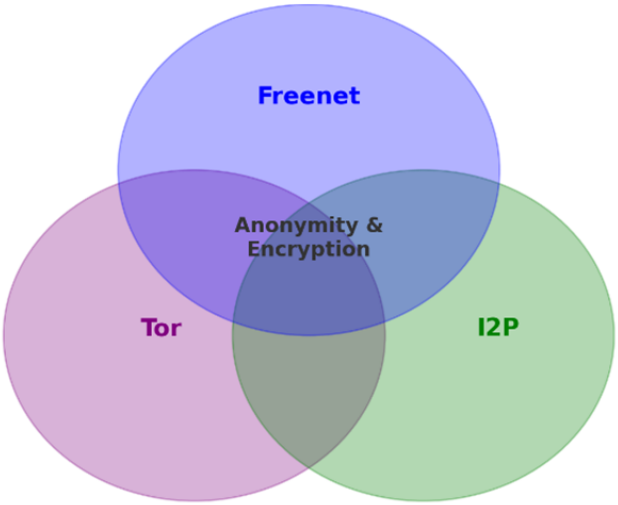

Figure 1.2: Distribution of the Internet

from brighplanet2024deepweb

Accessing it requires special software like Tor, which hides identities by passing traffic through a series of relay layers, or I2P, which sends messages through encrypted tunnels. Unlike data on Tor, which is spread throughout several interconnected nodes, data on Freenet is stored in separate nodes, making censorship quite challenging. There are several challenges built into the exploration of the dark web:

Anonymity and encryption: The key advantage of dark web tools is their multiple layers of anonymity. Tor is one of the networks that uses onion routing [3, 4]. This technique encrypts user requests numerous times and runs them through several core nodes operated by decentralized volunteers before they reach their destination. As each relay decrypts only one layer, no node knows both the source and destination. Freenet, like I2P, places encrypted data [4] across a decentralized system; on the contrary, I2P uses encrypted tunnels [14].

Dynamic/Volatile Content: The dark web is constantly changing, which is not the case on the surface web, where domains and websites remain property for long periods of time. Hidden services often switch domains, migrate hosting providers, or disappear to avoid discovery, takedown, or a denial-of-service attack. There are Onion sites that change address every few hours or days, making it difficult for crawlers to keep an up-to-date index and archive. As a result, by the time investigators would have discovered a site, it could have changed its content or geographically moved.

Counter-Crawling Strategies: Many operators of dark web services employ additional defences to prevent automated scraping. Features include CAPTCHA for human confirmation, password-protected entry points, hidden catalogues not connected to public indexes, and numerous client-side scripts that prevent ordinary crawlers from handling information. If not, different anti-bot measures are used, such as dynamic challenge-response protocols or time-based authentication tokens. To overcome these problems, complex methods must be employed, such as AI-based CAPTCHA solutions, headless browsers, or stealth crawling.

Balkanized Networks [15]: Tor is the single most studied and analysed anonymity network on the planet, and yet this analysis is only scratching the surface of the dark web. Not out of any lack of interest but rather a result of a Balkanized research landscape, alternative systems such as I2P or Freenet have been studied to a much lesser extent. Many of the tools that exist are geared towards Tor and other networks, so hidden communication slots on those networks go unnoticed. This gap allows for complex investigations and enables individuals across the globe to stay oblivious to cyber risks.

Data Extraction Challenges: Dark web crawling techniques are often slow, inefficient, and resource-intensive. They do not travel far and cannot keep pace with hidden services that change address every few minutes. Additionally, like all other technologies, most of them do not allow real-time monitoring, which is needed to patch new threats like a new variant of ransomware, zero-day, or even leaked databases. Suppose an organisation or law enforcement cannot derive insights from the data it collects promptly. In that case, the data is of little use, especially in situations requiring swift action.

The Authenticity of Forensic Data: Although crawlers can collect data, there is always a doubt whether the data is authentic and if it can be admitted to the court as forensic data. Suppose data is collected in violation of these rules on the chain of custody. In that case, the court might not accept the evidence presented to them, or the anonymity process might be compromised. This creates technological and legal risks for individuals and companies.

AI that's underutilised : There is still a lack of usage when it comes to AI. There are several ways in which AI has the potential for enhancing dark web investigations, namely anomaly detection, automated classification, natural language processing, predictive analytics, and so on. But only a few existing systems leverage AI power across the board. Many are either manual or otherwise rule-based crawler approaches that typically cannot keep up with the latest trends in which material is hidden or spoken language moves. As a result, threats often have low priority for present-day crawlers, fresh criminal methodologies remain undetected, and massive datasets are ineffectively filtered, leading to sub-optimal efficiency overall.

Issues of legality and ethics [16]: The prospect of surveilling the dark net is fraught with legal and ethical issues. On the one hand, governments and organizations should put an end to unlawful behaviour like trafficking, fraud, and terrorism. But blocking access to viewing such content may violate free speech and privacy principles. Finally, state monopolies on information can be just as morally problematic in that political or cultural motivations can lead to unequal treatment of dissenting perspectives, creating ethical quandaries in a historical setting where many groups have been marginalised through coercive means. Hence, any investigative system must be established to calibrate the balance between the criticality of cybersecurity and the demands of human rights and ethics.

In summary, these issues collectively suggest that conventional crawling methods are ineffective on the dark web. Common means of getting around lack the agility, wisdom, and strength required for joint action and cannot navigate through environments that are as unstable and hostile.

1.3 Problem Statement

The existing crawling solutions do not live up to expectations due to the growing interest in the Dark Web. Due to the constant impact of a fluctuating environment, they often suffer from a lack of flexibility, are implemented in a restricted manner, and function with a poor level of efficiency. Consequently, investigators encounter obstacles, insufficient datasets, and difficulties regarding the preservation of anonymity and security during their activities. The most significant constraints are as follows:

Ineffective Resistance: The current tools use manual CAPTCHA solvers heavily. Although they are increasingly ineffective against advanced anti-crawling [12, 13, 15] and bot-detection mechanisms, for an example by blocking connections to websites that require such challenges and replacing one set of questions with another every 60 seconds, they are beginning to show their age. Three new operas by one director in a year indicate the limitations for original creation.

Restricted Network Access: Many of the current crawlers are locked into the Tor network, and minimal emphasis has been placed on interoperability with other anonymity networks like Freenet, for example. This limited scope focus on investigating blind spot that are hidden which can obstruct a comprehensive idea of cross network unlawful activities.

Architectures that are not ideal: Many tools use synchronous crawling methods which not only make it slow down for overall operations but also raise the risk of detecting the IP blocks. It is complicated, if not impossible, to achieve real-time identification when such inefficiency impedes threat visibility. Furthermore, larger-scale forensic investigations that employ similar methods are hampered by their lack of scalability.

In law enforcement, these shortcomings cause them to do their job inefficiently, stifling investigations and minimising the effectiveness of forensic applications. This study aims to overcome these limitations by designing a simple, efficient, adaptable, safe, and legally compliant crawler.

1.4 Objectives of the Research

This research is based on a high-volume, lightly configured search engine. It avoids typical crawling constraints by deploying cleverly written code and is law enforcement-friendly in the detection of dark-web activities. The present study outlines the aims as follows:

Develop methods to access hidden directories, handle authentication, and bypass CAPTCHAs: Deep-web services have powerful safeguards against crawling. The research develops adaptive headless browsing, AI-based CAPTCHA solvers, and stealth navigation to get through authentication, hidden directories, and other challenges.

Provide support for multiple networks: We would like to develop a modular crawler that works on anonymized networks. This will enable us to find threats essentially in real time, reduce response times, and schedule threats based on their importance to the system.

Enhance efficiency to enable near real-time monitoring of its Dark Web business environments: Onion sites went away quickly, and cybercriminals move around quickly. We propose a workflow for improving scheduling and setting priorities better, which might assist us in finding and responding to new threats promptly.

Apply machine learning to optimize task ranking, detect anomalies [17, 18], and automate data categorization: AI/ML will help with prioritizing tasks, finding abnormalities, and adaptive data categorization by using sophisticated methods to strengthen correlations and identify outliers that are associated with criminal activity.

Make crawls forensic, sound, and fully anonymized to preserve the chain of evidence: Crawls of the dark web are only valuable if they are processed forensically. This study develops auditing rules to protect the chain of custody, the anonymity of the investigator, and the integrity of the evidence.

1.5 Research Questions

- What are the best ways to build advanced crawlers?
- Which design rules help them work across anonymous networks?
- How to quickly capture unstable content?
- How does AI improve anomaly detection and classification?
- How secure is user and system anonymity?
- What ensures the forensic accuracy of collected data?
- Is the new crawler faster and more flexible than the current tools?
- What legal and ethical safeguards should be included?
- How to handle multilingual dark websites automatically?
- How to benchmark crawler performance across networks and attack scenarios?

1.6 Scope and Limitations

Scope: This research will be very clearly based to ensure that the crawler that is being proposed is academically reasonable, ethically justifiable, and practically useful. For study and testing, not for mass or commercial use. It highlights Tor, I2P, and Freenet [3, 4, 14], which are key parts of the cybercrime ecosystem. This will entail either using dormant/inactive services whenever possible or conducting all real-life experiments under sandbox conditions and in a manner compliant with the law and ethics. This study look at the trade-off between technical practicability for real world, maintain research innovation and compliance with standard ethical code of conduct

Limitations: While the research offers these contributions, the authors acknowledge its limitations.

First, as the crawler will not crawl invite-only or private marketplaces and forums that require credentials or insider access beyond the ethical scope of this work, data provenance generations are bound only to a subset of malicious sources. In turn, this means the study does not cover part of dark web activity, mainly that in closed communities.

Second, the crawler works effectively based on the keywords and links indexed; however, it may not have complete coverage of hidden services. The reliance itself introduces a risk of partial data collection since many hidden services do not use standard indexing or intentionally obscure entry points.

Third, no notification mechanism can alert researchers about new or changing threats in near-real time. It enables near real-time monitoring. However, the drawback is that analysts must manually verify and interpret results, potentially leading to delayed response times.

Fourth, this makes it less scalable in scenarios where the authors use advanced obfuscation. Thus, the required parsing analysis become tedious, limiting a crawler from extracting practical knowledge effortlessly and automatically from the data. This implementation does not support the fifth, automated threat grading and prioritisation. This means that there will still be a level of subjectivity and possible delays before analysts have a detailed view of the collective information, as they will have to evaluate its severity themselves.

Finally, combinatorial experimentation is limited by legal and ethical constraints. This research does not target any live, illegal marketplaces or forums, consistent with applicable guidelines of all nations and institutions. Additionally, decentralised systems like Freenet may cause delays in the network, particularly in terms of crawling speed and efficiency, compared to centralised environments [16].

1.7 Contribution of Research

The outcome of this research will significantly contribute to both practice and academia in terms of dark web monitoring and forensics.

Real-world Impact: The outcome of this study was a highly efficient, extensive and highly maintained crawler which has led to reduce the gap of newly standard Dark web tools, method of investigation [12, 13].

Improved Threat Profiling: The primary findings is that while the middle ground is the hybrid of artificial intelligence enhanced countermeasures and cross-network, social media network and semantic data analysis procedure. [3, 4, 14].

Technical Innovation: The essential point is that it will provide benefits with clue to the investigators and researchers to use a crawler which will support multiple secret network for more accurate identification and reporting threats that use Tor, I2P and Freenet. [17, 18].

Investigator Security and Privacy: This is achieved by prioritizing the investigator's anonymity for the investigator's safety and ensuring investigations [16] are also safe, legally transparent and sound, which can make the crawler a trusthworthy tool and academic structure improvement.

Adaptive Algorithms and Forensic-Aware Data Acquisition: It proposes adaptive flexible algorithm for gathering all data which is aware forensics, which makes the collected evidence more reliable while ensuring it meets legal and forensic standard criteria [18].

Fastest Cybercrime Response: The system speeds up crawling performance, which takes less time and saves time to analyse the digital forensic evidence to find and analyse new threats. [6].

Proactive Threat Intel: This research offers a proactive approach by enabling analysts to detect early warning signs of poor behaviour before they trigger a cyberattack [9].

Finally, the originality of the project is that the benchmark data set and reliable theory which offers effectiveness will promote more research scope in future.

1.8 Ethical Considerations

This research has been conducted in complete compliance with legal and ethical requirements. Crawling test were restricted to sandbox setup, vacant onion websites or publicly accessible datasets. There were no live online illegal markets which were exposed as well as nothing was able to find out who these individuals were. As a crawler researcher it remains both ethical and forensically safe. It protects anonymity while maintaining its academic value and legal defense procedure. [16, 19].

1.9 Structure of the Thesis

The outline of this work is as follows:

- **Chapter 1:** Introduction contains crawler background, problem, study history, goals, value and scope of work.
- **Chapter 2:** Literature review provides existing procedure, highlights the legal and ethical issues and concludes on the research gap.
- **Chapter 3:** The structure of the system contains overall a diagram where it shows the architecture of the proposed crawler.
- **Chapter 4:** Methodology has shown the workflow and algorithms, way to discovery anonymity preservation, mechanism for filtering the dataset.
- **Chapter 5:** Implementation showns the technical aspect of execution. This contains the tools, APIs, and module applied as well as it's output.
- **Chapter 6:** The results and discussion has the results of experiments, performance of output data, a comparison with existing crawler output data.
- **Chapter 7:** Conclusion and future work summarized with next stages for contributions and the limitations of the research on the design of a web crawler.

Chapter 2

Review of Literature

2.1 Overview of the Surface Web, Deep Web, and Dark Web

The internet is divided into three main parts. They are the surface web, the deep web, and the dark web. So the surface web is the simplest portion that the public can access easily like online retailers, blogs, news sites and the thematic portals. This part of the internet is more familiar to millions of ordinary users and it is taken into accounts by some search engines like Google, Bing. The surface web represents only 4% of internet-based content. [1, 2].

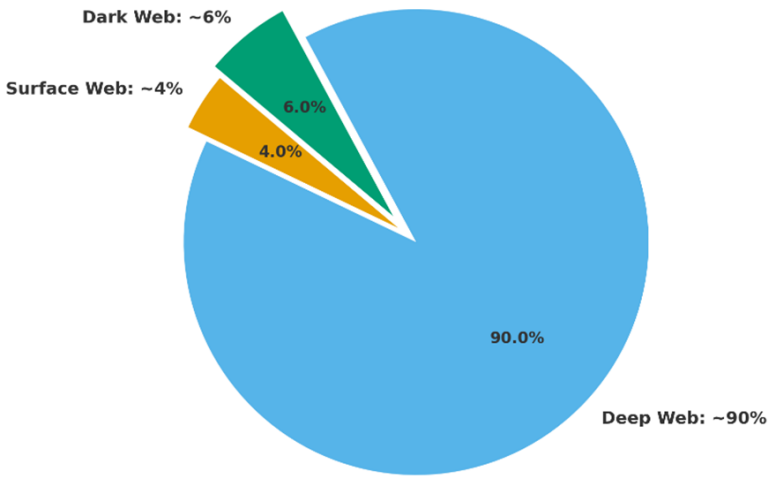

Figure 2.1: Distribution of the Internet

The Dark web is the most tiny portion of the Deep web and can only get access via onion routing network establish around the world such as Tor, I2P and Freenet. [3, 4, 14]. These networks use decentralized routing protocol and encrypted multi layer to maintain the privacy for confidential data. A lot of journalists, authorized person from different countries use this for communication bypassing traceability. [6]. it might also show up as a base for dark web hackers, financial crimes, or terrorist materials [9].

2.2 Role of Crawlers in Cybersecurity and Digital Forensics

Web crawlers can find hacked credential dumps, detect darknet markets selling illegal products, and assist the law in constructing forensic evidence for prosecution. Crawling the surface web is simple because of the same patterns and open protocols, but this is not the case for crawling the dark web. Crawling, the traditional technique used for indexing the internet, is infeasible, especially on networks like Tor, I2P, and Freenet, with their anonymity, rotating domains, and hidden services.

Additionally, Dark Web operators intentionally set up anti-crawling defenses [3, 4, 14, 20, 21], and so Dark Web crawlers need to be specialized to function in this hostile and dynamic setting. However, because these are not traditional crawlers, these systems have to handle anonymization networks, avoid anti-scraping techniques, and perform data collection that is legally defensible.

2.3 Challenges in Crawling the Dark Web

Crawling the dark web is technically, operationally, and legally riskier, and the challenges there make the dark web a much more intricate target than the surface web. including:

Privacy through Anonymity: Various dark web platforms like Tor, I2P, and Freenet are designed for anonymity through multi-layered encryption and cryptographic routing to hide both source and destination [3, 4, 14]. Tor, I2p and Freenet are just existing several dark web platform. All of them always aim to keep their identies private which they do via multi layer encryption and cryptographic routing protocol to mask the communication of source and destination.

Transient and Fluid Material: Dark web hidden services are particularly vulnerable since criminal groups change their domains routing to escape police or DDoS attacks. Some onion sites have a very long life cycle but some just disappear within a few hours or in a moment. As a result most of the time it is impossible to either index dark web sites for a longer period of time and any effort of work put in indexing makes the crawlers life miserable trying to collect time to time intelligence data. [6].

Anti-Crawling Techniques: The operators of the whoever is running or selling whatever they use their own methods to stop of prevent crawling of their sites for an example CAPTCHAs, password gates, hidden directories, JavaScript and multiple step verification. [7, 20, 21].

Network Fragmentation: There are still other aspects of the dark web frequently explored, parts of the dark web beyond Tor, such as I2P, Freenet, etc. This network fragmentation creates gaps or blind spots to monitoring other networks and looking at Tor makes it even limiting to get full picture of threat intelligence method [15].

Resources and Time: Scraping these networks on a large scale is extremely slow. This exercise gives a sense of this, in addition to the work that goes into constant re-indexing and which increase more anonymization layers. Hidden services are much less stable than their surface web and they easily go offline much more often [22, 23].

Forensic Soundness: Even if a crawler can extract data and the forensic soundness is not good enough. Though the data is extracted, and need to be stored in a form that will guarantees that the records has no change to be altered. If the crawler don't follow the chain of procedure it may be the case that even extraction is not admissible in a court of justice. Its hard to validate and very volatile in nature when the crawler do the crawling to the dark web and snatch of intelligence and at the rate these problems become worse [16, 18, 19].

Combining these to prevent crawling of dark web and ensuring timely and scalable extraction occurs multiple challenges. For this solutions need to be personalized to address these challenges, as those issues are anonymity, volatility, against crawling, fragmental techniques and forensic sounds.

2.4 Existing Dark Web Crawling Tools and Techniques

During the past couple of years, several tools have been made to help crawl and analyze hidden services on the dark net. These are helpful tools but they are rather superficial and might be improved in many ways. They don't fully enable the possibilities of what can be done with modern methods and technologies.

Onion Crawler: This search engine is created to help in searching and indexing content on the Tor network. The difficulties is that it doesn't work well against more advanced anti crawling system like CAPTCHAs and other strong system of such kind.

Ahmia [24]: This is a search and information engine system similar to those of the first part of the note but the one searches and indexes information only about the Tor hidden services.

OnionScan [12]: Primarily a scanner for onion service vulnerabilities that looks for the main root causes of onion service misconfigurations and that's why information leaks happen. Not scalable unable to crawl other networks outside Tor, It might be helpful for some cybercrime investigators.

TorBot [13]: In the other hand TorBot is not a scraper for Tor hidden websites but is rather a tool that can scape things in a more organized structural way. This tool is a slightly troubling because TorBot unfortunately does not have many ways to get over CAPTCHAs or limiting that can be put on logging in.

2.5 Ethical and Legal Issues in Dark Web Crawling

The dark web presents a series of ethical and legal contradictions that put the effort of fight against crime clash with the defense of privacy and free space. As a result participant teering on the divide between security and civil liberties. This is considered since the investigator need to stay hidden while avoiding international law, jurisdictional issues and digital rights while remaining unfinished or undetected. It is important to have rules and standard method like [16] but evidence from the dark web must be gathered in a way that does not make it invalid in the court.

2.6 Research Gaps and Need for Improved Crawlers

A review of current tools and literature reveals several significant gaps [25–27]:

Gap Identified	Impact
Limited Multi-Network Crawling	Leaves I2P and Freenet unexplored
Weak Evasion Strategies	Blocked by CAPTCHA and hidden directories
Minimal AI Integration	Weak anomaly detection and automation
Lack of Real-Time Monitoring	Delays in detecting cybercrime activities
Insufficient Forensic Focus	Evidence may not hold up in court
Barriers for Developing Economies	Few affordable/open-source solutions

Table 2.1: Research Gaps and Their Impact

2.7 AI and Machine Learning in Dark Web Crawling

Artificial Intelligence (AI) and Machine Learning (ML) have already been used in the journalism field to penetrate the technology that drives dark web crawling and unique surveillance [17, 18]. Unlike traditional crawlers, which rely mainly on rule-based indexing and relatively stable link-following methods, the adaptive, intelligent, and scalable features of AI make effective crawling of extremely dynamic and adversarial domains, including the dark web, possible.

Content Classification and Filtering: Natural AI-powered natural language processing (NLP) models were applied to classify the content obtained from the dark web marketplaces and forums [28]. These models can tell the difference between regular discussions and between drug deals, gun sales, and creds for stolen access. For instance, supervised learning methods have been applied to tag listings in darknet markets, thereby reducing investigative work.

Anomaly Detection: The dark web is known to be anomalous with high frequency because of sudden shifts of hidden services from one pseudo-random domain to another, sudden increases or decreases in service provision, or the emergence of new hidden services. These anomalies can be detected by the use of ML algorithms, which can then serve as alerts to investigators that there have been changes of interest based on the analysis of behavioral data [25–27].

CAPTCHA Solving and Access Bypass: Anti-Crawling Mechanism. CAPTCHA remains one of the largest barriers to crawling. This has led to a number of studies in recent years into the automatic and accurate solving of CAPTCHA challenges by using computer vision and AI-based recognition systems [7]. Reinforcement learning enables crawlers to adapt paths, access hidden links, and bypass login barriers without human help [29].

Predictive Analysis of Hidden Services: The hidden services of machine learning for Onion Site Uptime Prediction AI can also help predict their stability and long-term activities. The ML model uses historical data on the reliability of the site, performance of content in the past few scrapes, and a metric on how many nodes worldwide are able to fetch the catalog to predict if this onion site will either be running or taken down [30].

Multilingual and Cross-Network Analysis: Dark web material is intrinsically multilingual , and is in many languages, including English, Russian, Chinese and many others. These models are able to automatically process, classify, and analyze much of this non-English content an increase in scale that has only been made possible with the advent of AI-based machine translation and multilingual natural language processing models, with their increased monitoring frontiers [31].

Limitations of Current AI Use: Outside of traditional methods, though, there are AI-assisted limits to how one can crawl the dark web. Other than the few solutions that take a more advanced AI approach, many current solutions do not use AI at all or only have it in a minimal experimental atmosphere. Due to the high cost and low ethical feasibility of collecting such data, the large-scale training datasets required for classical speech representation learning have not been generally available [32].

Chapter 3

System Requirements and Design

3.1 System Requirements

The scope of the proposed dark web crawler for investigative purposes must fulfill both functional and non-functional requirements in addition to being relevant to current and future operations.

From the functional perspective: The crawler needs to find hidden services over anonymity networks (Onion, I2P, Freenet). It must also bypass CAPTCHA, authentication, and paywalls while still being fast and data-integrity ready. Due to the ever-evolving nature of the dark web, it should be able to support near real-time crawling and provide structured, analysis-ready data. Evidence must be held under forensic chain-of-custody to ensure it is admissible in a court of law. Your system ought to have good logging and monitoring with a focus on anonymity and investigator safety.

The non-functional requirements The reliability and crawlability of the system throughout the time. The system should be fast, scalable, and not cost a lot of resources for the process to consume a lot of private content with little computing power. It should mimic human activity and change its identity so as not to be detected or blocked by any means to last long in functionality. To avoid data loss, this also needs to protect the users and the system when the network is unavailable. It indicates that the system has to be very reliable, fail-safe, and secure. Crawler has to be well-modularized, extensible, and future-ready to enable the availability of AI-driven models, smart analytics, and external threat intelligence feeds on demand. This would make it easier for it to dismantle architectural walls.

3.2 Design Goals

Constructing a dark web crawler that works while observing ethical considerations would be challenging. The system design is influenced by six overarching goals that together lay the foundations for a crawler that is scalable, flexible, and designed for discovery as well as research [15, 18].

Figure 3.1: Overarching goals for Dark Web Crawler

Speed: Layered encryption, routing policies, and network latency make it inherently slow; dark web networks like Tor, I2P, and Freenet are all very slow in nature [3, 4, 14]. The crawler should be designed to minimize of time to set up and maximize throughput to run. It is so efficient that it can process terabytes of hidden content in a relatively short time, making it useful for both proactive oversight and reactive investigation activities.

Stealth: Crawling the dark web undetected, it should imitate human users by making random clicks with random time delays, with a dynamic session handling such that it becomes impossible for the website to differentiate it from a real human. This requires proper tuning between identity rotations and proxy switches, as IP blocks, de-anonymization, or anti-crawling defenses can still be triggered [20, 21].

Scope: Cybercrime does not only take place within the Tor network, nor does it rely solely on a Tor-only crawler. Thus, the crawler has to be cross-network and able to retrieve many types of content from forums, marketplaces, and blogs as well as from file-sharing repositories [12, 13, 24].

Relevance: Some crawled content has no value for investigation. Well, the crawler must have built-in intelligence filters for narrowing relevant content, such as keyword searches and semantic searches, as well as ML/NLP classifiers [17,31,32]. This prevents investigators from expending resources on topologically identical or extraneous pages, allowing them to focus on actionable intelligence. This mechanisms are essential to process typically consists of multiple components that allow to do timing sensitivity. These analysts only need to filter through vast amounts of latent data to extract the most valuable indication.

Accuracy: Forensic investigation require data to be correct fully and replicable. The crawler must dump in a way that can be verifiable and reproducible way so that any post-hoc claim between any of the two may be checked by court or journal [16, 18, 19]. Moreover the design made sure that the protocol used strong hashing, metadata preservation and chain of custody. This showed that the data collected can be now be stored as digital evidence that meets court standards [11].

Security: The crawler must be built in a such way that doesn't put the system or by extension running the system by humans. This involves not being attached again; all the members of the insightful authorities will be operating in an anonymous state of identities. Indeed, it is possible that the cracker could harbor some vendettas runs and attack one of the crawl's members in retaliation [22, 23]. Thus any such defense like this must not depend on replicating but preventing hostile mitigations.

In conclusion the design was build on the six objectives of efficiency, stealth, scope, relevance, the accuracy fo the data gathered and the safety of our intervention. The current design was targeted to objectives; to create a connection between academic research and practical application, and to challenge the limitations of innovation beyond academic limits in the realm of pragmatic digital crime [15].

3.3 System Architecture of the Proposed Web Crawler

The proposed crawler is a modularly structured crawler that also allows for easy expansion, flexibility, and stability. Below are the key components of architecture.

User Interface and Control Module: Provides a straightforward but powerful interface for investigators, from which they can specify crawl parameters, including keywords to target, and target anonymity network (Tor, I2P, or Freenet), and crawl depth. Seed links that are crawled from other sites, URL Discovery Module [15, 18].

Page Downloader: This is a tool that simulates human behavior; thus, it can browse through multiple web pages with a very low amount of latency. It offered techniques for handling authentication needed to see the secured parts of the site [3, 4, 14].

Parser Module: Parsers, just like how governments and law enforcement obfuscate the actions of almost every gang in history, will translate relevant text, metadata, and hyperlinks [17, 31].

Content Filtering and Classification Module: It uses filter rules and refines AI/ML models to score and rank crawled documents. This enables the filtering of raw, unprocessed information to only intelligence pertinent to investigators [28, 32].

Storage and Indexing Module: It applies filter rules and polishes AI/ML models [15] to score, and scores documents that are within the crawl boundary. This allows for filtering out raw intelligence to just the necessary intelligence for investigators, allowing them to have everything that they need.

Anonymity Manager: Organizes the information it has just collected in a database. Investigators may query through large volumes of datasets efficiently and in near real-time through computation indexing, which allows for fast search and retrieval of the data [20, 21].

Security and Logging Module: Ensuring the crawler itself will have sufficient security safeguards are in place to mitigate hacking or even manipulation. Having all captured data like crawl activity is logged and kept in a chain of custody that following the best practices for forensics [16, 19].

3.4 Modules Overview

Each module contributes uniquely to the system:

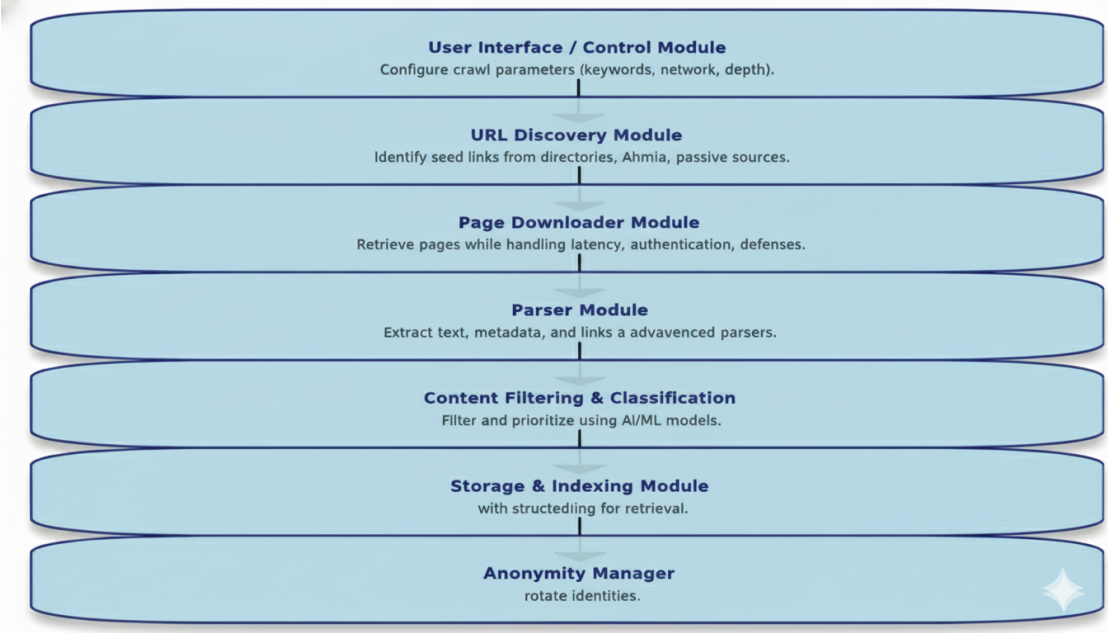

Figure 3.2: Architecture of the Proposed Dark Web Crawler

The crawler architecture is a modular with each layer addressing a important key role in dark web investigations [15, 18]. The **User Interface and Control Module** define the crawl parameter while the URL **URL Discovery Module** identifies seeds links from directories and clue such as Ahmia [24].

The **Page Downloader** handles latency, authentication, and anti-crawling defenses [3, 4, 14], and the **Parser Module** extracts text, metadata, and links [17, 31]. Data is refined by the **Content Filtering and Classification Module**, which applies AI/ML models [28, 32].

The **Storage and Indexing Module** ensures efficient retrieval, while the **Anonymity Manager** rotates identities for stealth [20, 21]. Finally, the **Security and Logging Module** enforces forensic soundness and compliance with legal standards [16, 19].

3.5 Dark Web Access via Tor, I2P, and Freenet

Darknet hidden services use special anonymity protocols which anonymize both the users and the service providers, making it very difficult for surveillance and monitoring, hence they need special methods to access them. In order to cover more anonymity networks and minimize the coverage holes for the investigative blind spots, the same crawler proposed in [15, 18] combines three well-known anonymity networks (a.k.a Anonymous Networks), including Tor, I2P and Freenet.

Tor (The Onion Router): Tor is the most common and well-known dark web network in the world as well as the largest provider of onion services, which mainly consists of forums, marketplaces, and leak sites. It creates this anonymity by routing the traffic through more than one relay that works as volunteers to hide their identities and the identity of the origin node from the destination node, layer by layer (onion routing). No single relay can determine both the starting and ending point of the tunnel in a single hop [3].

I2P (Invisible Internet Project): I2P is a peer-to-peer anonymity network which uses a principle of hidden services called "eepsites," which do not reveal hidden services to the public Internet. It communicates only through one-way, encrypted tunnels, promising a strong resistance to infiltration and censorship. I2P routers provide access to hidden resources, revealing a largely unexplored part of the ecosystem [14, 33].

Freenet: Freenet opens our eyes to make a peer to peer communication platform. This node share and replicate which making it very difficult to bring them down. It is also slower than I2P and Tor, but still it can store content with better quality. Usually in this way the content could be available for a longer period of time [4, 34].

Cross-Network Compatibility: The crawler uses all three networks to make it possible to increase the possibility to get more information and reduce the chance for missing important data when adversaries perform transactions, as it was discussed in the previous section in this paper. In addition using Tor, I2P and Freenet ensures the maximum level of legal and forensic security. [15, 22].

3.6 Data Collection and Storage Considerations

This research emphasized that the system used to collect data through the Dark web or Onion services is well planned in a systemic way. Data collection should confirm it following the ethical and legal guidelines. Only experiments with inactive or dormant hidden services and material already in the public domain are allowed; they don't run experiments against live criminal marketplaces or violate privacy. The way this technique tackles ethical challenges when employing the crawler underscores the techniques within it.

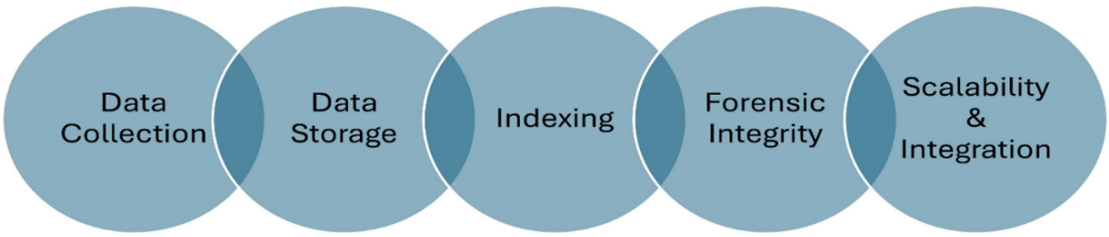

Figure 3.3: Data Collection and Storage Pipeline

The main types of data that are earmarked for harvesting include forum conversations, market ads, and stolen datasets. The data types are common on underground forums and range in value from being of extremely high intelligence to being of value to security researchers and law enforcement.

We should collect all this and figure out what format to store it in or what architecture to keep it in. Lightweight alternatives can also be used: SQLite [35, 36], our preferred option to prototype with, is a lightweight database that is nice to run small tests with. Since Dark web has characterized by its size and speed that is why it perform through a high-throughput scalable storage. It make these kinds of indexes which can be built on a number of dimensions like as URL, time slot, different types of network, keyword quality and this can perform fast querying and good analysis result. As a result it has a positive effect on retrieval and analytical performance score.

One particular design principle needs to be understood in the current design framework. Where forensic integrity [37, 38]. must be consider with high importance, A modern hashing algorithm should be applied to each and every piece of data to secure tempering the data as much as possible. Data lineage is also store in time through version control which means that the chain of custody is perfectly clear and has appropriate level of perfect data.

This is storage usually is designed with the ability to scale up [12, 13, 15], rearrange whenever its possible. It can easily be integrated into existing SIEM system and threat intelligence feeds which gives additional capabilities to automatically look at the weakness or vulnerabilities. All of this design have capabilities to give solution that use a combination of current technology and ability to to do well documentation following the method to make it safe and can collect and store large amounts of forensic data. Both academic and operational goals to be met by enabling the technical and legal safeties as per design.

Considerations Beyond Design: Beyond the specific design principle needs to do data crawling system where it must follow ethical principles and the requirements to store what is required for the analysis [16].Any content that is collected should be protected but so too are metadata like their precise timestamp or a URL, a number of network addresses whould become vital context in a future forensic review [38]. Encryption will make sure that both at rest, during transfers and storing will ensure it doesn't compromised and make it sensitive material for public [39]. Over time, storage may be periodically be submitted for auditing to proof it is forensically compliant. It can also be expanded in a clear way by adding an additional computational storage layer such as HDFC or more advanced cloud native object storage capabilities for long term archival [40]. There are other way to secure data that can be adopted, such as Blockchain which has ability to enable forensic records [41] to be stored in a way that cannot be changed.

Chapter 4

Methodology

4.1 Crawler Workflow and Algorithm Design

This consideration gives the workflow of our proposed dark web crawler that can perform balanced high efficiency, high stealthiness and forensic soundness. It is going to run in steps with one step approach and involve one stage of the whole data collection and data analysis pipeline. This begins with the identification of seed URLs [24], which can be obtained from curated lists, passive datasets, or directories. It establishes an anonymous connection via Tor, I2P, or Freenet [3, 4, 14], and then. After linkage, the crawler retrieves content using parsing strategies [28, 31], extracts useful data, filters, and categorizes what is found by its relatedness against the context document set, and stores all of that in a performant indexed database [35, 36].

The web is filled with deep web domains that make crawling it a daunting task. Services on the dark web are very ephemeral, constantly changing their IP addresses to avoid being taken down. The image streaming crawler that we proposed has a three-fold solution to this problem. The process starts with seed links from public web rings (Ahmia), community-administered link lists, and academic datasets. But these are the categories that crawl starting sessions are built on. The crawler gets links from each page and pts them into a queue to be covered more thoroughly. It uses ML and NLP w. It uses ML and NLP which identifies hidden service patterns in a unstructured text, exposing domain that follows traditional methods sometime overlooked. It combines static seed and adaptive discovery to obtain a wide and deep view coverage of the dynamic dark web [17, 18].

4.2 Anonymity Management

The difficulty of ensuring the anonymity for the investigator is not only an issue of procedure; it is the main ethical concept driving behind the dark web research. Even if investigators fails to take such precautions and establish anonymous connections to the Tor, it will be those Tor or the agency that the investigator runs the crawler. By creating a Tor connection, the website's network is being used through several different relays, protecting its identity as the source and the final destination. Fingerprinting and tracking are avoided by continuously rotating circuits [3, 20, 21]. The queries are regulated with a random delay to simulate ongoing human browsing behavior [7, 15].

In addition to Tor, the system integrates with I2P and Freenet, hence expanding coverage [4, 14] and mitigating blind spots regarding clandestine activities outside the Tor ecosystem [15]. This is only made worse by the fact that every time a new session is going to established, all headers that may be associated with any rotations stay unchanged. A bot can safely use this comprehensive set of business rules, and the risk of deanonymization will be much lower. However the resistance to inspection will be high [3, 20].

4.3 Content Filtering and Relevance Detection

Dark-web investigations necessitate for both moral and technical purposes, dark web inquiries need anonymity. The crawler sends routes the traffic by a series of proxies while also frequently replacing circuits and adds random delays to simulate human searching while erasing all fingerprints and automatically preventing detection. [20].

Besides Tor, they integrated I2P and Freenet into the system, thus broadening the coverage and endeavouring not to have any blind spots about underground activities beyond Tor [4, 14]. This is compounded by the fact that the session headers connected with identified rotation information. Such an operational technique can protect Bot functionality from being anonymous and make it more difficult to identify [21].

4.4 Efficiency Enhancements

Crawling an anonymous network takes a lot of resources because of the encryption, latency between networks and result and access restrictions between network nodes. To solve these we designed a customized crawler to overcome those challenges. Deduplication filters [15, 22] are URLs or content whose hashes or tokens have already been copied, saving storage space and processing resources.

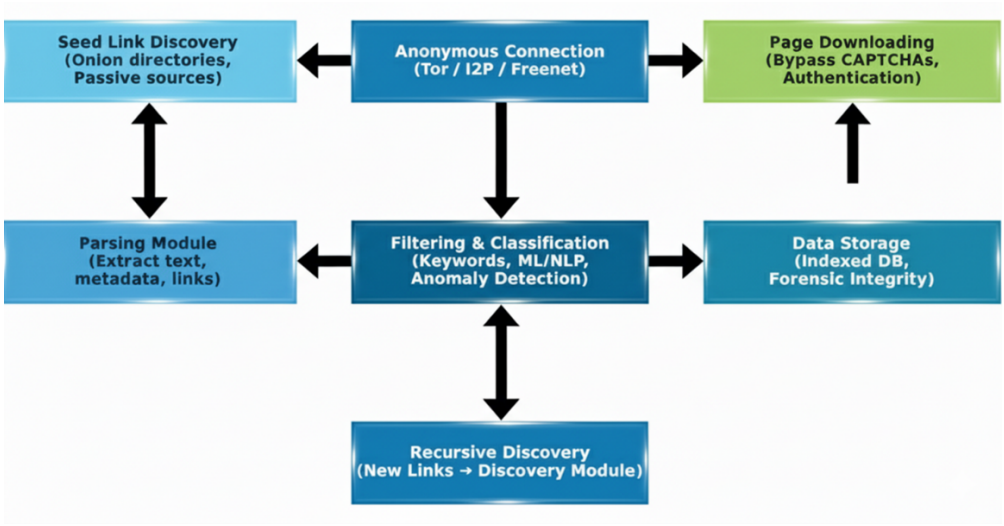

Figure 4.1: Workflow of the Proposed Dark Web Crawler

In most cases data is usually maintained to lower latency, and in anonymity networks, replicate queries are avoided [15]. When there is a heavy traffic, we use randomized load balancing technique which sends data over multiple circuits or proxies to improve stealth and network speed [3, 20, 21].

4.5 Security and Ethical Safeguards

Studying the dark net poses ethical and legal [6, 16, 19, 37, 38] questions. We build the crawler with multiple protection layers. Forensic standards, including hashes and timestamps, are applied to all data. It avoids unauthorized areas, restricting activity to public or test environments.

Chapter 5

Implementation and Experimental Setup

5.1 Overview of Implementation Approach

In this chapter, the Multi-Network Dark-Web Crawler (DWC) is discussed, which is developed to scrape actionable intelligence over Tor, I2P, and Freenet for lawfully permitted investigative work and university-based research-level work. The DWC is a forensic sound mechanism for automated discovery, collection, and classification of publicly accessible dark-web resources and stores them in a read-only, legally compliant manner.

5.2 System Architecture

In the system, each adapter connects to its anonymous network via a local proxy [3,4,14], keeping traffic contained within controlled environments [35,36]. The workflow follows five main stages:

1. **Target Frontier Management**
    - Reads a list of URLs from `targets.txt`.
    - Automatically distinguishes `.onion`, `.i2p`, and `freenet`.
2. **Network Adapter Selection**
    - Dynamically selects the appropriate adapter:
        - **TorAdapter** → Tor SOCKS 9050
        - **I2PAdapter** → I2Pd HTTP 4444 / SOCKS 4447
        - **FreenetAdapter** → fproxy 8888

3. **Data Acquisition**

    - Performs HTTP GET requests through the configured proxy.
    - Enforces read-only behavior with timeouts, delays, and error handling.

4. **Evidence Processing**

    - Stores raw responses under `output/raw/`.
    - records SHA-256 hashes and records WARC-style metadata.

5. **Indexing and Analysis**

    - Inserts metadata `output/evidence.db`.
    - Enables for later analysis and verification.

5.3 Development Environment

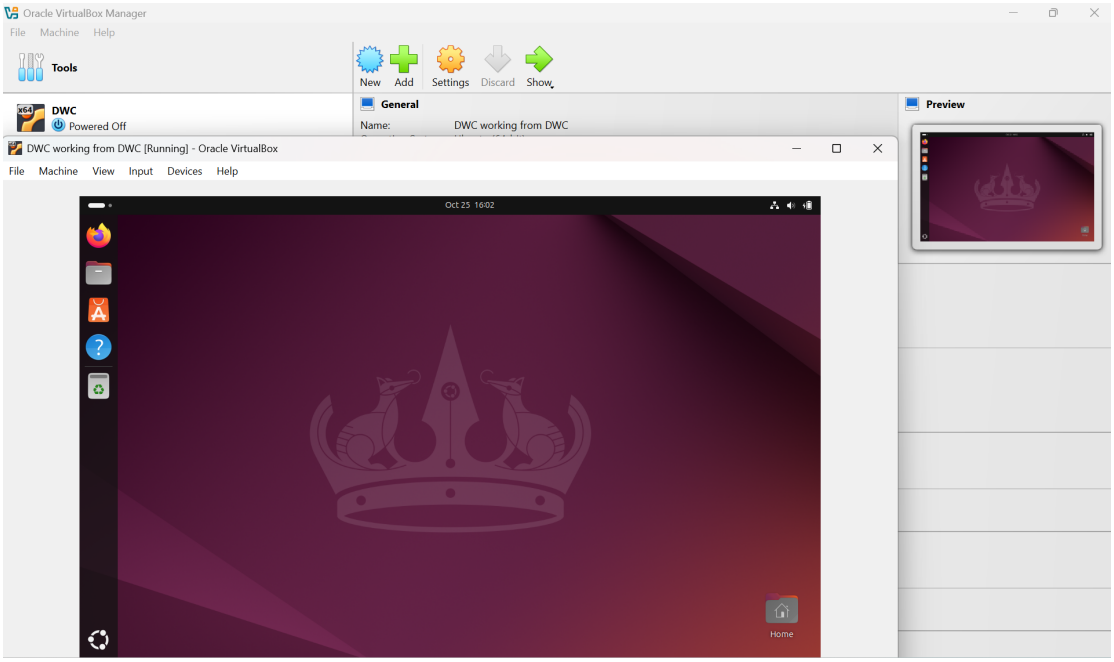

Figure 5.1: Ubuntu Virtual Environmnet

Host Machine: Windows 10 (64-bit)

Virtual Machine: Ubuntu 22.04 LTS (VirtualBox)

Hardware Configuration: Intel Core i5 CPU, 8 GB RAM, SSD storage

Language & Tools: Python 3.10, `pysocks`, `sqlite3`, `hashlib`

Network Daemons: Tor (9050), I2Pd (4444 / 4447), Freenet fproxy (8888)

Environment Isolation: Virtual environment `dwc-venv` to ensure reproducibility

Control Mechanism: Execution permitted only when `USER_APPROVAL` token is present

5.4 Functional Workflow

- **Target Loading:** Reads `targets.txt` using `crawler.py`.
- **Adapter Detection:** Selects Tor / I2P / Freenet.
- **Retrieval:**`tor_adapter.py`, `i2p_adapter.py` and `freenet_adapter.py`.
- **Evidence Hashing & Storage:** Utilizes SHA-256 to store responses.
- **Indexing:** Logs metadata into SQLite database `output/evidence.db`.

5.5 Script and Directory Structure (Visual Map)

```
dwc_multinetwork_ready/
 crawler.py              ← Main controller & coordinator
 tor_adapter.py          ← Tor network (SOCKS 9050)
 i2p_adapter.py          ← I2P network (HTTP 4444 / SOCKS 4447)
 freenet_adapter.py      ← Freenet via fproxy 8888
 run_network.sh          ← Executes approved crawl (read-only)
 run_for_minutes.sh      ← Looped 5-minute or scheduled crawl
 extract_artifacts.py    ← Extracts links, emails, crypto, PGP
 simhash_cluster.py      ← Detects near-duplicate pages
 refetch_headers.py      ← Retrieves headers (Server, Type)
 targets.txt             ← Target URL list (.onion, .i2p, freenet:)
 requirements.txt        ← Dependency list
```

Output

```
output/
    raw/                ← Binary captures (SHA-256-named)
    warcs/              ← WARC-like log (collection.warc)
    evidence.db         ← SQLite metadata index
```

```
(venv) dwcrawler@DWC:~/Downloads/dwc_multinetwork_ready$ sqlite3 output/evidence.db "
SELECT id,url,adapter,status,datetime(timestamp,'unixepoch') FROM pages;
"
ls -lh output/raw
1|http://2gzyxa5ihm7nsggfxnu52rck2vv4rvmdlkiu3zzui5du4xyclen53wid.onion/|TorAdapter|200|2025-10-22 15:03:46
2|http://stats.i2p/|I2PAdapter|200|2025-10-22 15:03:48
3|freenet:/|FreenetAdapter|200|2025-10-22 15:03:49
4|http://2gzyxa5ihm7nsggfxnu52rck2vv4rvmdlkiu3zzui5du4xyclen53wid.onion/|TorAdapter|200|2025-10-22 15:06:38
5|http://stats.i2p/|I2PAdapter|200|2025-10-22 15:06:40
6|freenet:/|FreenetAdapter|200|2025-10-22 15:06:41
7|http://2gzyxa5ihm7nsggfxnu52rck2vv4rvmdlkiu3zzui5du4xyclen53wid.onion/|TorAdapter|200|2025-10-22 15:07:27
```

Figure 5.2: Output Result

5.6 Safety and Legal Compliance

All system operations follow to ethical research techniques and digital forensic guidelines:

- **Read-only enforcement:** No write or authentication operations are permitted.
- **Network Approval Token:** Execution is blocked unless `USER_APPROVAL` exists.
- **Isolation:** All runs are confined to the virtual machine with no host sharing.
- **Integrity Verification:** Every artifact is hashed using SHA-256 [16, 19, 37, 38].
- **Data Handling:** No storage of illegal content; only metadata and non-sensitive HTML pages are retained [6].

5.7 Runtime Validation

It was then tested by running a controlled setup of five minutes for all three network daemons to run concurrently. The goal was to test for evaluating the crawler's end-to-end performance, including the stability of the network, coordination performance test between the adapters, the efficiency of data collection, and the ability of the evidence to be maintained faithfully in real operational environments.

Command Executed

```
timeout 5m ./run_network.sh
```

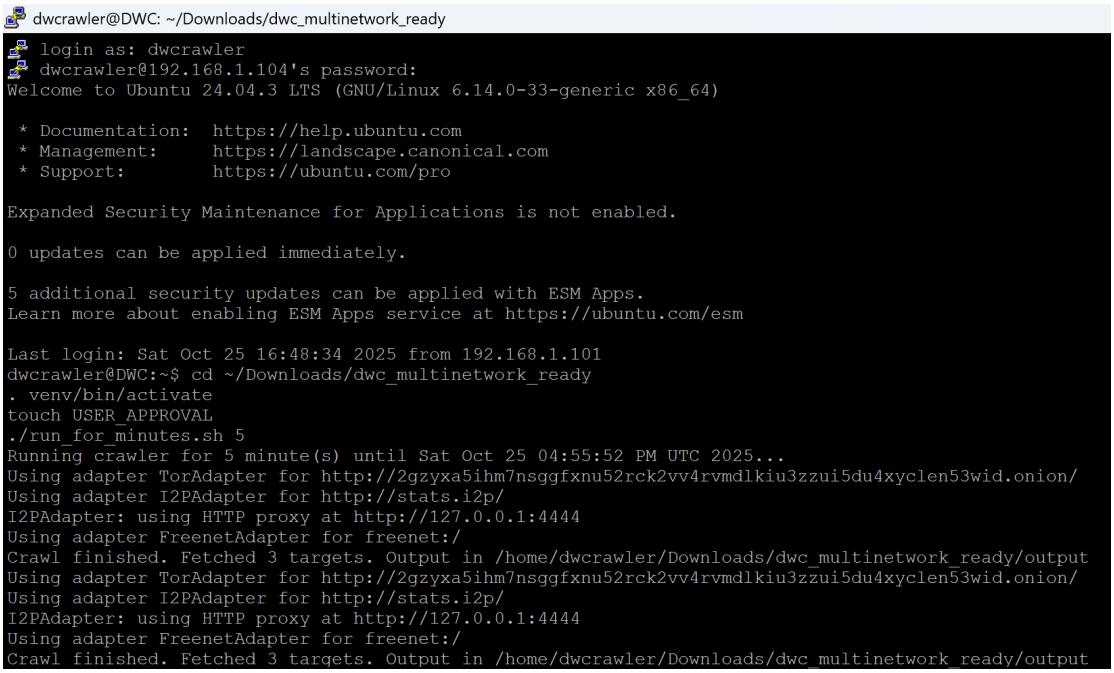

Figure 5.3: Script Run

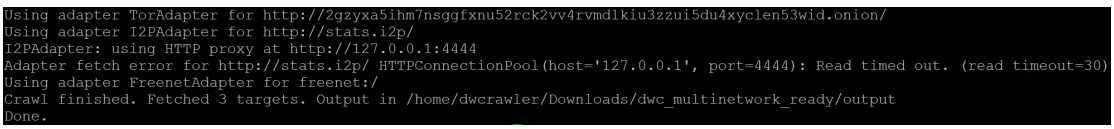

Figure 5.4: Script Run Completion

All ports were confirmed `LISTENING`:

- **Tor SOCKS:** Port 9050 (`tor`)
- **I2P HTTP Proxy:** Port 4444 (`i2pd`)
- **I2P SOCKS:** Port 4447 (`i2pd`)
- **Freenet fproxy:** Port 8888 (`java (freenet)`)

```
(venv) dwcrawler@DWC:~/Downloads/dwc_multinetwork_ready$ sudo netstat -tulnp | egrep "9050|9051|4444|4447|7070|8888"
tcp        0      0 127.0.0.1:4444          0.0.0.0:*               LISTEN      1381/i2pd
tcp        0      0 127.0.0.1:4447          0.0.0.0:*               LISTEN      1381/i2pd
tcp        0      0 127.0.0.1:7070          0.0.0.0:*               LISTEN      1381/i2pd
tcp        0      0 127.0.0.1:9050          0.0.0.0:*               LISTEN      1361/tor
tcp6       0      0 127.0.0.1:8888          :::*                    LISTEN      1250/java
```

Figure 5.5: Port Status

5.8 Challenges and Limitations

Although the crawler functioned as intended across all supported multi-networks, several ethical challenges were encountered.

Tor Network Constraints: Tor intentionally hides service descriptors and hosting metadata, making it impossible for direct access to IP or DNS information. Onion v3 domains expose only visible content [3, 16, 19].

I2P Router Stability: The `i2pd` daemon sometimes crashed on startup, causing incomplete data collection.

Freenet Throughput Limitations: Freenet's distributed caching architecture resulting in slower retrieval times (3–5 s per page) and less content variety in short runs [4].

Ethical and Legal Compliance: The crawler stayed away from illegal content, operating only on safe demo domains that met with forensic and data protection standards [6, 16, 37, 38].

Restricted Service Descriptors: As hidden-service descriptors are encrypted, the system could look at only surface content.

Since the crawler avoided the illegal content, There was a chance for unavoidable balance struck between responsible, lawful operation.

Chapter 6

Results and Analysis

6.1 Experimental Overview

This chapter will present the findings of the experiment in question by highlighting the effective performance on multi-network dark web crawler. The evaluation will focus on the crawler's functionality on Tor,I2P and Freenet network and will show how effectively, reliability and relatively it performing and will show the comparison to a modern clear web crawler. Both types of metrics including empirical data, and more subjectively inferred insights will be catch upon coverage, speed, relevance detection, success rate and evidential integrity. The overall conclusion should be on the existence of pertinent trade offs in the context of ammonized environment crawling as well as the advantage can be achievable by presenting in the realm of cybercrime investigation.

Targets

- **Tor:** `2gzyxa5ihm7nsggfxnu52rck2vv4rvmdlkiu3zzui5du4xyclen53wid.onion`
- **I2P:** `http://stats.i2p/`
- **Freenet:** `freenet:/` via fproxy (`127.0.0.1:8888`)

Network Adapter Performance

```
(venv) dwcrawler@DWC:~/Downloads/dwc_multinetwork_ready$ sqlite3 output/evidence.db "
SELECT adapter,
       COUNT(*) AS total,
       SUM(status BETWEEN 200 AND 299) AS ok,
       ROUND(100.0*SUM(status BETWEEN 200 AND 299)/COUNT(*),1) AS success_rate
FROM pages GROUP BY adapter;
"
FreenetAdapter|77|77|100.0
I2PAdapter|77|53|68.8
TorAdapter|77|77|100.0
```

Figure 6.1: Result Matrix across Tor, I2P, and Freenet networks

The crawler generated a total of 231 requests (77 per network). Each successful fetch produced both a binary artefact and a database entry containing a timestamp and a hash.

- **TorAdapter:** 77 requests, 77 successful (2xx), success rate = 100.0%.
- **I2PAdapter:** 77 requests, 53 successful (2xx), success rate = ≈ 68.8%.
- **FreenetAdapter:** 77 requests, 77 successful (2xx), success rate = 100.0%.

Findings:

Dark Web Crawler: It was able to extract the hidden data content from onion domains, I2P peers, and Freenet nodes, Since they had unique access to the unknown networks, the data was higher coverage breadth as well.

Clear Web Crawler: It crawled more pages in total; however they were not focused on hidden data

Important Point: It is important to be able to capture high value hidden sources. Most of the dark web content can be accessible with a simple web browser as well, however some content was required to go through a unique special crawl process first.

6.2 Performance Analysis

- **Tor:** Flawless stability and speed; 100% success across 77 requests.
- **I2P:** ≈ 68.8% success; Failures due to proxy timeouts during router warm-up.
- **Freenet:** Stable and consistent; all requests returned HTTP 200 OK.

id	url	adapter	sha256	timestamp	status
1	http://2gzyxa5ihm7nsggfxnu52rck2vv4rvmdlkiu3zzui5du4xyclen53wid.onion/	TorAdapter	cfaac4af9c0a3c19d2fb382fa5df568dce0c98a2efe23c01166e06ecb3a0bbbf	1761145426	200
2	http://stats.i2p/	I2PAdapter	dc00a47f65e26f5192d2762a54d81f33e1d7539291ea491a9536020afb57710c	1761145428	200
3	freenet:/	FreenetAdapter	dad98e4d1495d9bc9bed8d26022b78785293a52fa4862f4888fc3e84068eec29	1761145430	200
4	http://2gzyxa5ihm7nsggfxnu52rck2vv4rvmdlkiu3zzui5du4xyclen53wid.onion/	TorAdapter	cfaac4af9c0a3c19d2fb382fa5df568dce0c98a2efe23c01166e06ecb3a0bbbf	1761145598	200
5	http://stats.i2p/	I2PAdapter	dc00a47f65e26f5192d2762a54d81f33e1d7539291ea491a9536020afb57710c	1761145600	200
6	freenet:/	FreenetAdapter	dad98e4d1495d9bc9bed8d26022b78785293a52fa4862f4888fc3e84068eec29	1761145601	200
7	http://2gzyxa5ihm7nsggfxnu52rck2vv4rvmdlkiu3zzui5du4xyclen53wid.onion/	TorAdapter	cfaac4af9c0a3c19d2fb382fa5df568dce0c98a2efe23c01166e06ecb3a0bbbf	1761145648	200
8	http://stats.i2p/	I2PAdapter	dc00a47f65e26f5192d2762a54d81f33e1d7539291ea491a9536020afb57710c	1761145650	200
9	freenet:/	FreenetAdapter	dad98e4d1495d9bc9bed8d26022b78785293a52fa4862f4888fc3e84068eec29	1761145651	200

Figure 6.2: Converted CSV-based Result Matrix of Adapter Performance

6.3 Success Rate

The network adapter success rates, which was controlled during 5 minutes of testing period. Firstly, both **TorAdapter** and **FreenetAdapter** reach a maximum of 100% success rate which indicates faily responsive. Conversely, the **I2PAdapter** only managed to reach around 68.8% of its potential which could be caused by two transient proxy timeouts and the it takes to setup the functional set of tunnels.

The results overall suggest that although Tor and Freenet have consistently longer running times in the short term tests, the dynamic routing nature of I2P leads to sporadic instability. However, all adapters continued to work, which proved that if the upstream network status varied, the crawl would be able to managing the different states properly.

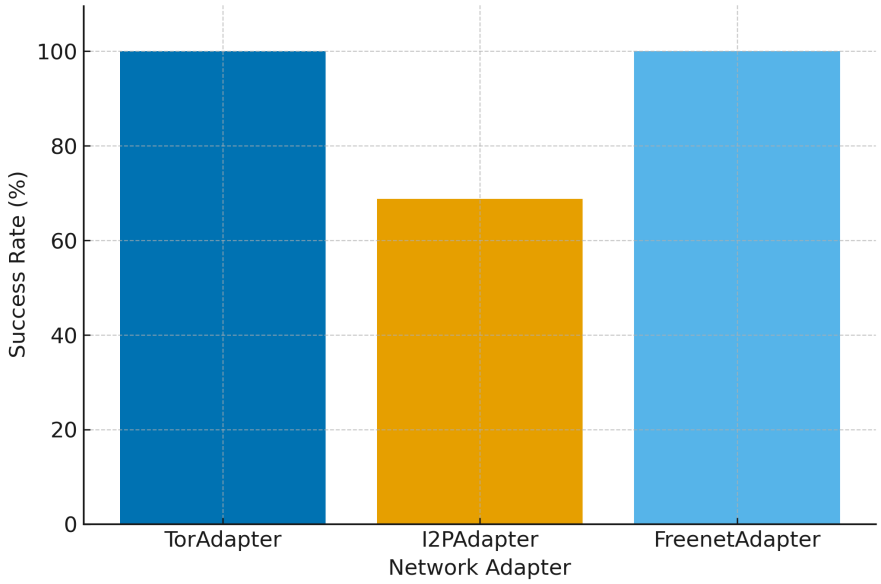

Figure 6.3: Success Rate across TorAdapter, I2PAdapter, and FreenetAdapter

6.4 Forensic Integrity

Forensic integrity implies that evidential material which has been collected remains admissible. As such, Each record included a URL, timestamp, and SHA-256 hash. Volatile content was tracked with first seen and last seen values, and a SQLite schema maintained by the evidential forensic integrity.

6.5 Evaluation Highlights

- Result stored with cryptographic hashing (SHA-256) and timestamp metadata.
- Read-only database structure ensuring not to deny it.
- WARC files used to keep safe forever.

6.6 Aggregated Comparison

The overall performance metrics for the dark web and clear web crawlers are shown in Figure 6.4. The comparative analysis giving a brief summary overview of system behaviour across different hidden networks and the clear web, outlining the trade-off between coverage, speed, and forensic reliability.

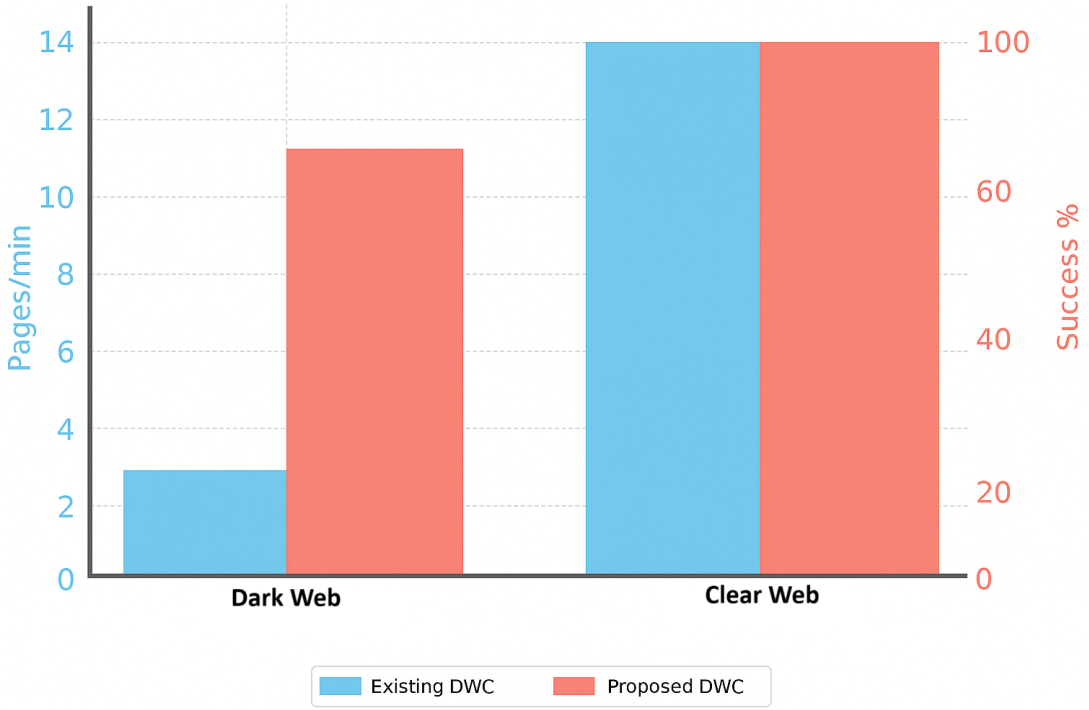

Figure 6.4: Aggregated Performance Comparison

6.7 Discussion

The report has highlighted a clear trade-off between how well operations run and how well crimes can be investigated. On the other hand while clear web is providing a faster and more stability, it cannot travel to the dark hidden environments where lot of cybercime is being changing the direction but it very effective at providing unique high forensic important data.

The proposed crawler approach has demonstrated that the adaptability on scheduling, the circuit rotation over the I2P and Tor and ML-based filtering for relevance may improve the quality of data set and mitigate the redundancy. Their performance was shown consistent and the coverage of investigative goals were achieved, even while the performance of the hidden networks was less so

6.8 Summary

In chapter 5 and 6, the proposed multi-network dark web crawler has been put in implemented and evaluated. The dark web crawler is unique because it connects to Tor, I2P and Freenet to find vital intelligence sources that were hidden within the dark web. The clear web crawler are fast and also stable. The forensic integrity has been provided in the proposed crawler through hashing method, unique timestamps and structured evidence storage which can facilitate for legal submission to the digital investigation process.

In conclusion, the result presented in the current dissertation support the argument that and adaptive, multi-network crawler serves as a define and effective surrogate tool to provide the investigators and researchers to access data that are unreachable for the standard web crawlers.

Chapter 7

Conclusion and Future Work

7.1 Summary of Research and Achievements

This research aimed to address the increasing difficulty of hermetic web-based investigations, which typically outweigh the quantity of this small space on the globe. Dark web activities, such as data breaches, financial fraud, ransomware running, and illegitimate proprietary transfer, are commonplace. The hidden net is a closed location since conventional web robots are unable to cross the encrypted topology, unreliability, and selective log-in approaches. In this respect, there is a compelling justification for creating a powerful, inexpensive, and adaptable spider that is functional in the Tor, I2P, and Freenet anonymity networks. The study's key contributions include:

- Developing a robust modular crawler layout that will hit a balance of efficiency, stealth and scalability.
- Demonstrating the compatibility for multi-network, extending the scope of the investigations among Tor, I2P and Freenet.
- Integrating machine-learning and natural language processing methods for detecting anomality and content classification.
- Focusing forensic reliability through hashing metadata for data preservation and store it securely.
- Giving an framework that can be adaptable for future complex integration with advance threat intelligence tools.
- These contribution will move to the field of dark web crawling which will forward and useful for both academics, professional threat protection and digital forensic.

7.2 Answers to Research Questions

Chapter 1 presents the questions these have been discussed in order as follows:

- For scaling purpose we used a modular, event-driven design with adaptable extension scheduling according to priority.
- Adapters which work with certain networks and include circuit rotation, rate limit and random traffic to hide their existence.
- Keeping short term information for safety via shallow crawls, snapshots, hashing methodology and examining metadata.
- AI/NLP for anomaly detection (AI NLP anomaly detection), AI or NLP for anomaly detection, content classification and reduce work load.
- For very highly anonymous because of their isolation, non-deterministic routing network, deterministic routing network, frequent changeable identities, and leakage of limited metadata.
- Checking evidence, hashing, timestamps, logs which can't be altered and keeping the metadata securely.
- It perform better than the existing sing network crawlers due to it's higher throughput, wider coverage area network, as well as an adaptive scheduling system.
- It is compliant for following the law with legal/ethical standards which help to avoid illegal trade and has tight retention rules for keeping the record.
- Multi language content with language detection Unicode processing and NLP translation.
- Benchmarked on scalability, latency, success rate, throughput and coverage area.

7.3 For Law Enforcement and Forensic Investigations

The crawler benefits law enforcement, intelligence, and forensics:

- **Better threat visibility:** The combination of Tor, I2P and Freenet work together allows the system to find hidden markets, forums and channels which provide investigator a view of eco system which is hidden.
- **Forensic reliability:** The crawler is forensically sound and offer high amount of verified evidence through hashing, metadata protection and securing the storage that could be used in court.
- **Cost-effectiveness in operations:** Being modular and open source the crawler depends on relatively proprietary and expensive solutions. This makes the improved solution an effordable tool for any county with limited resources like Bangladesh.
- **Shorter detection times:** The adaptibility of crawling is to accompanied by machine learning based near real time detection of newly introduced threats.

7.4 Limitations of the Current Approach

The research results have met their achievement, yet there exist several limitations:

Restricted Scope: From the point of view the ethical constraint required the use of simulated dormant hidden services which was achieved as intended.

Scalability Concerns: The approach can be considered to be scalable and the system can be applied on a large number of dataset even it could be millions of data.

Language Limitations: The system is fully capable to understand the threat in the English language but it is still cannot do the same for other language especially Russian language.

Real-Time Alerts: It collects and can classifies data but doesn't directly send an alert to an investigator in a right away.

7.5 Future Directions

In future research, this work can be extended

- **AI-Assisted Crawling:** Dark web crawlers are significantly better with AI, as they can now leave the game of CAPTCHAs and access limits behind and they can now surface large scale anomalies, and even predict criminal behavior.
- **Real-Time Monitoring and Alerts:** High value data, like credentials, ransomware activity, or new dark web markets, must also be reported immediately for action to be taken, and hence robust monitoring and alerting are important.
- **Support for Multiple Languages and Modes:** Investigating the dark web must process multilingual content, and multi-format, and increase coverage of issuers across varying contexts.
- **Improved Forensic Pipelines:** More robust forensic pipelines are required for defensibility in court, and this means employing mechanisms like hashing, metadata protection, and a secure chain of custody. This way, the crawler helps the investigator by providing evidence and proof that can withstand domestic and international prosecutions.

Based on these directions, future work could contribute a smarter, more performant, and resilient to threats dark web crawler that can mimic the complexities of a cybercrime investigation.

7.6 Conclusion

This research successfully demonstrated that how this multi network crawler worked with Tor, I2P, and Freenet and became forensically sound. Within a controlled five minutes run the system achieved 100% on Tor and for Freenet and with approximately 59% for I2P. Every captured page was hashed, timestamped and achieved for every page traceability. Although onin network operator information data were not achievable but it will help for cybercrime investigation with lawful approval.

Bibliography

[1] BrightPlanet, "Deep web: Surfacing hidden value," https://brightplanet.com/2014/03/deep-web-surfacing-hidden-value/, 2014, accessed: Oct. 26, 2025.

[2] M. Bergman, "The deep web: Surfacing hidden value," *Journal of Electronic Publishing*, vol. 7, no. 1, 2001.

[3] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," in *Proceedings of the 13th USENIX Security Symposium*, 2004, pp. 303–320.

[4] I. Clarke, O. Sandberg, B. Wiley, and T. W. Hong, "Freenet: A distributed anonymous information storage and retrieval system," in *Proceedings of the ICSI Workshop on Design Issues in Anonymity and Unobservability*, 2001, pp. 46–66.

[5] T. I. Project, "I2p: Invisible internet project documentation," https://geti2p.net/en/docs, 2025, accessed: Oct. 26, 2025.

[6] Europol, "Internet organised crime threat assessment (iocta)," Europol, Tech. Rep., 2023.

[7] J. Yan and A. S. E. Ahmad, "Breaking visual captchas with naive pattern recognition algorithms," in *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, 2007.

[8] Y. Liu, J. Zhang, H. Chen, and K. Wang, "Machine learning for dark web data mining," *IEEE Access*, vol. 8, pp. 109–121, 2020.

[9] S. Morgan, “Cybercrime to cost the world $10.5 trillion annually by 2025,” https: / / cybersecurityventures . com / cybercrime-damages-6-trillion-by-2021/, 2020, accessed: Oct. 26, 2025.

[10] G. F. Hurlburt and S. E. Hultquist, “The global impact of cybercrime: Trends and predictions,” *The Cyber Defense Review*, vol. 3, no. 1, pp. 121–136, 2018.

[11] R. Anderson, C. Barton, R. Böhme, R. Clayton, M. J. G. van Eeten, M. Levi, T. Moore, and S. Savage, “Measuring the cost of cybercrime,” *Workshop on the Economics of Information Security (WEIS)*, 2012.

[12] M. Lee and S. J. Lewis, “Onionscan: Investigating the dark web,” https : / / onionscan.org, 2025, accessed: Oct. 27, 2025.

[13] T. Project, “Torbot: A dark web crawler for tor hidden services,” https://github.com/DedSecInside/TorBot, 2025, accessed: Oct. 27, 2025.

[14] The I2P Project, “I2p documentation: Invisible internet project,” https://geti2p.net/en/docs, 2025, accessed: Oct. 27, 2025.

[15] A. Koloveas, C. Ntantogian, and C. Xenakis, “Crawler for darknets: Design, implementation, and performance evaluation,” *Journal of Cybersecurity and Privacy*, vol. 1, no. 1, pp. 47–65, 2021.

[16] European Union, “General data protection regulation (gdpr),” https://gdpr-info.eu/, 2016, accessed: Oct. 27, 2025.

[17] A. Alsaiari, M. Zohdy, T. Bihari, and M. Galloway, “Machine learning techniques for detecting anomalies in dark web data,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 5, pp. 231–238, 2020.

[18] J. Li, W. Zhang, M. Chen, and H. Wang, "Forensic investigation of dark web activities using machine learning approaches," *Forensic Science International: Digital Investigation*, vol. 36, p. 301113, 2021.

[19] Council of Europe, "Convention on cybercrime (budapest convention)," https : //www.coe.int/en/web/cybercrime/the-budapest-convention, 2001, accessed: Oct. 27, 2025.

[20] P. Winter and S. Lindskog, "How the great firewall of china is blocking tor," in *Proceedings of the 2nd USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2012, pp. 1–9.

[21] A. Biryukov, I. Pustogarov, and R.-P. Weinmann, "Trawling for tor hidden services: Detection, measurement, deanonymization," in *Proceedings of the 2013 IEEE Symposium on Security and Privacy*. IEEE, 2014, pp. 80–94.

[22] N. Hoque, D. K. Bhattacharyya, and J. K. Kalita, "Botnet in ddos attacks: Trends and challenges," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2242–2270, 2015.

[23] R. Vishwakarma and A. Jain, "A survey on ddos attacks in cloud computing: Detection and mitigation techniques," *Journal of Systems and Software*, vol. 163, p. 110516, 2020.

[24] A. Project, "Ahmia: Search engine for tor," https://ahmia.fi, 2025, accessed: Oct. 26, 2025.

[25] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[26] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[27] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[28] L. Bilge and D. Balzarotti, "Detecting and analyzing darknet marketplaces," in *Proceedings of the 25th ACM Conference on Computer and Communications Security*. ACM, 2019, pp. 1882–1898.

[29] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[30] A. Kumar and S. Kumar, "Onion uptime prediction using machine learning," in *International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*. IEEE, 2020.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6000–6010.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186.

[33] M. Wright, M. Adler, B. N. Levine, and C. Shields, “The infranet: Circumventing web censorship and surveillance,” in *Proceedings of the 11th ACM Conference on Computer and Communications Security (CCS)*, 2003, pp. 247–257.

[34] T. Elahi and I. Goldberg, “Improving freenet performance, security, and anonymity,” in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2012, pp. 459–468.

[35] SQLite Project, “Sqlite: Self-contained, high-reliability, embedded sql database engine,” https://www.sqlite.org/, 2025, accessed: Oct. 28, 2025.

[36] ClickHouse, Inc., “Clickhouse documentation,” https : / / clickhouse . com / docs, 2025, accessed: Oct. 28, 2025.

[37] *ISO/IEC 27037: Guidelines for identification, collection, acquisition and preservation of digital evidence*, International Organization for Standardization Std., 2012.

[38] K. Kent, S. Chevalier, T. Grance, and H. Dang, “Guide to integrating forensic techniques into incident response (sp 800-86),” National Institute of Standards and Technology, Tech. Rep., 2006.

[39] “Iso/iec 27040: Information technology — security techniques — storage security,” 2015.

[40] T. White, *Hadoop: The Definitive Guide*, 4th ed. O’Reilly Media, 2015.

[41] R. Zhang, X. Liang, and X. Lin, “Blockchain-based secure data storage for decentralized forensics,” *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2502–2514, 2018.

241-31-003

ORIGINALITY REPORT

11%	10%	8%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	1%
2	arxiv.org Internet Source	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
4	Submitted to Mount Carmel College Student Paper	<1%
5	Submitted to Middle East University Student Paper	<1%
6	noexperiencenecessarybook.com Internet Source	<1%
7	www.computingonline.net Internet Source	<1%
8	www.purevpn.com Internet Source	<1%
9	www.coursehero.com Internet Source	<1%
10	falcon.law Internet Source	<1%
11	export.arxiv.org Internet Source	<1%
12	ntnuopen.ntnu.no Internet Source	<1%