MIT | Academy of Engineering

School of Computer Engineering

Presentation for TY Mini Project (Review 1)

AI Web Crawler

Students Name:

Dheeraj Singh (97)

Amey Satone(31)

Viraj raina (18)

Ayush Koul (02)

Guide:

Mrs Neha Hajare

Index

- Introduction
- Literature survey (in short)
- Problem Statement and Objectives
- Proposed Block Diagram and it explanation
- References

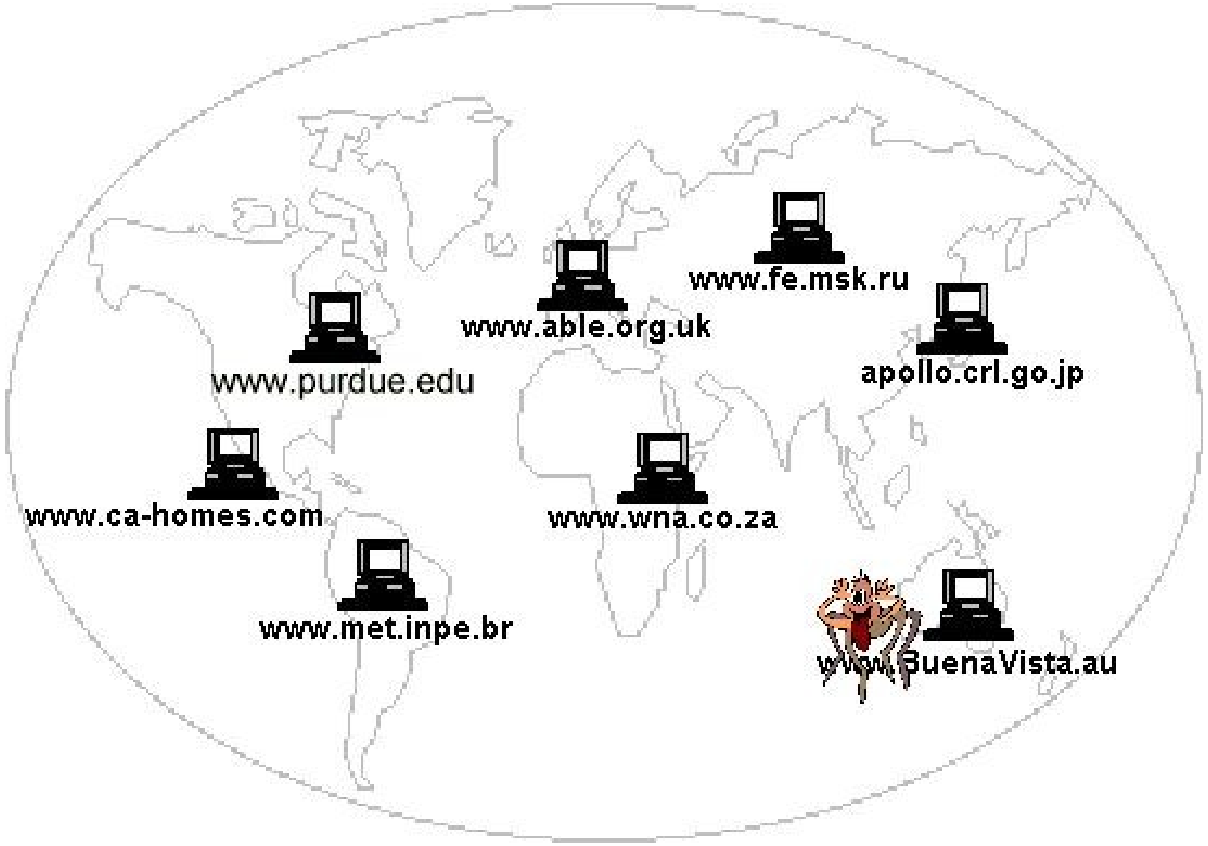
Introduction

Addressing Online Threats:

Tackling illegal activities, including weapons dealing, on the deep and dark web.

Advanced Technology:

Utilizing web crawling and NLP for effective monitoring and analysis.



Lit Survey

1. "Mining the Dark Web: Drugs and Fake IDs" (2016)

Authors: A. Baravalle, M. S. Lopez, and S. W. Lee

Summary: This paper explores the challenges and methodologies involved in mining the dark web for information related to drugs and fake IDs.

It emphasizes the need for specialized web crawlers to collect data from dark web marketplaces and forums where illegal activities often occur.

The research discusses techniques for content extraction and data analysis, shedding light on the role of web crawling in monitoring illicit activities on the dark web.

Continue..

2. "Exploring the Topological Properties of the Tor Dark Web" (2021)

Authors: A. Alharbi, M. Faizan, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan

Summary: This recent paper investigates the topological properties of the Tor dark web, focusing on network structure, connectivity, and website interconnections.

The authors use network analysis techniques to study the behavior of websites hosted on the Tor network.

This research provides valuable insights into the structural aspects of the dark web, which can be essential for security and research purposes.

Continue..

3. "Phishing with a Darknet: Imitation of Onion Services" (2020)

Authors: F. Barr-Smith and J. Wright

Summary: This paper discusses the use of darknet websites (onion services) for phishing attacks and explores the imitation of onion services.

It examines the challenges associated with detecting and preventing phishing activities on the dark web.

The authors propose strategies for addressing this issue, including content analysis and user education. The research highlights the importance of web crawlers and monitoring tools in identifying malicious activities in hidden parts of the web.

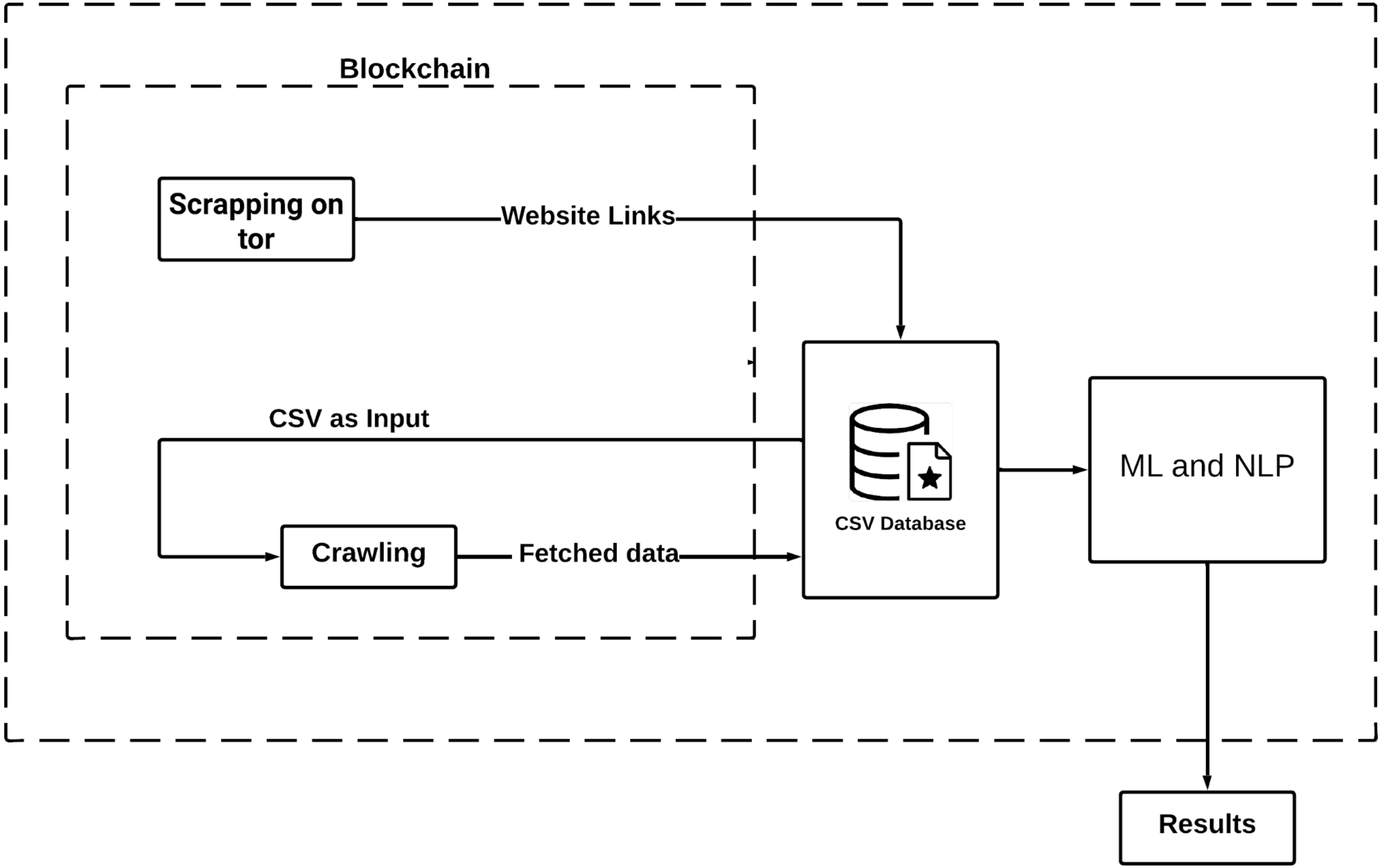
Problem statement

- AI Web Crawler For Dark Web

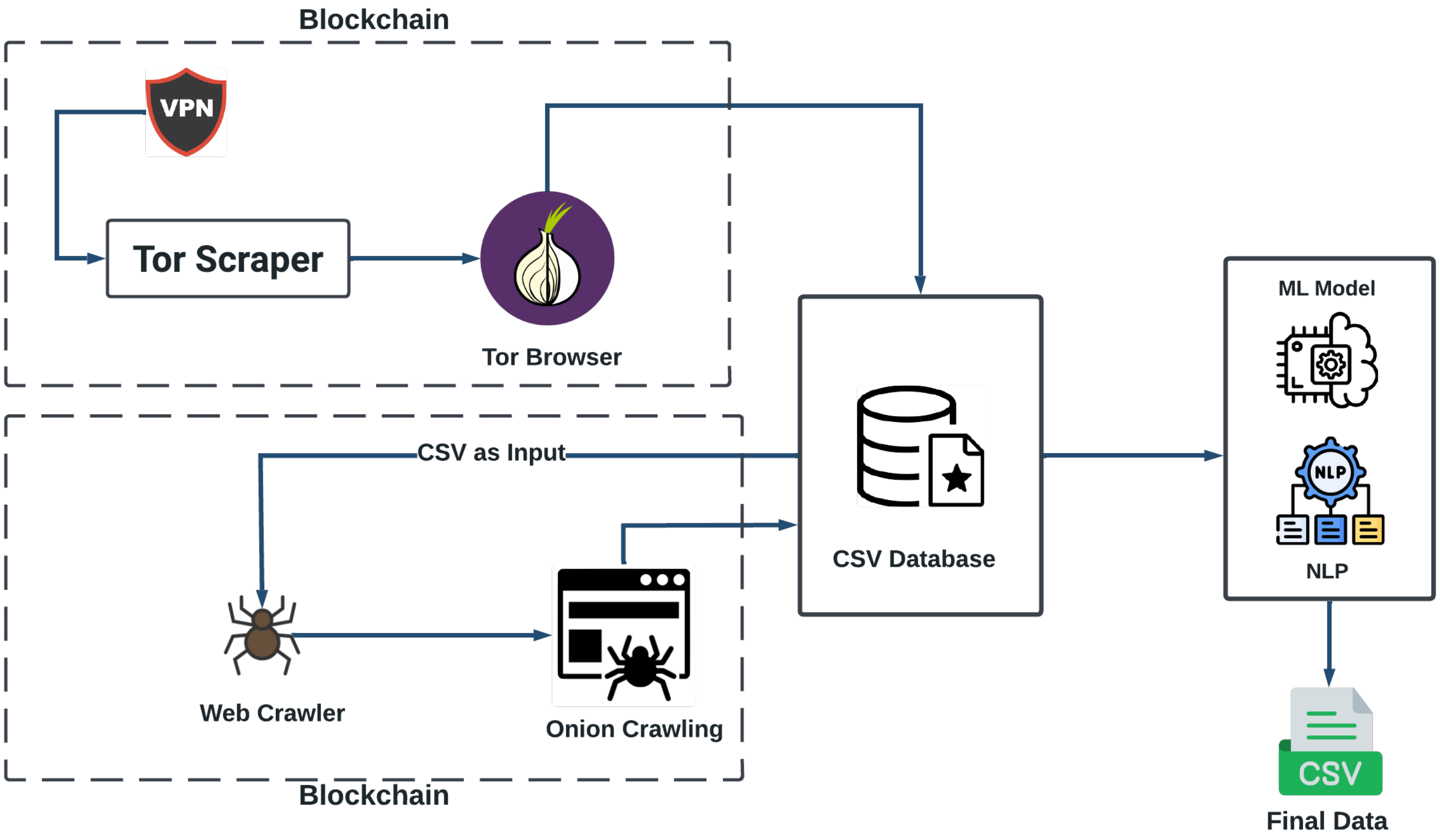
Objectives to be achieved

1. **Illegal Activity Detection**
2. **Data Privacy**
3. **Keyword Analysis**

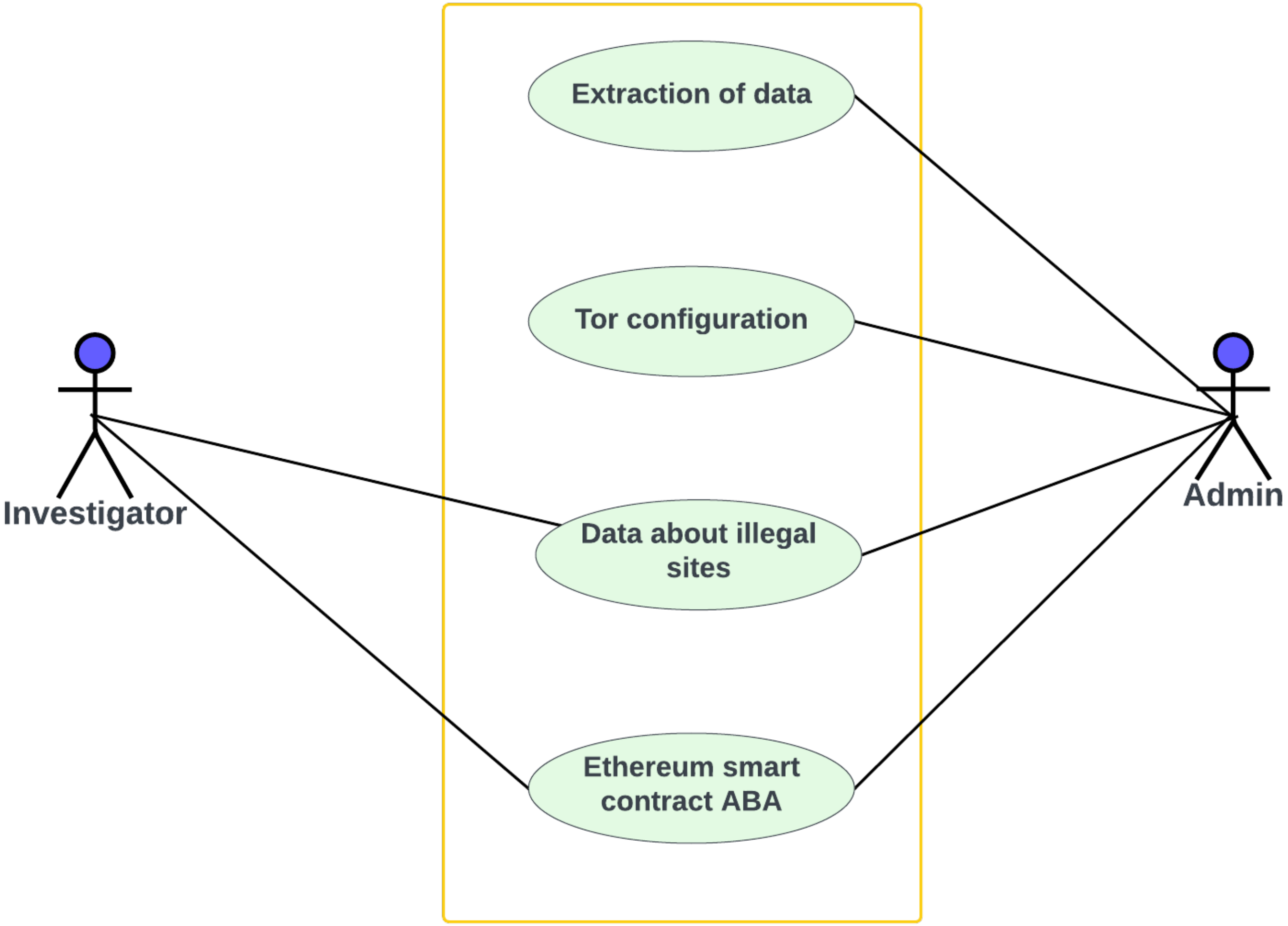
Proposed Block System



System Architecture



Proposed Use Case Diagram



Methodology

- **Data Collection and Web Crawling**

  Collect data from deep and dark web sources using web crawling.

- **Data Preprocessing**

  Clean and prepare data for analysis.

- **NLP Analysis**

  Analyze text data using NLP techniques like BERT.

- **Machine Learning**

  An AI solution could incorporate machine learning algorithms to learn past incidents.

Mathematical Modelling

Regular Expression= ” \W +\.onion”

This regular expression is designed to find strings in the content variable that match the pattern of a typical ”.onion” domain on the dark web. Let’s break down the components of this regular expression:

1. \W + : Matches one or more word characters (alphanumeric characters plus underscore).

2. ˙: Matches a literal dot character.

3. onion: Matches the literal string ”onion”.

Combined, this regular expression looks for sequences of word characters followed by a dot and the string ”onion,” which is a common pattern for dark web domain names.

Selected References

1G. Acar and M. Juarez. (2020). Individual Contributors. TOR-BrowserSelenium TOR Browser Automation With Selenium. [Online]. Available: https://github.com/webfp/tor-browser-selenium

2J. Aldridge and D. Décary-Hétu, "Hidden wholesale: The drug diffusing capacity of online drug cryptomarkets," Int. J. Drug Policy, vol. 35, pp. 7–15, Sep. 2016, doi: 10.1016/j.drugpo.2016.04.020.

3A. Alharbi, M. Faizan, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan, "Exploring the topological properties of the Tor dark web," IEEE Access, vol. 9, pp. 21746–21758, 2021, doi: 10.1109/access.2021.3055532.

4K. Avrachenkov, B. Ribeiro, and J. K. Sreedharan, "Inference in OSNs via lightweight partial crawls," ACM SIGMETRICS Perform. Eval. Rev., vol. 44, no. 1, pp. 165–177, Jun. 2016, doi: 10.1145/2964791. 2901477.

5A. Baravalle, M. S. Lopez, and S. W. Lee, "Mining the dark web: Drugs and fake ids," in Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW), Dec. 2016, pp. 350–356, doi: 10.1109/ICDMW.2016. 0056.

Continue…

6F. Barr-Smith and J. Wright, "Phishing with a darknet: Imitation of onion services," in Proc. APWG Symp. Electron. Crime Res. (eCrime), Nov. 2020, pp. 1–13, doi: 10.1109/ecrime51433.2020.9493262.

7J. Bergman and O. B. Popov, "The digital detective's discourse—A toolset for forensically sound collaborative dark web content annotation and collection," J. Digit. Forensics, Secur. Law, vol. 17, no. 5, pp. 1–25, doi: 10.15394/jdfsl.2022.1740.

8M. Bouchard, K. Joffres, and R. Frank, "Preliminary analytical considerations in designing a terrorism and extremism online network extractor," in Computational Models of Complex Systems. Cham, Switzerland: Springer, 2014, pp. 171–184, doi: 10.1007/978-3-319- 01285-811.

Thank You!!

Any questions.