Unravelling the Information Asymmetry in Threat Intelligence

Daniel Safavi-Zadeh
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
d.safavi-zadeh@student.utwente.nl

ABSTRACT

In the digital era, the cyber threat landscape has evolved significantly, posing a major concern for stakeholders such as governmental agencies, law enforcement, and companies. The understanding of the prevalence, spread, and trends of these cyber-attacks is of paramount importance. However, the primary source of insights into cybercrime numbers often stems from private entities. This leads to potential information asymmetry, with the public sector relying on data provided by private companies to comprehend current trends and threats. Furthermore, the frequent referencing of each other's threat reports by private entities complicates the assessment of the original information source. This research addresses the information imbalance in cyber threat intelligence by experimenting with the collection and analysis of threat data from various government reports. The study poses two research questions: (1) To what extent can the extraction of sources from Cyber Threat Intelligence (CTI) reports be automated? (2) How can the methodologies used in Research Question 1 (RQ1) be applied to identify the organizations that significantly contribute to providing threat references from Cyber Intelligence Data resources, such as threat reports or blog posts, and to determine if there are indications of an oligopoly? The study offers valuable insights into the fields of Cyber Threat Intelligence, reference extraction methodologies, and the understanding of information asymmetry. The identified patterns and observations provide a foundation for future research, guiding the refinement of CTI threat reference extraction practices and fostering a more comprehensive understanding of the cyber threat landscape. The study's findings reveal the most influential sources of threat intelligence, including Mitre Corporation, CISA, and ENISA, potentially indicating an oligopoly in the industry which still has to be investigated.

Keywords

Cyber Threat Intelligence, Oligopoly, Information Asymmetry, Threat Reports, Information Retrieval, Automated Web Search.

1. INTRODUCTION

In the era of digitalization, the landscape of threats and risks has significantly evolved, with cyber threats becoming a prominent concern for various stakeholders, including governmental agencies, law enforcement, and private companies [1]. The understanding and management of these threats require comprehensive and accurate information, often derived from cyber threat intelligence (CTI) reports [3]. However, a significant challenge in this domain is the potential information asymmetry that arises due to the reliance on private entities for insights into cybercrime trends and threats [1]. Private entities often generate threat reports that reference each other, creating a complex web of information that can be difficult to trace back to the original source [1]. This lack of transparency and traceability can hinder the assessment of the severity of information asymmetry and the determination of necessary countermeasures [1]. Moreover, the increasing complexity of network attacks necessitates an active defense strategy based on intelligence sharing [3]. In this research, we will delve deeper into the related work in this field, outline our proposed methodologies for addressing the research questions, and discuss our findings after implementing and evaluating this research. A key part of our methodology involves analyzing government reports such as the "ENISA Threat Landscape 2023"[4]. Published by the European Union Agency for Cybersecurity (ENISA), these reports are particularly suitable for our research as they provide comprehensive and authoritative insights into current cyber threats and trends. The main objectives of this research are to create an overview of sources cited in these reports, assess their influence, and trace the origin of the initial information. This information will aid in assessing the severity of the information asymmetry and the necessity for further investigation into an oligopoly.

1.1. Research Questions

This research aims to achieve two primary objectives: first, to develop a reliable methodology for extracting and evaluating threat information from various sources, and second, to investigate the most influential sources of threat intelligence to decide if a search for signs of oligopoly in the field is needed. To achieve our research objectives, we will formulate the following research questions (RQ) to guide our investigation:

RQ1: To what extent can we automate the extraction of sources from Cyber Threat Intelligence (CTI) reports?

Answering RQ1: To answer this question, we will design a methodology that can efficiently and accurately extract references from Cyber Intelligence Data resources such as threat reports and blog posts, thereby contributing to the automation and scalability of Cyber Threat Intelligence operations. A method for automatically extracting threat actions from unstructured CTI reports using multimodal learning by Zhang et al. [2] uses a combination of natural language processing techniques and machine learning algorithms to extract threat actions from unstructured CTI reports. The method involves several steps, including data preprocessing, feature extraction, and classification. The authors use part-of-speech tagging and semantic analysis to identify relevant features in the text, such as verbs and noun phrases, and then use a machine learning algorithm to classify these features as threat actions or non-threat actions. The proposed method is evaluated using a dataset of CTI reports, and the results show that it can achieve a certain balance between accuracy and information completeness in the action extraction. However, the authors also note some limitations of the method, such as overreliance on part-of-speech tagging and failure to recognize pronoun referents. The main objective of this research is to develop a methodology more limited in scope for extracting and evaluating threat information from various sources, specifically focusing on the extraction of sources from CTI reports in PDF format. This is a departure from previous methods such as the multimodal learning approach by Zhang et al. [2], which focused on extracting threat actions from unstructured CTI reports.

RQ2: How can the methodologies used in Research Question 1 (RQ1) be applied to identify the organizations that significantly contribute to providing references from Cyber Intelligence Data resources, such as threat reports or blog posts, and to determine if there are indications of an oligopoly?

Answering RQ2: Therefore, the second goal of RQ2 is to identify the most influential sources of threat intelligence and assess the degree of information asymmetry in threat intelligence products market. This will involve methodologies employed in answering Research Question 1 (RQ1) followed by a detailed analysis of the sources cited in these reports and an assessment of their influence and signs of oligopoly. While there is limited research on the oligopoly in the threat intelligence products market, a recent paper by Conyon et al. (2022) [5] explored the oligopolistic nature of the private sector and the need for better regulation and inclusive industrial policy. However, this paper did not provide a comprehensive analysis of the threat intelligence products market or identify the most influential sources of threat intelligence. Nonetheless, this paper can still be useful in assessing the degree of oligopoly in the threat intelligence products market by providing a basis for comparison if evidence of oligopoly is found later on. By achieving this, the research provides valuable insights into the most influential sources of threat intelligence and assesses the severity of information asymmetry in the field. This will aid in understanding whether evidence of oligopoly should be investigated in further research.

1.2. Contributions

This research paper aims to achieve the following contributions:

- Provide insights into the prevalence of information imbalance in cyber threat intelligence, focusing on the asymmetry between public institutions and private entities. The paper highlights the challenges posed by disparate access to information between attackers and defenders, emphasizing the complexities faced by organizations in collecting, analyzing, and interpreting cyber threat data.
- Explore methodologies for automating the extraction of sources from Cyber Threat Intelligence (CTI) reports, aiming to enhance the efficiency and accuracy of threat intelligence operations. The study delves into the development of structured, machine-readable formats for processing vast amounts of CTI data, emphasizing the importance of taxonomies, sharing standards, and ontologies in deriving actionable intelligence from threat information.
- Investigate the most influential sources of threat intelligence and assess the degree of information asymmetry in the threat intelligence products market. By analyzing government reports and identifying key sources cited in CTI reports, the research aims to provide valuable insights into the severity of information imbalance and potential indications of an oligopoly within the industry.

These contributions collectively contribute to a deeper understanding of information asymmetry in cyber threat intelligence, offer practical insights for enhancing threat intelligence operations, and pave the way for further research into the dynamics of the threat intelligence landscape.

2. RELATED WORK

(I) Cyber Threat Intelligence (CTI) field:

Cyber Threat Intelligence (CTI) is a crucial resource in identifying and combating cybersecurity threats. CTI is defined by Wlosinski (2021) [6] as evidence-based knowledge about adversary motives, intents, capabilities, enabling environments, and operations. It is used by various industries, including governments, financial services, banking, insurance, retail companies, ecommerce, healthcare, manufacturing, telecommunication, and energy enterprises [6]. CTI is a proactive extension to incident response, leveraging the output from existing cybersecurity monitoring tools to prepare for, prevent, and identify cybersecurity threats that are trying to take advantage of valuable data [6]. Despite the importance of CTI, there are gaps in the existing literature regarding the standard practices employed by public and private institutions for the collection and publication of CTI. This gap is significant because it limits the ability of organizations to effectively leverage CTI to combat cybersecurity threats. One reason for this gap is the limited resources available to both government and private sector organizations, which restricts their ability to collect a comprehensive set of foreign-based malicious cyber activity [7].

(II) Information Asymmetry in CTI: A Focus on Public and Private Entities

It is crucial to note that in this paper the concept of "information asymmetry" in the context of Cyber Threat Intelligence (CTI) bears a distinct meaning compared to its usage in literature. While literature often refers to information asymmetry between attackers and defenders, this paper specifically addresses the information asymmetry between public institutions and private entities. As noted by Conti et al. [8], CTI involves the collection, analysis, and interpretation of data related to cyber-attacks in order to detect and defend against them. However, this process is complicated by the fact that attackers and defenders often have disparate access to information. Attackers may have access to a wide range of tools and techniques that are not publicly known, while defenders must rely on publicly available information and their own expertise to identify and respond to threats. This information asymmetry can create significant challenges for both public and private institutions that are responsible for collecting and publishing CTI. For example, it may be difficult to determine which sources of information are reliable and which are not, or to identify emerging threats before they become widespread. Additionally, there may be concerns about sharing sensitive information with other organizations or with the public, particularly if doing so could compromise ongoing investigations or reveal vulnerabilities in existing security measures.

(III) The Importance of Structured, Machine-Readable Formats in CTI Reports:

Vasileios et al. [9] evaluated the effectiveness of various taxonomies, sharing standards, and ontologies in the field of cyber threat intelligence. The study found that structured, machine-readable formats such as ontologies, schemas, and taxonomies are required to process, correlate, and analyze vast amounts of threat information and data and derive intelligence that can be shared and consumed in meaningful times. Based on their findings [9], structured, machine-readable formats are essential for processing, correlating, and analyzing vast amounts of cyber threat intelligence (CTI) data. By leveraging existing methodologies and tools for CTI collection and publication, we can develop a more efficient and accurate methodology for extracting references from CTI data resources such as threat reports and blog posts. This research provides valuable insights into the use of structured CTI data and its potential to improve prevention, detection, and response capabilities.

(IV) The Role of STIX in Standardizing CTI Information and Enhancing Data Processing and Analysis

The paper "Cyber Threat Intelligence: Challenges and Opportunities" by Conti, Dehghantanha, and Dargahi [8] suggests that the use of Structured Threat Information eXpression (STIX) in standardizing CTI information can be an effective practice for facilitating the extraction of references from Cyber Intelligence Data resources. STIX is a language developed by the MITRE Corporation to standardize the representation and exchange of CTI. STIX provides a common language for describing cyber threat indicators, tactics, techniques, and procedures (TTPs), and other relevant information about cyber-attacks. STIX is designed to be machine-readable, which enables automated processing and analysis of CTI data. STIX has been adopted by various public and private institutions, including the US Department of Homeland Security (DHS), the National Institute of Standards and Technology (NIST), and the Cyber Threat Alliance (CTA). By adopting STIX as a standard language for describing cyber threat indicators and TTPs, public and private institutions can improve the efficiency and accuracy of their CTI data processing and analysis, which can help in identifying relevant references and extracting them more effectively.

(V) Enhancing Reference Extraction from CTI Reports: Standardization and Sophisticated Analytics:

The standard practices employed by public and private institutions for the collection and publication of CTI can be utilized to design a more efficient and accurate methodology for extracting references from Cyber Intelligence Data resources such as threat reports or blog posts. One approach is to standardize the format and content of threat reports. As noted by (Samtani et al., 2018, pp. 4) [10] "intelligence communication standards (e.g., STIX)" can be used to facilitate the dissemination of CTI. Another approach is to develop more sophisticated analytics tools for analyzing CTI data resources "well-refined analytics such as malware analysis, event correlation, and forensics are utilized to derive the intelligence needed for CTI professionals to deploy appropriate security controls" (Samtani et al., 2018, pp. 4) [10].

(VI) (VI) Emerging Trends in CTI Landscape: AI, OSINT, and Unanswered Questions

Emerging trends in CTI include the use of machine learning and artificial intelligence (AI) to improve the accuracy and speed of threat detection and response. These technologies have the potential to address the issue of information asymmetry in the field of CTI by automating the analysis of large volumes of data from various sources. Additionally, the use of open-source intelligence (OSINT) and social media monitoring has become increasingly popular in recent years. However, despite these advancements, the question of which sources of CTI are the most influential and whether there is evidence of oligopoly in the industry remains unanswered. Further research is needed to explore these issues and their implications for CTI professionals and organizations (Samtani et al., 2018, pp. 4,20) [10].

3. METHODOLOGY

This section outlines the methodology adopted to address the research question. A comprehensive overview is offered for both dataset creation and the construction of the dataset processing pipeline used for threat reference extraction.

3.1. Dataset Creation

A comprehensive dataset comprised of a series of fifty-one cybersecurity threat intelligence (CTI) reports was gathered from the official websites of prominent governmental cybersecurity agencies. A detailed overview of the collected threat reports is offered in a Table 3 of the Appendix section (9). The dataset was carefully selected encompass a wide range of countries and timelines. All collected reports are in the Portable Document Format (PDF), to ensure consistency and streamline both the extraction process and subsequent visual inspection testing. Next, we are discussing the rationale behind selection criterion of the datasets and grouping in subsets. The grouping was made in practice with the "Filtering Step" of the pipeline. The steps included in the dataset processing pipeline will be detailed in the "System Overview" section (3.2) where also Figure 1 is provided, depicting an overview of the process.

The gathered dataset of threat reports was grouped in three different subsets by origin of the authors and composition:

(i) ENISA Threat Landscape Reports (2012-2023): A set comprised of eleven annual reports (ID range 1-11 in Table 3 of appendix) from the European Union Agency for Cybersecurity (ENISA), spanning from 2012 to 2023. The 2019 and 2020 editions were consolidated in one edition by the author. These annual reports provide a comprehensive overview of the cyber threat landscape, focusing on the most prevalent threats and emerging trends. They cover a wide range of countries in the European Union over a significant timeline (eleven years), offering a broad perspective on global cybersecurity issues.

(ii) Joint Cybersecurity Advisories (CSAs) exclusively involving U.S. Agencies:

A collection of twenty-one distinct joint cybersecurity advisories (CSAs) [11] authored by prominent U.S. institutions (ID range 11-32 in Table 3 of appendix), including the Cybersecurity and Infrastructure Security Agency (CISA) [12], the National Security Agency (NSA) [13], the U.S. Federal Bureau of Investigation (FBI) [14] and the Multi-State Information Sharing and Analysis Center (MS-ISAC) [15]. These advisories reflect the U.S.'s internal efforts to combat cybersecurity threats.

(iii) Joint Cybersecurity Advisories (CSAs) authored by U.S. Agencies with Collaborations:

Contains fourteen joint cybersecurity advisories (CSAs) featuring between U.S. agencies mentioned also above and counterpart agencies from multiple countries (ID range 34-49 in Table 3 of appendix), including members of the Five Eyes Alliance [16] and others like Poland, Israel, Japan, and Norway. These advisories represent a collaborative effort between the U.S. and other countries to address global cybersecurity threats.

These three subsets were also combined into a comprehensive dataset for separate analysis, offering an overview of the entire dataset. Table 1 provides a detailed overview of the final datasets and obtained results:

Dataset	Number of threat reports	Number of pages	Number of (unique) extracted URLs	Number of (unique) domain names
ENISA	11	1070	4108	1832
Joint CSAs (only US)	21	335	746	99
Joint CSAs (multiple countries)	15	262	696	120
Combined	47	1667	5352	1945

Table 1 Dataset grouping results.

It is notable that three samples were excluded from analysis. Two of these were omitted due to their lack of alignment with other samples; they were specifically reserved for testing purposes. Another reason for their exclusion was our consideration that most of the samples in 2nd and 3rd datasets are joint Cybersecurity Advisories (CSAs) from North American agencies collaborating with others from different countries. All CSAs were obtained from a single source, the CISA website for tools and resources [17], and they represent joint cybersecurity advisories led by the United States, as they were published by U.S. entities. However, these two were identified as CTI reports authored by national agencies from France (ID 49 in Table 3 of appendix) and the UK (ID 50 in Table 3 of appendix). The third sample, a CSA report (ID 33 in Table 3 of appendix), was also excluded during the “Filtering Step” because it contained broken hyperlinks that rendered them inaccessible and was not offering any new information after the extraction was completed.

3.2. System overview

This section describes the pdf link extraction methods used in the dataset processing pipeline, including the quantitative research design, data collection techniques, and data analysis methods.

Figure 1 illustrates the comprehensive data processing pipeline employed in this study. The pipeline encompasses various stages, starting from “Data Extraction” step, followed by a cleaning and labeling procedures. Subsequently, the pre-processed data can undergo an intermediary step of “Data Collection” or lead to its ultimate culmination in visualization and then analysis. The “Data Collection” step (Figure 1) serves as intermediary phase for understanding the origin of the initial information collected in pre-processing. Although we only tested in one iteration due to time constraints, the intermediary phase could be applied iteratively multiple times to improve the understanding and refinement of the collected data. Between each step, and for every PDF file, the resulting Pandas “DataFrames” are converted and saved in CSV format using Pandas built in functions for facilitating visual inspection and portability between all steps of research

The study was conducted using Python programming language [18] along with Jupyter Notebook environment [19] combined with Numpy[20], Pandas[21] and Matplotlib[22] libraries.

Numpy numerical operations in conjunction with Pandas for data analysis and manipulation on “DataFrames” provide an efficient and powerful toolkit which supported all steps of the study. Matplotlib was used selectively in the visualization step for generation and export of graphs and pie charts. Each separate step is explained in detail with its own section. The detailed setup and configuration information, including code and configuration details for the computer used in this research, can be accessed on GitHub [46].

3.2.1 Extraction Step

The core part of our data processing pipeline is the URL extraction step, executed using a combination of three Python libraries: PyMuPDF[23], PDFMiner[24], and URLExtract[25]. These tools were selected for their efficient parsing and accurate text extraction capabilities, their ease of integration with Python, and their comprehensive support for complex PDF structures. PyMuPDF is used for data extraction, analysis, conversion, and manipulation of PDF documents. PDFMiner, on the other hand, is utilized for text extraction from PDF documents, particularly through its extract_text() function. Lastly, URLExtract is employed for collecting URLs from given text based on locating TLD. The extraction process commences with the conversion of threat reports into plain text using PDFMiner’s extract_text() function. This is followed by reformatting and removal of new line spaces and escape characters to enhance the accuracy of URL detection, especially for URLs split across multiple rows. To further improve URL detection, we use a carefully constructed regex pattern in conjunction with the URLExtract method. The regex pattern is

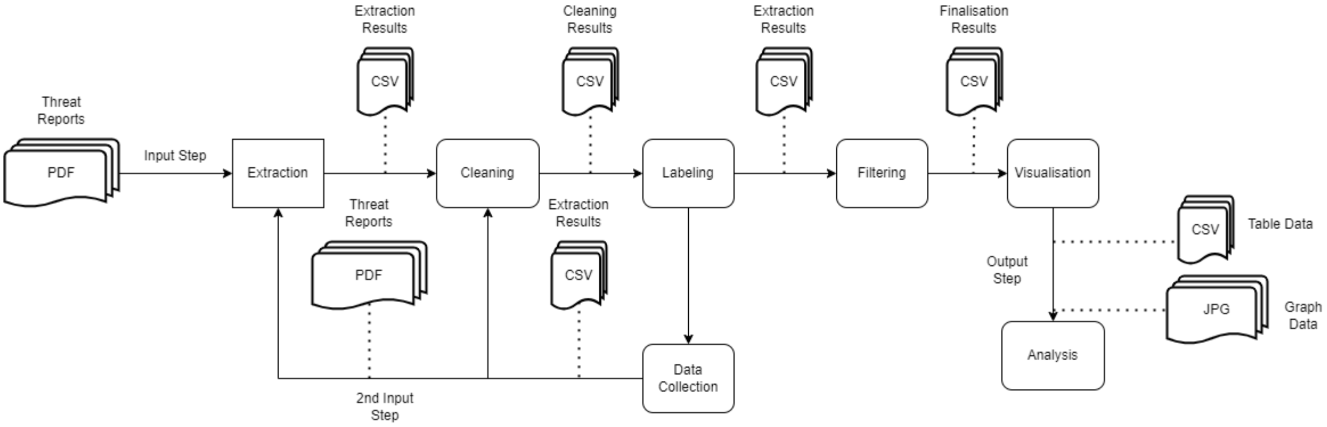

Fig. 1. Diagram of the data processing pipeline used for threat reference extraction.

compiled with the extracted text, and links are identified using the findall() method of the re Python module. This method returns all non-overlapping matches of a pattern in a string, as a list of strings or tuples. A subsequent cleaning step removes undesired links, such as subsets of already found links from alternative extraction methods. This step is executed per page to remove duplicate instances found in a page.

In addition, we leverage PyMuPDF's robust rendering engine for hyperlink detection. Given that URL references can appear as hyperlinks embedded in the PDF file, in plain text, or a combination of both, we employ dedicated methods for text links and hyperlinks per page to increase extraction accuracy. Despite the necessity of a cleaning process to remove duplicates and partial links, no instance is lost in this process.

3.2.2 Cleaning Step

The Cleaning step was necessary for preparing our data for feature creation in the subsequent Labeling Step. This process involved several key removal procedures:

- **Non-Alphanumeric Characters:** We removed non-alphanumeric characters that were present at the end of the extracted link strings to ensure the uniformity and accuracy of our data. The built-in Python method "rstrip()" was used without any arguments to remove trailing whitespace characters from strings representing URIs.

- **Unwanted Punctuation:** Punctuation marks found after spaces ('/') and at the end of strings were also removed using "rstrip()". A specific method was developed to eliminate everything following the last forward slash if it was immediately succeeded by punctuation. These issues arose during the extraction step when the removal of newlines caused the URI string to merge with the following sentence. This was an accepted compromise made in the previous step to enable the extraction of links that were split between multiple rows. Additionally, a heuristic was incorporated into this method to only consider instances where fewer than 10 characters follow a punctuation mark, thereby helping to circumvent unusual cases.

- **Duplicates:** Duplicates were removed following these steps to speed up the comparison function that will be used in the removal of subsets. The "drop_duplicates()" method of Pandas library was used on values with same page number. Duplicate data can skew the results of our analysis, so it was crucial to ensure that each entry in our dataset was unique.

- **Subsets of Other Links:** Finally, a comparison function was used to only remove subset strings of other links per page. This was a necessary step because the extraction process has an inherent tendency to create duplicates per page in cases where text links are also considered as hyperlinks.

The first 3 initial steps consisted in removal of links extracted incorrectly and helped to reduce the execution time of the final step which includes a comparison function that has a linear complexity given by the number of links compared.

By performing this cleaning procedure, we were able to refine our dataset and ensure that they were ready for the subsequent stages of our research. This rigorous cleaning process helped us to improve the reliability and validity of our findings.

3.2.3 Labeling Step

After the cleaning process, a Labeling Step was applied to the resulting links to prepare the dataset for visualization and analysis. The following features were added for each extracted link. For clarity the procedure is detailed after feature descriptions are offered:

- **Domain Name:** A domain name, which is assigned to an entity through a process known as domain registration, represents a distinctive digital identity. In this study, the domain name from every URL was collected. This was done to identify the author of the resource or threat reference to which the link points. The data from the domain registration was crucial in our investigation, as it provided valuable insights into the origins and ownership of the domains, thereby aiding in the identification process.

- **Top-Level Domain (TLD):** A TLD is the last segment of a domain name – the part that comes after the final dot. Examples include .com, .org, .net, etc. The Internet Assigned Numbers Authority (IANA) [26] officially recognizes three types of TLDs: Generic Top-Level Domains (gTLD), Sponsored Top-Level Domains (sTLD), and Country Code Top-Level Domains (ccTLD). IANA is a standards organization that oversees global IP address allocation, autonomous system number allocation, root zone management in the Domain Name System (DNS), media types, and other Internet Protocol-related symbols and numbers. It ensures the global uniqueness of these Internet identifiers, which are crucial for the functioning of the Internet. The TLD was separated from the URL to identify the sponsored domains which point to government entities.

- **WHOIS Info:** WHOIS is a protocol that is used for querying resources' registered users or assignees. These resources include domain names, but it is also used for a wider range of other information. The protocol stores and delivers database content in a human-readable format. WHOIS protocol was used to retrieve registered info for the specific domain name.

- **HTTP Request and HTTP Request Code:** HTTP is a client-server protocol that controls how the client formulates a request and how the server responds to them. An HTTP request was executed for each URL to assess its validity. The resulting HTTP code and the success state were recorded and stored in distinct labels. This step is crucial in ensuring that the links being analyzed are active and accessible.

- **MIME Types:** MIME (Multipurpose Internet Mail Extensions) types indicate the nature and format of a document, file, or assortment of bytes. They are defined and standardized in IETF's RFC 6838 [27]. Examples include text/plain for textual files and application/octet-stream for binary data. MIME types were collected to understand the type of resource and to decide further steps.

The domain name of each URL was extracted and recorded in a separate column. The extraction of domain names and TLDs was achieved with the help of the URLparse[28] and tldextract[29] libraries, respectively. The validity of each domain was assessed using a combination of Python's built-in "socket" library [30] and the "python-whois" [31] package. The "python-whois" is simple importable Python module that produces parsed WHOIS data for a given domain. It offers several other advantages, including the ability to extract data for all popular TLDs (such as .com, .org, .net, etc.), direct querying of the WHOIS server (bypassing the need for an intermediate web service), and compatibility with both Python 2 and Python 3. [31].

The gethostbyname() method from the socket library was employed to locally validate the domain. This method checks if a domain is valid and can be resolved to an IP address. This local validation is advantageous as it does not necessitate a server query, thereby conserving resources. It was used to determine whether feature collection for a specific link should proceed. The results of the link validation were recorded in another column.

As such, for every valid domain, additional information was collected. This began with querying the WHOIS database for each domain to obtain documentation and identification. The Top-Level Domain (TLD) of each domain name was also verified against the official database of valid TLDs maintained by the Internet Assigned Numbers Authority (IANA).

An HTTP request was then executed to assess the validity of the URLs, and the resulting HTTP code and success state were recorded and stored in distinct labels. In the following figure (Figure 2), top 4 HTTP codes obtained for all valid links are shown below as extracted in the Visualization Step.

For every link that returned an HTTP code of 200, representing successful request, an additional step was undertaken to collect the MIME type of the accessed resource for subsequent analytical purposes. The MIME type, which specifies the nature and format of a document, was obtained to enhance the comprehensiveness of the data. The Python "requests" package [32] was instrumental in constructing HTTP requests and capturing MIME types.

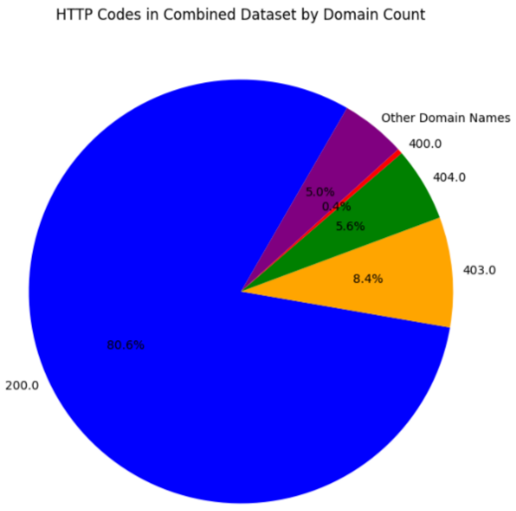

Figure 2 Distribution of retrieved HTTP Response Codes.

While most websites responded positively to the Python scripts, however, a subset subsequently enacted preemptive access restrictions through Internet Service Providers (ISPs) as a deterrent measure against web scraping. To minimize network traffic and avoid this situation, caution was exercised to execute requests only after validating the domain of the URL locally as explained above. Furthermore, MIME types were collected only after the HTTP request was successful, ensuring that resources were not wasted on unsuccessful requests. This approach ensured the efficient use of resources and the integrity of the collected data. Considering that a request took a few seconds on average this approach reduced the duration of the experiment significantly. The distribution of MIME types obtained for all links within the Combined Dataset is illustrated in Figure 3 below. Variation of the same MIME type categories text/html and application/pdf are combined for easier visualization.

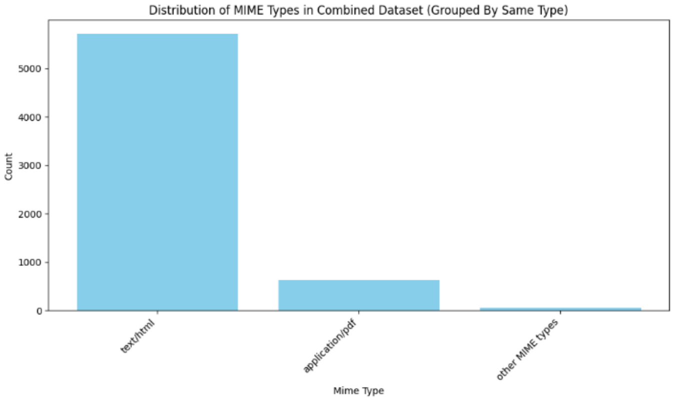

Figure 3 Distribution of MIME Types in Combined Dataset by Domain Count.

3.2.4 Data Collection Step

The "Data Collection" step successfully downloaded potential threat reports targeting valid links with an HTTP status code of 200 and a MIME type of application/pdf. Simultaneously, additional links were extracted from websites associated with links that presented an HTTP status code of 200 and a MIME type of text/html.

Despite the incomplete implementation, numerous links were identified from a total of 272 websites. Furthermore, 308 PDF files were downloaded using threat references from 19 PDF reports. This stage of threat report and link collection, although incomplete, provides a foundation for future investigations into CTI resources of different formats, such as blog posts, and for expanding the dataset used in the pipeline.

Web scraping was conducted using the BeautifulSoup Python package [33]. BeautifulSoup is a Python library that facilitates the extraction of data from HTML and XML documents, including those with malformed markup [33]. It creates a parse tree from these documents, which can be used to extract data, making it a valuable tool for web scraping [33].

PDF retrieval was executed using the Python requests module. The module retrieved only the content of the file, necessitating the extraction of the PDF file name, typically stored in inaccessible metadata. A simpler solution was developed, parsing the URL and extracting the commonly located file name at the end. For URLs that did not provide the file name in this manner, a pdf extension was appended to the string to properly assign a file name to the downloaded threat report.

3.2.5 Filtering Step

The Filtering Step was designed to respect and implement the subset grouping described in Section 1.3, "Dataset Creation". The secondary objective of this step in the pipeline was to filter out invalid URLs or those that lacked the necessary features for subsequent steps. The filtering criteria were based on the presence of values in the columns representing the labels for "Domain Validity" and the "WHOIS Database Info" labels. As such only entries with a valid domain and existent domain name registration information were retained. The table below (Table 2) illustrates the proportion of the total, as well as the subsets, that were retained. As can be observed, over 95% of threat references in all instances was preserved.

Dataset	Count of extracted URLs	Accuracy (Percent of working URLs)
ENISA threat reports	5393	96.9405
Joint CSAs (multiple countries)	2670	96.5169
Joint CSAs (US only)	1371	95.9154
All threat reports (total)	8604	96.7570

Table 2 Accuracy of the data processing pipeline

3.2.6 Data Visualization Step

In this phase of our research, we aimed to identify the most influential authors of threat references, thereby addressing our second research question. To this end, we focused on a select set of variables that encapsulated our findings. Matplotlib[22] was selected as the primary tool for generating all graphical outputs in this study due to its versatility and simplicity. The versatility of Matplotlib is evident in its comprehensive range of plot types, including but not limited to, bar charts, line graphs, histograms, and scatter plots. Furthermore, the simplicity of Matplotlib made configuration of various aspects of the graphs easier, such as color schemes and label positions.

We constructed a pie chart using the 'Domain Name' variable from each dataset to determine the top 10 most influential sources of information in threat reports. The decision to use pie charts, as opposed to other chart types such as bar charts, was driven by the presence of thousands of unique domain names in the largest dataset. Given this large number, a bar chart would not be suitable for effectively comparing the proportion of each domain name. Additionally, we used the "TLD" values from the Combined and ENISA datasets to create two other pie charts. The use of "TLD" values provided a clearer understanding of the distribution of threat references between government and private institutions, especially considering that we found that domain names outside the top 10 accounted for more than 50% of the shares for both datasets.

However, we could generate scatterplots only in the case of the ENISA dataset, being the only case where the time variable was relevant as the dataset is comprised of annual reports. These plots allow us to track the evolution of top domain names over the years. Some of the generated items which were excluded from the analysis are offered in the appendix section. For example, detailed bar charts representing the top 50 domain names were placed in the appendix section of this paper, aiming to offer the reader an in-depth display of the study's variables. A comprehensive analysis of each visualization output is provided in Section 4, 'Experiments and Results'. An overview of our general analysis strategy is explained in the subsequent section.

3.2.7 Analysis Step

The strategy employed for the analysis begins with a description of the characteristics of each dataset. This includes highlighting the key variables that are represented in the graphical representations. For each graphical representation, a concise description is provided, detailing the variable it depicts, the dataset it is derived from, and any notable observations. Following this, an interpretation is given to explain how each item answers the research question. Any observed patterns, trends, or anomalies are discussed in the context of answering Research Question 2: "how can these practices (methodology employed to answer RQ1) can be utilized to assess which organizations play a major role in providing these references from Cyber Intelligence Data resources such as threat reports or blog posts?". If applicable, items from different datasets are compared and contrasted, with discussions on any similarities or differences and their implications for RQ2. The section concludes with a summary of the findings and their implications, a discussion on how the graphical analysis helped answer RQ2, and any limitations or areas for further research.

4. EXPERIMENTS AND RESULTS

The Analysis Step of the dataset pipeline succeeded the previous Filtering and Visualization Steps. The categorical variables deemed most relevant and meaningful for this Analysis Step included domain names, TLDs, and MIME types. For each distinct dataset, a pie chart was incorporated into the analysis to display the proportion of each domain name within the dataset, featuring the top ten domain names. Furthermore, for both the Combined Dataset and the ENISA Dataset, a pie chart was used to illustrate the distribution of TLDs. The analysis identified several domain names representing government agencies as the most prominent. Due to their frequent mention in this section, a brief description of each is provided herein, others which are less prominent can be visible in the bar charts (Figures 7,8,9,10) offered in appendix:

Government affiliated agencies:

- MITRE Corporation (mitre.org) - The MITRE Corporation is a not-for-profit organization that operates research and development centers sponsored by the federal government [34]
- ENISA (enisa.europa.eu) - The European Union Agency for Cybersecurity (ENISA) is the Union's agency dedicated to achieving a high common level of cybersecurity across Europe [35]
- CISA (cisa.gov) - The Cybersecurity and Infrastructure Security Agency (CISA) is a U.S. government organization dedicated to protecting the nation's critical infrastructure from physical and cyber threats [12]
- NIST (nist.gov) - The National Institute of Standards and Technology (NIST) is a U.S. government agency that promotes innovation and industrial competitiveness by advancing measurement science, standards, and technology [36]
- FBI (fbi.gov) - The FBI is the principal investigative arm of the U.S. Department of Justice, responsible for intelligence and law enforcement duties. [14]

- NSA (nsa.gov) - NSA leads the U.S. Government in cryptology, providing signals intelligence and cybersecurity services. [13]
- US Dept. of Defense (defense.gov) - The U.S. Department of Defense (DoD) is a government agency that provides the military forces needed to deter war and protect the security of the United States [37]

Private entities:

- Microsoft (microsoft.com) - a key player in the Cyber Threat Intelligence (CTI) market, providing critical and actionable data to help companies navigate the evolving threat landscape. [38]
- Github (github.com) - a widely-used platform from a Cyber Threat Intelligence (CTI) perspective, hosting various CTI resources such as data visualization tools, threat intelligence platforms, and collections of CTI fundamentals. It is owned by Microsoft. [39]
- Trend Micro (trendmicro.com) - in the context of Cyber Threat Intelligence (CTI), Trend Micro (trendmicro.com) is a prominent American-Japanese public cybersecurity company that provides comprehensive CTI resources, including the Cyber Risk Index [40]
- Securelist (securelist.com) – a comprehensive cybersecurity resource owned by Kaspersky, providing threat intelligence reports, malware research, and analysis1. [41]
- Bleeping Computer (bleepingcomputer.com) - Is a technology news website that provides free computer help via its forums, focusing heavily on cybersecurity, and is owned by Bleeping Computer LLC [42]
- Symantec (Symantec.com) – a cybersecurity software and services provider, formerly owned by Symantec Corporation and now part of Gen Digital Inc. [43]
- Mandiant (mandiant.com) – a cybersecurity services provider, known for its dynamic cyber defense, threat intelligence, and incident response services, and is a subsidiary of Google [44]
- ZDNET (zdnet.com) - a business technology news website that provides professionals with news and advice to embrace innovation, and is owned and operated by Red Ventures [45]

Combined Dataset

Figure 4.1 reveals the distribution of domain names in the “Combined Dataset”, where it is evident that mitre.org, accounting for 21.7% ranks first among the top 10 domain names.

Moreover, in the TLD table (Figure 4.2) the .org TLD with a share of 28.3% is dominated by mitre.org. This suggests that MITRE plays a significant role in providing references in the CTI reports used in this study. In contrast, cisa.gov (8.5%), nist.gov (2.5%), and fbi.gov (0.8%) are the most prominent .gov domains, collectively representing 15.8% of the .gov values according to the TLD distribution depicted in Figure 4.2. This indicates that these government organizations are also key contributors to the cyber intelligence data. Similarly, enisa.europa.eu, with a 5% share, makes up the majority of the .eu TLDs (5.9%), suggesting its importance among European organizations. In terms of domain names associated with companies, microsoft.com (2.3%), github.com (1.7%), trendmicro.com (1.5%), securelist.com (1.4%), and bleepingcomputer.com (1.0%) are the most frequently referenced .com domains, indicating their prominence in the cyber intelligence data, while the ".com" TLDs account for 44% of the values in the TLD table (Figure 4.2).

However, it is crucial to note that these findings are based on the “Combined Dataset”. To draw a comprehensive conclusion, each dataset must be evaluated individually, considering its unique characteristics and context. This will ensure a more accurate understanding of the organizations' roles in providing references from Cyber Intelligence Data resources.

ENISA Dataset

The analysis of the data from Figures 4.2 and 4.3 reveals significant insights into the organizations providing references for ENISA Threat reports. Government agencies, specifically mitre.org, enisa.europa.eu, and cisa.gov, emerge as major contributors, providing around 17% of all references. These agencies, despite being only three in number among the top 10, have a substantial impact.

On the other hand, commercial entities also play a crucial role. Companies such as Securelist (securelist.com), Trend Micro (trendmicro.com), Microsoft (microsoft.com), Bleeping Computer (bleepingcomputer.com), Symantec (symantec.com), Mandiant (mandiant.com), and ZDNET (zdnet.com) contribute just under 11% of the references. These seven commercial entities, all with the “.com” TLD, are also part of the top 10 contributors.

Interestingly, while government agencies contribute a higher percentage of references in ENISA threat reports, the “.com” TLD dominates, accounting for 61.9% of the total. This underscores the significant role of commercial entities. The TLDs associated with the top three government agencies make up around 30% of the total, with “.gov” contributing 3.9%, “.eu” 9.4%, and “.org” 16.6%.

Furthermore, Figures 5 and 6 illustrate the evolution of domain name shares in ENISA reports from 2012 to 2023 using scatterplots. These shares were computed individually for each domain name across all threat reports in the dataset. Figure 6 uses the count of domain names on the while figure 5 uses the percentage to depict the shares in each year on the y axis. To not overcrowd the scatterplots filtering conditions were used. In scatterplot of Figure 6 only domain names with count larger than 25 were included while in the one of Figure 5 only those with percentage larger than 5% were shown.

Lastly, it is observed that companies ceased to have a relatively large share from the 2021 edition onwards, with securelist.com being the last company represented.

Analyzing Figures 5 and 6 in isolation can lead to misinterpretations. For instance, the large percentage of ENISA references in Figure 5 might be misconstrued as an anomaly. However, Figure 6, which shows a steady increase in ENISA references over the years, contradicts this. Interestingly, there is a decrease in these references during 2019-2020, coinciding with the perceived anomaly in Figure 5.

The threshold set in Figure 5 to include only entities with shares larger than 5% has resulted in the omission of several companies.

For example, microsoft.com, despite having similar occurrence counts and percentages as securelist.com, is excluded. This omission becomes more significant considering that microsoft.com surpassed securelist.com in occurrence counts, a fact not discernible from Figure 5 due to both having shares smaller than 5% from 2021 onwards.

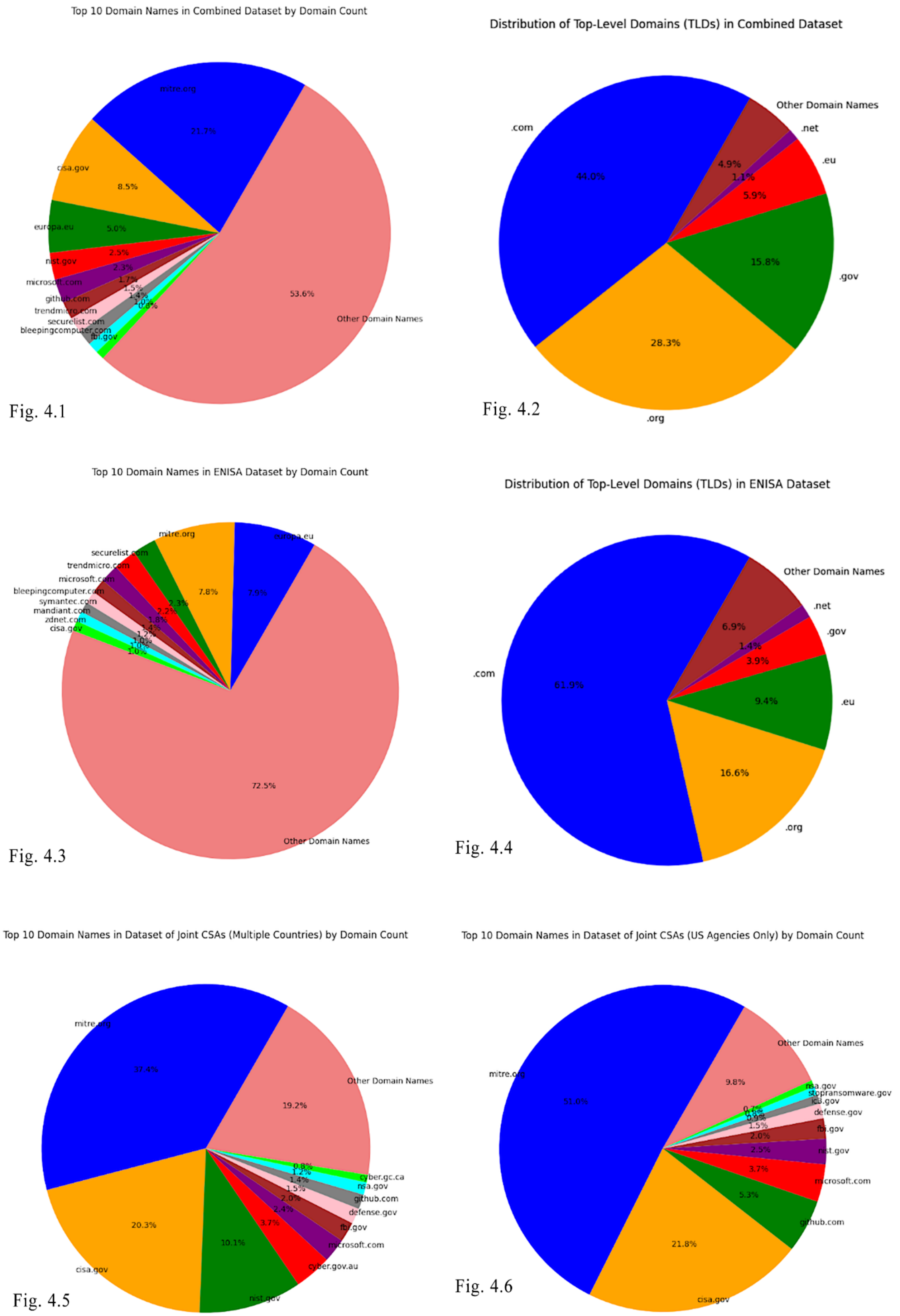

Figure 4 A collection of pie charts depicting the distribution of domain names and TLDs for every dataset numbered from Top to bottom (4.1-4.6)

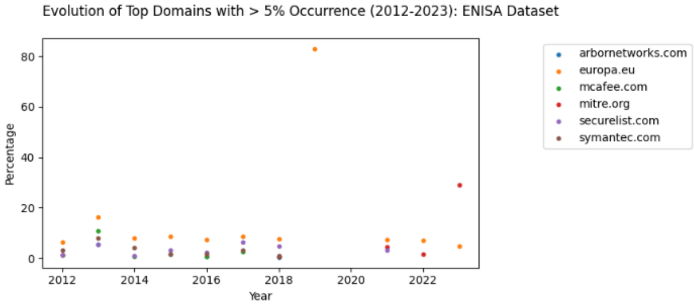

Figure 5 Evolution of top Domain Names (by Percentage) ENISA Dataset

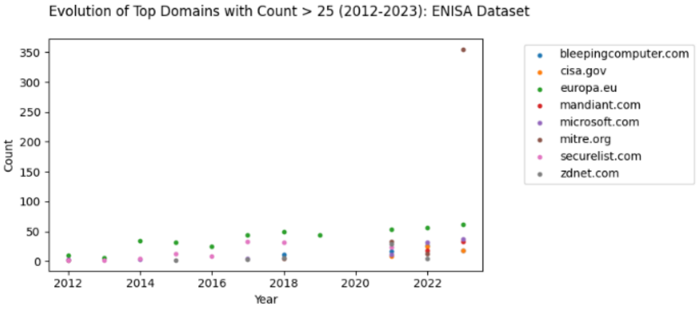

Figure 6 Evolution of top Domain Names (by Count) ENISA Dataset

An apparent anomaly in Figure 6 is the surge in occurrences for mitre.org in 2023, which is also reflected in Figure 5. This raises several questions. Why did MIRE Corporation experience such a high growth rate during 2021-2023, eventually to surpass ENISA in the final year? MITRE is the only entity to do so during the entire period of 2012-2023. Concurrently, why did the percentage of ENISA references, the largest during the years except 2023 according to Figure 5, seem to decrease in the final year? These queries warrant further investigation.

This analysis provides valuable insights into the evolution of domain name shares in ENISA reports. However, it is important to consider these findings as part of a broader context, taking into account other remaining datasets with CSAs and factors to draw comprehensive conclusions.

Joint CSAs Datasets

Figures 4.5 and 4.6 present the top 10 domain distributions in Joint CSAs for two distinct datasets: one that includes joint CSAs from multiple countries, and another that comprises national CSAs from the US. The rationale for this separation becomes evident upon examining the pie charts, as it facilitates a comparative analysis between these similar datasets and underscores the influence of prominent US agencies relative to their collaborators.

The top two reference authors, mitre.org and cisa.gov, feature in both datasets. As anticipated, their shares are more substantial in the national CSAs than in the collaborative ones. Mitre.org accounts for 51% in national CSAs, compared to 37.4% in collaborative CSAs, while cisa.gov constitutes 21.8% and 20.3%, respectively. Interestingly, the share of other domains, which fall outside the top 10, is nearly double in the collaborative CSAs (19.2%) compared to the North American CSAs (9.8%), indicating a greater diversity.

Notably, nist.gov has a significantly larger share in national CSAs (10.1%) compared to collaborative CSAs (2.5%). In contrast, fbi.gov, defense.gov, and nsa.gov maintain similar shares in both datasets. Github.com holds a more prominent position in US CSAs (3rd place, 5.3%) compared to collaborative CSAs (8th place, 1.4%). Its parent company, microsoft.com, ranks 4th in US CSAs and 3rd in collaborative CSAs. The only references that appear in the top 10 of the collaborative CSAs are ciber.gov.au and cyber.gc.ca, representing Australian and Canadian authorities, respectively.

5. FUTURE WORK

The current study, while insightful, was confined to the analysis of 47 threat reports in PDF format. More data was successfully retrieved from 308 associated websites as well as thousands of links from an additional 272 websites. However, due to resource constraints, the pipeline for in-depth analysis and further data collection was not fully executed. Future research should aim to extend the dataset size for a more comprehensive examination of the cyber threat landscape. This would involve a meticulous implementation of the entire analysis pipeline to reveal patterns, trends, and relationships within the expanded dataset. The methodology used in this study, which focused on general PDF file formats, has potential applicability to other domains. This opens up a broader scope for future investigations

Furthermore, the observed trends and their implications for cyber intelligence warrant further exploration. Specifically, future studies should investigate potential signs of an oligopoly in the cyber threat intelligence market.

6. CONCLUSIONS

In conclusion, taking into account the findings from all datasets together both government agencies and commercial entities make significant contributions to providing references from Cyber Intelligence Data resources. The data suggests a more substantial role for government agencies, despite their fewer numbers. Further research could explore the reasons behind these trends and their implications for cyber intelligence. However, this doesn't directly suggest evidence of an oligopoly in the CTi market. The the paper by Conyon et al. (2022) discusses the concept of oligopoly within the context of Big Tech companies, suggesting that evidence of an oligopoly can be inferred from the rise in corporate power and market monopolization. However, this effect was not observed in the analyzed dataset. Notably, the recent rise in the share of references for the MITRE Corporation, a nonprofit entity, is an exception, rendering the effects of this rise not applicable to the oligopoly discussion. While the overall share of corporate entities in the ENISA dataset is large, it is also diverse and has changed over time. Therefore, it remains unclear if the CTI market exhibits characteristics of an oligopoly, warranting further research.

The methodology employed in this study has provided significant insights into the landscape of Cyber Threat Intelligence (CTI) data resources. The analysis section of the pipeline, comprising Filtering, Visualization, and Analysis steps, was instrumental in identifying key contributors to the CTI reports.

The study revealed that both government agencies and commercial entities play substantial roles in providing references for CTI data resources. Government agencies, despite their fewer numbers, appear to have a more significant influence. This is evident from the prominence of domain names such as mitre.org, enisa.europa.eu, and cisa.gov in the datasets.

Commercial entities also contribute significantly, with domain names like microsoft.com, github.com, and trendmicro.com frequently referenced. However, the diversity among these entities and the temporal changes in their contributions suggest a dynamic landscape rather than a static oligopoly.

The study also highlighted the importance of considering each dataset individually due to their unique characteristics and contexts. For instance, the Combined Dataset and the ENISA Dataset showed different distributions of domain names and TLDs. Similarly, the Joint CSAs datasets revealed differences between national and collaborative CSAs.

Interestingly, the study observed a recent surge in references for the MITRE Corporation, a non-profit entity. This trend, along with the decrease in the percentage of ENISA references in 2023, raises questions that warrant further investigation.

In conclusion, this study provides a comprehensive examination of the threat references in CTI data resources landscape. However, it also underscores the need for further research to explore the reasons behind the observed trends and their implications for cyber intelligence. It remains unclear if the CTI market exhibits characteristics of an oligopoly, and this question forms an intriguing direction for future research. The findings of this study, therefore, not only contribute to our understanding of the current CTI landscape but also pave the way for deeper explorations in the future.

7. REFERENCES

[1] Cremer, F., Sheehan, B., Fortmann, M., Kia, A. N., Mullins, M., Murphy, F., & Materne, S. (2022). Cyber risk and cybersecurity: a systematic review of data availability. The Geneva Papers on Risk and Insurance - Issues and Practice, 47, 698–736. 1

[2] Huixia Zhang, Guowei Shen, Chun Guo, Yunhe Cui, Chaohui Jiang, "EX-Action: Automatically Extracting Threat Actions from Cyber Threat Intelligence Report Based on Multimodal Learning", Security and Communication Networks, vol. 2021, Article ID 5586335, 12 pages, 2021. https://doi.org/10.1155/2021/5586335 2

[3] Johnson, C., Badger, M., Waltermire, D., Snyder, J., & Skorupka, C. (2016). SP 800-150, Guide to Cyber Threat Information Sharing. National Institute of Standards and Technology. 3

[4] ENISA Threat Landscape 2023. (2023). European Union Agency for Cybersecurity. 4

[5] Conyon, M., Ellman, M., Pitelis, C. N., Shipman, A., & Tomlinson, P. R. (2022). Big Tech Oligopolies, Keith Cowling, and Monopoly Capitalism. Cambridge Journal of Economics, 46(6), 1205–1224. 5

[6] Wlosinski, L. G. (2021). Cyberthreat Intelligence as a Proactive Extension to Incident Response. ISACA Journal, 2021(Volume 6). 6

[7] Dusek, L. (2020). Key Challenges in Cyber Threat Intelligence. Office of the Director of National Intelligence. Retrieved February 11, 2024. 7

[8] Conti, M., Dehghantanha, A., & Dargahi, T. (2018). Cyber Threat Intelligence: Challenges and Opportunities. In Cyber Threat Intelligence (pp. 1–6). Springer. 8

[9] V. Mavroeidis and S. Bromander, "Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies within Cyber Threat Intelligence," 2017 European Intelligence and Security Informatics Conference (EISIC), Athens, Greece, 2017, pp. 91-98, doi: 10.1109/EISIC.2017.20. 9

[10] Samtani, Sagar & Abate, Maggie & Benjamin, Victor & Li, Weifeng. (2019). Cybersecurity as an Industry: A Cyber Threat Intelligence Perspective. 10.1007/9783-319-90307-1_8-1. 10

[11] Cybersecurity and Infrastructure Security Agency. (2023). Joint Advisories: Fact Sheet 11

[12] Cybersecurity and Infrastructure Security Agency. (n.d.). Cybersecurity and Infrastructure Security Agency | Home Page 12

[13] National Security Agency. (n.d.). National Security Agency | Central Security Service 13

[14] U.S. Federal Bureau of Investigation. (n.d.). Homepage 14

[15] Multi-State Information Sharing and Analysis Center (MS-ISAC) 15

[16] Five Eyes Intelligence Oversight and Review Council. (n.d.). Five Eyes Intelligence Oversight and Review Council (FIORC) 16

[17] Cybersecurity & Infrastructure Security Agency. (n.d.). Resources and Tools. Retrieved [2023] 17

[18] Python Software Foundation. (n.d.). Homepage 18

[19] Project Jupyter. (n.d.). Project Jupyter | Home 19

[20] NumPy developers. (n.d.). NumPy: The fundamental package for scientific computing with Python 20

[21] Pandas developers. (n.d.). Pandas: Python Data Analysis Library 21

[22] Matplotlib developers. (n.d.). Matplotlib: Visualization with Python 22

[23] PyMuPDF developers. (n.d.). PyMuPDF: A high performance Python library for data extraction, analysis, conversion & manipulation of PDF (and other) documents 23

[24] PDFMiner developers. (n.d.). PDFMiner: A text extraction tool for PDF documents. 24

[25] URLExtract developers. (n.d.). URLExtract: A tool for extracting URLs from text. 25

[26] Internet Assigned Numbers Authority. (n.d.). Internet Assigned Numbers Authority. 26

[27] Freed, N., Klensin, J. C., & Hansen, T. (2013). Media Type Specifications and Registration Procedures. Internet Engineering Task Force. RFC 6838. 27

[28] Python Software Foundation. (2024). urllib.parse — Parse URLs into components. Python 3.12.1 documentation. 28

[29] Kurkowski, J. (2023). tldextract. PyPI. 29

[30] Python Software Foundation. (2023). socket — Low-level networking interface. 30

[31] Penman, R. (2022). python-whois 0.8.0. 31

[32] Reitz, K. (2022). Requests: HTTP for Humans™ 32

[33] Richardson, L. (2024). Beautiful Soup Documentation — Beautiful Soup 4.12.0 documentation 33

[34] MITRE Corporation. (2024). MITRE Corporation Official Website 34

[35] ENISA. (2024). ENISA Official Website 35

[36] National Institute of Standards and Technology. (2024). 36

[37] U.S. Department of Defense. (2024). DoD Official Website 37

[38] Statista. (2023). Topic: Cyber Threat Intelligence (CTI). 38

[39] GitHub. (2024). GitHub. 39

[40] Trend Micro. (2024). Trend Micro. 40

[41] Securelist. (2024). Securelist. 41

[42] Bleeping Computer LLC. (n.d.). Bleeping Computer. 42

[43] Gen Digital Inc. (n.d.). Symantec. 43

[44] Google LLC. (n.d.). Mandiant. 44

[45] Red Ventures. (n.d.). ZDNET. 45

[46] PDF URL Extractor 46

8. APPENDIX

Figure 3 Showing an overview of the used CTI threat reports in research:

ID	CTI report Name (with references where available)	Dataset
1	ENISA Threat Landscape 2012	ENISA
2	ENISA Threat Landscape 2013	ENISA
3	ENISA Threat Landscape 2014	ENISA
4	ENISA Threat Landscape 2015	ENISA
5	ENISA Threat Landscape 2016	ENISA
6	ENISA Threat Landscape 2017	ENISA
7	ENISA Threat Landscape 2018	ENISA
8	ENISA Threat Landscape 2019	ENISA
9	ENISA Threat Landscape 2021	ENISA
10	ENISA Threat Landscape 2022	ENISA
11	ENISA Threat Landscape 2023	ENISA
12	aa23-025a-protecting-against-malicious-use-of-rmm-software	Joint CSA (US only)
13	aa23-039a-esxiargs-ransomware-virtual-machine-recovery-guidance	Joint CSA (US only)
14	aa23-059a-cisa_red_team_shares_key_findings_to_improve_monitoring_and_hardening_of_networks_1	Joint CSA (US only)
15	AA23-061A: Stopransomware Royal Ransomware Update	Joint CSA (US only)
16	AA23-074A: Threat Actors Exploit Progress Telerik Vulnerabilities in Multiple U.S. Government IIS Servers	Joint CSA (US only)
17	AA23-075A: Stop Ransomware LockbitAA23-131A: Malicious Actors Exploit CVE-2023-27350 in PaperCut MF and NG	Joint CSA (US only)
18	AA23-158A: Stopransomware CL0P Ransomware Gang Exploits MOVEit Vulnerability	Joint CSA (US only)
19	AA23-193A: Joint CSA Enhanced Monitoring to Detect APT Activity Targeting Outlook Online	Joint CSA (US only)
20	AA23-201A: CSA Threat Actors Exploiting Citrix-CVE-2023-3519 to Implant Webshells	Joint CSA (US only)
21	AA23-242A: Identification and Disruption of Qakbot Infrastructure	Joint CSA (US only)
22	AA23-250A: APT Actors Exploit CVE-2022-47966 and CVE-2022-42475	Joint CSA (US only)
23	AA23-284A: Joint CSA Stopransomware Avoslocker Ransomware Update	Joint CSA (US only)
24	AA23-289A: Threat Actors Exploit Atlassian Confluence CVE-2023-22515 for Initial Access	Joint CSA (US only)
25	AA23-319A: Stopransomware Rhysida Ransomware	Joint CSA (US only)
26	AA23-320A: Scattered Spider	Joint CSA (US only)
27	AA23-339A: Threat Actors Exploit Adobe ColdFusion CVE-2023-26360	Joint CSA (US only)
28	AA23-349A: Enhancing Cyber Resilience: Insights from the CISA Healthcare and Public Health Sector Risk and Vulnerability Assessment	Joint CSA (US only)
29	AA23-353A: #StopRansomware: ALPHV Blackcat	Joint CSA (US only)
30	JOINT CSA: Top Ten Cybersecurity Misconfigurations	Joint CSA (US only)
31	Joint Cybersecurity Advisory: #StopRansomware: Snatch Ransomware	Joint CSA (US only)
32	AA23-131A: Malicious Actors Exploit CVE-2023-27350 in PaperCut MF and NG	Joint CSA (US only)
33	AA23-129A: Snake Malware	Removed
34	AA23-136A StopRansomware: BianLian Ransomware Group	Joint CSA (Multiple Countries)
35	AA23-165A Understanding TA LockBit 0	Joint CSA (Multiple Countries)
36	AA23-187A Increased Truebot Activity Infects U.S. and Canada Based Networks	Joint CSA (Multiple Countries)
37	AA23-208A Joint CSA Preventing Web Application Access Control Abuse	Joint CSA (Multiple Countries)
38	AA23-213A Joint CSA Threat Actors Exploiting Ivanti EPPM Vulnerabilities	Joint CSA (Multiple Countries)
39	AA23-215A Joint CSA 2022 Top Routinely Exploited Vulnerabilities	Joint CSA (Multiple Countries)
40	AA23-325A LockBit 3.0 Ransomware Affiliates Exploit CVE-2023-4966 Citrix Bleed Vulnerability	Joint CSA (Multiple Countries)
41	AA23-335A IRGC-Affiliated Cyber Actors Exploit PLCs in Multiple Sectors	Joint CSA (Multiple Countries)
42	AA23-335A IRGC-Affiliated Cyber Actors Exploit PLCs in Multiple Sectors - 1	Joint CSA (Multiple Countries)
43	AA23-347A Russian Foreign Intelligence Service (SVR) Exploiting JetBrains TeamCity CVE Globally	Joint CSA (Multiple Countries)
44	AA23-352A StopRansomware: Play Ransomware	Joint CSA (Multiple Countries)
45	APT28 Exploits Known Vulnerability to Carry Out Reconnaissance and Deploy Malware on Cisco Routers UK	Joint CSA (Multiple Countries)
46	CSA BLACKTECH HIDE IN ROUTERS TLP-CLEAR	Joint CSA (Multiple Countries)
47	CSA PRC State Sponsored Cyber Living off the Land v1.1	Joint CSA (Multiple Countries)
48	CSA RANSOMWARE ATTACKS ON CI FUND DPRK ACTIVITIES	Joint CSA (Multiple Countries)
49	CERTFR-2022-CTI-004	Removed
50	Threat Report 24th March 2023	Removed

Bar charts showing a comprehensive overview of the top 50 Domain Names for every Dataset (Figures 7-10)

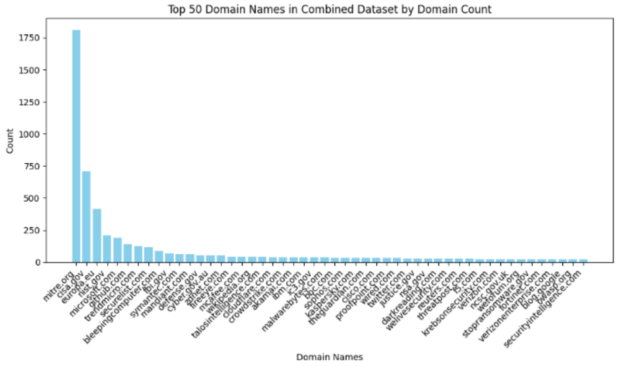

Figure 7 Distribution of top 50 domain names by occurrence in Combined Dataset

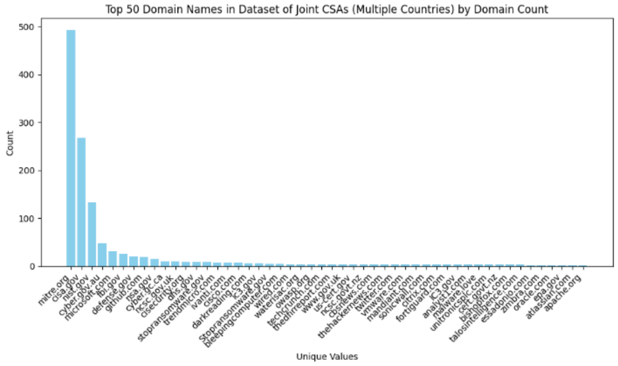

Figure 8 Distribution of top 50 domain names by occurrence in Joint CSAs (Multiple Countries)

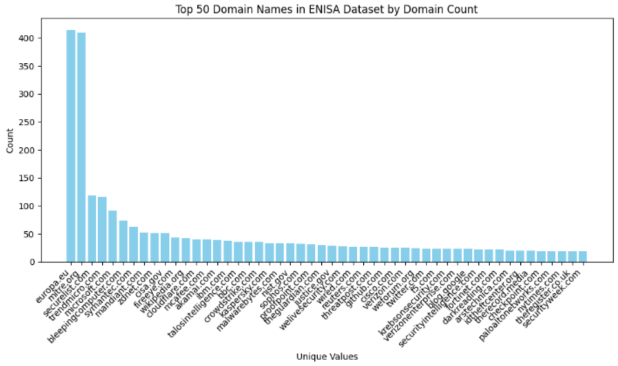

Figure 9 Distribution of top 50 domain names by occurrence in ENISA Dataset

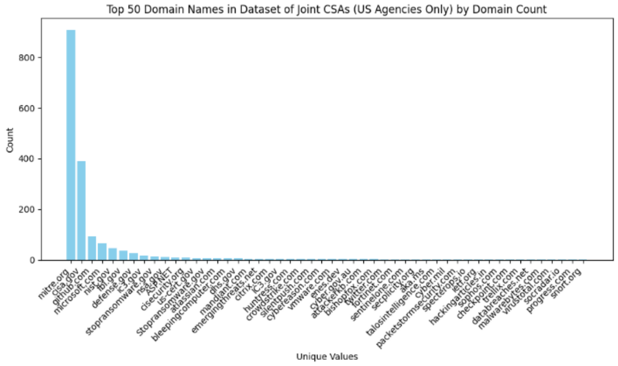

Figure 10 Distribution of top 50 domain names by occurrence in Joint CSAs (US agencies)