Cyberspace Mapping System Based on Multi-Source Data Aggregation

Yingjie Wang [1], Zhe Zhang [1], Hongjie Fan [2], and Songtao Ye [1,*]

[1] School of Computer Science, Xiangtan University, Xiangtan, China
[2] The Department of Science and Technology Teaching, China University of Political Science and Law, Beijing, China
Email: 202205566502@smail.xtu.edu.cn (Y.W.); 202205566510@smail.xtu.edu.cn (Z.Z.); hjfan@cupl.edu.cn (H.F.); yesongtao@xtu.edu.cn (S.Y.)
*Corresponding author

***Abstract*—In the context of the increasing complexity of cyberspace, accurately and efficiently mapping network assets has become paramount for effective cybersecurity measures. This study presents a novel cyberspace mapping system that leverages multi-source data aggregation to enhance the discovery of network assets. To resolve inconsistencies in query syntax across various tools, the system employs a unified query framework utilizing the Aho-Corasick algorithm, which significantly enhances syntax conversion and matching efficiency. The system amalgamates data from multiple cyberspace mapping engines and conducts rigorous credibility assessments to ensure the reliability and quality of the data. Real-time synchronization, achieved through multithreading, ensures data remains current and comprehensive, thereby facilitating accurate and timely cybersecurity analyses. Experimental evaluations conducted on 2000 entries yielded a high credibility score of 99.9337% and demonstrated an average accuracy improvement of 3.33% compared to individual tools. Furthermore, the system incorporates advanced visualization tools, such as radar charts and tree diagrams, which support effective data interpretation and aid in decision-making processes. This comprehensive system not only optimizes query standardization, data aggregation, and credibility assessment but also enhances visualization capabilities, creating a strong foundation for future integrations of deep learning technologies and enabling real-time responses to evolving cybersecurity challenges.**

***Keywords*—cyberspace surveying and mapping, multi-source data aggregation, credibility assessment, network assets**

I. Introduction

A. Background and Motivation

Mapping cyberspace is a crucial step in the process of information gathering, offering detailed insights into the structure and composition of network assets [1]. Tools such as Shodan, ZoomEye, FOFA, Hunter, and Quake are instrumental in identifying a diverse range of network components, thereby laying a solid foundation for further analysis and decision-making [2]. However, each tool exhibits specific limitations when utilized independently. For instance, Shodan is adept at detecting Internet of Things (IoT) devices but offers limited visibility into certain industrial systems, whereas ZoomEye primarily targets Industrial Control Systems endpoints and might overlook other types of network assets. FOFA, Quake, and Hunter, employing unique scanning methodologies and focusing on different geographical regions, may create blind spots if deployed singularly. To achieve a more comprehensive and reliable overview of the cyberspace environment, it is imperative to amalgamate the strengths of these platforms through a coordinated, multi-source strategy. By harnessing their complementary capabilities—such as expansive scanning coverage, frequent updates, and varied analytical emphases—organizations can address the limitations inherent in each tool and construct a more fortified cyberspace mapping framework.

B. Technical Challenges

Despite advancements, formidable challenges persist in the development of a multi-source cyberspace surveying and mapping system. A primary challenge is the inconsistency in query syntax across different tools. Recent studies have sought to overcome this obstacle by developing methods for query conversion and syntax mapping. Although some approaches are capable of translating queries into a uniform format, they frequently necessitate ongoing adjustments to accommodate the rapid evolution of these tools.

Another significant challenge is the aggregation of data from multiple sources while maintaining data quality. Each mapping tool offers only a partial view of network assets, resulting in inconsistencies. These discrepancies can lead to conflicts during the process of data aggregation. Recent research has introduced algorithms that evaluate data reliability based on the credibility of the source and real-time conditions to preserve data quality. These algorithms proficiently eliminate low-quality data, thereby enhancing the overall integrity of the data [3]. Nevertheless, current integration techniques continue to encounter obstacles when dealing with large datasets and real-time updates [4].

Manuscript received November 18, 2024; revised January 23, 2025; accepted February 14, 2025; published July 15, 2025.

Furthermore, multi-source visualization remains a challenge due to the need to manage data inconsistencies and address the heterogeneity across different tools. Visualization techniques such as radial tree diagrams, heat maps, and radar charts have proven effective in representing complex network data, facilitating comparisons of features across different cyberspace surveying and mapping tools [5]. However, dynamic data still present challenges to visualization and real-time analysis. Additionally, manual analysis of cyberspace mapping is notably limited in efficiency and scalability as network assets expand. Moreover, the absence of objective evaluation in manual analysis heightens the risk of bias in assessments.

C. Contributions and Novelty

In response to the identified challenges, this study proposes a comprehensive solution that incorporates data aggregation, consistency verification, and advanced visualization tools. The principal contributions and innovations of this research are delineated as follows:

(1) Unlike traditional single-tool approaches or basic aggregation methods, this study introduces a standardized query mechanism that ensures consistency across multiple platforms. This system not only reduces the manual workload and minimizes human errors but also promotes seamless interoperability among diverse platforms such as Shodan, ZoomEye, FOFA, Quake, and Hunter. The proposed mechanism significantly streamlines query operations, addressing a gap often neglected in previous research.

(2) To guarantee data freshness and completeness, we implement a real-time synchronization mechanism that continuously updates the aggregated dataset to reflect changes across platforms. This approach diverges from conventional batch-based methods and is essential for rapid threat detection and analysis, particularly vital in the dynamic context of cyberspace.

(3) Addressing the shortcomings of simplistic data merging, this work develops a framework to assess the credibility of data, considering various factors including historical performance and consistency across tools. This robust framework not only filters out low-quality or conflicting data but also ensures the accuracy and reliability of the integrated dataset.

(4) This research integrates sophisticated visualization tools that facilitate effective analysis and comparison of complex network data. Unlike traditional approaches that merely display raw scan results, our system employs visual analytics to elucidate data patterns, overlaps, and discrepancies across multiple platforms, thereby transforming large-scale network scans into actionable cybersecurity insights.

D. Structure of the Paper

The organization of this study is as follows: The literature review section provides a synthesis of existing methodologies in data aggregation, credibility assessment, and visualization, laying the groundwork for the system proposed. The system design section elaborates on the architecture and functionalities of the system, including its query syntax conversion mechanism, credibility assessment framework, and data visualization capabilities. The experiment and results section assesses the system's effectiveness through experimental setups and results, concentrating on coverage, credibility, and accuracy. The conclusion highlights the main findings, contributions, and future research directions, underscoring the integration of sophisticated data fusion techniques and real-time capabilities to enhance cybersecurity measures.

II. Literature Review

This section reviews contemporary developments, emphasizing how modern techniques have enhanced the efficiency, accuracy, and reliability of cyberspace mapping.

A. Data Aggregation Technology

Data aggregation forms the cornerstone of cyberspace surveying and mapping. Extensive research has explored multi-source aggregation methods aimed at constructing a comprehensive view of network assets. Typically, these methods involve several critical steps [6] such as data deduplication, format standardization, and consistency checks [7]. Over time, data aggregation techniques have evolved from traditional batch processing and Extract, Transform, Load (ETL) processes to contemporary real-time data processing within distributed computing frameworks [8]. In recent years, technologies such as Hadoop and Spark have revolutionized the processing of large-scale network data [9], with ongoing research focusing on optimizing data fusion to enhance data accuracy and efficiency.

B. Cyberspace Surveying and Mapping Technology

Cyberspace surveying and mapping technology, which plays a crucial role in the field of cybersecurity, involves a range of techniques that span from network probing and port scanning to the construction of network topologies [10]. This domain typically utilizes both active scanning and passive listening strategies, each offering distinct benefits and inherent limitations [10]. Active scanning, which includes tools such as Nmap and Zmap, enables rapid acquisition of information regarding the status of network hosts and open ports but also carries an increased risk of detection by the target network [11]. Conversely, passive listening minimizes detection risks by capturing network traffic data, although its effectiveness is contingent upon the traffic activity of the target network [12].

C. State of the Art

Prominent tools in the realm of cyberspace mapping include Shodan, ZoomEye, and FOFA, which are

extensively employed for the detection and characterization of internet-exposed devices and services. Shodan, for instance, conducts scans of open ports and services, yielding comprehensive details about devices, including operating systems, protocols, and manufacturers. This tool is particularly valuable for security assessments involving the Internet of Things (IoT) and Industrial Control Systems (ICS) [13]. ZoomEye excels in conducting deep scans across various protocols and services, identifying a broad spectrum of devices. It is especially adept at handling IoT devices and ICS, providing detailed asset information through protocol analysis and service identification [14]. FOFA leverages fingerprint recognition technology to swiftly identify exposed devices, with a focus on IoT. It offers detailed device fingerprints and protocol analysis, aiding security researchers in pinpointing potential vulnerabilities [15]. These tools are integral to vulnerability detection and asset risk assessment. With advancements in deep learning, researchers are exploring its application across various domains. For instance, in the context of brain tumor classification, Patro *et al.* [16] have utilized deep learning techniques to effectively differentiate among tumor types by integrating multiple models. This breakthrough serves as a catalyst for enhancing the classification and identification capabilities of cyberspace mapping technologies through deep learning, potentially increasing the automation of these processes [17]. Furthermore, in the arena of Cyber-Physical Systems (CPSs), Yao *et al.* [18] have developed a Prescribed-Time Output Feedback Control (PTOFC) algorithm, which tackles the challenges posed by output constraints and malicious attacks. This innovative approach could be adapted for cyberspace mapping, integrating advanced control strategies to optimize data aggregation, enhance system response times, and bolster the reliability of mapping tools against network threats and data irregularities.

D. Credibility Assessment

Credibility assessment is essential for enhancing the reliability of data, particularly in the context of multi-source data environments. Evaluating the credibility of mapping results from various sources presents a significant challenge. Existing literature has introduced diverse methodologies for this purpose, including weighted calculations, probabilistic graphical models, and trend analysis through time series data [19]. These approaches determine credibility based on the historical performance and the reliability of data provided by each cyberspace mapping tool [20]. Additionally, the integration of deep learning technologies in cyberspace mapping has facilitated the adoption of uncertainty estimation methods based on neural networks, which aid in identifying the most reliable data [21]. The incorporation of expert knowledge and prior information further enhances both the reliability and applicability of these evaluations [22].

In summary, notable progress has been achieved in the fields of cyberspace surveying and mapping, particularly in data aggregation, mapping technology, and credibility assessment. The development of data aggregation techniques has enhanced the efficiency of integrating multi-source data, while advanced technologies have increased the precision of cyberspace mappings. Credibility assessment, in particular, plays a pivotal role in improving data reliability. Building on these advancements, future research should focus on unifying query syntax, assessing data credibility, and efficiently aggregating data to overcome challenges in cyberspace surveying and mapping [23].

III. System Design

A. Credibility Assessment

In a multi-source cyberspace mapping system, the quality and consistency of data collected from different tools can vary significantly due to variations in scanning methodologies, data coverage, and update frequencies. Such discrepancies often result in conflicting or incomplete results, complicating the assurance of data reliability. To mitigate these issues, the system incorporates a credibility assessment algorithm designed to identify and prioritize the most trustworthy data. Fig. 1 illustrates the credibility calculation model.

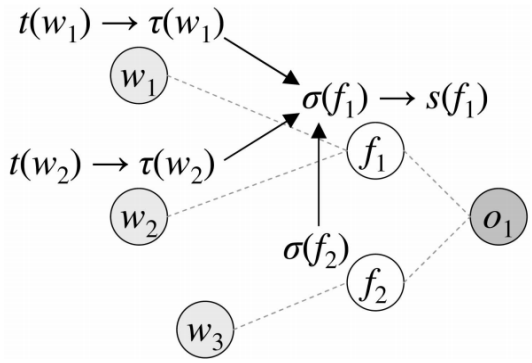

Fig. 1. Credibility calculation model.

The algorithm's foundation lies in two definitions and four heuristic rules that collectively guide the selection and aggregation of reliable information from diverse data sources:

1) Foundational definitions

Credibility of a Fact: The credibility level of a fact represents the probability of its accuracy, considering the current knowledge base.

Reliability of a Cyberspace Mapping Tool: The reliability of a mapping tool w is gauged by the expected credibility level of the facts it produces.

2) Heuristic rules

(1) An object typically possesses only one verifiable fact concerning its attributes.
(2) Facts that are identical or highly similar across different tools are more likely to be accurate.
(3) False facts rarely appear identical or similar across various tools.
(4) Tools that consistently provide accurate facts about numerous objects within a domain are more likely to yield reliable information for other objects in the same domain.

By adhering to these rules, the algorithm methodically assesses the credibility of both the tools and the facts they provide. This systematic evaluation prioritizes data with higher reliability during the aggregation process, effectively minimizing inconsistencies and enhancing overall data accuracy.

3) Credibility calculation framework

To quantify the process effectively, the reliability of a tool, denoted as w, is assessed by calculating the average of the credibility scores associated with all facts it provides. This relationship can be mathematically represented as shown in Eq. (1).

$$t(w)=\frac{\sum_{f\in F(w)} s(f)}{|F(w)|} \tag{1}$$

Here, $F(w)$ signifies the collection of facts yielded by the cyberspace surveying and mapping tool *w*, and $s(f)$ denotes the credibility score assigned to each fact *f*.

The reliability of the tools that deliver it and the consistency of the associated data are critical in determining the credibility of the data. For example, if f_1 is the only fact available for the object $o(f)$ and is furnished by tools w_1 and w_2, an error in *f* implies potential inaccuracies in both w_1 and w_2. Assuming w_1 and w_2 operate independently, the probability that both tools concurrently produce errors can be calculated using the following equation:

$$(1-t(w_1))\cdot(1-t(w_2))$$

In general, if a fact *f* acts as the sole source of information for a specific object, its credibility level, $s(f)$, can be determined as illustrated in Eq. (2):

$$s(f)=1-\prod_{w\in W(f)}(1-t(w)) \tag{2}$$

Here, $W(f)$ represents the set of cyberspace surveying and mapping tools that furnish *f*. Given that $1-t(w)$ is typically minimal, its product may lead to numerical underflow. To circumvent this issue, a logarithmic transformation is applied, as shown in Eq. (3):

$$\tau(w)=-\ln(1-t(w)) \tag{3}$$

In this equation, $\tau(w)$ ranges from 0 to $+\infty$, with higher values indicating a greater reliability of the tool *w*.

Accordingly, the credibility score of a fact *f* is calculated as follows:

$$\sigma(w)=-\ln(1-s(w)) \tag{4}$$

This key principle asserts that the credibility score of a fact *f* is determined by aggregating the reliability scores of all cyberspace surveying and mapping tools that contribute data for *f*, as mathematically depicted in Eq. (5).

$$\sigma(f)=\sum_{w\in W(f)}\tau(w) \tag{5}$$

When different tools provide conflicting descriptions of the same fact, an influence function $imp(f'\to f)$ is introduced. This function, which ranges from −1 to +1, indicates the impact of one fact on the credibility of another. If f' is accurate, *f* may also be accurate; conversely, if f' is accurate, *f* might be inaccurate. To adjust the credibility of a fact based on the influence of a related fact, the adjusted credibility score of fact *f* is defined as:

$$\sigma^*(f)=\sigma(f)+\rho\cdot\sum_{o(f')=o(f)}\sigma(f')\cdot imp(f'\to f) \tag{6}$$

Here, ρ is a parameter between 0 and 1, utilized to modulate the influence of related facts. Thus, σ^* represents the aggregated credibility score of fact *f*, wherein the credibility score of each related fact f' is weighted by its influence on *f*. When conflicting information exists for *f*, this influence is considered to adjust its credibility level accordingly.

$$imp(f'\to f)<0$$

The credibility of *f* can be computed using $\sigma^*(f)$ in the same manner that $\sigma(f)$ is used to compute the credibility of *f*. This adjusted credibility is denoted as $s^*(f)$.

$$s^*(f)=1-e^{-\sigma^*(f)} \tag{7}$$

We propose a methodological approach for evaluating the credibility of facts by utilizing the reliability of cyberspace surveying and mapping tools. This method considers the interconnections among related facts. Our system evaluates both the credibility of informational sources on the network and the veracity of the facts they provide. These evaluations are then mathematically formulated through basic matrix operations, capitalizing on the efficiency and parallelism offered by such computations. To demonstrate the computational process, we employ vector notation to represent the reliability of both the cyberspace tools and the facts in question. The vectors are defined as follows:

$$\begin{aligned}\vec{t}&=(t(w_1),...,t(w_M))^T\\ \vec{\tau}&=(\tau(w_1),...,\tau(w_M))^T\\ \vec{s}&=(s(f_1),...,s(f_N))^T\\ \vec{\sigma^*}&=(\sigma^*(f_1),...,\sigma^*(f_N))^T\end{aligned} \tag{8}$$

Subsequently, we introduce an $M \times N$ matrix A, which is employed to deduce the reliability of cyberspace surveying tools based on the credibility of facts. In a reciprocal manner, an $N \times M$ matrix B is constructed to infer fact credibility from tool reliability. The definitions of these matrices are as follows:

$$\begin{aligned} \vec{t} &= A\vec{s} \\ \overrightarrow{\sigma^*} &= B\vec{\tau} \end{aligned} \tag{9}$$

Vectors $\vec{t}$ and $\vec{\tau}$ can be converted using Eq. (3), while vectors $\vec{s}$ and $\vec{\sigma}$ undergo transformations as per Eq. (9). From Eq. (1), matrix A is described:

$$A_{ij} = \begin{cases} \dfrac{1}{\left|F\left(w_i\right)\right|}, & \text{if } f_j \in F\left(w_i\right) \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Matrix B, in comparison to matrix A, incorporates additional complexities due to the interdependencies among multiple facts pertaining to the same entity. As outlined in Eq. (6) and Lemma 1, the element B_{ij} of matrix B is determined as follows:

Case 1: If tool w_i provides fact f_j, then B_{ij} is assigned to a value of 1.

Case 2: If tool w_i provides fact f_k, the object associated with f is the same as the that associated with f_j $(i.e)$, $o\left(f_k\right) = o\left(f_j\right)$), then B_{ij} is set to $\rho \cdot imp\left(f_k \to f_j\right)$.

Case 3: In all other scenarios, the value of B_{ij} is set to 0.

Building upon these definitions and formulas, we propose an optimization framework as shown in Eq. (11). The framework encompasses:

$$\max_{\{t(w)\},\{s(f)\}} \left[\begin{array}{l} \underbrace{\sum_{f \in F}\left(s(f) \times \sum_{w \in W(f)} t(w)\right)}_{(A)} + \\ \underbrace{\alpha \sum_{(f,f') \in P} s(f) \cdot s(f') \cdot imp(f' \to f)}_{\text{(B) Positive Synergy}} - \\ \underbrace{\beta \sum_{(f,f') \in N} s(f) \cdot s(f') \cdot \left|imp(f' \to f)\right|}_{(C)\text{Conflict Penalization}} \end{array} \right] \tag{11}$$

Here,

(A) Tool–Fact Consistency, which rewards facts corroborated by reliable tools,

(B) Positive Synergy, which grants additional credibility when pairs of facts mutually reinforce each other, and

(C) Conflict Penalization, which imposes penalties when conflicting facts are both deemed highly credible.

The parameters α and β ($\alpha, \beta \geq 0$) modulate the relative weight of synergy versus conflict, with a higher α emphasizing the promotion of cooperative, corroborative facts, while a higher β focuses on minimizing contradictions.

By maximizing this objective, our system simultaneously evaluates tool reliability and fact credibility, thereby ensuring a coherent and consistent alignment between trustworthy tools and accurate facts, while also mitigating factual contradictions.

B. Computational Complexity

This section delineates the analysis of computational complexity, focusing on four integral components: the query statement parsing module, the data acquisition module, the data credibility assessment algorithm, and matrix operations.

The query statement parsing module exhibits a time complexity of $O\left(n+m+k\right)$, where n denotes the length of the query, m represents the total length of all patterns, and k corresponds to the number of patterns matched.

The data acquisition module expedites the process by concurrently dispatching queries to multiple cyberspace mapping tools, yielding a time complexity of $O\left(t \cdot Ttool\right)$, where t is the number of tools engaged, and $Ttool$ is the response time per tool.

The data credibility assessment algorithm is designed to calculate the reliability of the tools and assess the credibility of facts, resulting in a time complexity of $O\left(n^2 \cdot m\right)$, with n as the number of facts and m the number of tools involved.

Furthermore, the matrix operations, pivotal for calculating the relationships between tools and facts, also manifest a time complexity of $O\left(n^2 \cdot m\right)$. This complexity significantly contributes to the overall computational burden. Given these complexities, it is imperative that the system workflow is meticulously engineered to optimize both efficiency and scalability. These considerations are elaborated upon in the subsequent section.

C. System Workflow

The workflow of the system is designed to integrate multiple processes, thereby enabling users to query network assets and obtain reliable, visually represented outcomes. Fig. 2 illustrates an overview of this workflow, which progresses through the following stages:

(1) User Query Submission: Users input query parameters using a unified query syntax, which are subsequently transmitted to the backend.

(2) Query Syntax Conversion: The backend deconstructs the query into ‘key fields’ and ‘field values’, rectifying naming discrepancies across different tools through identifier matching.

(3) Field Mapping and Query Dispatch: The system aligns the key fields with the formats required by various cyberspace surveying and mapping tools and dispatches the queries via their respective Application Programming Interface (API).

(4) Data Standardization: Data returned from these tools is standardized to ensure consistency, thereby facilitating subsequent processing.
(5) Credibility Calculation: The system evaluates the credibility of the data based on historical performance and consistency metrics, filtering out results of lower quality.
(6) Data Aggregation: Following credibility assessment, the system amalgamates data from multiple sources, resolves discrepancies, and standardizes the outcomes. The aggregated data is then conveyed to the frontend for visualization, aiding in user analysis and decision-making.

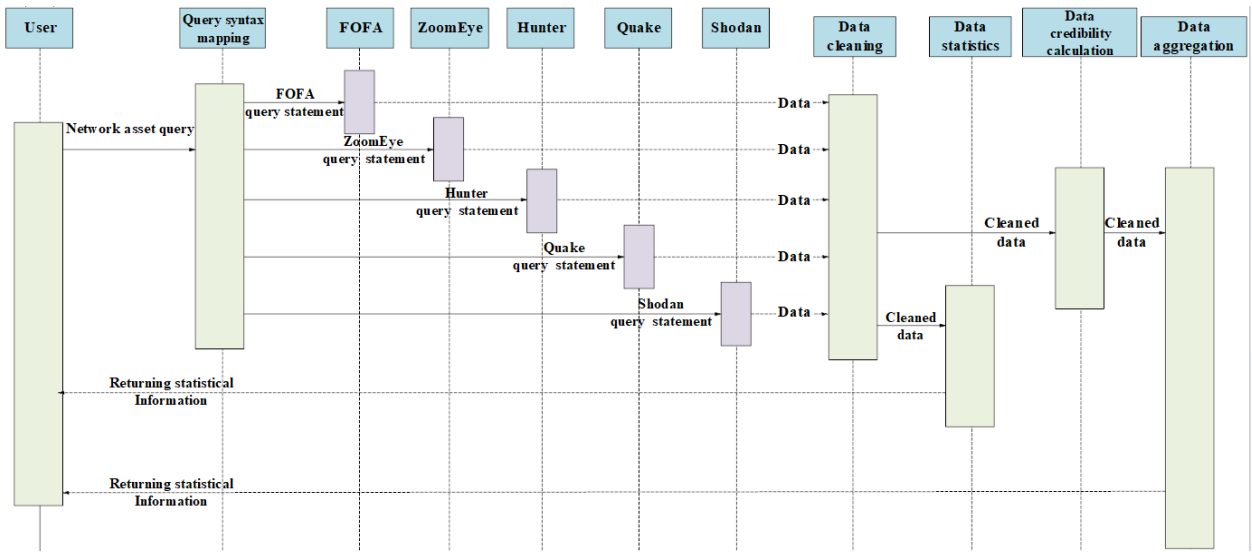

Fig. 2. System workflow overview.

D. Illustration of the Data Aggregation Process

Based on the preceding discussion, it is apparent that data aggregation plays a pivotal role in the system's functionality. To elucidate this, consider a hypothetical scenario involving a query for open ports on network assets located in Taipei, China. The system harvests information from multiple sources, such as FOFA, ZoomEye, and Shodan, which yield varied results for the IP address 114.32.164.140:

(1) FOFA identifies port 80 (HTTP) as open.
(2) ZoomEye detects ports 80 (HTTP) and 443 (HTTPS) as open.
(3) Shodan confirms that port 443 (HTTPS) is open.

Initially, the system standardizes the data into a consistent format to ensure a uniform representation across different tools. For example, while FOFA reports port = "80" and ZoomEye denotes protocol = "HTTP", both entries are normalized to port = 80 (HTTP).

Subsequently, the system assesses the credibility of the data based on historical reliability scores of the respective tools: FOFA is assigned a reliability score of $t(FOFA)=0.90$, ZoomEye receives $t(ZoomEye)=0.85$, and Shodan has $t(shodan)=0.80$. The credibility of each piece of information is computed, resulting in $s(f80)=1-(1-0.9)\times(1-0.85)=0.985$ for the data point "Port 80 open" and $s(f443)=1-(1-0.80)\times(1-0.87)=0.97$ for "Port 443 open."

Following this, the system aggregates the results, resolving discrepancies by prioritizing data with higher credibility scores. For the IP address 114.32.164.140, both ports 80 (HTTP) and 443 (HTTPS) are retained due to their substantial credibility, while data of lower credibility is excluded.

Ultimately, the aggregated data is transmitted to the frontend interface, where it is depicted through various visual aids. These include heatmaps that display geographic concentrations of IP addresses, radar charts that compare the reliability of different tools, and detailed tables that list IP information, open ports, protocols, and credibility scores. This comprehensive presentation enables users to make well-informed decisions based on robust data aggregation processes.

E. System Architecture

To ensure seamless integration and efficient execution of workflow steps, the system architecture has been meticulously designed with a three-tier structure. This layered approach effectively addresses the challenges associated with processing complex, multi-source data, thereby enhancing modularity, scalability, and maintainability. The architecture, as illustrated in Fig. 3, assigns specific functions to each layer to support the overarching system workflow.

The system begins by acquiring raw network asset data from various cyberspace mapping tools through the Data Source Layer, which includes the Query Statement Parsing Module and the Data Acquisition Module. The use of multiple cyberspace mapping tools, each requiring distinct query syntax, introduces a level of complexity that can render the process cumbersome and error-prone for users. To mitigate this, the system is equipped to standardize query fields.

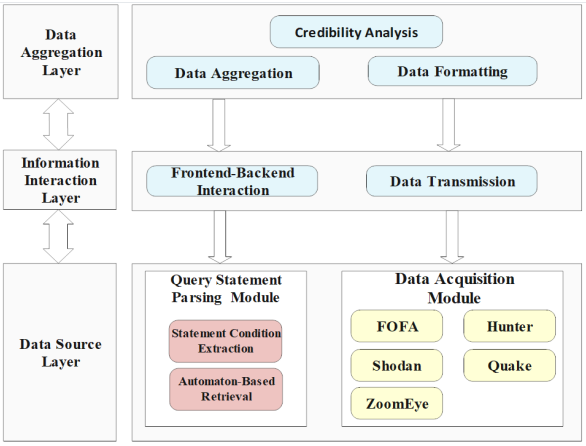

Fig. 3. System architecture overview.

The Query Statement Parsing Module is pivotal in transforming user queries. It scrutinizes user inputs, extracting essential fields and values while discerning the intent of the query. This ensures compliance with each tool's specific syntax requirements and defines the scope of data retrieval.

To enhance efficiency, the system utilizes automata technology, specifically the Aho-Corasick (AC) automaton—A renowned multi-pattern string matching algorithm that is highly effective in text processing and adept at managing multiple patterns simultaneously. The choice of the AC automaton over alternatives such as Deterministic Finite Automata (DFA) and Pushdown Automata (PDA) is justified by its superior performance in multi-pattern matching. It processes user queries by aligning key fields with the specific syntax of each tool, thus facilitating efficient syntax conversion across various tools and enabling precise data retrieval.

Upon parsing a query, the system deconstructs user-submitted queries (e.g., "os = Windows") into three components: field name, operator, and value. It then employs a mapping table to translate field names (e.g., "os") into the formats required by each specific tool (e.g., FOFA as 'os' and Hunter as 'ip.os'). The AC automaton processes inputs character by character, efficiently matching and converting fields through optimized state transitions.

The Data Acquisition Module retrieves data from diverse cyberspace mapping tools via API. It leverages multithreading to dispatch queries concurrently, thereby minimizing response times.

The subsequent section elaborates on the design and implementation of the system's user interface. The front end of the system is developed using Vue3, which, compared to its predecessors, enhances both functionality and performance, significantly improving user experience. Network requests are managed by axios, which facilitates asynchronous operations.

F. User Interface

The user interface functions as the system's primary interface, seamlessly integrating backend data processing with user-friendly visualization tools to provide actionable insights. It is composed of several essential elements, including the search interface, data overview, and detailed data views. These elements are developed using a modular design, enhancing both maintainability and reusability. The search interface enables users to enter queries, offers query syntax assistance, and features a search history component. For instance, a user wishing to locate network assets in Taipei, China might enter the query city = "Taipei", as illustrated in Fig. 4 The interface provides guidance on search syntax to facilitate ease of use and maintains a log of previous searches for user reference.

Fig. 4. Search interface overview.

The data overview interface displays the outcomes of user queries, as depicted in Fig. 5 Using the query for Taipei, China as an example, the left and top portions of the interface deliver visualized data insights, allowing users to effectively discern the distribution of network assets. The key visual elements include:

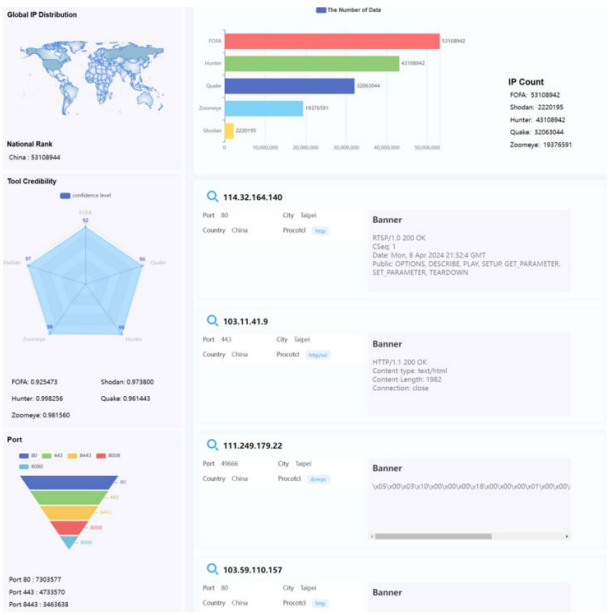

Fig. 5. Mapping results visualization.

(1) IP Distribution: A global heatmap accentuates the concentration of IP addresses in Taipei, China, providing a macroscopic perspective of network activities in the region.
(2) Country Rankings: A bar chart distinctly categorizes Taiwan, highlighting the relative

density of network assets in Taipei, China, in comparison to other areas.

(3) Tool Credibility: A radar chart evaluates the reliability of various tools such as FOFA, ZoomEye, Shodan, Hunter, and Quake for the specific query related to Taipei, China. For instance, FOFA and Quake are shown to offer superior dependability for this search.

(4) Top Ports and Protocols: A funnel chart enumerates frequently detected ports (e.g., port 443 and port 80) and the corresponding protocols (e.g., HTTPS and HTTP), underscoring their significance in the assets located in Taipei, China.

In the bottom-right section, the interface furnishes a detailed list of queried assets in Chinese Taipei, which includes:

(1) IP: Addresses such as 114.32.164.140 and 103.11.41.9, among others relevant to the query.

(2) Port: Common ports like 80, 443, and 3389, typically associated with HTTP, HTTPS, and RDP services respectively.

(3) City and Country: All results pertain to Taipei, China.

(4) Protocol: Protocols such as HTTP and HTTPS are documented.

(5) Banner: Metadata offering detailed service information, exemplified by entries like Microsoft Remote Desktop or Nginx.

The query results pertaining to Taipei, China, facilitate the rapid acquisition of actionable insights by users. For example, users can identify high-risk open ports or evaluate the reliability of the sources providing data about the region in question.

To enhance the comprehension of the data, Fig. 6 presents an in-depth visualization of individual network assets associated with Taipei, China. This figure delineates several critical features:

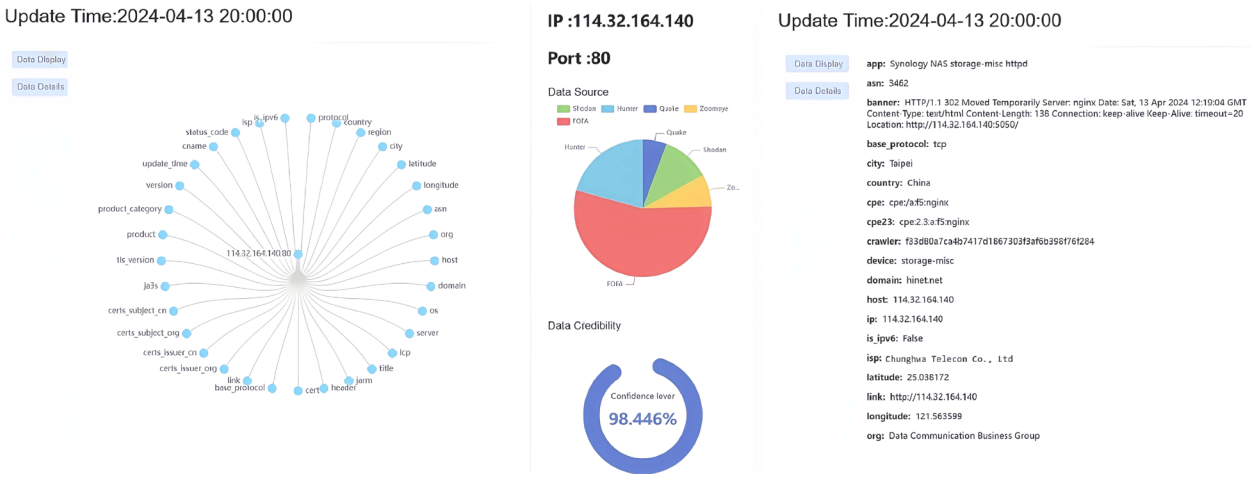

Fig. 6. Comprehensive data visualization and insights for Taipei, China query results.

Detailed Asset Information: Each IP address is accompanied by extensive attributes, such as the country, city, associated ports, protocols, and banner details. This detailed enumeration aids users in precisely assessing specific network elements.

Confidence Level Visualization: A pie chart delineates the confidence levels assigned to each tool contributing data on the queried assets. For instance, the data linked to IP address 114.32.164.140 indicates a confidence level of 98.446%, reflecting high reliability.

Attribute Mapping: A radial tree diagram portrays the relationships among various attributes of the queried assets. This diagram includes data points such as domain, host, protocol, longitude, and latitude, providing a comprehensive view of the asset's configuration and its environmental context.

IV. Experiment Result and Discussion

A. Experiment Objective

The experiment is designed to verify the accuracy and stability of the core functionalities of the system, identify potential defects, and test critical features such as query syntax mapping, data retrieval, and data aggregation. As outlined in Table I, these tests assess the system's capability to process user inputs, retrieve network assets, and aggregate data effectively. Comprehensive testing further seeks to identify issues across diverse use cases, thus laying the groundwork for system optimization and enhanced performance. For practical relevance, the experiment employed asset data from a specified region in Taiwan, China, providing insights into the system's real-world applicability.

TABLE I. System Functional Requirements Table

Function	Details		
	Functional Description	Input	Output
Query Syntax Conversion	Converts user query fields to each tool's fields	User's query fields	Corresponding query fields for each tool
Data Retrieval	Retrieves data from each tool using API	Query statements for each tool	tool-specific query statements
Data Aggregation	Aggregates results and selects the most credible data	Data to be aggregated	Field info with highest credibility

B. Experimental Environment and Setup

To enhance the reproducibility of the experiment, it is crucial to clearly define the project's dependencies, as detailed in Table II. This table enumerates the essential packages and configurations necessary for the experimental setup.

TABLE II. SYSTEM DEPENDENCY ENVIRONMENT

Package Name	Version
Python	v3.9.7
Baidu-aipl	v4.16.11
Pip	v23.1.2
Censys	v2.2.2
FOFA	v1.0.2
Requests	v2.30.0
ZoomEye	v2.2.0
ZoomEye-sdk	v1.0.6
Shodan	v1.28.0
Json	v2.0.9

C. Experimental Metrics and Formulas

1) Data coverage

Coverage is defined as:

$$Coverage = \frac{N_{retrieved}}{N_{total}} \times 100\% \quad (12)$$

where:

$N_{retrieved}$: Number of successfully retrieved entries.

N_{total}: Total number of entries mapped by this system.

This metric evaluates the data collection capabilities of the system by quantifying the extent to which it captures relevant data from various tools.

2) Data credibility

The computation of credibility, as detailed in Section 3, is employed in the experiment to compare the performance of different cyberspace surveying and mapping tools. It serves as a foundation for evaluating tool reliability, resolving discrepancies in data, and assessing overall data quality, thereby ensuring a thorough analysis of tool effectiveness.

D. Experimental Results and Conclusions

The system utilizes an automaton-based optimization for transforming query syntax, which accelerates the conversion speed compared to the conventional loop method. As illustrated in Fig. 7, this optimization enhances the data processing capabilities, particularly in large-scale data retrieval operations.

Fig. 7. Speed comparison between automaton and loop method.

From Fig. 7, it is evident that the advantages of using an automaton for field conversion increase as the data volume grows. With 1000 entries, the performance difference between the automaton and the loop method is negligible; however, with 10,000 entries, the automaton significantly outperforms the loop method.

To evaluate the data collection capacity, experiments were conducted with 2000 entries from a specific region, utilizing five different tools, as demonstrated in Fig. 8.

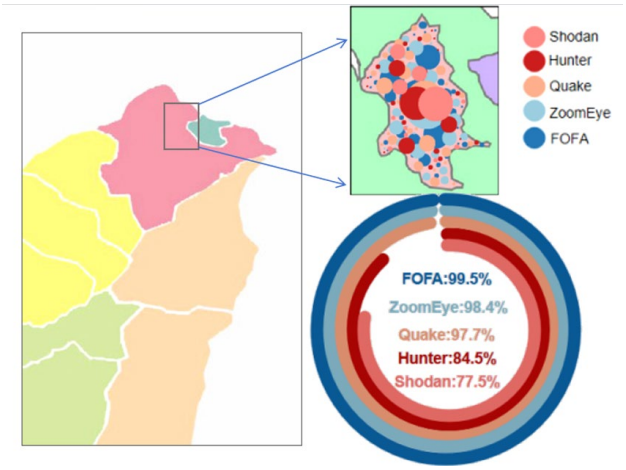

Fig. 8. Comparison of data collection capabilities across five mapping tools.

As depicted in Fig. 8, Fofa achieved the highest retrieval rate, detecting 1999 out of 2000 entries, which corresponds to a coverage rate of 99.95%. The tools were ranked in terms of coverage as follows: Fofa, ZoomEye, Quake, Hunter, and Shodan. Fofa's high detection rate highlights its robustness in data collection volume, while Shodan recorded the fewest entries, likely due to policy restrictions that limit its scanning scope.

To assess the system's ability to evaluate the credibility of cyberspace mapping tools and achieve optimal data aggregation, an analysis was conducted using 2000 entries from the aforementioned region across the five tools. The results underwent a credibility analysis, followed by optimal data aggregation based on calculated credibility scores. As shown in Fig. 9, this process was applied to the 2000 data entries, resulting in credibility scores for both the system and each individual tool. The integrated system achieved the highest credibility score of 99.9337%, surpassing all five individual tools. The rankings among the tools were Hunter, ZoomEye, Shodan, Quake, and FOFA, with FOFA scoring the lowest at 92.5473%.

Notably, although the Hunter system attained an impressive credibility score of 99.33%, its coverage rate was only 84.5%, which did not meet the performance metrics of our system. Furthermore, the aggregation process significantly enhances the overall credibility across all tools, with an average increase of 3.33%. By synthesizing data from multiple sources, our system not only achieves high credibility but also broadens its coverage.

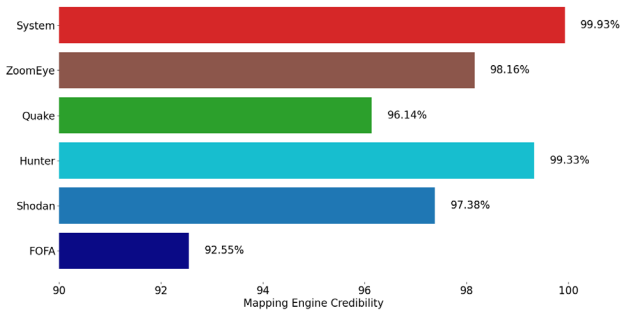

Fig. 9. Comparison of credibility across different mapping tools.

The experimental outcomes corroborate the system's capability to amalgamate data from diverse tools, thereby augmenting both data coverage and credibility. By harnessing the complementary strengths and resolving discrepancies among data sources, the integrated system surpasses individual tools in terms of both accuracy and reliability.

However, these experiments also exposed certain limitations. The process of credibility analysis during large-scale data aggregation demands substantial computational resources, potentially compromising performance in scenarios involving continuous data streams. The scope of evaluation was confined to a specific region and dataset size, which leaves the performance of the system on a larger scale and in varied global contexts uncertain. Moreover, the dependence on third-party API introduces an element of instability, as variations in data retrieval may occur due to tool-specific policies and network conditions.

These challenges underscore the necessity for optimizing computational efficiency and broadening the scope of testing to include diverse datasets. Future experiments aim to extend evaluations on a global scale and to enhance the system's adaptability and robustness to ensure consistent performance under varied conditions.

V. Conclusion and Future Work

The proposed system effectively addresses several challenges in cyberspace surveying and mapping, utilizing an integrated multi-source approach to significantly improve query syntax consistency, data credibility assessment, and energy efficiency. By employing data fusion techniques, the system mitigates the limitations associated with using individual tools and combines broader scanning coverage with more frequent updates, culminating in a more reliable cyberspace mapping framework.

Nevertheless, certain limitations persist. The real-time data credibility assessment mechanism can be computationally demanding, which might affect performance, especially when processing extensive datasets or managing continuous data inflows.

Future research will focus on the incorporation of deep learning models for dynamic data credibility assessment to enable real-time adaptability to fluctuations in data quality. Further optimization for handling large datasets will explore the use of distributed computing and advanced indexing techniques to alleviate computational burdens and enhance system responsiveness. Additionally, real-time data integration will be a key area of development, aiming to minimize delays and ensure seamless data flow.

Concurrently, we are seeking partnerships with industry professionals and annotators to validate our findings and evaluate their applicability in real-world settings. This validation process will provide critical insights into the practical challenges of deploying our system across various industry contexts, thereby refining its functionality. We are also considering partnerships for pilot studies and exploring industry-specific case studies to bolster the external validity of our results.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yingjie Wang wrote the paper; Zhe Wang performed the experiments; Hongjie Fan analyzed the data; Songtao Ye proposed the problem and solution. All authors had approved the final version.

Funding

This work was supported by the Outstanding Youth project of the Hunan Provincial Education Department (No.22B0149). This work is supported by the China University of Political Science and Law Research Innovation Project (Grant No. 24KYGH013), and the Fundamental Research Funds for the Central Universities. And this work is sponsored by Hunan Students' Platform for innovation and entrepreneurship training program (S202410530311, S202410530327), and the Hunan Higher Education Teaching Reform Project (HNJG-20230307).

References

[1] S. AlDaajeh *et al.*, "The role of national cybersecurity strategies on the improvement of cybersecurity education," *Computers & Security*, vol. 119, 102754, Aug. 2022.

[2] Z. Zou, J. Chen, W. Wu, B. Wang, and Y. Liu, "Survey of cyberspace surveying and mapping," in *Proc. 2024 3rd Int. Conf. Cyber Security, Artificial Intelligence and Digital Economy*, 2024, pp. 66–71.

[3] T. Yang, L. Chen, J. Liang, Q. Wang, and W. Zhang, "Cross-source data fusion in industrial automation," in *Proc. 2023 3rd Int. Conf. Mobile Networks and Wireless Communications (ICMNWC)*, 2023, pp. 1–7.

[4] F. Z. Rozony, M. N. A. Aktar, M. Ashrafuzzaman, and A. Islam, "A systematic review of big data integration challenges and solutions for heterogeneous data sources," *Academic Journal on Business Administration, Innovation & Sustainability*, vol. 4, no. 04, pp. 1–18, Oct. 2024.

[5] A. Protopsaltis, P. Sarigiannidis, D. Margounakis, and A. Lytos, "Data visualization in internet of things: Tools, methodologies, and challenges," in *Proc. 15th Int. Conf. Availability, Reliability and Security*, 2020, pp. 1–11.

[6] A. K. Shaw, "Next-generation cyber threat intelligence platform," Ph.D. dissertation, Marymount Univ., Arlington, VA, 2024.

[7] T. Wen, "Data Aggregation," in *Encyclopedia of Big Data*, Cham: Springer, 2020, ch. 1, pp. 1–4.

[8] A. Walha, F. Ghozzi, and F. Gargouri, "Data integration from traditional to big data: Main features and comparisons of ETL

approaches," *The Journal of Supercomputing*, vol. 80, no. 19, pp. 26687–26725, Sep. 2024.
[9] A. Gimaletdinova, "An in-depth comparative study of distributed data processing frameworks: Apache spark, apache flink, and hadoop mapreduce," *Вестник науки*, vol. 3, no. 4, 73, pp. 364–377, April 2024.
[10] M. S. Pour, C. Nader, K. Friday, and E. Bou-Harb, "A comprehensive survey of recent internet measurement techniques for cyber security," *Computers & Security*, vol. 128, 103123, May 2023.
[11] D. Everson and L. Cheng, "A survey on network attack surface mapping," *Digital Threats: Research and Practice*, vol. 5, no. 2, pp. 1–25, June 2024.
[12] M. Alhamed and M. H. Rahman, "A systematic literature review on penetration testing in networks: Future research directions," *Applied Sciences*, vol. 13, no. 12, 6986, June 2023.
[13] Y. Chen *et al.*, "Exploring shodan from the perspective of industrial control systems," *IEEE Access*, vol. 8, pp. 75359–75369, April 2020.
[14] Z. Zou, Y. Hou, and Q. Guo, "Research on cyberspace surveying and mapping technology based on asset detection," in *Proc. 2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conf. (IMCEC)*, 2024, pp. 946–949.
[15] R. Li *et al.*, "A survey on cyberspace search engines," in *Proc. Cyber Security: 17th China Annual Conf., CNCERT 2020*, 2020, pp. 206–214.
[16] S. G. K. Patro *et al.*, "Brain tumor classification using an ensemble of deep learning techniques," *IEEE Access,* vol. 12, pp. 162094–162106, Oct. 2024.
[17] S. E. H. Hassan and N. Duong-Trung, "Machine learning in cybersecurity: Advanced detection and classification techniques for network traffic environments," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 11, no. 3, pp. 1–22, July 2024.
[18] Y. Yao *et al.*, "Prescribed-time output feedback control for cyber-physical systems under output constraints and malicious attacks," *IEEE Transactions on Cybernetics*, vol. 54, no. 11, pp. 6518–6530, Nov. 2024.
[19] D. Verma and P. P. Dewani, "eWOM credibility: A comprehensive framework and literature review," *Online Information Review*, vol. 45, no. 3, pp. 481–500, May 2021.
[20] M. M. Yamin *et al.*, "Mapping tools for open source intelligence with cyber kill chain for adversarial aware security," *Mathematics*, vol. 10, no. 12, 2054, June 2022.
[21] A. Goel, A. K. Goel, and A. Kumar, "The role of artificial neural network and machine learning in utilizing spatial information," *Spatial Information Research*, vol. 31, no. 3, pp. 275–285, Nov. 2022.
[22] Y. Xu *et al.*, "Physics-informed machine learning for reliability and systems safety applications: State of the art and challenges," *Reliability Engineering & System Safety*, vol. 230, 108900, 2023.
[23] M. Ammi *et al.*, "Leveraging a cloud-native architecture to enable semantic interconnectedness of data for cyber threat intelligence," *Cluster Computing*, vol. 25, no. 5, pp. 3629–3640, April 2022.

Copyright © 2025 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).