Big Data Analysis of Threat Intelligence Based on IP

Fajian Xu* and Shuzhen Chen
Department of Computer and Information Security Management, Fujian Police College, Fuzhou, Fujian 350007, China

*Corresponding author e-mail: xufaxian@fjpsc.edu.cn

Abstract. This paper first discusses the concept of network threat intelligence and related resource integration, and proposes a three-tier architecture model of threat intelligence analysis system based on IP. Then, it discusses the diversified data acquisition technology, information index extraction technology, data mining and data modeling analysis technology, the key technology of threat intelligence credible intelligence analysis model based on machine learning and privacy protection. Finally, this paper considers and summarizes the legal standardization of threat intelligence analysis.

Keywords: Network Threat Intelligence, IP Based Threat Intelligence, Trusted Intelligence Analysis Model, Privacy Protection

1. Introduction
Threat intelligence in broad sense is a set of knowledge based on evidence to describe security threats, including situation, mechanism, index and inference, providing operational suggestions and reference for organizations or decision makers to protect assets [1]. Threat intelligence mainly provides clues to restore or predict future attacks and other security issues, and integrates and shares threat information resources to counter the ever-changing security threats [2].

2. Cyber Threat Intelligence
The threat intelligence in cyberspace security can be understood as a basic and comprehensive network security prevention work. Combined with the threat intelligence data in the field of network security, the threat sources are collected, analyzed, classified and correlated in real time to evaluate the threats existing in the enterprise assets and provide suggestions for improving the security situation [3]. Threat intelligence is the knowledge to describe threats based on specific scenarios, mechanisms and other elements. It needs to be closely related to the actual network and application environment, so that users can fully evaluate the network security status [4].

2.1. Cyber Threat Intelligence is a Security Service
According to the past research and experience, many loopholes are not directly from the assets, but due to the leakage of personal information of enterprise employees. Through the employee and enterprise related mailbox, OA, VPN account intrusion, and then intrusion into the enterprise server, and even can use social engineering. Therefore, sans Institute defines Threat intelligence as: against security threats,
threats, exploits, malware, vulnerabilities and hazard indicators, and data sets collected for evaluation applications [5]. Vulnerability database, fingerprint database and IP reputation database are all part of threat intelligence. Threat intelligence can effectively identify potential threats in assets by combining global threat intelligence assessment and automatic detection of similar threat events [6]. Threat intelligence can associate, analyze and process the relevant contents of threat information, and provide the correlation and analysis results to the security equipment and relevant security personnel for use, so as to improve the efficiency and response speed of threat inspection [7].

2.2. Data Resource Integration of Network Threat Intelligence
The basis of network threat intelligence is the integration of data resources to form a big data threat intelligence platform. Specifically, it assets perspective (internal resources), global website search resources, social workers database, security personnel and incident database and other targeted threat intelligence. In addition, there are sub domain name database, URL database, email address database, global website fingerprint database, black sample file fingerprint database and other basic network security databases, which need to spend a lot of energy to accumulate and integrate, and constantly track and update.

The perspective of enterprise IT assets is very valuable. Many large enterprises do not know their own assets, such as how many servers, how many IP, what business each server carries, and so on. Once a 1day vulnerability is found, it is not so fast to patch your own server. It asset perspective enables enterprises to understand and master their own assets. By tagging each asset, data resource and important administrator, it is easy to retrieve and analyze the weak points and artificial information that cause enterprise information leakage, so as to lay a solid foundation for enterprises to prevent security risks and timely remedy risks.

We also need to integrate search resources of global websites like “Shodan” to obtain the fingerprint of server ports around the world, including the identification of HTTP or MySQL on this port, and the corresponding version information. Combined with the specific domain resources of the service object and the internal resources of some cooperative units, it can identify which port a website has which special applications and services you are concerned about, including which kind of web to use server, which programming language, which open source framework (such as CMS); through these functions, you can find the online squid proxy server in the field of concern, which websites use jQuery, which online H3C routers, which websites use Geo trust certificate, which websites hang a certain Trojan horse, which websites use cloudflare's CDN, etc. Only when these big data resources are integrated with the social worker database, security personnel and event database, can the analysis ability of network threat intelligence be gradually formed.

3. Architecture of Threat Intelligence Analysis System Based on IP
Our defense idea should evolve from the vulnerability based method to the threat centered method, and the threat intelligence analysis based on IP is the main work. Through the threat intelligence based on IP, we can predict the attack events, and the core of all this is to master the massive data and form a system analysis platform, so as to realize the powerful data analysis ability.

3.1. Threat Intelligence Analysis Based on IP
Accurate and timely labeling of unsafe data in the network can help users quickly deal with threats that have occurred or are occurring, such as hash of black samples, external connection of C & C and IP or domain name of downloader server; network boundary devices or agents running on the host can find and adopt automatic countermeasures by analyzing and matching these labeling data. When we try to analyze a suspicious event, threat intelligence can provide useful reference for users to determine the malicious nature of the suspicious event, such as whether the IP involved in the event is in some known blacklists, and whether the related domain name is used by known apt activities, etc. [8]. The information related to IP is the protagonist. It also needs to combine various application technologies,
app, tools and platform information related to IP, as well as the corresponding information resources at various levels such as events, organizations, personnel, etc.

In the analysis of threat intelligence based on IP, IP intelligence is the clue, and threat intelligence is all the clues needed to integrate resources in order to restore the attacks that have occurred and predict the attacks that have not occurred [9]. WAF, scanner and vulnerability management platform can exchange vulnerability information, antivirus and intrusion detection can exchange malicious sample information, business fraud and network attack and defense can exchange IP reputation information. All these are to achieve internal resources (scanner, WAF, IPS and other security components) and external resources (open source resource collection, manufacturer resource exchange). In order to have the ability to conduct in-depth analysis to find out the real valuable attack events and the advanced apt fixed-point attack events which are difficult to find out, only by the integration of collaborative defense effect can we have the ability to conduct in-depth analysis. Therefore, the threat intelligence analysis based on IP is to help us find all kinds of threats related to IP in the network, and provide the corresponding knowledge and methods to deal with them.

A complete threat intelligence cycle includes six links: plan making, information collection, processing, fusion analysis, intelligence application and result feedback [10]. Therefore, the general architecture of IP based threat intelligence analysis big data platform includes: data acquisition layer, access processing layer, storage management layer, mining analysis layer, shared output layer, application service layer, etc. [11]. The following will analyze the three core function modules and architecture.

3.2. System Architecture

By integrating the three modules of threat intelligence collection subsystem, security intelligence big data center and threat intelligence query and analysis subsystem (as shown in Figure 1), a threat intelligence platform based on big data is constructed, which can strengthen the system in advance, study and judge threats, preset defense strategies, and actively avoid threats. In the event, the detection and defense capabilities of security equipment can be enhanced in time. At the same time, it can sense the threat development situation and grasp the attack context in real time, so as to adjust the defense strategy and respond quickly; after the event, it can support the comprehensive tracing, forensics and countermeasures of the attack behavior and the attacker.

Fig 1. Three layer architecture of threat intelligence system

3.2.1. Threat intelligence collection subsystem. Threat intelligence collection subsystem includes intelligence query collection agent, internet intelligence monitoring collection and IOC data extraction. The subsystem is one of the main inputs of threat intelligence data for threat intelligence platform. The security intelligence big data center obtains external threat intelligence data through this subsystem.

The subsystem uses crawler and manual collection to collect data and extract IOC data into database. Then, the audit system is used to manually audit, edit, modify and publish the data to the
The 2nd International Conference on Computing and Data Science (CONF-CDS 2021)  
Journal of Physics: Conference Series 1881 (2021) 032067  
doi:10.1088/1742-6596/1881/3/032067

business portal for display. The intelligence data collected by the crawler is reported to the security intelligence big data center. Combined with the existing IT asset perspective (internal resources), global website search resources, social worker database, security personnel and event database and other threat intelligence, the network security threat intelligence big data resources are formed.

3.2.2. Network threat intelligence big data center. It also supports massive data collection, data security and threat analysis by using large amount of data security indicators such as big intelligence, big data storage and data warehouse Consistency definition and management. It adopts distributed data flow management engine, distributed computing resource elastic scaling architecture, supports deployment in private cloud, public cloud and hybrid cloud, and supports seamless migration between different cloud platforms to meet the guarantee of high availability and high reliability.

3.2.3. Threat intelligence service and decision support subsystem. After the threat intelligence collection subsystem preprocesses the massive heterogeneous data, it forms a unified and integrated Threat Intelligence basic database. Through the establishment of a multidimensional analysis model, it provides a variety of data analysis functions, such as adding calculation indicators, slicing and slicing data, data linkage between components, data early warning, data rolling and drilling down, data dimension rotation, etc.

The subsystem is a window for threat intelligence platform to provide external intelligence services, assist security personnel to understand and trace intelligence knowledge, complete online risk assessment services of target objects, realize the monitoring and display of threat intelligence news information, global Internet side security incident report and early warning information, and provide threat intelligence online application services, including threat intelligence query analysis and threat It provides various data transmission interfaces to provide threat intelligence query for internal and external systems, and supports threat intelligence query and use intelligence query of internal and external systems in different business and application scenarios.

4. Key Technologies of Threat Intelligence Analysis
The process of threat intelligence analysis includes five levels (as shown in Figure 2).

Fig 2. Flow of threat intelligence analysis
Whether you use FPC / FPI, DPI / DFI, SOC, SIEM / Sandbox / Honey net and other technologies, you can use them as machine automatic learning and analysis; for some large service objects, if policy
allows, you can also use desktop applications to collect and obtain more detailed privacy data for machine learning and automatic analysis.

There are many unnecessary or duplicate data in the collected threat intelligence, and the definitions of threat intelligence by various intelligence source manufacturers are also different. Therefore, how to identify the same or related threat intelligence information becomes a difficult problem in the process of intelligence collection.

Due to the different classification of threat intelligence, the requirements of different business requirements on the field are relatively complex. Therefore, it is necessary to build a standardized threat intelligence data classification and inclusion specification, classify the data to be accessed, and provide the data access field and format requirements, data examples, to provide a standardized basis for intelligence collection and platform storage database construction. The key technologies are analyzed from four aspects.

4.1. Diversified Data Acquisition Technology
In order to ensure the comprehensiveness and reliability of threat intelligence information, it should be able to support multiple sources of data, provide different types of data acquisition methods, and meet the technical requirements of supporting intelligence query and collection agent, Internet intelligence monitoring and collection, etc.

Comprehensive collect internet security intelligence blog information, open threat intelligence sources and other information, to provide support for threat intelligence risk analysis. On the one hand, data collection is carried out by deploying probes; on the other hand, active and passive technologies are used to collect heterogeneous massive intelligence data in the whole network, and log normalization processing and information are conducted through extraction, cleaning, conversion, comparison, enrichment and eliminate duplication report data classification to provide rich data sources for the platform.

4.2. Information Index Extraction Technology
It is necessary to obtain the latest malware means, malicious domain name, website, IP address and host based violation indicators from the collected network threat information to provide intelligence feedback related to security threats. For the existing regular expression matching, it is often unable to capture these things, such as IP address: 127 [. ] 0 [. ] 1. It is necessary to use specific regular expression and anti obfuscation monitoring technology according to the characteristics of Threat Intelligence indicators to extract these confused or damaged intelligence indicators, and even restore them to the original state, so as to provide guarantee for the effectiveness of intelligence [11].

4.3. Threat Intelligence Data Mining and Data Modeling Analysis Technology
Through multi-dimensional association, statistics and mining analysis of threat information data, new threat information can be found continuously, and high-level threat information can be generated to support the application services of the upper level. On the one hand, association analysis can be carried out on threat information from various dimensions of threat intelligence, and the correlation between threat intelligence can be found, and threat intelligence relation map can be generated. On the other hand, statistical analysis is carried out on the quantity, distribution and frequency of all kinds of threat intelligence information from different dimensions such as source, time, threat target, etc., which can comprehensively grasp the traceability information of threat intelligence occurrence, and predict the development trend, which can be used for the judgment, interpretation and decision-making of security situation.

4.4. Credible Intelligence Analysis Model of Threat Intelligence Based on Machine Learning
The IOC is used to extract the collected factual intelligence data to form a series of character relationship and fact description, involving the information of the person, time, place and event of intelligence occurrence, and to construct the knowledge map basic triple data of entity, relationship and other
attribute information. Machine learning is used to construct the knowledge identification model, and the continuous bottom dimensional vector space mapped by IP based threat intelligence knowledge is obtained, and the credibility evaluation of intelligence is carried out on this basis. In order to establish a credible analysis model, a large number of threat intelligence data sets are needed for model training. By constantly changing the training data sets, the average absolute error in each case is calculated to improve the credibility of threat intelligence.

5. Legal Standardization of Threat Intelligence Analysis

In the big data environment, with the rapid development of mobile Internet, the use of big data analysis tools can facilitate the collection and sharing of network logs and mobile terminal data, which inevitably involves user personal information. Therefore, the large-scale and deep data mining of threat intelligence analysis will aggravate the contradiction with the protection of personal privacy information. It is necessary to analyze the causes of conflicts from three aspects: privacy disclosure caused by large-scale data transfer, threat intelligence collection and processing technology, and imperfect legislative protection of personal privacy information, and gradually propose solutions to the causes of conflicts [12]. How to strengthen the control and privacy protection of threat intelligence data will become the key issue of threat Intelligence Research and application in the future [9].

6. Summary

Threat intelligence is an important way for enterprises to know themselves and their competitors. It is against the times to do information security behind closed doors. Technology changes every day, network threat intelligence can synchronize enterprise information with hacker information, and provide great help for information security construction. Threat intelligence can take different measures according to different types of intelligence, which can repair loopholes, make enterprises know about hostile hacker teams and prevent in advance; some threat intelligence can also let enterprises change passwords, change servers, and strengthen defense system. Threat intelligence is not only a technical problem, but also a strategic issue in its analysis and application. For example, in the execution process of apt confrontation, is it protection or monitoring? We also need to choose the timeliness of attacks and the continuity of protection, because most threat intelligence has time limits. I believe that with the development of technology, threat intelligence can provide more and more help, but we should also pay attention to privacy protection issues.

Acknowledgements

1. "Development and industrialization of threat intelligence platform based on big data intelligent analysis" (2020H6024), a science and technology project of university industry cooperation in Fujian Province in 2020;
2. "Research on governance and legal issues of intelligent mobile terminal" FLS(2019) A08, a key project of Fujian law society in 2019.

References

[1] Weimin Wei, etc. Based on large data security gateway rib intelligence research. Shanghai electric power college journal, 2018, 1:53-58.
[2] Bo Mingxia, Feng miaodong. Research on technology architecture of security information analysis and sharing platform based on big data. Telecommunication technology, 2019, 11:5-9.
[3] Liang Guo, Zeus yu. Threat analysis and intelligent assets management platform Information technology and standardization, 2018, 9:81-84.
[4] Zhang Chen. Discussion on the implementation of threat intelligence analysis in the era of "Internet of things". Telecommunication network technology,2017,4(4):56-57.
[5] Chen Jianfeng Fan Hang bo. Facing the threat of cyber security intelligence ontology sharing research. Communication technology, 2018, 1:171-177.
[6] Yanfang hou, Jinhua wang. Based on the updated threat information library of data security analysis method. Telecom science, 2018 (3)

[7] Shi Zhixin et al. Review of threat intelligence standards. Information security research, 2019, 7:560-569.

[8] Peian Yang, etc. Cyberspace threats intelligence sharing technology review. Computer science, 2018, 6:9-18.

[9] Sun Hui et al. Research on the application of big data technology in information network threat intelligence. Information technology and network security, 2020, 5:28-32.

[10] Li Chao. Threat Intelligence analysis in big data environment. Intelligence journal, 2017, 36 (9): 24-30.

[11] Hua Berlin, Li Guangjian. Construction of intelligence methodology system oriented to intelligence process. Journal of information science, 2016, 35(2):177-188.

[12] Huang Wei. Anti-terrorism intelligence work and personal privacy information protection in the era of big data. Library and intelligence, 2018, 4:43-50.