PROCEEDINGS OF SCIENCE

A comprehensive security operation center based on big data analytics and threat intelligence

Jiarong Wang,[a,*] Tian Yan,[a] Dehai An,[a] Zhongtian Liang,[a] Chaoqi Guo,[a] Hao Hu,[a] Qi Luo,[a] Hongtao Li,[a] Han Wang,[a] Shan Zeng,[a] Caiqiu Zhou,[a] Lanxin Ma[a] and Fazhi Qi[a]

[a] *Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, P.R.China*

E-mail: wangjr@ihep.ac.cn,yant@ihep.ac.cn

The continued growth of cybersecurity incidents calls for effective cybersecurity monitoring solutions. The operation of security operation centers (SOCs) is the recommended best practice to which large and medium-size organizations rely for the detection, notification, and ultimately response to cybersecurity incidents. However, current SOCs face several issues, such as inferior defense against specific types of attacks, low-quality threat intelligence, low speed of response and low level of automation.

In this paper, a comprehensive SOC is introduced to mitigate above mentioned issues of current SOCs. First, the SOC collects a wide variety of data including network traffic, server logs, security incidents logs. The collected data is preprocessed and stored in a big-data storage platform. Secondly, the SOC provides multi-perspective behavior analysis which can combine the detection performance of multiple behavior detectors. Different detectors can analyze different and specific types of attack based on the data on the big data storage platform. Besides, threat intelligence is collected accurately from unstructured open-source cyber threat intelligence reports by using deep learning model and is correlated with incidents detection to identify attacks rapidly. Finally, the SOC can uniformly manage and automatically respond the incidents identified from multi-perspective behavior analysis and threat intelligence. At the same time, visualization is adopted to reveal the cybersecurity situation of entire organizations or enterprise. The framework of the SOC is derived from the CERN design, and is customized to make it is practical and deployable for the Institute of High Energy Physics to discover, identify, understand, analyze, and respond to cybersecurity incidents from a comprehensive perspective.

International Symposium on Grids & Clouds 2021, ISGC2021 22-26 March 2021
Academia Sinica, Taipei, Taiwan (online)

*Speaker

PoS(ISGC2021)028

1. Introduction

Organizations suffer frequent attacks by adversaries with different agendas, and these attacks, if successful, result in severe consequences to the organizations, their clients, and their partners. To defend against these attacks, Security Operations Centers (SOCs) have been created by many organizations or enterprises as effective cybersecurity monitoring solutions.

SOCs are centralized defense groups in medium or large organizations employing people, processes, and technology to provide continuous monitoring operations [1, 2, 5, 6]. SOCs have been shown to help improve an organization's security posture while preventing, detecting, analyzing, and responding to cybersecurity incidents [3, 4, 7]. However, serious security incidents remain frequent and rampant [8, 9], and the current SOC setup is insufficient to defend against these cyber-attacks.

In a recent study, Kokulu et al. [1] interviewed security analysts and mangers, who explicitly identify the issues facing current SOCs, such as inferior defense against specific types of attacks, low-quality threat intelligence and poor speed of response and level of automation.

In this paper, we propose a comprehensive SOC and try to address and mitigate above mentioned issues by creating several critical components at the SOC to effectively defend against specific types of attacks, acquire high-quality threat intelligence, and achieve faster automatic response.

Firstly, a multi-perspectives behavior analysis module is set up against the different and specific types of attacks. The multi-perspectives behavior analysis module combines the detection performance of multiple behavior detectors which can analyze abnormal behaviors from different perspectives, such as DNS, SSH and network traffic behavior, to identify different types of attack correspondingly.

Secondly, threat intelligence collection module is created for collecting high-quality threat intelligence at the SOC. On the one hand, security related blogs and news are crawled and indicators of compromise (IOCs) are accurately extracted from these threat descriptions based on natural language processing and deep learning techniques. On the other hand, IOCs can be collected by the results of manual investigation for security incidents.

Thirdly, a fast and automatic response mechanism is established for handling the detected threats from multi-perspectives behavior analysis and threat intelligence. The detected threats are filtered by using whitelist that can decrease false positives significantly, and if the remaining threats don't match with the pre-defined rules or thresholds, these threats are reported for manual investigation, otherwise the matched threats can be quickly input into blacklist and be automatically blocked.

In addition, we also create other modules to build an integrated SOC platform, such as for data collection and big data storage. The integrated SOC will be introduced in the following paragraphs.

2. Framework of the proposed SOC

For defending against specific types of attacks, acquiring high-quality threat intelligence, and achieving faster automatic response, we propose the SOC framework as has been deployed at IHEP(Institute of High Energy Physics, Chinese Academy of Sciences). The proposed framework includes data collection, data preprocessing, data storage and incident response as similar with the

CERN design [18]. Moreover, the proposed framework also includes the multi-perspective behavior analysis module and multiple applications.

The framework can be seen in figure 1 .

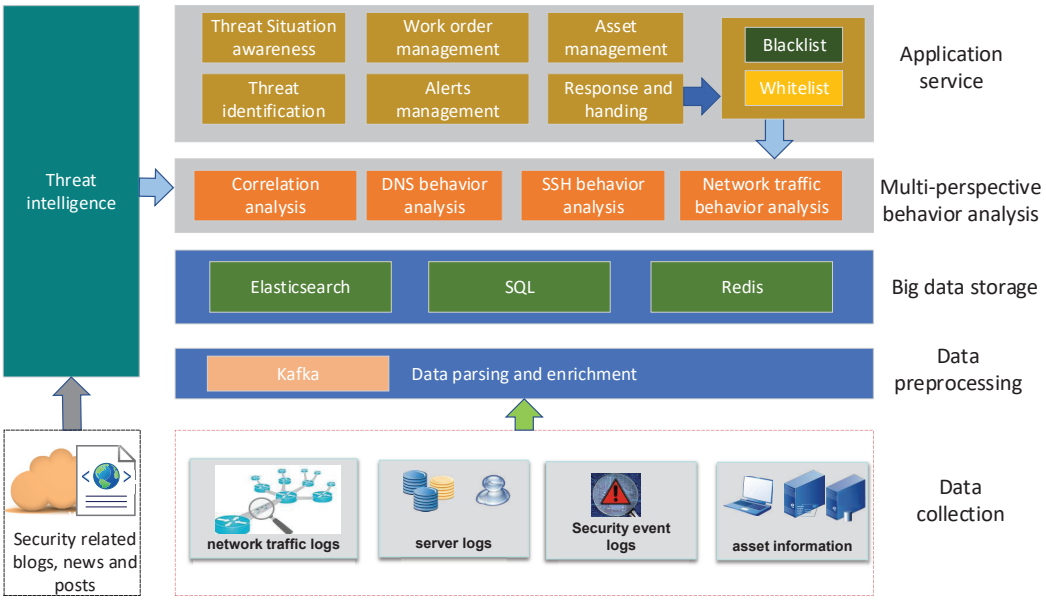

Figure 1: The framework of the proposed SOC

It contains several layers:

- Data Collection: The SOC collects different types of data from multiple objects, such as network devices, servers, security devices and asset databases within IHEP. Unstructured open-source cyber threat intelligence reports on the Internet also have been crawled.
- Data Preprocessing: The collected raw data is parsed and enriched based on a well-defined structured format in real time by using Kafka.
- Big Data Storage: Big data storage provides different types of databases for large-scale data, such as a relational database MySQL, non-relational database Elasticsearch and in-memory database Redis.
- Multi-perspective Behavior Analysis: In this module, several behavior analysis methods are deployed to model and detect various types of attacks from different perspectives, such as DNS behavior, SSH behavior and network traffic behavior.
- Threat intelligence: Threat intelligence layer extracts IOCs from security related blogs and news by utilizing natural language processing and a deep learning model.
- Application Services: Application services provide multiple applications, such as alerts management, response and handling.

Based on the framework, some key techniques are employed. The technology roadmap can be seen in figure 2. From the figure 2, we can see the collected raw data is input into data preprocessing,

and then the multi-perspectives behavior analysis module combines different behavior analysis methods to identify attacks. The detected threats and security events logs are input into teh threats and response module to handle different kinds of attacks. According to the pre-defined rules or thresholds, some attacks are input into blacklist and some results are added into whitelist. At the same time, some detected attacks are also added into threat intelligence. Besides, security related blogs are input into the IOC extraction model, and the IOC extraction model outputs threat intelligence to the threat intelligence storage platform. The threat intelligence can be correlated with the network behavior to identify attacks quickly. In addition, the behavior analysis and threats and response can be visualized and controlled in a centralized way. The detailed description of key techniques will be presented in Section 3.

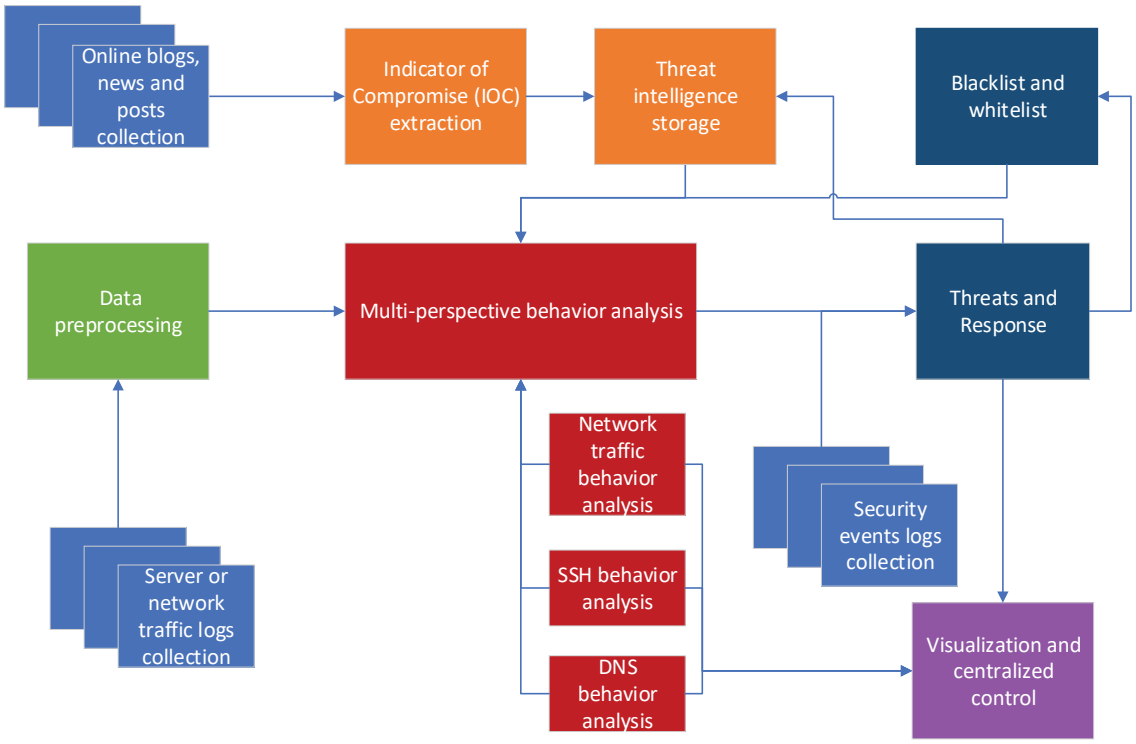

Figure 2: The technology roadmap of the proposed SOC

3. Key techniques of the SOC at IHEP

3.1 Data collection

We collect different network logs from diverse network devices. For network traffic, we have deployed a network traffic sensor at an egress router and the network traffic is resolved to network connection logs by Zeek. For network servers, the DNS logs are collected from internal DNS server and SSH logs are collected from the major login servers. For security equipment, the security event logs are collected from the commercial security device. For the assets, the related IP address and responsible person information is collected from the management system at IHEP. In addition,

security-related reports from blogs, hacker forum posts, security news, and security bulletins are automatically captured in real time. The figure 3 shows the data sources for data collection.

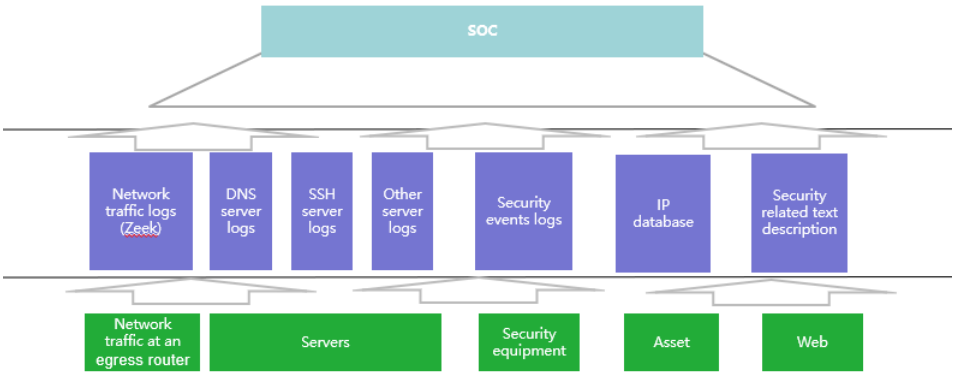

Figure 3: Data sources for data collection

3.2 Data preprocessing

In order to analyze network behavior for convenience, the data preprocessing module will structure the collected raw data in real time based on Kafka. Firstly, the different structured formats of logs are defined by considering the requirements of behavior analysis algorithms. According to the well-defined structured formats, the different log data are read from Kafka in real time and are parsed to structured data respectively. And then, these structured data are enriched by correlating external data sources, such as geographical IP database. Finally, the raw logs are structured to our defined formats and saved into Elasticsearch. The data preprocessing procedure can be seen in figure 4.

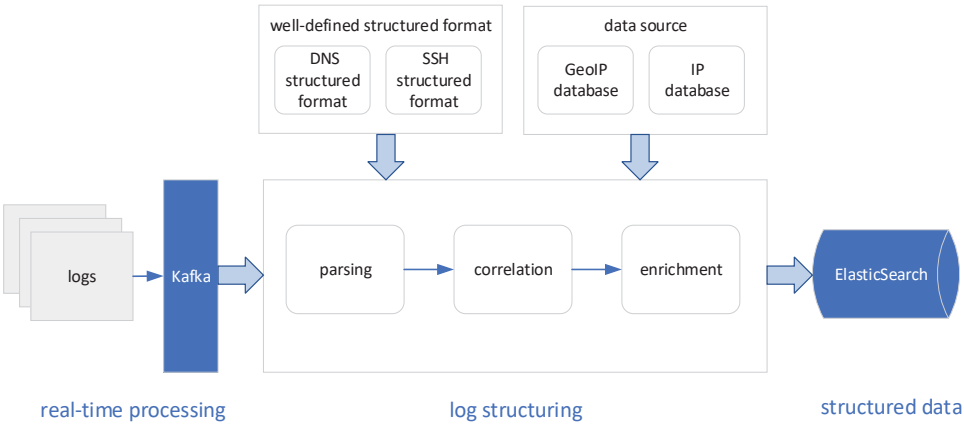

Figure 4: Data preprocessing

3.3 Multi-perspective behavior analysis

After data collection and preprocessing, the different network behavior data is stored in the big data platform and can be accessed by behavior analysis methods. The multi-perspective behavior

analysis module attempts to model different network behaviors against the different types of attacks and three behavior analysis methods have been proposed to identify attacks, i.e. network traffic behavior analysis, DNS and SSH behavior analysis.

For the network traffic, researchers [10] have validated that incomplete sessions of a host can indicate the existence of attacks such as Botnet, DDoS and Worm. Following this theory, we model the network traffic of different hosts as a directed bipartite graph by analyzing the network traffic logs as shown in figure 5 . The vertices in the directed bipartite graph denote the hosts and the edges represent the communication between internal hosts and external hosts. Most of attacks generate incomplete flows due to their running mechanisms, which only contain the one-way communication. As shown in figure 5, the edges between the internal host I_3 and the external hosts denote incomplete sessions. To detect whether the internal hosts generate incomplete sessions, two connection degrees are defined: (1) the In Connection Degree of a specific internal host(ICD), which represents the number of edges that direct to this internal host; (2) the Out Connection Degree of a specific internal host(OCD), which is the number of edges that point to external hosts from this internal host. Moreover, the Symmetry Degree of a specific internal host(SDI) is defined, which can be calculated using the radio between the ICD and OCD, to characterize the incomplete sessions effectively. If the SDI is larger than three standard deviations, a potential threat is considered significant and classified as real. The procedure can be seen in figure 6.

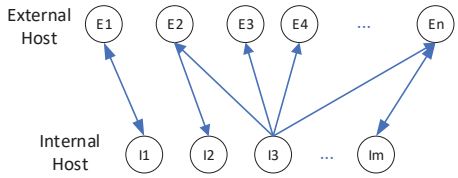

Figure 5: Communication model based on directed graph

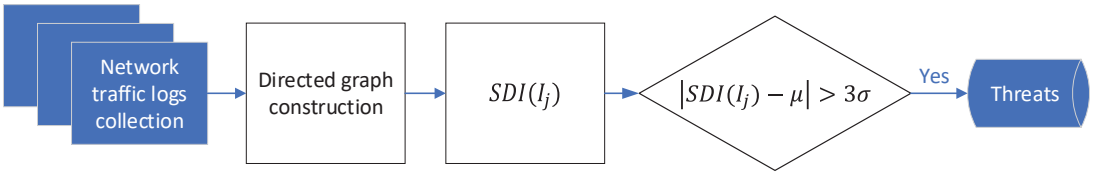

Figure 6: Network traffic behavior analysis

In addition to network traffic behavior analysis, DNS and SSH behavior analysis methods use feature engineering and machine learning to model DNS and SSH behavior and detect specific types of attacks. Based on the DNS logs, the behavior features are extracted and the behaviors are modeled as the high-dimension feature vector, and feature vectors of DNS are input into iForest and Local Outlier Factor(LOF) algorithms to detect the abnormal DNS behaviors such as Domain Generation Algorithms(DGA) and DNS tunnel. Based on SSH logs, on the one hand, feature vectors of ssh are input into Long Short-Term Memory(LSTM) to identify abnormal ssh behavior such as brute force, and on the other hand, we have built ssh behavior rules to identify whether the

ssh behavior of a user is abnormal or not, such as credential dumping. Figure 7 shows the procedure of methods.

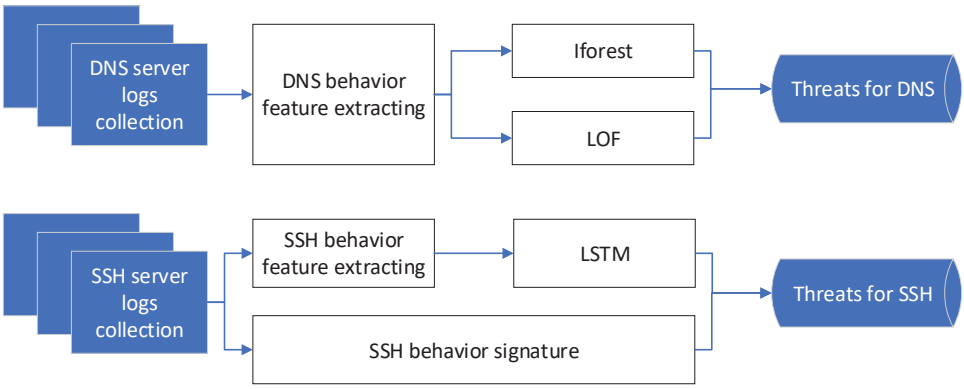

Figure 7: DNS and SSH behavior analysis

3.4 Threat intelligence collection

Cyber Threat Intelligence (CTI), as a collection of threat information, has been widely used by security researchers and industries to defend against prevalent cyber attacks. CTI is commonly represented as Indicators of Compromise (IOC) for formalizing threat actors. We set up the threat intelligence module to collect high quality threat intelligence.

Existing IOC extraction methods [11–13] use regular expression to collect IOC from the unstructured descriptive texts. However, these methods often produce high false positive rate by misjudging legitimate entities as IOCs [14] and the accuracy of IOC extraction is low. For acquiring high quality threat intelligence, we employ a specialized natural language processing (NLP) pipeline to accurately extract IOCs. Figure 8 gives the pipeline:

Step 1: Texts Preprocessing: The descriptive texts have been captured automatically as described in Section 3.1. The texts preprocessing removes all punctuations and segment sentences of these texts.

Step 2: IOC Recognition: We firstly set up Brat [15] text tagging tool and tag IOC from texts such as malicious file hashes, IPs/domains of botnets. And then the Bidirectional Long Short-Term Memory + Conditional Random Fields (BiLSTM+CRF) model is learned based on the tagged texts for IOC recognition.

Once new descriptive texts are captured, the learned BiLSTM+CRF model can recognize the IOC accurately from these texts. According to the extracted IOCs, we can identify cyber threats quickly from numerous logs.

Moreover, IOCs can also be collected by cyber security staff when they investigate security incidents and discovery the malicious IP address or port.

3.5 Threats and response

Generally speaking, the cyber attacks can be classified according to the lifecycle for an entire attack scenario. The multiple stages can be compacted from most of the attack scenarios as follows:

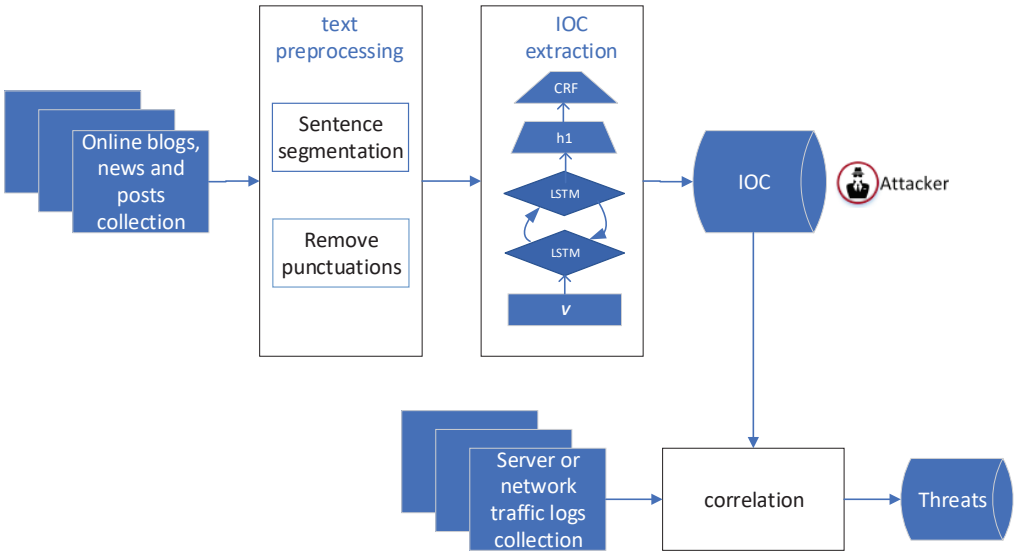

Figure 8: Threat intelligence collection

Phase 1: Reconnaissance: It marks the beginning of any successful attack. In order to understand the target, attackers generally use scanning, brute force or phishing.

Phase 2:Establish the foothold: In order to achieve the goal, attackers need to establish a foothold in the target's network. They may build the botnet, use DGA or exploit the vulnerability.

Phase 3: Lateral movement: Attackers may use the worm, brute force and credential dumping to laterally move within the target's network in search of critical components or data.

Phase 4:Exfiltration/Impediment: Attackers steal data by C&C server and DNS tunnel or destruct critical components by DDoS.

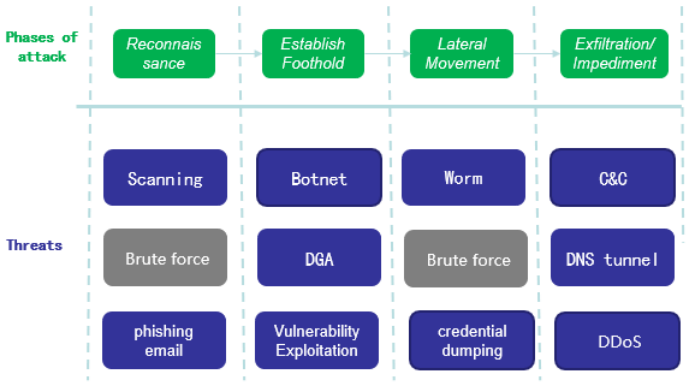

Figure 9: Phases of attack

Figure 9 present the phases of attack and threats. These different kinds of threats from several phases of attack can be detected from multi-perspective behavior analysis and threat intelligence. For example, multi-perspective behavior analysis can detect Scanning, DDoS attacks and botnets based on network traffic behavior analysis, can identify DGA and DNS tunnel based on DNS

behavior analysis, and can identify brute force and credential dumping based on SSH behavior analysis.

Facing the detected threats, cyber security staff need to handle these threats. To make a quick and automatic response, we employ the response mechanism based on the blacklist, whitelist and greylist, and it can be seen in Figure 10.

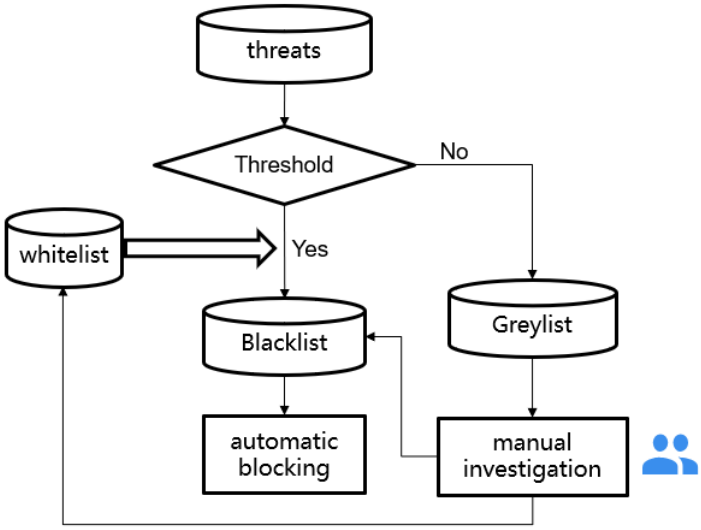

Figure 10: Response mechanism

Firstly, the detected threats can be filtered based on the pre-defined thresholds or rules. If any detected threat satisfies the threshold or conforms the rule, and it doesn't match the whitelist, this threat will be input into blacklist, and automatic blocking process starts, otherwise, this detected threat will be added into greylist and manual investigation begins. After investigation, if the detected threat is not a false positive, it will be appended into blacklist and be blocked.

3.6 Visualization and centralized control

In order to achieve a high level of situational awareness, we build a front-end web application that provides visualization and centralized control. The web application is implemented in Python. The (HTTP) API is based on Django framework [17]. The web-frontend is built on top of the API using vue.js [16]. In the web application, we adopt visualization technology to show detected threats based on different views such as threat trends, origin of attack, types of attacks, number of attacks, so that cyber security staff can aware security situations at IHEP. Besides, the web application offers several operations. For example, blacklist/whitelist management can add and search IP or domain, and if an IP is added to blacklist, then this IP is blocked automatically. Once an IP is input into whitelist, this IP will not be detected as a threat.

4. Deployment of the SOC at IHEP

We firstly deploy a network traffic sensor at the egress router and a log collection probe on the log server. Furthermore, the data captured by traffic sensor and log collection is input into the

analysis platform to detect threats and provide threat hunting, investigation, response and handling ability. At the same time, cyber security staff have situational awareness. The deployment can be seen in figure 11.

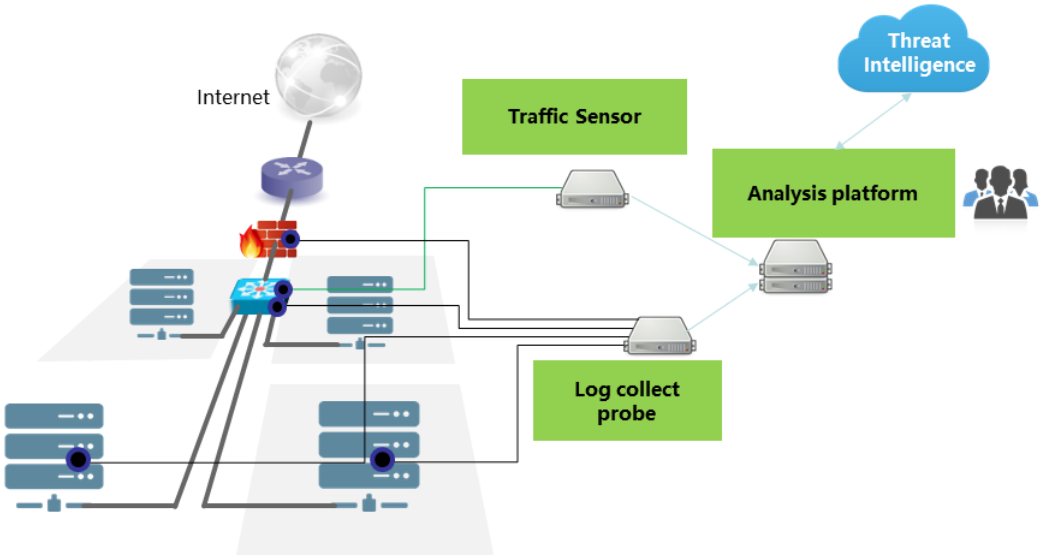

Figure 11: Deployment of SOC

The SOC combines the detection performance of multiple behavior detectors and every detector can generate alerts of the specific type of attacks every day by using the rule-based method, anomaly detection techniques or graph theory. Moreover, the SOC can extract about 5,000 IOCs every day from the open-source cyber threat intelligence. In addition, the SOC can block about 50 external IP addresses and a few domains every day by automatically correlating the analysis platform with the egress router.

5. Conclusion

This paper proposes a comprehensive SOC. It firstly collects and preprocesses different data sources and stores these data into the big data platform. Furthermore, the multi-perspective behavior analysis module can identify different kinds of threats that cover the lifecycle for most of the attack scenarios. Moreover, the threat intelligence platform accurately extracts IOCs from blogs or news by the deep learning model to quickly identify threats. Finally, the detected threats can be quickly blocked by matching to the pre-defined thresholds or rules. In the future, more behavior analysis components will be deployed to enhance the detection capability of the SOC at IHEP.

Acknowledgments

This work was supported in part by the Cybersecurity Protection Project for Large Scientific Facilities, and National Natural Science Foundation of China (NSFC) under Grant 11675199, Grant 61901447.

References

[1] Faris Bugra Kokulu, Ananta Soneji, Tiffany Bao, Yan Shoshitaishvili, Ziming Zhao, Adam Doupé, and Gail-Joon Ahn. 2019. Matched and Mismatched SOCs: A Qualitative Study on Security Operations Center Issues. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019. ACM, 1955–1970. https://doi.org/10.1145/3319535.3354239

[2] Joseph Muniz, Gary McIntyre, and Nadhem AlFardan. 2015. Security Operations Center: Building, Operating, and Maintaining Your SOC. Cisco Press, Hoboken, NJ, USA.

[3] C. Hill. Security operation center (soc). https://www.nascio.org/wp-content/uploads/2020/09/NASCIO-IL-2018-Cybersecurity-SOC.pdf, 2017.

[4] M. Kan. Boeing's wannacry run-in is a reminder to patch your systems. https://www.pcmag.com/news/boeings-wannacry-run-in-is-a-reminder-to-patch-your-systems, 2018.

[5] https://www.mcafee.com/enterprise/en-us/security-awareness/operations/what-is-soc.html . Retrieved 2021-04-22.

[6] J. De Groot. What is a Security Operations Center (SOC)? https://digitalguardian.com/blog/what-security-operations-center-soc , 2020.

[7] D. Ritchey. Creating the gsoc: 4 leading examples of successful security operations centers. https://www.securitymagazine.com/articles/87849-creating-the-gsoc-4-leading-examples-of-successful-security-operations-centers, 2017.

[8] https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents , Retrieved 2021-04-22

[9] Sharma. Why do data breaches happen? https://www.marshall.usc.edu/blog/why-do-data-breaches-happen, 2017.

[10] Qin T , Liu Z , Wang P , et al. Symmetry Degree Measurement and its Applications to Anomaly Detection[J]. IEEE transactions on information forensics and security, 2020, 15:1040-1055.

[11] IBM X-Force, https://exchange.xforce.ibmcloud.com/. Retrieved 2021-09-05.

[12] Threat crowd, https://www.threatcrowd.org/. Retrieved 2021-09-05.

[13] Phish-Tank, https://phishtank.org/. Retrieved 2021-09-05.

[14] Xiaojing Liao, Yuan Kan, Xiao Feng Wang, Li Zhou, and Raheem Beyah. Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In ACM Sigsac Conference on Computer Communications Security, 2016.

[15] Stenetorp P, Pyysalo S, Topic G, et al. BRAT: A Webbased Tool for NLP-assisted Text Annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012: 102-107.

[16] Evan You. [n.d.]. vue.js. Retrieved 2020-06-13 from https://vuejs.org/

[17] Django, https://www.djangoproject.com/. Retrieved 2021-09-05.

[18] David Crooks. Building a minimum viable Security Operations Centre for the modern grid environment. International Symposium on Grids and Clouds, 31st March - 5th April, 2019. Academia Sinica, Taipei, Taiwan.