Hindawi
Security and Communication Networks
Volume 2019, Article ID 5483918, 9 pages
https://doi.org/10.1155/2019/5483918

Research Article

HeteMSD: A Big Data Analytics Framework for Targeted Cyber-Attacks Detection Using Heterogeneous Multisource Data

Ankang Ju[1], Yuanbo Guo[1], Ziwei Ye[1], Tao Li[1], and Jing Ma[2]

[1]*Zhengzhou Institute of Information Science and Technology, 450001, China*
[2]*Science and Technology on Information Assurance Laboratory, 100071, China*

Correspondence should be addressed to Ankang Ju; jusissp@yeah.net

Received 25 January 2019; Accepted 10 April 2019; Published 2 May 2019

Guest Editor: Pelin Angin

In the current enterprise network environment, multistep targeted cyber-attacks with concealment and advanced characteristics have become the main threat. Multisource security data are the prerequisite of targeted cyber-attacks detection. However, these data have characters of heterogeneity and semantic diversity, and existing attack detection methods do not take comprehensive data sources into account. Identifying and predicting attack intention from heterogeneous noisy data can be meaningful work. In this paper, we first review different data fusion mechanisms of correlating heterogeneous multisource data. On this basis, we propose a big data analytics framework for targeted cyber-attacks detection and give the basic idea of correlation analysis. Our approach will offer the ability to correlate multisource heterogeneous security data and analyze attack intention effectively.

1. Introduction

In the current network environment, network attacks are more covert and systematic. Advanced Persistent Threat (APT) brings huge economic losses and security risks to governments, enterprises, and other institutions [1]. Network security administrators do not realize that their network has been compromised until weeks, months, or even years later. Traditional detection techniques cannot ideally handle targeted cyber-attacks with characteristics of complexity and customization. Multistep targeted cyber-attacks have become the most critical factor affecting network security [2].

The significance of timely detection and analysis of targeted cyber-attacks lies in two aspects [3]. First, timely detection of intrusion behaviors: illegal system services can be cleaned up and recover from a crash in time after the system has been attacked. That can reduce losses caused by less normal service hours. Second, targeted cyber-attacks are often implemented in multiple stages, and the abnormal behaviors detected currently may not rebuild the whole attack sequence. Timely detection of preattack steps can help us understand the intention of an attack. Detecting and predicting follow-up attacks in time can help us take protective measures to prevent further damage caused by follow-up attacks.

In the current research of intrusion detection technology, it can be divided into host detection and network detection [4]. Host-based intrusion detection method analyzing procedure behavior to find payload is represented by the malicious code. Network-based intrusion detection mainly analyses network flow to identify unknown network attacks. Threat intelligence, combined with big data analysis method of massive logs and traffic data, provides a more comprehensive security perspective.

However, the intrusion detection method based on single source data cannot reflect the relationship of seemingly irrelevant events. Besides, traditional security solutions (such as intrusion detection system, antivirus software, etc.) generate a lot of intrusion alerts. These alerts still need to be examined and cross-checked with other available data (such as host log and network communication data) in order to eliminate false positives and identify any legitimate attacks [5].

Illegal events affected by attack behavior are hidden in the dispersive log. Existing detection methods are insufficient in recognizing anomaly events to support sophisticated attacks detection. Targeted cyber-attacks detection based on heterogeneous multisource data has become a consensus. At present, the limitations of targeted cyber-attacks detection fall in three major categories as follows.

(1) Heterogeneous multisource security data is complicated and the semantics of expression is more abundant. There are various attack detection methods for different data sources. But the combinatorial relationship between research questions is not clear enough. There is a lack of systematic discussion in academic research.

(2) Existing methods are unable to quickly and efficiently locate anomalies related to network attacks. The problem of data correlation has not been well solved. Data association method in heterogeneous multisource still constrains the development of targeted cyber-attacks detection.

(3) The intelligence and automatization of existing detection technology are not satisfactory. Semantic information about secure data is not effectively expressed. Manual analysis is still the main way in attack recognition.

In order to address the limitations discussed above, we propose a novel layered framework for targeted cyber-attacks detection based on the integration of heterogeneous multisource data. The proposed framework utilizes attack investigation through correlation analysis, which can efficiently cope with targeted cyber-attacks detection in a big data environment. Based on this framework, inner-layer and cross-layer analysis approach for targeted cyber-attacks is proposed. Follow-up researchers can carry out further research on this basis.

This paper concentrates on the issue of the targeted cyber-attacks detection from the practical perspective. We propose a novel framework that uses big data correlation analysis techniques to analyze cyber incidents for enterprise network in order to improve attack detection ability. The proposed framework is named as *HeteMSD: A Big Data Analytics Framework for Targeted Cyber-Attacks Detection* that uses heterogeneous multisource data. We aim to develop an efficient framework that can assist security analyzers to reduce the blindness of data analysis from heterogeneous data sources without reducing the level of digital security guarantees.

Currently, both of experimental and theoretical works for targeted cyber-attacks detection have been in a groping stage, and there are still some significant discrepancies among the research of targeted cyber-attacks detection based on heterogeneous multisource data. Our proposed framework has a great significance as it intends to solve the conflicts between advanced network attack protection and big heterogeneous data burdens. The main contributions of our work are as follows.

(1) This paper comprehensively summarizes the existing data sources and points out the existing problems from the view of expression difference. Then we summarize and classify the multisource heterogeneous security detection data from three aspects: nonsemantic data, semantic data, and security knowledge data. Thus, we can have a more intuitive understanding of heterogeneous multisource security data.

(2) This paper proposed a novel framework designed for eliminating data redundancy while enhancing data relevance. We divide the research problem into five layers: sensing layer, event layer, alert layer, context layer, and scenario layer. We've made classification and integration of data sources and provide a dataflow diagram of the data process.

(3) This paper combines cyber-attacks analysis methods with big data correlation techniques to improve security levels and realize the comprehensive network security situation awareness. The research issues in related fields were investigated from perspectives at different levels.

The remainder of this paper is organized as follows. Section 2 gives a definition of targeted cyber-attacks and analyses existing targeted cyber-attacks detection technologies. Section 3 summarizes heterogeneous data sources and gives a novel classification perspective. Then we present the proposed framework based on the integration of heterogeneous multisource data and explain the brief dataflow in practical application. Section 4 discusses the application of correlation analysis for targeted cyber-attacks detection. Finally, the last section concludes the paper and describes the direction of future work.

2. Related Work

Targeted cyber-attacks are a class of dedicated attacks that aim at a specific user, company, or organization to realize specific intention, such as stealing sensitive data from a back-end database or paralysis system service. Targeted cyber-attacks have a characteristic of discrimination and are not random in nature. It means attackers involved in targeted attacks differentiate the targets and wait for the appropriate opportunity to realize the attack plan. Targeted cyber-attacks usually require several stages to achieve the goal. The lifestyle of a successful implemented targeted cyber-attack process usually includes gathering, infecting targets, system exploitation, data exfiltration, and maintaining control. [6]. Each of these steps is the key factor of targeted cyber-attacks. To successfully implement targeted cyber-attacks, all the above stages must be successful.

The difference between the targeted cyber-attack and a traditional attack is that targeted cyber-attacks are more complex and usually with strong intrusion motivation. Attackers spend more time choosing the target, find vulnerabilities, and customize malware. Targeted cyber-attacks are usually implemented by professionals instead of simply using attack tools.

The example attack scenario described in Figure 1 is as follows.

(i) Step 1: Attacker utilizes social engineering tools such as phishing mail to bypass firewall and infiltrates application server (Server1).

(ii) Step 2: Attacker takes Server1 as a springboard to identify and penetrate hosts in the internal network.

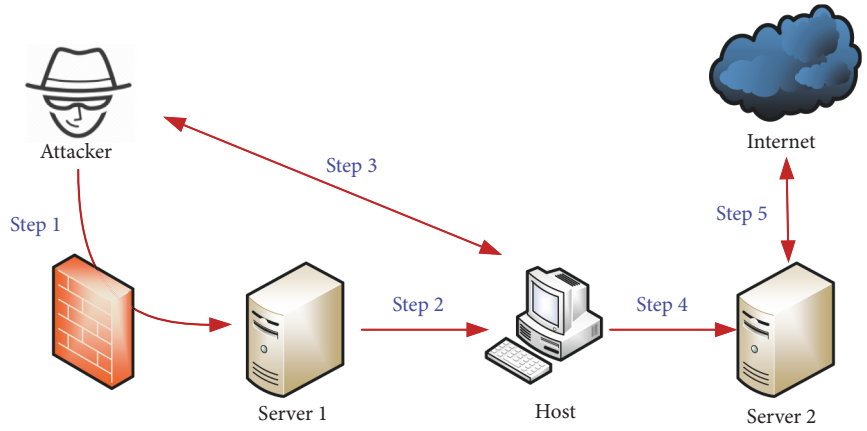

Figure 1: An example of targeted cyber-attack.

(iii) Step 3: Attacker establish C&C channel between the Attacker and host to control the host of the intranet to make further intranet penetration.

(iv) Step 4: Attacker uses the controlled host to log into the internal data server Server2 to collect sensitive data.

(v) Step 5: Send the collected sensitive data to the Internet to achieve the purpose of data theft.

When talking about targeted cyber-attacks, another concept Advanced Persistent Threat must be involved. APT can be seen as a subset of targeted cyber-attack [7]. Commonly speaking, APTs are targeted cyber-attacks with higher-level attack means. APT is implemented through a variety of different attack paths. It exists for a long time without being discovered in a real network environment. Generally, when talking about advanced complex and targeted network attacks, targeted cyber-attacks and APTs are interchangeable. There are few differences between targeted cyber-attacks and APTs [6].

The implementation process of targeted cyber-attacks can be described by the Kill Chain model [8]. As shown in Figure 2, the Kill Chain model summarizes the attack process as target reconnaissance, weapon customization, delivery, exploitation, installation, C&C channel establishment, action implementation, and other attack steps. Kill Chain attack model is as follows.

Targeted cyber-attacks have brought enormous threat to enterprise network security. In recent years, targeted cyber-attacks detection represented by APTs detection has attracted great attention from academic researchers and security industry. A variety of detection schemes have been proposed. Generally, it can be divided into host-based detection and network-based detection. At present, more detection methods are combined with big data analysis method and machine learning technology. Researchers also proposed correlation analysis method of audit log and network data based on threat information [9].

Host-based detection method: the representatives are antivirus software and HIDS. The main idea of malware detection is to detect malicious programs through a monitoring system call, network access, file operation, process creation, and memory modification. Through static analysis and dynamic analysis, malicious programs can be detected and APT attacks can be prevented [10]. The HIDS method based on pattern matching can effectively identify known attacks. But unknown attacks cannot be detected and the rule base needs to be updated regularly. Then behavior-based detection method is proposed. In recent years, with the rise of data mining and machine learning technology, researchers [11] have proposed a variety of detection technologies based on abnormal behavior recognition.

Network-based detection method: the pattern of command and control channel by malware has certain regularity (such as attack payload signature, sequential characteristics of network communications, and the generated domain name). Network traffic generated by targeted cyber-attack is different from those in normal business environments. Therefore, it is feasible to find the attack load of attack process by the network detection method.

Due to increase in number of sophisticated threats and the great increase in the volume of security data, the landscape of analyzing heterogeneous security data has drastically changed, as now working with security data has entered in the category of Big Data problem [4].

Parth Bhatt [12] proposed a research framework to handle complex attacks and it was tested with simulated attack data. The central basis of the framework consists of an Intrusion Management System and a multistage attack model. The multistage attack model is used to identify prevention and detection controls that provide logs used by the Intrusion Management System, and it is also used as a guide to logs correlation activities.

Mirco Marchetti [13] designed and evaluated a novel framework that is tailored to support security analysts in detecting APTs. The proposed framework uses multifactor approaches where big data analytics methods are applied to internal and external information to support human specialists so that those specialists can focus their security and intelligence analyses on the subset of hosts that are most likely to have been compromised. The proposed approach represents a step forward with respect to the state of the art

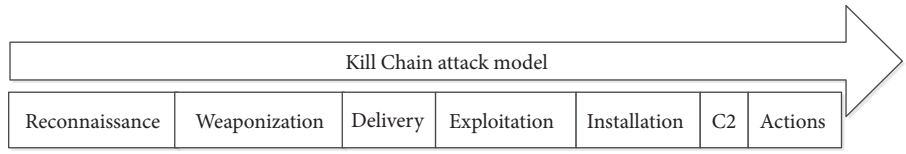

Figure 2: Kill chain attack model.

and paves the way to novel methods for early detection and mitigation of APTs

At present, traditional detection schemes focus on one or more stages. They cannot achieve comprehensive attack detection and have a high false negative rate and false positive rate. Although there have been a lot of research and great progress is made, targeted cyber-attacks are still occurring constantly. The shortcoming is that existing method still depends on manual analysis. It cannot identify and respond quickly. Here we want to reduce human participation and make the detection process as intelligent as possible. Similar attacks should be detected according to the attack pattern. Data parsing should be as automatic as possible and be detected in an adaptive way. The intelligence of attack detection is mainly embodied in several research points.

(1) Accuracy. A large number of false alerts pose a big challenge in intrusion detection. Even if the false alarm rate is merely 1%, a large amount of data can bring a lot of alerts, which will bring enormous burdens to security managers. Reducing false positive and false negative is an important content in intrusion detection.

(2) Efficiency. Fast identification of attack behavior is also important in practical application. Shortening attack detection time will minimize the damage caused by the attack. Thus it can reduce security risk for the system and handle security accidents in time.

(3) Intelligence. Automated attack inference process. Reducing manual labor costs on irrelevant information can help human analysts put artificial judgment in the key part.

The ability to cross-correlate events or alerts from various sources in the network is key for detecting sophisticated multistage attacks at an early stage, which would allow a reasonable time to stop the attack or mitigate the damage.

In summary, existing targeted cyber-attacks detection technologies for network security threats recognition are not fully adapted. These methods cannot meet the needs for network security situational awareness. Current utilization of multisource heterogeneous data is inadequate. The purpose of this paper is to better correlate security data from different sources and improve situational awareness ability. The related research issues and the next research directions are discussed in the following content.

3. Concepts and the Proposed Framework

3.1. Heterogeneous Multisource Data. Data is the premise of attack detection and threat analysis. However, security data come from diverse sources, such as IDS alerts, firewall alerts, NetFlow data, Linux syslog, Event Tracing for Windows, etc. Heterogeneous data formats from various sources have diverse semantic meanings. It is difficult to identify the relationship in practical application. Firstly, we investigate and summarize the heterogeneous multisource security data from threat hunting aspect. Security data has the following characteristics.

(1) Heterogeneous Format. Comprehensive security data come from local hosts, servers, routers, firewalls, IDSs, and other security devices. Data acquisition and storage from different sources are quite different. Data formats include structured, semistructured, and unstructured.

(2) Diverse Semantic. Different types of data represent different levels of security knowledge. For example, network data and host logs represent information in different fields. IDS alerts, firewall logs, and other security logs represent information at different semantic levels from host logs and network data. Different security data calls for different treatment in practical application.

(3) Correlation across Data Sources. A complex targeted cyber-attack hides its behavior in multisource heterogeneous data. A single attack will bring records of data from multiple sources. These data can be linked together with the hosts, users, location, etc. as the associated elements. And all these data as a whole can be restored to a complete attack scenario.

It is an important research issue to classify security data from various sources. This paper summarizes and analyses the multisource heterogeneous security data in targeted cyber-attack detection. Firstly, we classify them into three categories according to the different semantic levels of security data expression.

(1) Nonsemantic Data. Nonsemantic data includes detailed descriptions of the running process and logs reflecting the attacks, but it does not contain security semantic information. Nonsemantic data includes operation logs, system calls, NetFlow data, user behavior logs, etc. Analysts can analyze malicious behavior from their data and generate semantic data.

(2) Semantic Data. Based on pattern matching and other technologies, we can get security alerts from the nonsemantic data. This kind of data has security semantics information,

which indicates the violation of security rules or the abnormal operation of the system such as IDS alerts, firewall logs, OS security logs, and so on. There are many different security systems deployed in the current network environment, including firewall, antivirus software, intrusion detection system, and traffic analysis system. The key technologies corresponding to these security systems have made great progress in the past decades. But these independent systems can only solve local problems but cannot reflect the overall situation of network security. The limitations of traditional security systems are becoming more and more obvious.

(3) Security Knowledge Data. At present, most network attacks rely on specific vulnerabilities and services, such as operating system vulnerabilities, software vulnerabilities, or protocol vulnerabilities. If vulnerability information is integrated into the process of attack detection, the accuracy of attack detection can be improved. Besides, with the development of security technology, threat intelligence technology has become an important data source for detection [14]. Security knowledge data includes vulnerability database, attack pattern database, and another external threat intelligence.

3.2. Framework Design. In the previous section, we analyze and summarize three kinds of security data. But how to carry out targeted cyber-attack detection based on multisource heterogeneous data? In this section, we present a novel framework for targeted cyber-attacks detection after an in-depth study of targeted cyber-attack detection technology, According to the different analytical perspective, we divide the detection into five levels: sensing layer, event layer, alert layer, context layer, and scenario layer. Figure 3 shows a high-level architecture of our framework.

(1) Sensing Layer: sensing layer includes the initial data source of targeted cyber-attack detection. In this layer, heterogeneous multisource security data acquisition and data sensing is an important basic research issue. Summarizing all kinds of data sources comprehensively and systematically and collecting and transmitting security data efficiently from many collectors is still a challenge to the current security research. In addition, as an important source of perceived data, how to quickly and timely response to new security threats combined with security knowledge is also a problem that needs to be considered. The main research area in the sensing layer includes data acquisition, data aggregation, data parsing and preprocessing, data fusion, feature extraction, and feature selection.

(2) Event Layer: event layer handles with nonsemantic data. As mentioned above, the nonsemantic data of network security are various and the attributes are quite different (for example, time sequential data, spatial data, trajectory data, and NetFlow data). How to extract specific features from these data and express large-scale heterogeneous data will be a big challenge. Especially in targeted cyber-attack detection, only by establishing the relationship between different data can the later analysis process be conducted efficiently. In this layer, the main research area includes event aggregation, content security, event correlation, anomaly detection, etc.

(3) Alert Layer: data input of alert layer includes anomaly detection results and semantic data. The semantic data include alerts coming from event layer data after analysis and detection and alerts generated by firewall, intrusion detection system, antivirus software, and other security devices. The difference from the event layer is that the alert layer contains all kinds of alert data with security semantics, such as exception score, alarm level, etc. In this layer, the main research area includes alert expression, alert clustering, alert fusion, security information, and event management.

(4) Context Layer: due to the wide range of alert data sources, the result of alert correlation is only attack fragments or preliminary attack steps. How to obtain reinforcing knowledge from security data is still a challenging problem. Traditional machine learning technologies focus on a single data source rather than heterogeneous multisource data. It is still a challenging problem in targeted cyber-attack detection to extract attack context and attack fragments from security data of different sources. Human participation is an important factor in this layer. Reducing manual analysis is a problem to be solved at this level. In this layer, the main research area includes attack modelling, attack pattern mining, attack scenario reconstruction, attack inference, etc.

(5) Scenario Layer: In this layer, the attack scenario is reconstructed based on attack modelling and attack scenario reconstruction. Attack scenarios are expressed in a way consistent with human cognition. In this layer, the main research area includes threat intelligence expression, attack causality analysis, visual analysis, and attack planning.

Other existing research issues can be introduced into our framework. The research contents of each layer and across layers are analyzed respectively. Current research issues in this field are shown in Figure 4.

Discovering targeted cyber-attacks from heterogeneous multisource data is a systematic analysis process. The problem is more complicated because of the diversity of security data. As data continues to be added, data processing will provide a deeper understanding of attack scenarios. As shown in Figure 4, there are data dependencies across layers. The lower layer provides input data for the upper layer, while the upper layer analyses the input data for further analysis. More specifically, the data flow process is shown in Section 3.3

3.3. Dataflow Diagram. In the current research, intrusion detection technology needs further refinement and modification. The key issue is to analyze the relationship between different levels of security data. However, even if the detection framework is given, it is difficult to find the behavior traces of attacks from heterogeneous data. This paper presents a general attack detection method. The dataflow diagram discussed in Figure 5 gives the description of the process

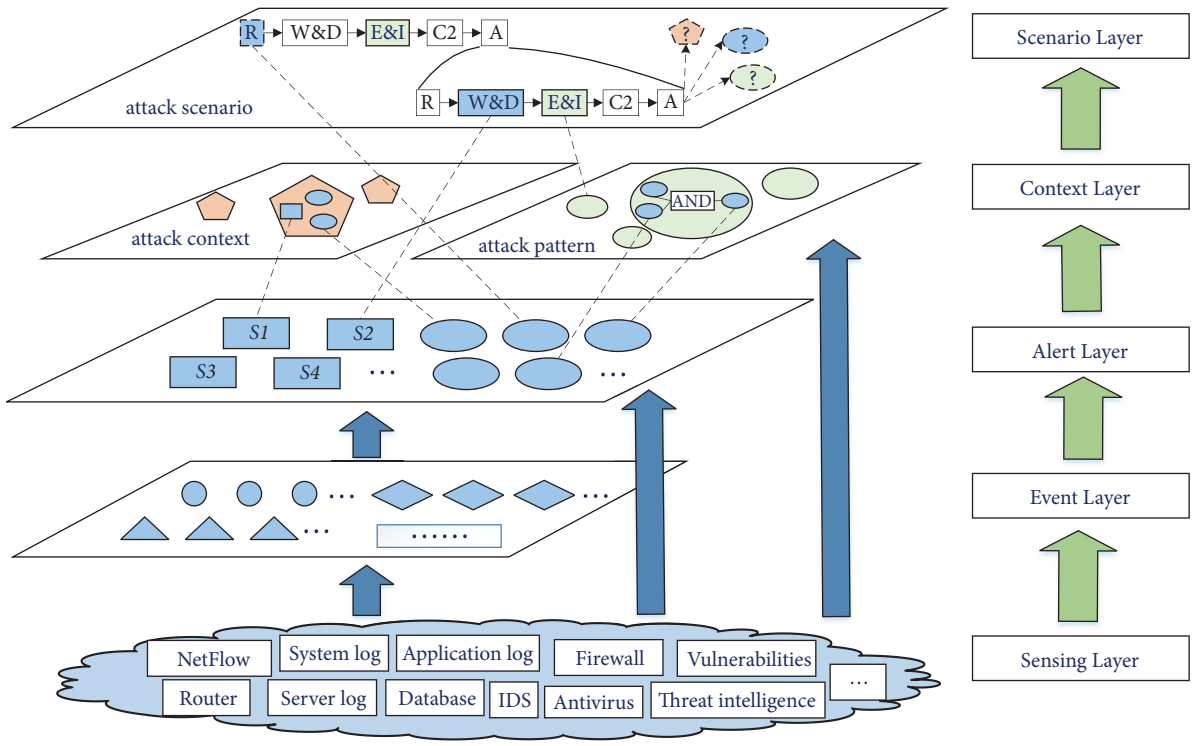

Figure 3: HeteMSD: a big data analytics framework for targeted cyber-attacks detection.

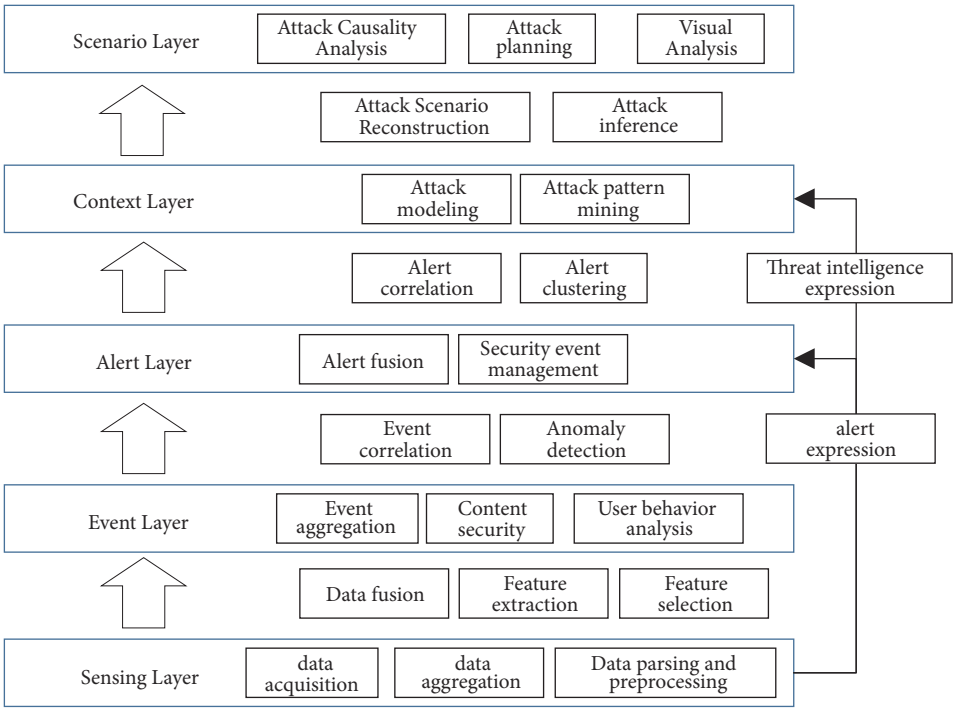

Figure 4: Research issues in this field.

from the original security data to the generated attack scenario.

Three kinds of data summarized in Section 3 are inputs for different stages in the dataflow diagram. The initial input data is nonsemantic data. Raw data is preprocessed to generate security event with security semantics. Preprocessed security events, combined with semantic data, are the input of the data analysis module. After data fusion analysis, these data are analyzed and generate security alerts with higher confidence. Furthermore, security data supplemented by

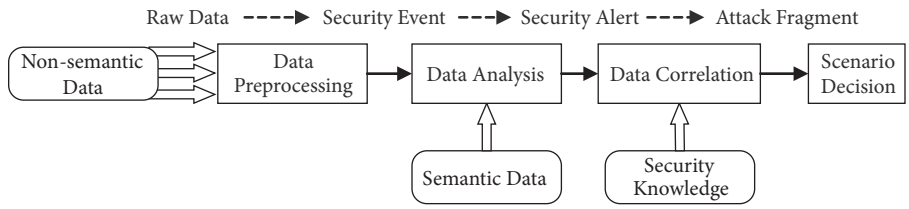

Figure 5: Dataflow diagram.

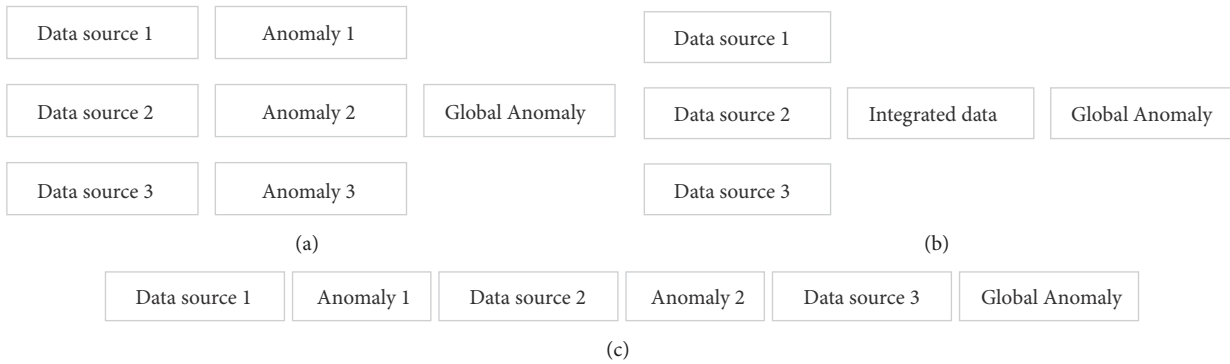

Figure 6: Anomaly detection technology based on data fusion.

security knowledge will generate attack fragment with more cognitive information. Finally, these fragments are filled into the attack inference model to form a more complete attack scenario. Attack scenario reconstruction is implemented with attack correlation and attack investigation. Specifically, the data analysis process is shown in Figure 5.

The main functions of each module in Figure 5 are as follows.

Data preprocessing module: the input of data preprocessing module is nonsemantic data. It mainly processes raw data into security events for subsequent analysis, where security events are with certain security semantics. This module is the key to find abnormal traces of targeted cyber-attacks. The main methods used in this module are feature extraction, feature selection, and anomaly detection.

Data analysis module: the function of this module is extracting security alerts with security semantic data from basic security incidents such as low-level alarms and anomalies. And that can eliminate false alarms of alarms produced by security devices to a large extent. By combining analysis of alarms and anomaly detection results generated by security equipment, a more reliable result is obtained. Ranking algorithms are commonly used in this module.

Data correlation module: this module generates fragmented information that expresses attack patterns based on multiple correlation analysis methods. That can establish the relationship between different levels of data and enhance security situation awareness. This module is designed to aggregate alerts and security knowledge representing human cognition. Attack scenarios will eventually be formed through knowledge extracting from the original data.

In summary, we introduced the framework and explain the flow of data processing in the application. The following section will give the main correlation analysis methods used in this framework.

4. Correlation Analysis

The key to targeted cyber-attack detection lies in how to integrate and establish the correlation between multisource heterogeneous security data. In the previous section we have discussed the framework design. Furthermore, there is a need to correlate security measures with attack phases for better understanding. In this section technical aspects of correlation analysis and defence measures are discussed, more concretely, mainly in the following aspects.

4.1. Event-Event Correlation. The correlation between event and event can be achieved by multisource data fusion. For raw input data, abnormal detection results represent event correlation. Data fusion can reduce data redundancy and collaborative information acquisition through complementarity [15]. The features of original heterogeneous data need to be extracted as the formats of the original data are inconsistent. The global abnormal results are obtained by utilizing the comprehensiveness association of multisource data. At present, the main method of heterogeneous event fusion is to extract features from different security data and form a global feature vector. The relationship between different sources of data can better reflect the global anomaly that cannot be expressed by a single data source.

The anomaly detection method of multisource heterogeneous data is different from that of single source data. Existing multisource heterogeneous data anomaly detection technologies based on data fusion can be divided into three categories (as shown in Figure 6). First, anomaly detection

algorithms are applied to different data sources, and the global anomaly is obtained after aggregation analysis of anomaly detection results. Second, different data sets are merged into a unified data source with the same data pattern. In this way, the multisource data anomaly detection is transformed into the traditional single-source data anomaly detection problem. Thirdly, more data sources are added in the process of anomaly detection to supplement and strengthen the anomaly detection results.

In this field, extracting feature vectors from multisource heterogeneous data is a key factor. And anomaly detection of multisource heterogeneous data based on ensemble learning needs further research.

4.2. Alert-Alert Correlation. An attack behavior may trigger many different alert events, so it is necessary to associate the original security events to higher-level alerts. Alert correlation is an important technology to cope with large number of alerts generated by intrusion detection systems. And it has become a new research trend in the field of attack detection. In current attack analysis methods, manual analysis is still a key factor in identifying attack scenarios. The automatic method is to establish a link between alerts to simplify the analysis of security experts. At present, there are mainly three kinds of alerts correlation analysis methods.

(1) Similarity-based method. Alerts are clustering by calculating the similarity of attribute values between alerts.

(2) Condition-based method. Alerts are correlated based on the precondition and postresult of an attack. The attack correlation is established by matching the premise and result of different attack types.

(3) Scenario-based method. The attack scenarios are established by matching the alarm with the known attack pattern.

Through alert correlation, redundant information of alert events can be deleted. On the other hand, the higher-level semantic information of alarm events can be extracted and improve the accuracy of attack final detection results. It reduces the large number of alerts generated by intrusion detection system and can help analysts to better grasp and analyze the attacker's motivation. Establishing alarm correlation model by adjusting probability inference results with optimization algorithm needs further study.

4.3. Pattern-Knowledge Correlation. In recent years, representation learning technology has gained wide attention in the fields of speech recognition, image analysis, and natural language processing. Representation learning aims at representing the semantic information of the object as a dense low-dimensional real-valued vector. In this low-dimensional vector space, the closer the two vectors are, the higher their semantic similarity is. Knowledge representation learning is a representation learning method oriented to entities and relationships in the knowledge base. Recently, a series of important advances have been made in this field [16]. It can efficiently calculate the semantic relationship between entities and relationships in low-dimensional space. This method can effectively solve the problem of data sparsity and significantly improve the performance of knowledge reasoning.

The vectors learned by the representation learning algorithm can find the similarity between the descriptive text data by calculating the distance or tag the document further. They are usually used in the field of emotional analysis. In the field of cyber security, the result of the representation learning algorithm is used to express the similarity between attack pattern and vulnerabilities. By matching the evaluation results with the attack context detection results, we can find the closest attack pattern and reduce the cost of manual analysis.

4.4. Alert-Context Correlation. After discovering attack steps or fragments by correlation analysis, the reconstruction of attack scenario often relies on manual analysis of security experts. However, manual analysis cannot cope with the change of targeted cyber-attacks and the dramatic increase of big security data. Therefore, it is necessary to remove irrelevant factors from fragmented attack data by attack modeling method. Correlating alerts and attack patterns can simplify the judgment process by using the intelligent method. That will extract more cognitive attack scenarios from the results and leave them to the analysts to judge, so as to improve the detection efficiency and shorten the attack detection time.

Attack modeling is an important technology to analyze attacks in cyberspace and to guide the detection and investigation of targeted cyber-attacks. Attack modeling can further help security analysts in dealing with targeted cyber-attacks. Researchers have proposed a variety of targeted cyber-attack analysis models, such as attack graph, attack chain, diamond model, pyramid model, and so on. Targeted cyber-attack scenarios can be reconstructed based on cognitive analysis model.

5. Conclusions and Perspectives

This paper discusses cyber-attacks detection from an attacker's perspective to help security professionals to have in-depth understanding of targeted cyber-attack. In this paper, we proposed a novel framework, HeteMSD, in order to enhance data relevance using heterogeneous multisource data. The proposed research work was a novel attempt in targeted cyber-attacks detection field and provided the theoretical fundamental for future research. HeteMSD is explained with technical components and corresponding methodologies. Moreover, correlation analysis against attack investigation is discussed with effective implementation using existing solutions. Various techniques for mitigation, prevention, and detection at different layers are also discussed in detail to help the defender in implementing effective defenses.

This work represents a first step in the definition of a comprehensive framework for the investigation of targeted cyber-attacks. HeteMSD still needs to be complemented with more features for the integration of the human expert, who, beyond being a simple observer, also has the knowledge required to enrich the preliminary analyses proposed by the framework. Further work must be done to anomaly detection based on multisource data fusion. More research will be done

on the extension of proposed correlation method. Lastly, security knowledge reasoning is likely to become a major topic of interest for security investigation in the near future.

Data Availability

The relevant data related to this paper is in https://github.com/kbandla/APTnotes.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Authors thank the support from The National Natural Science Foundation of China no. 61501615 and no. 61602515. The paper is also supported from the Foundation of Science and Technology on Information Assurance Laboratory (no:. 614211203010417).

References

[1] Y. Li, W. Dai, J. Bai, X. Gan, J. Wang, and X. Wang, "An intelligence-driven security-aware defense mechanism for advanced persistent threats," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 646–661, 2019.

[2] J. Navarro, A. Deruyver, and P. Parrend, "A systematic survey on multi-step attack detection," *Computers & Security*, vol. 76, pp. 214–249, 2018.

[3] Y. Liu, M. Zhang, D. Li et al., "Towards a timely causality analysis for enterprise security," in *Proceedings of the Network and Distributed System Security Symposium*, 2018.

[4] R. Zuech, T. M. Khoshgoftaar, and R. Wald, "Intrusion detection and Big Heterogeneous Data: a Survey," *Journal of Big Data*, vol. 2, no. 1, pp. 1–41, 2015.

[5] I. Herwono and F. A. El-Moussa, "A system for detecting targeted cyber-attacks using attack patterns," in *Proceedings of the International Conference on Information Systems Security and Privacy*, Communications in Computer and Information Science, pp. 20–34, Springer, 2017.

[6] A. Sood and R. Enbody, *Targeted Cyber Attacks: Multi-staged Attacks Driven by Exploits and Malware*, Syngress, 2014.

[7] A. K. Sood and R. J. Enbody, "Targeted cyberattacks: a superset of advanced persistent threats," *IEEE Security & Privacy*, vol. 11, no. 1, pp. 54–61, 2013.

[8] M. S. Khan, S. Siddiqui, and K. Ferens, "A cognitive and concurrent cyber kill chain model," *Computer and Network Security Essentials*, pp. 585–602, 2017.

[9] L. Qiang, Y. Zeming, L. Baoxu, J. Zhengwei, and Y. Jian, "Framework of cyber attack attribution based on threat intelligence," in *Proceedings of the International Conference on Interoperability in IoT*, pp. 92–103, Springer, 2016.

[10] P. Gao, X. Xiao, Z. Li et al., "AIQL: enabling efficient attack investigation from system monitoring data," *USENIX Security*, pp. 113–126, 2018.

[11] S. R. Snapp, J. Brentano, G. Dias et al., "DIDS (distributed intrusion detection system)-motivation," *Architecture, and an Early Prototype*, pp. 167–176, 2017.

[12] P. Bhatt, E. T. Yano, and P. Gustavsson, "Towards a framework to detect multi-stage advanced persistent threats attacks," in *Proceedings of the 8th IEEE International Symposium on Service Oriented System Engineering, SOSE 2014*, pp. 390–395, IEEE, UK, April 2014.

[13] M. Marchetti, A. Guido, F. Pierazzi, and M. Colajanni, "Countering Advanced Persistent Threats through security intelligence and big data analytics," in *Proceedings of the 8th International Conference on Cyber Conflict, CyCon 2016*, pp. 243–261, Estonia, June 2016.

[14] Z. Syed, A. Padia, T. Finin, M. L. Mathews, and A. Joshi, "UCO: a unified cybersecurity ontology," *AAAI Workshop: Artificial Intelligence for Cyber Security*, 2016.

[15] Y. Zheng, "Methodologies for cross-domain data fusion: an overview," *IEEE Transactions on Big Data*, vol. 1, no. 1, pp. 16–34, 2015.

[16] J. Navarro, V. Legrand, S. Lagraa et al., "HuMa: A multi-layer framework for threat analysis in a heterogeneous log environment," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 10723, pp. 144–159, 2018.