A Visualization Method Based on Graph Database in Security Logs Analysis

Xinyu Tang[1, 2], Chunguang Ma[1*,a], Min Yu[2, 3*,b], Chao Liu[2]

[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Institution of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[a] machunguang@hrbeu.edu.cn, [b] yumin@iie.ac.cn
*corresponding author

Keywords: Network Forensics; Log Analysis; Association rule mining; Graph Database

Abstract. Network security logs can provide evidence for forensic investigators. However, the network logs are facing the problems of high repetition rate, high false alarm rate, uniform format and so on, which make it difficult for forensic researchers to find useful information. In this paper, the association rules mining algorithm is used to analyze the network security logs, so as to eliminate the redundant data and to find the implied association between the logs data. Then the graph database is used to visualize the log information. Forensic analysts can obtain effective evidence by observing the graph database, which can improve the efficiency of the forensic analysts to discover the sensitive event information.

1. Introduction

Nowadays, network forensics is becoming more and more important because the number of network crime is increasing day by day. On a global scale, there were 974 data leaks that had been exposed in the first half of 2016, involving more than 554 million data records. In early June 2016, a hacker, code-named "peace" claimed to get 360 million user account and 427 million password from the world's second largest social networking site MySpace's. So the network security products cannot be completely protected network. However, the network administrator in the analysis of the network log will encounter many problems. Forensic personnel often have a large number of network security logs, but cannot get valuable information related to security incidents.

Network forensics is a branch of electronic forensics, which was first put forward by Marcus Ranum [1], and it is the technology of obtaining, analyzing and recording data packets or network logs for the purpose of investigation. As early as 1969, the United States has submitted evidence of the invasion of the site to the court. After then, the Wenke Lee Research Group [2-3] participated in the Intrusion Detection Evaluation Program, for the first time to apply data mining to the analysis of network security data. Through cluster analysis, we can find the abnormal data, classify the abnormal data, and produce the evidence [4-6]. In visualization, Richar A. Becker first proposed the concept of visualization of network data [7]. Girardind used a variety of visualization techniques to

analyze firewall logs [8]. Since 2004, industry and academia have held an annual International Conference on Visualization for Computer Security (VizSEC) [9], indicating that network security information visualization technology is gradually becoming a hot emerging technologies. Tudumi [10] is one of the earliest systems that use log data to visualize network security, Tudumi system uses 3D icons to represent system nodes, and lines are used to represent connections. Alsaleh et al. [11] designed an open source system for intrusion detection and protection for PHP-based Web applications. At present, for the network security data visualization, many methods are used, such as matrix graph, tree diagram, folding map, map, thermal map, plane coordinate, three-dimensional graphics, scatter plot and radar map [12-14]. This article uses the graph database to visualize the event information to facilitate the forensic analysis of the event.

The contributions of this article are summarized as follows, the paper releases the association rule mining algorithm in order to provide a more intuitive and more convenient interface for analysis, and improve the efficiency of the network attack case. The coordinate graph can only represent the relationship between the two attributes, and cannot represent the relationship between the attributes. So we present the graph database to visualize the data and the graph database can show the relationship between multiple attributes than other method.

The rest of the paper is organized as follows, Chapter 2 introduces the current situation of network security log processing and visualization research. Chapter 3 introduces the visualization technology used in this paper. Chapter 4 describes the network security log visualization scheme designed in this paper. In the fifth chapter, the proposed scheme is verified by real data sets. The sixth chapter summarizes the design of this paper.

2. Visualization Methods

If the intrusion detection log for a large network is monitored, there are tens of thousands of risks per day. If these logs are analyzed manually, the workload is too large and the results are unreliable. Therefore, this paper uses the Apriori algorithm to mine the log, in order to remove redundant information and get relevant evidence. We use the graph database to visualize the log after data mining. This method can provide a more intuitive and convenient interface for network security analysis system, and improve the efficiency of network attacks.

2.1. Association Rule Mining

The clustering algorithm relies on the prior knowledge, and the result is quite different from the actual result. Therefore, this paper chooses the association rule mining algorithm. Association rule mining is one of the most widely used methods in data mining. It is a method to discover useful rules from the original data. A large source for network forensics is firewall or intrusion detection system log. The data includes attribute information such as source, destination IP, source port, destination port, flow size and record type. Each frequent item represents a set of attributes that appear multiple times in the input network records. After frequent item mining, we find that the frequent item features include fixed source IP, destination IP and stream size, which can be inferred that the system may suffer port attack in a certain period of time. Collecting the correspondence records of the corresponding IPs in the frequent item sets can depict the flow of the port scanning attack events. Classical association rules mining algorithm includes Apriori algorithm and FP-TREE algorithm, and it does not need to input too much prior knowledge.

2.2. Graph Database

Some large log data processing and storage system using text files or traditional relational database to save the log data, this method has low storage space utilization, low data reliability. This way cannot effectively store and manage the massive log data. And some use of distributed storage as the underlying storage devices. These distributed storage systems use multi copy technology to ensure the reliability of the data, resulting in low utilization of storage space. And because the log data whether the original data or the final analysis of the results can only be stored in files on the distributed storage device. Access to these data can only be accessed by the file. When the file gradually increases, access to the file will also be more time-consuming. As a general data structure, graphs can be used to represent various complex relationships between data objects. Due to the limitation of scalability, traditional relational database has been unable to meet the requirements of large-scale data.

Graph database is a special kind of No SQL database, which stores the data in the graph (graph) model. The graph model can be seen as a branch of the key-value model. The difference is that it has a wide range of data association, forms a graph structure, and supports some graph structure algorithms [15-17]. Graph database is extensible and flexible, and is suitable for complex relational management and relational query reasoning. Most graph databases provide a query language which is suitable for expressing graph structure and graph query. This paper uses the graph database Neo4j to store the log data.

The graph database model is a general data structure that is modeled by digraphs and stores data in the graph [18, 19]. The graph model (GM) can be described in a four-tuple. $GM := \langle N, R, L, f \rangle$, where N denotes a finite number of no empty node set. Where R denotes the relation set of ordered node pairs (n_i, n_j) in vertex set N and satisfies, $R \subseteq N \times N$. Where L is a set of node labels and relational labels. Where f represents a function that assigns a label to a node and a relationship, namely.

Neo4j [20] uses the graph-related concept to describe the data model, saving the data as nodes in the graph and the relationships between nodes. The data consists of three parts:

- Nodes. A node represents an object instance. Each node has a unique ID which is used to distinguish with other nodes, and the node has attributes.
- Relationship. Relationship is the edge of the graph that connects two nodes. The relationship is directional and has attributes.
- Properties. Properties are key-value pairs that exist in nodes and relationships.

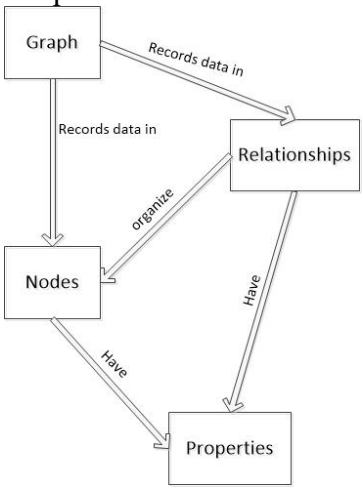


Figure 1 Structural unit of Neo4j

The nodes and relations of the graph database model correspond to the vertices and edges of the normal graph. The nodes are organized by relations. Nodes and corelations can form complex graph models in a certain direction. The nodes and relationships have multiple properties. Each different type of node or relationship is distinguished by a label. The relationship between nodes in the graph database is very flexible. The direction can be either unidirectional or bidirectional. The number of attributes can also be dynamically expanded. When using relational databases, we often encounter some complex design problems. For example, a film not only has the protagonist and supporting role, as well as director, field staff and so on. Typically, these people are often abstracted as Person types, corresponding to the same database table. However, one of the actors can be a director of the film, the host of other programs, singers, and even another investor in television works. This relationship is very complex, and there are often multiple relationships between the two entities.

In traditional relational databases, when we need to represent many-to-many relationships, we need to create an association table to record the relationship. If there are multiple relationships between the two entities, then we need to create multiple association tables between them. Based graph database, we only need to indicate that the relationship between two entities has different attributes.

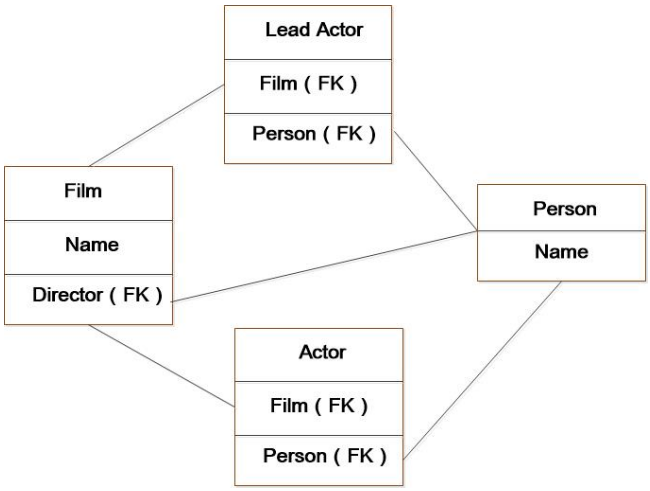


Figure 2 Relational representation of relational database

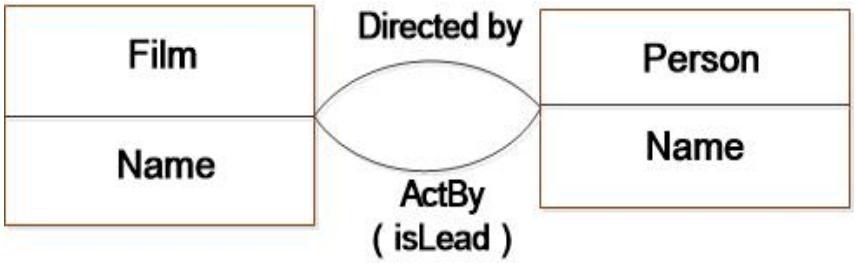


Figure 3 Relational representation of graph database

By using Neo4j to model the mentioned examples, the visualization results are shown as below:

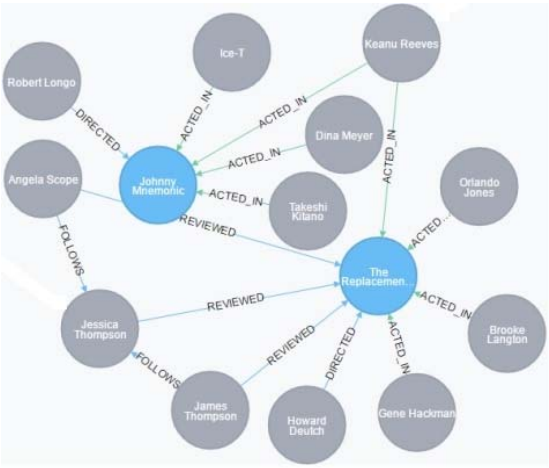


Figure 4 The relationship between characters and movies using graph database

In this paper, we use the IP as nodes, and shows the relationship between each IP address. The graph database is used to visualize the network security log event information, which is convenient for the analysis of the network security events and the subsequent forensics.

3. Visualization Scheme

First, we collect the firewall logs and IDS logs. Then select the source IP, destination IP, source port, destination port, time, and event type attribute for formatting the logs in the dataset. Through the preprocessing of the log, the other useless attribute information is removed, and only the data needed for the subsequent forensic analysis are retained. After the data is processed, we will analyze and visualize it. In this paper, association rules mining and graph database are used to process and visualize the network security log.

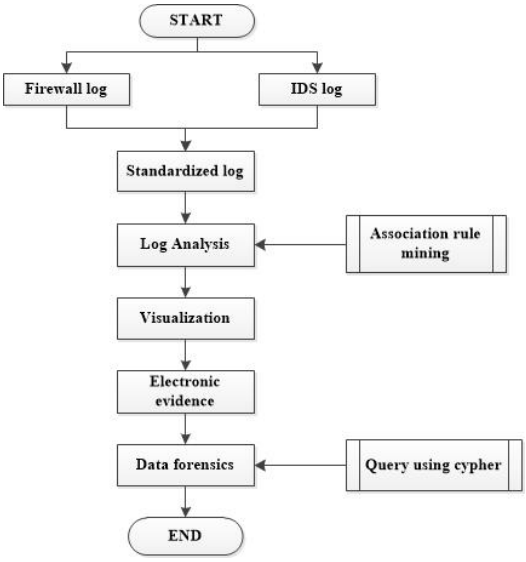


Figure 5 Visualization scheme

a. Data Preprocessing

In this paper, the firewall and IDS log data are used as the input of the system. However, the attributes of the two kinds of security devices to record event information are different, and the format is not uniform. It is necessary to preprocess the logs before visualization. The data will stored in the graph database after preprocessing.

b. Association Rule Ming

Because of the high repetition rate and false alarm rate, it is necessary to make a further filtering for the unified log before visualization. Due to the need to increase a lot of prior knowledge, there is a big gap between the aggregated data and the real situation. The association rule mining does not need to input a lot of prior knowledge in the processing, which can maximize the retention of the original data between the situations. Association rules can be used to predict the hidden rules of data sets. Association rule mining algorithms need to be processed by the following two processes, namely: to find out all the high frequency correlation item sets from the data set, and to generate association rules from the high frequency items.

c. The Visual Display

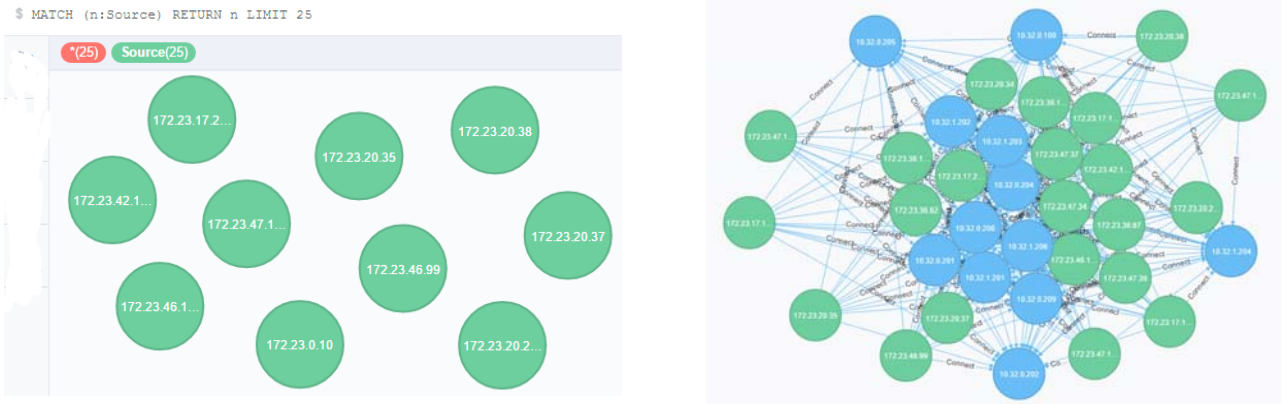
After mining association rules, we can filter out redundant and incomplete data, leaving useful log information. At present, for the network security data visualization, many methods are used, such as matrix graph, tree diagram, folding map, map, thermal map, plane coordinate, three-dimensional graphics, scatter plot and radar map. In order to overcome the relationship between multiple attributes cannot be analyzed by various visualization methods, this paper uses the graph database to visualize the associated frequent itemsets. As a general data structure, graphs can be used to represent various complex relationships between data objects. In this paper, each IP address is used as the node of the graph database.

d. Forensic Analysis

After the network security log information is displayed by the graph database, the forensic analyst can clearly observe the relationship between the various events hidden in the data. By observing the results of visualization, forensic analysts can quickly identify the types of attacks in the network. Visualization can reduce the range of attack records, so that the analyst can track the attacker's information and reconstruct the attack process.

4. Experimental Verification

This section uses the VAST2012 network security data set to validate the proposed log analysis and visualization scheme. Through the pretreatment of the data, we store the processed data into the graph database. Through the preprocessing of the log, the other useless attribute information is removed, and only the data needed for the subsequent forensic analysis is retained. Each IP address stored in the graph database and the relationship between the nodes are shown in Figure 6:



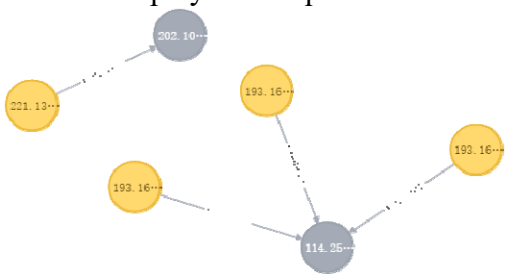
(a)The form of information storaged in graph database

(b)Connection between source / destination IP address

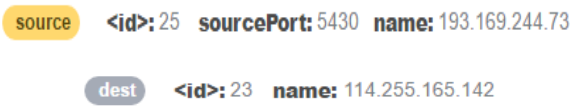
Figure 6 Data stored in graph database

Using the graph database to visualize the frequent item sets that have been extracted by frequent items, the source/destination IP addresses are represented by nodes, and the connections between

the two nodes indicate the communication between the IP addresses. According to the results of frequent items mining, Cypher is used to display the frequent items in Neo4j:



(a)Frequent itemsets display



(b)Information of each node

Figure 7 Visualization

Based on the plane coordinate method can only infer the relationship between the two attributes, but cannot determine the association between multiple attributes. At the same time, when the association is large, the use of plane coordinate will cause the association relationship is not easy to find. In this paper, we can not only get the association between IP address by using the graph database, but also show the association between ports through selecting nodes. This can effectively detect attacks, and facilitate further forensic work.

5. Conclusions and Future Work

Due to the large number of records and redundant data, it is difficult to understand the data and find valuable information. In this paper, we use association rule mining and graph database to redesign the visual analysis scheme, which is convenient for forensic personnel to quickly retrieve valuable event information. And for the network security log visualization system, real-time display and processing large-scale data is imperative. How to achieve prior warning and defense is the focus of future work.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 61472097, 61173008, 61402124, 61303244), Strategic Pilot Technology Chinese Academy of Sciences (No. XDA06010703), Young Scholar Foundation of Institute (No. 1104005704).

References

[1]Pilli, Emmanuel S., R. C. Joshi, and R. Niyogi. "Network forensic frameworks: Survey and research challenges." Digital Investigation 7.1-2(2010):14-27.

[2]Lee W, Stolfo S J. Data mining approaches for intrusion detection[C]// Conference on Usenix Security Symposium. USENIX Association, 1998:291-300.

[3]Lee W, Stolfo S J, Eskin E, et al. Real Time Data Mining-Based Intrusion Detection[C]// DARPA Information Survivability Conference & Exposition II, 2001. DISCEX '01. Proceedings. 2001:89 - 100.

[4]Rui M E, Rui P, Rong C. K-means Clustering in the Cloud -- A Mahout Test[C]// IEEE Workshops of International Conference on Advanced Information NETWORKING and Applications. 2011:514-519.

[5]Elshoush H T, Osman I M. Alert correlation in collaborative intelligent intrusion detection systems—A survey[J]. Applied Soft Computing, 2011, 11(7):4349-4365.

[6] Oyanagi S, Kubota K, Nakase A. Application of Matrix Clustering to Web Log Analysis and Access Prediction[J]. Webkdd —mining WebLog Data Across All Customers Touch Points Third International Workshop, 2001:13--21.

[7]Cook, Kristin, et al. "VAST Challenge 2012: Visual analytics for big data." 2012 IEEE Conference on Visual Analytics Science and Technology (VAST) IEEE, 2012:251-255.

[8]Takada T, Koike H. Tudumi: information visualization system for monitoring and auditing computer logs[C]//International Conference on Information Visualisation.IEEE Xplore,2002:570-576.

[9]Alsaleh M, Alqahtani A, Alarifi A, et al. Visualizing PHPIDS log files for better understanding of web server attacks[C]//Tenth Workshop on Visualization for Cyber Security. 2013:1-8.

[10]Gugelmann D , Gasser F , Ager B , Lenders V , Hviz: HTTP(S) traffic aggregation and visualization for network forensics[J]. doi: 10.1016/j.diin.2015.01.005. DIGITAL INVESTIGATION, 2015 :S1 - S11 .

[11]Mansmann F, Göbel T, Cheswick W. Visual analysis of complex firewall configurations[J]. Vizsec Proceedings of the Ninth International Symposium on Visualization for Cyber Security, 2012, 9(4):1-8.

[12]Koike H, Ohno K, Koizumi K. Visualizing cyber attacks using IP matrix[C]// Visualization for Computer Security. IEEE, 2005:91-98. Le Mal&#, Cot E, Kohara M, et al. Interactively combining 2D and 3D visualization for network traffic monitoring[C]// The Workshop on Visualization for Computer Security. 2006:123-127.

[13]Fischer F, Mansmann F, Keim D A, et al. Large-Scale Network Monitoring for Visual Analysis of Attacks[C]// International Workshop on Visualization for Computer Security. Springer-Verlag, 2008:111-118.

[14]Robinson I, Webber J, Eifrem E. Graph Databases[M]. O'Reilly Media, Inc. 2013.

[15]Qiu Shenghai,Wang Yunxia,Fan Shuhai,Jia Xiaolin.Study on application of modeling technology of graph database on cloud environment[J].Application Research of Computers,2016, 33(3):794-797.

[16]Vicknair C, Macias M, Zhao Z, et al. A comparison of a graph database and a relational database:a data provenance perspective[C]// Southeast Regional Conference, 2010, Oxford, Ms, Usa, April. DBLP, 2010:42.

[17]Diestel R. Graph theory[J]. Mathematical Gazette, 2011, 173(502):67-128.

[18]Cormen T H, Leiserson C E, Rivest R L, et al. Introduction To Algorithms[M]. MIT PRESS ; MCGRAW, 2002.

[19]Cheng M, Xu K, Gong X. Research on audit log association rule mining based on improved Apriori algorithm[C]// IEEE International Conference on Big Data Analysis. IEEE, 2016:1-7.

[20]Neo Technology.Neo4j[EB/OL].https://neo4j.com/,2012-12-19.

[21]Qiu Shenghai, Wang Yunxia, Fan Shuhai, Jia Xiaolin. Study on application of modeling technology of graph database on cloud environment[J]. Application Research of Computers. 2016, 3(3):794-797

[22]Xu Xuetian, Zou Tonghao. Design and Implementation of Network Forensics Data Acquisition System Based on Weak Association Mining[J]. Computer Measurement&Control. 2017, 25(1): 123-126