International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 11 | Nov -2017 www.irjet.net p-ISSN: 2395-0072

Using Data Mining for Discovering Anomalies from Firewall Logs: a comprehensive Review

Hajar Esmaeil As-Suhbani [1], Dr. S.D. Khamitkar [2]

[1] *Research Scholar, Yemen Nationality, School of Computational Sciences, S. R. T. M. University, Nanded , India.*
[2] *Associate Professor, School of Computational Sciences, S. R. T. M. University, Nanded , India.*

Abstract - *Firewall function as an interface of a network to one or more external networks. It is the de facto core technology of today's network security and the first line of defence against external network attacks and threats. However, due to the continuous growth of security threats, the management of firewall rules is very complex, costly and error-prone. These firewall rules are often custom designed and hand written by the human policy writer of an organization and modified to accommodate continuous changing market and business demands of the Internet. Well designed policy rules can increase the security defence effect to against security risk. Therefore, these rules are in a constant need of tuning, updating, well organized and validating to reflect the current characteristics of network traffics and to optimize the firewall security. This review covers various approaches proposed by researchers for creating, modifying the rule sets of firewall in such a way to make the correct firewall policy rule optimal, generalized and anomaly free.*

Key Words: **Firewall, logs, policy rules, anomaly detection, data mining.**

1. INTRODUCTION

Firewall is an element in networks that controls the flow of packets across the boundaries of a network based on a specific security policy. Ideally, firewall is a separate computer, which exists between a network and the Internet. It usually functions as a router which connect different network segments together [1]. It is a perimeter defense element/device that splitting network environment into internal and external network for controlling and filtering incoming and outgoing network traffic. In general, to analyze traffic and log of the packets, the first step is to bridge the gap between what is written in the firewall policy rules and what is being observed in the network [2]. Since logs have a magnificent and periodic data stored, they can be converted into datasets and be an applicant for data mining techniques.

A firewall security policy is a list of ordered filtering rules that define the actions performed when matching packets. Firewall rules are usually in the form of a criteria, and an action will be taken if any packet matches th criteria [3]. Testing a packet by the Firewall means that, the header of the incoming and/or outgoing packet is tested against all the rules in the list, which are stored in the Firewall rule set. The rules in the Firewall rule set consists all the header information like protocol type, source IP address, destination IP address, source port and destination port, and the related action to performed i.e. whether to accept or deny any packet which matches all the fields of any rule in the rule set. The rules stored in the rule set could consist of up to seven attributes. These attributes are in the following format [2]:

"<order> <prtcl> <S_ip> <S_port> <D_ip> <D_port> <action>"

Where, order indicates the number at which the rule is stored in the rule set, prtcl is the type of the protocol, s_ip and s_port are the source IP address and port number respectively. Similarly, D_ip and D_port are the IP address and port number of the destination. In the last, action field which can be either ACCEPT or DENY [3]. It defines the action to be performed on the packet which matches all the previous fields. Hence, a packet is accepted/allowed or denied/dropped by a specific rule if the packet header information matches all the network fields of this rule. An example of typical firewall rules is shown in Figure 1.

R.#	Proto	Direction	SRC IP	SRC Port	DST IP	DST Port	Action
1	TCP	INPUT	129.110.96.117	1160	129110.96.80	22	DENY
2	TCP	INPUT	129.110.96.117	1170	129110.96.80	22	DENY
3	TCP	INPUT	129.110.96.117	1160	129110.96.80	80	DENY
4	TCP	INPUT	129.110.96.117	1162	129110.96.80	80	DENY
5	UDP	INPUT	10.110.49.115	67	10.110.96.255	1211	DENY
6	UDP	INPUT	129.110.96.187	48668	10.110.96.255	67	DENY
7	UDP	INPUT	129.110.10.7	53	129.110.96.80	32790	ACCEPT
8	UDP	INPUT	129.110.10.1	53	129.110.96.80	32799	ACCEPT
9	UDP	OUTPUT	129.110.96.80	32799	129.110.96.92	53	ACCEPT
10	UDP	OUTPUT	129.110.96.80	32789	129.110.96.187	53	ACCEPT
...	...	...	...	...	...	...	...

Figure 1. Atypical Firewall rules

When the rule set becomes huge, it will be difficult to check all the rules for any redundancy, further more the updating of rule set may generate wrong set of rules. These errors in the rule set are called anomalies/outliers that have to be detected and removed from rule set for the efficient working of any firewall. Till date, existing literature shows

that five types of anomalies are discovered and studied; Shadowing Anomalies, Correlation Anomalies, Generalization Anomalies, Redundancy Anomalies, and Irrelevance Anomalies:

1. Shadowing anomaly: A shadowed rule will never be applied to any packet, when a previous rule matches all the packets that match this rule.
2. Correlation anomaly: A rule is correlated with another rule, if two rules have different filtering actions, also, the first rule matches some packets that match the second rule and the second rule matches some packets that match the first rule.
3. Generalization anomaly: Two rules, which are in order one of them is said to be in generalization of another if the two rules have different actions, and the first rule can match all the packets that match the second rule.
4. Redundancy anomaly: A rule is said to be redundant if it performs the same action on the same packets as another rule, so there is no effect on the firewall policy if one of redundant rules will be removed from the rule set [4].
5. Irrelevance anomaly: Any rule is said to be irrelevant if it does not match any of the packets either incoming or outgoing. Thus if any type of the packets do not match a rule then it is irrelevant. Therefore, there is no need to put that rule in the rule set [3].

Anomalies are defiend as a pattern in the data that do not conform to a well defined normal behavior. The cause of anomaly may be some kind of intrusion or a malicious activity. The main problem that arises in firewalls is that anomalies that generated during updating the rules in the rule set. The anomalies cannot always be categorized as an attack but it can be a suddenly behavior which is not known previously. It may or may not be harmful. This abnormal behavior found in the dataset is interesting to the analyst and this is the most important feature for anomaly detection [4]. Anomaly detection is the process of discovering the patterns in a dataset whose behavior is not normal or expected. These unexpected behaviors are also termed as outliers or anomalies. Since this involves comparing unexpected behavior against existing behavior, the data mining technology has a role in anomaly detection. The anomaly detection provides very important and serious information in various applications, like identity thefts or Credit card thefts [5]. So the most important part of research is the detection and removal of firewall anomalies. There are a number of approaches for this, which varies to each other in some implementation and performance.

2. DATA AND RULE GENERATION

In [2], Linux firewall was used to collect the logs. They processed Firewall Policy Rules to generate firewall traffic log file using Linux operating system firewall. The firewall log file contained about 33,172 packets filtered, initially with 10 firewall policy rules. However, the study used only the packet filtering log file and the application layer log files are not implemented in the prototype. Further to reduce the number of states for their research and prototype, they considered only the seven major fields in the firewall policy rules: protocol (TCP or UDP), packet direction (incoming or outgoing), source IP address and source port, destination IP address and destination port, and action. Each rule consists of seven attributes. These attributes are in the following format of "<direction> <protocol> <source-IP> <source-port> <dest-IP> <dest-port> <action>".

In [7], Snort [8]was used to log a large number of user's activities for Apriori training dataset. They used snort to collect information and activities of 10 people in period of two weeks. The collected information contained IP Addresses, Transport Layer protocols such as Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) which specify a source and destination port number in their packet headers. Snort is one of the most famous log analyzer, it is a successful light weight, open source network intrusion detector with log analyzer. It is considered as the core of Intrusion Detection System IDS, it the best alternative system to ensure network security. However, the problem is that snort system is not familiar with Windows Operating System. In [9] Intrusion Detection System IDS with snort has been implemented and configured with windows platform. The results in [9] showed that it is possible to configure snort IDS with Windows and it can be configured as a firewall. Moreover, GUI was created to allow snort user to create new rules. The GUI contsisted of two parts in which the user can select rule header information and specify the rule options to create specific functions such as controlling the access to the network or terminating the connections between two entities.

Although, Snort has been developed for relatively long time and has accumulated extensive research results, but Snort has announced on its website that it will not develop any new software on the Windows platform, and that left the current Windows users unable to use the Snort software and the various protective resources in the research communities. In [10], they used TWIDS [11] software along with Snort in Windows platform. TWIDS was designed using Windows Network Driver Interface Specification (NDIS) Filter drivers [12]. It installs the filter driver on the network interface card (NIC) drivers. After application installation, the TWIDS filter driver is displayed in the properties of the network connection. TWIDS software development is focused on the protective application transfer to Windows desktop and improvement of forensic technology. AS a gateway needs to be set up by a technician for Snort to bring together and analyze all the network traffic, TWIDS software does not require this process, because it directly transfer the protective applications to Windows desktop computers. Snort filters packets by the network gateway, but it is unable to obtain the internal process information of Windows. In

[10], the work showed that security software on Windows OS, using the snort rules to filter and drop malicious packet on Windows desktop computers. Also, they showed that TWIDS software could detect and block possible attackers in real time and alert to network technicians.

3. RULE FILTERING AND ANOMALY DETECTION

For filtering firewall rules a number of models have been proposed. In [13], ordered binary decision diagram is used as a model for optimizing packet classification. In [14] the model uses bucket filters indexed by search trees. In [15] another model using tuple space is developed, this model combines a set of filters in one tuple and stored in a hash table. The analysis in [4] showed that the tree-based model was simple and powerful enough to use for packet filtering. Other approaches in [16] propose using a high level language to define rules. This approach can avoid rule anomalies but it is not practical. In [17] E. Al-Shaer and H. Hamed developed a tool called the "Firewall Policy Advisor" for analyzing firewall polices. This tool discoverd all anomalies that could exist in a single or multi firewall environment, also it simplified the management of filtering rules and maintains the security of next generation firewalls. Also they presented a set of algorithms and techniques to automatically discover policy anomalies in centralized and distributed firewalls, but the drawback is this tool does not consider device logs or network traffic. Saboori et al. in [7] use Apriori Algorithm to detect anomalies in the network. This algorithm prophesies a novel attack and generates a set of real time rules for the firewall. This algorithm functions by extracting the correlation relationships among the large data sets.

Many studies have been done on detecting and resolving anomalies in firewall policy rules. In [18] they developed a method to detect and remove anomalies in network traffic. A comparison is done on current network traffic against baseline distribution, which gives a multidimensional view of network so that the administrator can easily detect anomalies that cause changes in the network traffic and it also provides the information about the detected anomalies. In [4,19,20], Al-Shaer et al. define the possible relations between firewall rules, and then define anomalies that can occur in a rule set in terms of these definitions. They also give an algorithm to detect these anomalies, and present policy advisor tools [17] using these algorithms and defentions. In [21] Fu et al., proposed a technique for anomaly detection in unstructured system logs that does not require any application specific knowledge. Also, a method have been included to extract log keys from free text messages. A work that focuses on detecting and resolving anomalies in firewall policy rules is implemented in [22], where they proposed a scheme for resolving conflicts by adding resolve filters. The limitation of thier technique is the algorithm requires the support of prioritized rules, which is not always available in firewalls. In[23], they outlined a novel anomaly management systematic detection and firewall policy to facilitate the resolution of discrepancies and proposed a partition rule based system and represent a grid based techniques to achieve the goal of efficient and effective discrepancy analysis were introduced to them. In [2]Golnabi et al.describe a Data Mining approach to the anomaly resolution. The analysis in [4] showed that the tree based model was simple and powerful enough to use for packet filtering. Instead diagrams and pre-processing of firewall rules are used in [24] to compact firewall policy rules and to resolve rule overlapped. However, this method cannot be used for anomaly detection. There are a few related work [25] address rule combination in filtering policies and a number of papers with emphasis on filtering performance [26].

4. DATA MINING TECHNIQUES

As the rules set of the firewall becomes enormous, the process of manually managing firewall policy rules becomes very difficult and time consuming. In general, there are some works proposing usage of data mining methods in log file analysis process. Data Mining is a predictive model and approach to explore large amount of datasets in search of finding an interesting pattern or trend in a large dataset to guide decision for further analysis [2]. It is a process of automatic knowledge extraction and it consists of four classes of tasks; association rule learning, clustering, classification and regression [27]. The main advantage of data mining is a fast way of gathering information. For that, there are a few approaches such as Association Rule Mining (ARM) [28] and decision tree [13]. Decision Tree algorithm learns a function represented as a decision tree, where each node in the tree tests as an attribute, and each branch corresponds a value comparison, and each leaf node assigns classification. The disadvantage of this approach includes handling of continuous attributes, growing tree, and computational efficiency [2]. Data mining is applied after standard log analysis in order to provide outcome of better quality, when considering the logging problem. A few recent advances especially with Association Rule Mining (ARM) technique, in the area of anomaly detection using data mining techniques, implemented by Lee [29] and Mahoney [30]. Association Rule Mining (ARM) algorithm searches the space of all possible patterns for rules that meet the user specified support and confidence thresholds. One example of an association rule algorithm is the Apriori algorithm which was designed by Srikant and Agrawal [31] and a complete survey for Association Rule Mining can be found in [32]. Apriori algorithm was used in [7], this algorithm extracts interesting correlation relationships among large set of data items. Also, they used Snort to record logs of user activities, and then Apriori algorithm have been used to create a model. This model can be used to create online rules for firewall based on current user activities.

A method to figure out firewall policy rules by mining its network traffic log with Association Rule Mining and mining firewall log using frequency was introduced in 2006 by

Golnabi et al., they described a Data Mining approach to the anomaly resolution in [2]. They have presented a new process of managing firewall policy rules which can detect many hidden and undetectable anomalies, consisting of anomaly detection, generalization and policy update using Association Rule Mining and frequency-based techniques. Also it identifies decaying rules and treats them accordingly. This prototype results in an anomaly free firewall rule set which is based on the dynamic network traffic logs' mining. In [33], they analyzed data mining methods for anomaly detection and proposed an approach for discovering security breaches in log files. Also, they employed Apache Hadoop framework to support parallel processing for distributed storage and distributed processing of data, so allowing computations to run in parallel on several nodes and therefore, speeding up the whole process. The outcomes of their testing showed potential to discover new types of breaches and plausible error rates. Also, rule generation and anomaly detection speeds are competitive to currently used algorithms, such as FP-Growth and Apriori. Another optimization was done in transformation of data into binary form, that making it more efficient to analyze particular transactions. The work in [34] described log file types, formats and contents and provided an overview of web usage mining processes. For the purpose of getting more information about users, they used data mining methods for analyzing web log files.

5. CONCLUSION

One of the most important factor of network security system are firewall policy rules. However, due to the continuous growth of security threats, the management of firewall rules is very complex, costly and error-prone. . Using firewall technology is the first important step toward securing networks. The management of policy rule is the main task for the network security. It plays the major role in management of any organization's network and its security infrastructure, but the firewall security effectiveness may be limited by a poor management of firewall policy rules. There have been a number of tools and techniques that used to perform rule editing and anomaly detection by utilizing a given set of existing policy rules. The reviewed literature has shown us an interesting problem "How much the rules are practical, up-to-dated, efficient and well-organized in order to reflect current characteristics and volume of network packets?". Most of the papers are intended to perform the anomaly detection and removal by using different techniques, to obtain the correct firewall policy rules generalized and anomaly free.

REFERENCES:

[1] Boddi Reddy Rama, Ch.SrinivasaRao, K.Naga Mani, " Firewall Policy Management Through Sliding Window Filtering Method Using Data Mining Techniques ", (IJCSES) Vol.2, No.2, May 2011.

[2] KoroshGolnabi, Richard K. Min, Latifur Khan, Ehab Al-Shaer, "Analysis of Firewall Policy Rules Using Data Mining Techniques ", IEEE, 2006.

[3] Rupali Chaure, Shishir K. Shandilya, " Firewall anamolies detection and removal techniques – a survey", International Journal on Emerging Technologies, 2010.

[4] Ehab Al-Shaer and HazemHamed, "Discovery of Policy Anomalies inDistributed Firewalls" in Proc. of IEEE INFOCOM'04, vol. 23, no. 1,March 2004 pp. 2605-2616.

[5] Dokas P., .Ertoz L., Kumar V., Lazarevic A., Srivastava J., Tan P. N., Data mining for network intrusion detection, In Proceedings of NSF Workshop onNext Generation Data Mining; 2002; p. 21-30

[6] Chandola V., Banerjee A. , Kumar V., Anomaly detection: A survey, ACM Computing Surveys (CSUR); 41(3); 2009;p. 15 .

[7] Saboori E, Parsazad S, Sanatkhani Y. Automatic firewall rules generator for anomaly detection systems with Apriori algorithm. 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE); 2010. p. 57–60.

[8] Snort. An open source network intrusion detection system. http://www.snort.org/.

[9] Moath Alsafasfeha, Abdel Ilah Alshbatatb, " Configuring Snort as a Firewall on Windows 7 Environment ", Journal of Ubiquitous Systems & Pervasive Networks, 2011.

[10] Shin-Shung Chen, Tzong-Yih Kuo, Yu-Wen Chen, " Security Software based on Windows NDIS Filter Drivers ", IEEE, 2013.

[11] TWIDS : http://twids.cute.edu.tw/en.

[12] MSDN, NDIS Filter Drivers (Windows Drivers), http://msdn.microsoft.com/en-us/library/windows/hardware/ff565501(v=vs.85).aspx .

[13] Mitchell, T.M., Machine Learning. 1997, Sydney: McGraw-Hill.

[14] Webb, G.I. Discovering Associations with Numeric Variables. In Proceedings of the International Conference on Knowledge Discovery and Data Mining. 2001: ACM Press.

[15] Piatetsky-Shapiro, G., Discovery, analysis, and presentation of strong rules. Knowledge Discovery in Databases, 1991: p. 229-248.

[16] Y. Bartal., A. Mayer, K. Nissim and A. Wool. "Firmato: A Novel Firewall Management Toolkit." Proceedings of 1999 IEEE Symposium on Security and Privacy, May 1999.

[17] E. Al-Shaer and H. Hamed. "Firewall Policy Advisor for Anomaly Detection and Rule Editing." IEEE/IFIP Integrated Management Conference (IM'2003), March 2003.

[18] Yu Gu, Andrew McCallum and Don Towsley. "Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation", Tech. rep., Department of Computer Science, UMASS, Amherst, (2005).

[19] E. Al-Shaer and H. Hamed. Firewall policy advisor for anomaly detection and rule editing. In IEEE/IFIP Integrated Management Conference (IM'2003), March 2003.

[20] E. Al-Shaer and H. Hamed. Taxonomy of conflicts in network security policies. IEEE Communications Magazine, 44(3), March 2006.

[21] Q. Fu, J.-G. Lou, Y. Wang, and J. Li. Execution anomaly detection in distributed systems through unstructured log analysis. In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, ICDM '09, pages 149{158, Washington, DC, USA, 2009. IEEE Computer Society.

[22] A. Hari, S. Suri, and G. M. Parulkar. Detecting and resolving packet filter conflicts. In INFOCOM (3), pages 1203–1212, March 2000.

[23] Rana Sabah Naser, YashwantraoMohite, "USING DATA MINING DETECTING AND RESOLVING FIREWALL POLICY ANOMALIES ", IJCET, 2014.

[24] J. Guttman. "Filtering Posture: Local Enforcement for Global Policies." Proceedings of 1997 IEEE Symposium on security and Privacy, May 1997.

[25] D. Eppstein and S. Muthukrishnan. "Internet Packet Filter Management and Rectangle Geometry." Proceedings of 12th!Annual ACM-SIAM

Symposium on Discrete Algorithms (SODA), January 2001.

[26] Z. Fu, F. Wu, H. Huang, K. Loh, F. Gong, I. Baldine and C. Xu."IPSec/VPN Security Policy: Correctness, Conflict Detection and Resolution." Proceedings of Policy'2001 Workshop, January 2001.

[27] Berkhin P., A survey of clustering data mining techniques;Grouping multidimensional data; Springer Berlin Heidelberg; 2006; p. 25-71.

[28] Agrawal, R., T. Imielinski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. in Proceedings of the 1993 Webb, G.I., Association Rules, in Handbook of Data Mining, N. Ye, Editor, Lawrence Erlbaum: To appear.

[29] W. Lee, "A Data Mining Framework for Constructing Features and Models for Intrusion Detection", Ph.D. Dissertation, Columbia Univeristy, 1999.

[30] M. Mahoney, ''A Machine Learning Approach to Detecting Attacks by Identifying Anomalies in Network Traffic'', Ph.D Dissertation, Florida Institute of Technology, 2003.

[31] Srikant, R., Q. Vu, and R. Agrawal. Mining Association Rules with Item Constraints. in Proceedings of the 3rd International Conference on Knowledge Discovery in Databases and Data Mining. 1997. Newport Beach, California.

[32] Agrawal, R. and R. Srikant. Fast Algorithms for Mining Association Rules. in Proceedings for the 20th Int. Conf. Very Large Data Bases.1994.

[33] Jakub Breier, Jana Brani_sov_ay, " A Dynamic Rule Creation Based Anomaly Detection Method for Identifying Security Breaches in Log Records ", ResearchGate, 2015.

[34] L.K.J. Grace, V. Maheswari, and D. Nagamalai. Web log data analysis and mining. In Natarajan Meghanathan, BrajeshKumar Kaushik, and Dhinaharan Nagamalai, editors, Advanced Computing, volume 133 of Communications in Computer and Information Science, pages 459{469. Springer Berlin Heidelberg, 2011.