SNORT LOG ANALYSIS WITH DATA MINING AND VISUALIZATION

[1]JAY GHOLAP, [2]SOURABH ARSEY, [3]JOSEPHINE M. NAMAYANJA

[1,2]Department of Information Systems, University of Maryland, Baltimore County, Baltimore, USA
[3]Management Science & Information Systems, University of Massachusetts, Boston, Boston, USA
E-mail: [1]jgholap1@umbc.edu, [2]arsey1@umbc.edu, [3]Josephine.Namayanja@umb.edu

Abstract— with the growing sophistication of cyberattacks, it has become necessary to combine techniques such as data mining into cyber security. However, the utilization of techniques such as association rule mining is still an open challenge in the context of cyber security. This study proposes the use of association rule mining to be applied to Snort logs before signature matching as primary check in order to detect intrusions. With association rules, it is possible to gain valuable insight within Snort logs in order to find key relationships. On the other hand, given that a large number of logs can be generated in Snort, this creates a possibility for identifying a large number of association rules which can make the process of analysis challenging for a user. Therefore, this study extends itself to integrate the process of association rule mining with data visualization to create a better representation of patterns discovered.

Index Terms— Association Rule Mining, Data Visualization, Intrusion Detection.

I. INTRODUCTION

Cyber-attacks have been increased significantly in last few years which has essentially increased the demand for improved cyber defense systems. With the growing sophistication of attacks, it has become necessary to combine techniques such as data mining into cyber security. Data mining is the process of finding the hidden information from a given data set. It is an analytic process in which we search for consistent patterns and systematic relationships between attributes and then validate the discoveries. One of the key patterns in data mining is association rules which identifies dependencies and relations among different attributes present in the data sets [1]. Association rules are generally represented as implications of the form x->y. Generating association rules is a two phase process. First phase involves derivation of frequent item sets based on threshold value of support. In second phase, possible rules are formed from the frequent item sets based on a given threshold value of confidence. Manageable sized rules can be obtained by increasing minimum support value, however it may remove interesting rules with less support. As a result, this often makes it inevitable to deal with large set of association rules derived from the data [2]. According to [2], there are two problems with association rule mining. First, too many rules make human exploration difficult. Another issue is that visualization of multiple item set rules is challenging as they are difficult to understand.

The utilization of association rule mining is still an open challenge in the context of cyber security. For example, Snort which is a packet sniffer and logger is a popular lightweight network intrusion detection system. It utilizes a rule based approach to perform content pattern matching and detect a variety of attacks and probes, such as buffer overflows, stealth port scans, CGI attacks, SMB probes, and much more [10]. In Snort, attacks are detected based on previously known signatures. However, extensive signature matching of incoming packets takes high computational costs in order to perform multilayer signature checks of sophisticated well-known attacks.
In this study, we propose the use of association rule mining to be applied to Snort logs before signature matching as primary check in order to detect intrusions. With association rules, it is possible to gain valuable insight within Snort logs in order to find relationships between key attributes such as source and destination of a specific type of attack. Given that association rules are an unsupervised technique, they can be used to execute basic checks in the form of 'if then else' statements which would essentially pose a lower cost of operation compared to signature matching which requires previously known information. Overall, this can significantly increase the performance of the intrusion detection process in Snort.

Also given that a large number of logs can be generated in Snort, this creates another possibility for identifying a large number of association rules which can make the process of analysis challenging for a user. Henceforth, we propose to integrate the process of association rule mining with data visualization. As such we present a graphic-based approach for mining Snort logs in order to make the discovered patterns more comprehensible for users. Our approach is tested using varying visualization techniques and parameters where we clearly demonstrate the benefit of graphic based pattern analysis over text-based patterns in association rule mining.

II. RELATED WORK

Snort logs have been used extensively to mine frequent item sets to improve performance of

anomaly based intrusion detection systems. However, a large number of logs poses challenges for efficient analysis. A study by [11] propose a visualization system of Snort logs called 'SnortView' where they utilize a heuristic based approach to identify key pieces of information that are critical in the investigation of an attack. While [11] is based on identifying anomalies in network traffic, interestingly [5] introduces the Minnesota Intrusion Detection System (MINDS) which uses a suite of data mining techniques to automatically detect attacks against computer networks and systems. First MINDS uses an anomaly detection technique that detects anomalous network connections based on an assigned score. Next, it utilizes an association pattern analysis based module that generates a summary of network connections with high anomalous ranking.

The association pattern analysis is used to identify associations in anomalous network traffic which are utilized as a base to generate new rules for known attacks. On the other hand, [4] presents Association Rule Explorer which allows users to visually explore association rules to find interesting relationships in the various attributes of the dataset. Here the authors propose a system that can make it easy to detect patterns with the help of visualizations rather than examining association rules independently. Comparatively, [3] propose a simple unified interface known as 'arulesViz' which provides a framework to visualize association rules with easily interpretable presentation. This framework supports state of the art visualizations such as scatter plots, graphs, matrix visualizations, mosaic plots among others. They also propose a novel method of clustering rules known as 'grouped matrix based visualization'. Given that a combination of data visualization and association rule mining is still an open challenge, this study aims to combine both techniques reviewed from the literature to mine association rules in Snort logs in order to identify critical components of an attack. Our objective is to create a comprehensive framework for visualizing the association rules identified in Snort logs in order to create an understandable view of attack patterns. Next we discuss our methodology.

III. METHODOLOGY

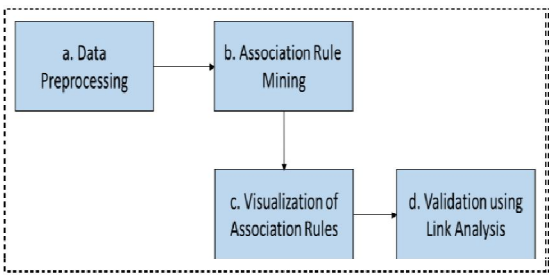

Figure 1: Overall Approach

Figure 1 shows the overall approach. In 1a), we perform data preprocessing on parsed Snort log data and then use Apriori algorithm to generate association rules and patterns in 1b). In 1c), we use RStudio to create visualization of the generated patterns. Finally, we validate the discovered patterns for graph-based visualization using link analysis in 1d).

A. Data Preprocessing

In this study, we examine our approach on CDX - Cyber Defense Competition 2009 Snort logs. This dataset was generated between November 8th, 2009 through November 12th, 2009 during a Cyber Defense Competition [12].This competition comprises of three teams. Specifically the red team is the attacker, the blue team is the defender and the white team is the traffic generator which is used to simulate the normal network traffic data. In this study, we focus on readily parsed and processed data collected from the prior experiments carried out by [6].

For our experiments, we selected the following attributes; Source_IP, Target_IP, Timestamp, Alert_Message, Classification, Priority and Protocol. Additionally, given that the Timestamp is captured as a continuous time attribute, we applied binning where we divided the timestamp into four daytime bins. Particularly 6 AM to 12 PM as Morning, 12 PM to 5 PM as Afternoon, 5 PM to 8 PM as evening and 8 PM to 6 AM as Night. Our objective was to create more defined time periods for discovering association rules. Hence, we renamed the Timestamp attribute as Daytime with the values of corresponding to Morning, Afternoon, Evening and Night respectively. We also select the date and time fields which we merged into a single column in order to divide the data into temporal bins of 15 minutes each. More so, the data set describes 4 protocols used (one at a time) which all have the TTL (Time to Live) field value. For purposes of this study, we convert these 4 different TTL fields into a single field as shown in figure 2. We thus create a single field named Protocol. We performed this data transformation using Talend data integration (ETL) tool [7].

TCP_TTL	UDP_TTL	ICMP_TTL	PIM_TTL	Protocol
63				TCP
	126			UDP
		125		ICMP

Figure 2: Example of Data Transformation for Network Traffic Protocols

Also, for those missing values associated with normal packets, we labelled those as "normal" in the classification attribute. In summary, a list of attributes used in this study is provided in table 1.

Table 1: Summary of Attributes

Attribute	Description
Datetime	Merged date and time
Source_IP	Source IP address
Target_IP	Destination IP address
Alert_Message	Alert generated by Snort
Classification	Type of attack or normal behavior
Priority	Level of priority
Daytime	Time ranges such as morning, afternoon, evening.
Protocol	Time to Live (TTL) protocols

C. Association Rule Mining

Next, we apply Apriori algorithm on the Snort logs. Specifically in this study we utilize arules package in R where we adjust the support and confidence values accordingly. Our objective is to determine a suitable number of rules that are representative for efficient analysis. A sample of our code is provided in figure 3 as used in the R.

Figure 3: R code for association rule mining.

D. Visualization of Association rules

Following the identification of association rules, we build graph and grouped matrix interactive visualizations using R package arulesviz. We use graph based visualization of association rules because it is very easy to interpret useful patterns compared to text based representation. Similarly, group based matrix of association rules facilitates the interpretation of associations, although it is effective for visualizing only a small number of rules. Given that we use varying support and confidence thresholds to identify association rules, we also create visualizations corresponding to the respective threshold values for support and confidence.

E. Data Validation using Link Analysis

We extend our approach by validating significant association rules using link analysis. Hence, we study the connections between the source and target IP addresses where we determine the degree centrality for given nodes. Our goal is to identify relevant association rules which consist of nodes or IP addresses that are highly central nodes on the network.

Next, we discuss our experimental results.

IV. EXPERIMENTAL RESULTS

A. Visualization of Association Rules

For our experiments, we extracted association rules with a minimum length of 2 item-sets. Given aconfidence as 0.5 and support as 0.001, figure 4 illustrates a textual representation of rules identified.

```
{priority=Priority: 3} => {DayTime=Afternoon}
{Classfication=Classification: Misc activity} => {DayTime=Afternoon}
{Source_IP=10.1.XX.XX}=>{Classfication=Classification: Attempted Administrator Privilege Gain}
```

Figure 4: Textual representation of rules

In figure 4, we present selected interesting association rules sorted by the confidence level. Here we see that the first rule indicates that most of the attacks with priority 3 were carried out in the afternoon time window. Similarly, the last rule indicates that source IP 10.1.XX.XX attempted administrator privilege gain. It should be noted that for our findings in this study, we mask the IP addresses to maintain anonymity. Now, we apply visualization on the rules identified as shown in figure 5.

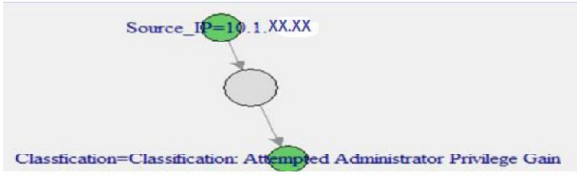

Figure 5: Graph based visualization of rules

In figure 5, the relationship between the Source_IP and attack type is clearly portrayed compared to the textual representation in figure 4. It is clear seen that the Source_IP 10.1.XX.XX is responsible for the administrator privilege gain. To further our analysis, we expand our graph visualization where we increase the display size for
association rules as well as varying support and confidence levels. We also evaluate our method using varying visualization tools, particularly R and NodeXL [9].

Hence in figure 6, our findings indicate that the ICMP protocol packets are associated with miscellaneous network activity in the afternoon time period. It can be argued that this is related to a possible Smurf attack which uses the ICMP protocol. We also see that different nodes have different sizes and colors where the size of
the node indicates the support of rule while the color indicates the lift of association rule.

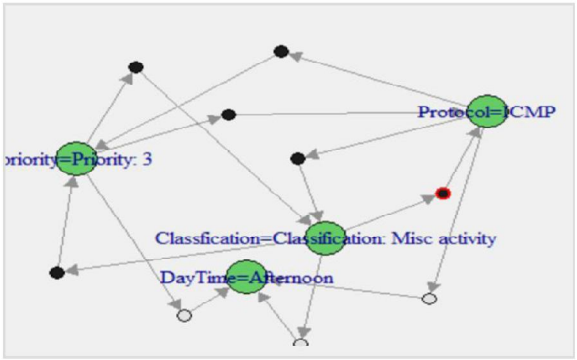

Figure 6: Top 10 association rules with support= 0.001,confidence = 0.5

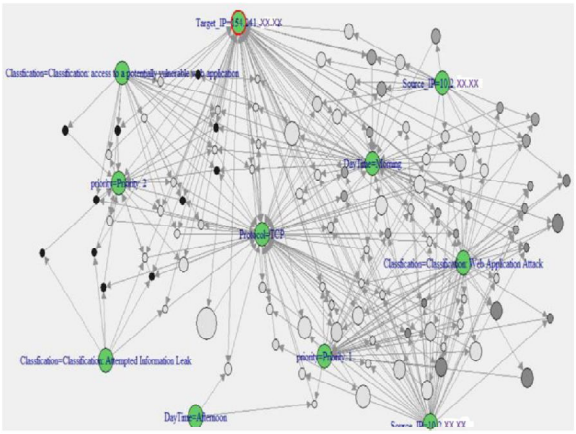

Figure 7: Graph based visualization of more than 100 rules using R

Alternatively, figure 7 shows that two Source_IPs 10.2.XXX.XXX and 10.2.XXX.XXX, try to access to a potentially vulnerable web application for web application attack on the same target node 154.241.XX.XXX. Also both attacks are taking place in the morning time period with priority 1. We could also see that there was an Attempted Information leak attack on the same target node. Therefore, this node proved to be highly vulnerable and destination of many attacks. It should be noted that these findings are based on support 0.1, confidence 0.6 and a minimum rule length 3.

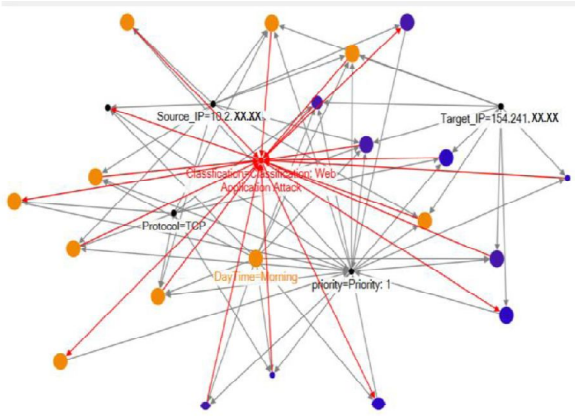

Figure 8: Graph based visualization of rules using NodeXL

In figure 8, using support 0.2 and confidence 0.8 we extracted rules and sorted by descending order of lift. We exported the graph generated with the help of arulezviz package in graphml format and imported it in NodeXL for better visualization. Here we took the top 20 rules into consideration and set vertex color as support and vertex size as confidence, where the size of the rule depicts confidence of that rule and color depicts the support. Interestingly, our findings indicate that the visualization graph with NodeXL was even more interpretable along with interactivity. It highlighted the same attack as discussed in figure 7 identified as priority 1 on node 154.241.XX.XXX. Comparatively, we use grouped matrix based visualization as shown in figure 9.

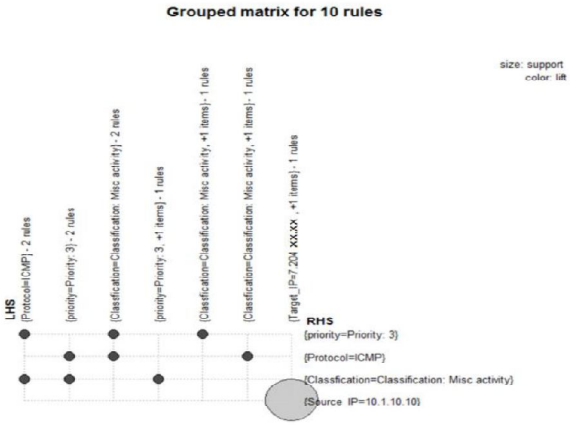

Figure 9: Grouped matrix-based visualization of association rules.

While the grouped matrix based visualization provides a good representation of the discovered association rules, it simply displays the rules with LHS (left-hand-side) and RHS (right-hand-side) as generated in text form which is not as effective as graph visualization shown in figures 6, 7 and 8. Overall, our findings for graph based visualization of association rules provides better representation and interactivity for rule analysis in comparison to text based analysis of rules.

B. Validation using Link Analysis

For our validation, we computed the degree centrality for all the nodes grouped by daytime. Our goal is to detect important nodes in the network by computing degree centrality of nodes. Our findings indicate that the node 154.241.XX.XX had the highest total degree centrality (greater than 20,000). This is seen across two daytime periods of morning and afternoon as shown in table 2.

Table 2: Summary of Total Degree Centrality for Validation

Node	Daytime	Centrality
154.241.XX.XXX	Morning	13527
154.241.XX.XXX	Afternoon	7057
224.0.XX.XX	Night	1149
7.204.XX.XX	Morning	887
7.204.XX.XX	Afternoon	646
224.0.X.XX	Evening	366
134.240.XX.XX	Night	347
3.75.XX.XX	Morning	306
3.75.XX.XX	Night	287
3.75.XX.XX	Afternoon	280

On the other hand, the total degree centrality level of other nodes was significantly lower than 1000. A further analysis of in and out degree respectively in table 3 indicates that 154.241.XX.XX had a high in-degree of 20584 compared to an out-degree of 403. Interestingly, our results visualizing association rules

clearly show that this node was a target of web application attack. On the other hand we also examine other nodes discovered in our visualization analysis. Particularly, we found that node 10.2.XX.XXX had the highest outdegree 6282. This node can also be a potential cyber threat given that the visualization results showed that this was the source node for the web application attack on the target node 154.241.XX.XX. We provide a summary for top in and out degrees in table 3 respectively.

Table 3: Summary of In and Out Degree for Validation

Node	Indegree
154.241.XX.XXX	20584
224.0.XX.XX	1595
7.204.XX.XX	1536
3.75.XX.XX	987
134.240.XX.XX	477
10.1.XX.XX	118
10.1.XX.XX	116
180.242.XX.XX	84
10.1.XX.XX	63
10.1.X.XX	28

Node	Outdegree
10.2.XXX.XXX	6282
10.2.XX.XXX	3825
10.2.X.XXX	1758
134.240.XX.XX	1595
10.2.X.XX	1458
10.2.XXX.XX	1323
31.154.XXX.XX	752
10.2.XX.XX	687
10.2.XXX.X	686
10.2.XX.XXX	631

It should be noted that our validation was conducted using Apache Drill [8] with single node.

C. Discussion

Our findings clearly indicate that data visualization provides a good representation of rules. Additionally, a further analysis of network connectivity using degree centrality analysis supports our results found using graph based association rules by identifying key sources and targets for an attack. This also suggests that degree centrality can also be a useful measure to detect the potential cyber threats. Most importantly, the derived association rules from Snort logs can be integrated with signature database of Snort in order to detect intrusions. Top N rules based on lift value associated with them can be picked up and merged with signature database. This will significantly reduce the processing overhead on Snort as these rules will quickly find intrusions with the reduced computational cost as compared to extensive signature matching.

CONCLUSION AND FUTURE WORK

This study evaluates the use of data visualization on association rules to identify interesting cyber-attack patterns from Snort logs. Our study concludes that graph based visualization is very useful in cases where there are several rules which are not feasible to read in textual format. Additionally, combining derived rules into signatures can significantly improve the performance of signature based IDS as it reduces signature matching of incoming packets. In this study, we were also successfully able to validate the results found using association rules and visualization using degree centrality analysis.

For our future work, we plan to explore sampling of association rules in order to reduce large number of redundant rules. We also plan to explore our approach in high dimensional datasets and large dataset using other association rule algorithms such as FP-growth where the Apriori algorithm.

REFERENCES

[1] Gosain, A., Bhugra, M. (2013). A comprehensive survey of association rules on quantitative data in data mining. IEEE Conference on Information & Communication Technologies (ICT), 2013, vol., no., pp.1003 - 1008, 11- 12 April 2013.

[2] Zaki, M., Phoophakdee, B. (2003). MIRAGE: A framework for mining, exploring and visualizing minimal association rules. In Computer Science Dept., Rensselaer Polytechnic Inst.

[3] Hahsler, M., Chelluboina, S. (2011). Visualizing association rules: Introduction to the R-extension package arulesViz. R project module. 223-238.

[4] Liu, Guimei, et al. (2012). AssocExplorer: an association rule visualization system for exploratory data analysis." Proceedings of the 18th ACM SIGKDD International conference on Knowledge discovery and data mining. ACM, 2012.

[5] L. Ertöz, E. Eilertson, A. Lazarevic, P.-N. Tan, V. Kumar, J. Srivastava, P. Dokas. (2004). The MINDS – Minnesota Intrusion Detection System. Next Generation Data Mining. MIT Press, Boston (2004).

[6] Song Chen, Vandana P. Janeja. (2014). Human perspective to anomaly detection for Cybersecurity. Journal of Intelligent Information Systems. 42, 1.

[7] https://www.talend.com/products/talendopen-studio

[8] https://drill.apache.org/

[9] http://nodexl.codeplex.com/

[10] Roesch, Martin. (1999) Snort - Lightweight Intrusion Detection for Networks. Stanford Telecommunications.

[11] Hideki Koike, Kazuhiro Ohno. (2004). SnortView: visualization system of Snort logs. In VizSEC/DMSEC '04: Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security (2004), pp. 143-147, doi:10.1145/1029208.1029232

[12] Dodge Jr, R.C., & Wilson, T. (2003). Network traffic analysis from the cyber defense exercise. In IEEE international conference on systems, man and cybernetics, 2003 (vol. 5, pp. 4317-4321). IEEE.

★★★