A Behavior-Based Method for Rationalizing the Amount of IDS Alert Data

Teemu Alapaholuoma, Jussi Nieminen, Jorma Ylinen, Timo Seppälä, Pekka Loula
Telecommunication Research Center
Tampere University of Technology, Pori Unit
Pori, Finland
teemu.alapaholuoma@tut.fi, jussi.nieminen@tut.fi, jorma.ylinen@tut.fi, timo.a.seppala@tut.fi, pekka.loula@tut.fi

***Abstract*—Intrusion detection systems typically rely on signatures. A signature describes a rule, which is realized as an alert whenever an IP packet matching the rule is observed in the network by an intrusion detection system. In the configuration phase of a signature based intrusion detection system, the operator usually activates the signatures considered interesting. Interesting typically refers to aberrant traffic and behavior in the network. The classification of signatures as interesting or uninteresting is typically based on prior knowledge about the characteristics of the monitored network. In this paper, we introduce a method based on network behavior for identifying, which alerts and signatures could be considered interesting. Based on the identification, only the signatures labeled as interesting should be activated, in order to rationalize the amount of alert data produced. The method is based on the K-means clustering of intrusion detection system alert data.**

Keywords-Alert; Detection; Intrusion; Clustering; Snort

I. INTRODUCTION

The unstoppable increase in data transferred on the Internet is reflected directly in the amount of measurement and maintenance data produced in operator systems. This is also the case with IDS (Intrusion Detection System), providing data about threats and intrusion attempts in the network. It is impossible for the operator to analyze every alert manually, and make conclusions about the severity or relevancy of the alerts. Usually the operator classifies the signatures as interesting or uninteresting. The classification can be based on prior knowledge about the target network characteristics, or simply on some specific point of interest. The operator might be interested only in a small amount of signatures, and activate only those. On the other hand, the operator might know that a certain service is never used in the network and the signatures related to it can be deactivated. This paper describes a behavior-based method for classifying signatures as interesting or uninteresting. The method is based on passive data analysis of alert data generated by an IDS system. In the data generation phase, all available signatures are activated. K-Means [1] clustering is utilized as a classification method. The method learns the normal behavior from the alert data, meaning that those alerts can be labeled as uninteresting.

As a data provider we have used Snort [2], an open-source signature-based IDS system. Snort was chosen because it is the leading open-source IDS system in the market. It is actively maintained, and the signature set is updated frequently. Snort is also known to cope with a large packet throughput rate, which is an essential feature in this particular analysis.

The target network was a campus area network. It is known that the information security policy in the target network is fairly free. From the data point of view, this is a benefit. If there were many restrictions in the network, the alert data would not be as generally applicable as it is in this case. The monitoring and analysis execution rely on anonymity. The IP address details of the alerting IP packet are anonymized in the packet capture phase, so the IDS system does not know the real IP addresses communicating in the network.

The data is pre-processed before the actual analysis. The arguments for clustering are computed from the raw alert data, and the data set construction is changed from a time-sorted list into a one-hour time series format. In the analysis phase, a data set of two days' total length is used.

This paper is divided into six sections as follows. In the next section, the research work related to reducing the intrusion detection alert data is presented. Section three presents the environment and processes of obtaining the data. The methods utilized are briefly presented in Section four. In Section five, data set pre-processing is presented together with the execution and results of the analysis. The conclusions and future plans are presented in Section six.

II. RELATED WORK

The rationalizing of the amount of intrusion detection alerts has been studied by many research centers and communities. Typically, the main goal in these studies has been the reduction of the amount of alerts, and presenting only the essential information to the network administrator or operator. In many studies, DARPA data sets have been used. They have been found suitable for information security related analysis, and also for evaluation of IDS systems. The data sets consist of both normal and aberrant network traffic, offering a convenient opportunity to compare observations under different circumstances. Another connective element between the various studies is the selected IDS software. The Snort IDS system is the primary element in many studies, when the intrusion detection is based on signatures. The main reason why the Snort IDS system is so popular among researchers is its good level of performance, affordability, and suitability for different environments. The Snort IDS system is freely downloadable from the Internet offering an effective solution for monitoring IP networks.

Alharby et al. have done related work in [3], where the reduction of false positive alerts is studied using continuous and discontinuous patterns. They achieved significant results where the amount of false positive alerts could be decreased by as much as 90%. Perdisci et al. [4], have collected alerts from different IDS systems in their study. They have used the clustering technique to form upper level alert classes. The creation of the alarm classes was not based on predefined definitions, as they were formed dynamically.

The above-mentioned journals are focused only on reducing the number of alerts. This study has also concentrated on the optimization of an active rule set. In some cases when a new rule set is adopted in the IDS system, some signatures might cause false positive alerts, because they are related to normal traffic and normal behavior in the network.

III. Obtaining The Data

The data set was collected in February 2011. The collection of the data set was carried out using the Snort IDS system (v2.8.6.1). Before starting the monitoring sequence, the Snort IDS was installed and configured. Since it offers only basic elements for network monitoring, the latest rule set was downloaded from the Snort support pages and extracted to the appropriate Snort subdirectory. The HOME_NET and EXTERNAL_NET variables were defined in a Snort configuration file. The first variable specified the IP address space of the campus area network and the second variable encompassed all IP address spaces, excluding the home network. In normal situations, the definition of the HOME_NET variable is a straightforward operation, but in this study an anonymization procedure was used to masquerade as the original IP address space. Before monitoring the campus area network with the Snort IDS system, network traffic is anonymized using the Libtrace library provided by the WAND Network Research Group from the University of Waikato (New Zealand) [5]. The anonymization process itself is performed real-time where original network traffic is captured from a physical interface, anonymized, and forwarded to a virtual interface. Instead of listening to the physical interface, the Snort IDS system monitors network traffic from the virtual interface.

The network traffic is monitored at a backbone link between the home network and the Internet. Unfortunately, the network traffic between different network segments could not be monitored, because it would have complicated the monitoring setup too much. Moreover, we did not want to disrupt the function of network switches by overloading them, so the monitoring phase carried out from a single point.

The home network under observation consisted of 4092 hosts. In order to temper the load of the Snort under high speed network traffic, two simultaneous Snort processes were used. Both processes analyzed a network segment of their own. Instead of storing observations in separate log files, both Snort processes stored the alerts in the same database. This simplified the analysis of the alerts, because we were able to use built-in SQL query clauses, instead of implementing a tool for parsing the log files.

IV. Utilized Methods

K-means clustering is utilized to be able to extract normal behavior from the alert data. K-means was selected as a clustering algorithm because of its tendency as a centroid based clustering algorithm to form round clusters, and for its applicability for clustering large data sets. Round clusters are desirable in this case since we operate in two dimensional argument space, where the clustering arguments are nearly on the same scale. The clustering process in general requires decision-making before execution. The main questions to be answered are:

1. What arguments do we choose for clustering?
2. Which distance metric should we use with this type of data?
3. What is the optimal amount of clusters for this data set?

In this paper, arguments were selected that described the amount of alerts. The time series format effects the selection of arguments, so counters have been used in the data pre-processing phase. The first argument indicates the amount of alerts per signature. This was an obvious choice because the aim in this study was to reduce the amount of alerts. The second argument clarifies how many hosts are causing an alert. This simplifies interpretation of the alerts in a situation where a single host produces a large amount of erroneous traffic that is triggered as alerts. The arguments that are used in the clustering are "*Total number of alerts per signature*" and "*Total number of alerting hosts per signature*", respectively. Both arguments are calculated in the data pre-processing phase.

It is an advantage if the clustering can be executed in a two- or three-dimensional space. Obviously, the execution of the clustering is less expensive from the computation point of view, and secondly, the visualization of the results is straightforward and simple. The visualization point is essential in our case, and therefore the target was to execute the clustering with a maximum of three arguments.

The Euclidean distance metric was chosen as it is typically applied in data analysis in general. It is suitable for centroid-based clustering algorithms operating in low-dimensional data spaces such as the data set in this study.

The optimal cluster amount for the data set is determined by calculating the Davies-Bouldin index [6] for cluster amount values of 2 to 10. The maximum index value indicates the best cluster amount. Additionally, visual interpretation was used in estimating the amount of clusters. Visual interpretation was easy to carry out because there were only two arguments used in the clustering analysis. If the number of arguments had exceeded the two-dimensional space, the visual interpretation would have been a much tougher task.

V. Analysis Execution And Results

A. Pre-processing the data set

There were over 117 million alerts during the four-week period, so it is clear that a method for rationalizing the amount of alerts is needed. A two-day period was selected from the four-week time period for the analysis. The total number of alerts was around 7 million. Approximately 4 million alerts originated from the home network, and the remaining 3 million or so alerts were from the external network. Although there was a huge amount of alerts, only 57 different home network bound alarm types were observed. Correspondingly, 79 alarm types caused alerts that originated from the external network.

Instead of using millions of rows in the data analysis, a one-hour time series was formed from the alert data. The time series data consist of eight arguments. Two of the eight arguments were selected as clustering arguments, and the remaining six were left aside at this point. They will be used in the result analysis phase, when the clustering results are back-traced to the original data set.

When the one-hour time series data was formed from the Snort alerts, the number of rows decreased drastically in comparison with the amount of raw data. After formation of the time series data, there were 1123 rows related to the home network and 1783 rows related to the external network, i.e., a total of 2906 rows.

A signature identification number was the key argument in the formation of the time series data. The identification numbers of the Snort alerts are system specific. A certain signature might be associated with a different identification number in different systems. This naturally complicates the comparison of the results between the various Snort systems. Instead of parsing the time series data from text-based log files, the data was formed from the database using SQL query clauses. Obtaining the alert distribution between the home and external networks was a relatively easy task, because the database supported query clauses, where an IP address range can be expressed by dotted-quad representation (IPv4). The data sets were stored in files of their own, where the values of the eight arguments were separated by a comma (CSV). Finally, we were able to proceed to the mathematical analysis phase.

B. Reduction of home network related alerts

To simplify and accelerate mathematical analysis, in this study the MATLAB program (2011b2) was used, which offered a comprehensive selection of tools for our purposes. Before beginning the statistical analysis, the removal of outliers has to be addressed. If the data set contains data points that are considered measurement errors, or otherwise differ significantly from the rest of the data, outlier removal is necessary. One possible means for removing outliers is a procedure where values that are three times larger than standard deviation are removed from the data set. In this study, outlier removal is not required because there is no method for distinguishing either the normal alert distribution or an aberrant one. We simply assume that all the data produced by the IDS system is valid.

The two arguments used as clustering arguments are the key factors in cluster analysis: the Home Count argument indicates the "*Number of alerts per signature*" and correspondingly, the Home Hit Count argument indicates the "*Number of alerting hosts per signature*". To minimize the effect of the different value ranges of the parameters, the values of the arguments were scaled in two phases, first with logarithmic scaling, and secondly with the MATLAB Zscore function. Logarithmic scaling simply takes a logarithm from the parameter values. Zscore subtracts the mean from every parameter value and divides the values by a standard deviation.

The alerts reduction of the IDS system is carried out in three phases, using K-means as a clustering algorithm, and the Davies-Bouldin index for determining the optimal cluster amount. Based on our earlier experience with the K-Means algorithm, we decided to use the algorithm in this study, too [7]. The decision was also favored by the performance of the algorithm. In the first step, the amount of clusters was estimated using the Davies-Bouldin index. This information was used in the clustering analysis, which was accomplished by using the K-means algorithm. This is a very straightforward process. In the second step, the outcome of the clustering analysis was interpreted. If an alert belongs to the same cluster over the time period, and it appears in every hour over the time period, it can be interpreted as normal behavior. Normal behavior is not interesting and it could be ignored from further analysis, or corresponding alerts could be removed from the Snort rule set. This decision requires human interpretation. In the third step, the alerts related to the normal behavior in the network were removed from the data set. This sequence was repeated until there were no alerts that matched our definitions, i.e., existed constantly and in all periods of the time series. In the final iteration round, alerts were scattered in different clusters and there were no alerts that appeared in every hour over the time period. Steps one to three were carried out separately for alerts, which originated from the home or external networks. This division simplifies the interpretation of the results, because the administrator or network operator clearly sees the initiator of a given alert, and can react to it in the appropriate way. In the following chapters, the mechanism for reducing and rationalizing alerts and signatures is described.

C. Results of the reduction

The results from the first iteration round are presented in a two-dimensional scatter-plot in Figure 1 and in the bar chart in Figure 2.

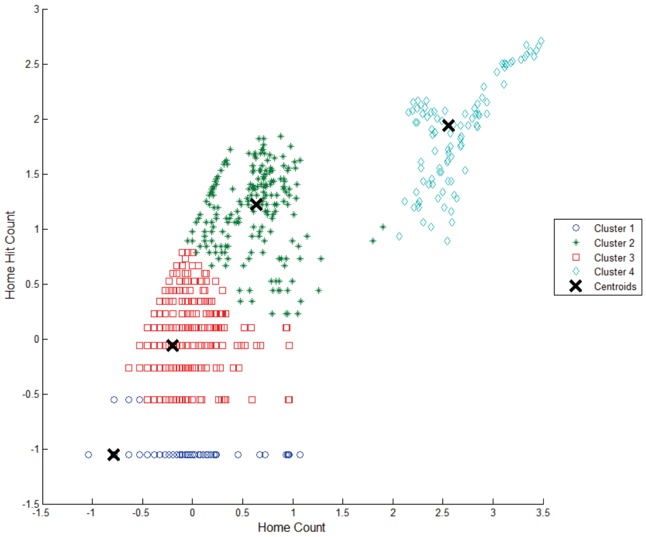

Fig. 1. Snort alerts clustering, home network initiated alerts, round 1

On the X-axis, the values of the argument Home Count are presented. The values of the argument Home Hit Count are presented on the Y-axis. In both cases the values of the arguments have been scaled. In the figure, four different clusters can be seen. The centroids of each cluster are marked by a bolded '*X*'. Clusters 1, 2, and 3 formed a super group on the left side, whereas cluster 4 is clearly separated from the other clusters. At this point we can present the hypothesis that cluster 4 consists of signatures that cause many of the uninteresting alerts, i.e., are related to the normal behavior in the network. In the previous paragraph, there was discussion about the handling of outliers. If the outliers had been removed from the data set, significant information would have been lost about the behavior of the signatures. Most likely the whole of cluster 4 would have been interpreted as outliers. The signature distribution between the clusters is illustrated in Figure 2.

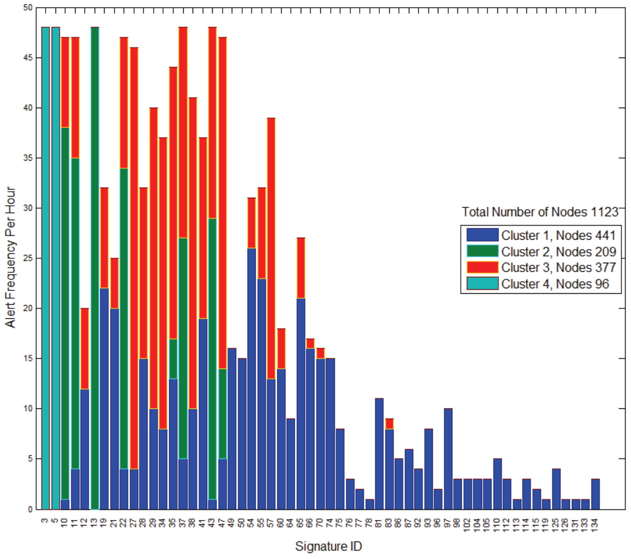

Fig. 2. Alert distribution into clusters, home network initiated alerts, round 1

Signature ID numbers are presented on the X-axis and one-hour time periods are presented on the Y-axis. The maximum value of the Y-axis is 48 hours, which refers to the two-day sampling period. When examining the graph more closely, it can be observed that only a few signatures appeared in every hour and existed in the same cluster over the time period. Signatures 3 and 5 in cluster 4, and signature 13 in cluster 2, fulfilled our requirements of existing in the same cluster in each one-hour period and over the whole 48 hours. Other signatures were distributed into the different clusters or they would not have appeared in every hour over the time period. It might be thought that those three signatures caused many uninteresting alerts. According to our assumption, these signatures relate to normal behavior in the network, and the alerts they caused can be ignored or removed from the rule set. To confirm this conclusion, we studied the nature of the alarms in question, and we found that the alarms in question did not violate the information security policy, so they can be omitted from further analysis. For reasons of privacy protection, only signature ID numbers are used in this paper, rather than using verbal signature identifiers. However, we will open the curtain a little to justify why some signatures can be ignored. Cluster 4 consisted of two signatures that caused many uninteresting alerts. The signatures were triggered from applications that use the IPv6 or BitTorrent protocol, so the hypothesis proved right. Correspondingly, cluster 2 consisted of eight signatures but only one signature can be ignored. The signature triggered form applications that use BitTorrent.

In two days the three signatures caused a total of 3877369 alerts, covering 99 percent of the total amount of alerts, which is a very high proportion. At this point it is clear that by using the method presented in this paper, the amount of alerts can be rationalized drastically.

In the second iteration round, signatures 3, 5, and 13 were removed from the data set in the preprocessing phase and after this procedure, the cluster estimation and analysis were performed again, as in the first iteration round. The results from the second clustering round are depicted in Figure 3.

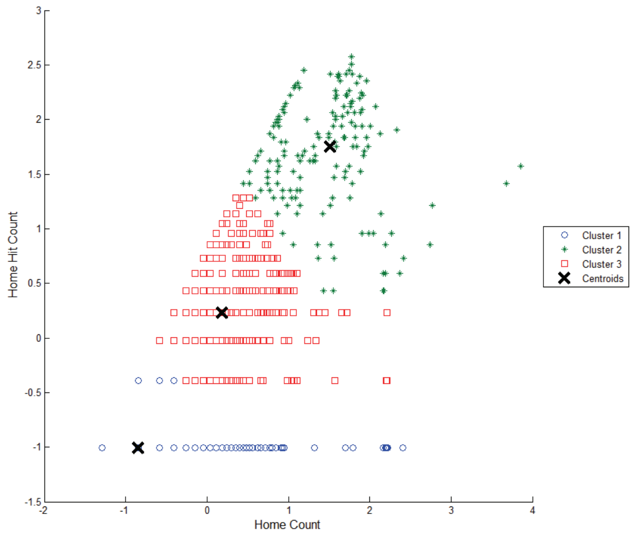

Fig. 3. Snort alerts clustering, home network initiated alerts, round 2

The number of clusters has been decreased from four to three, and there are no more clusters clearly separated from other clusters. Clusters 2 and 3 are strictly connected together, but most of the nodes of cluster 1 are separated from the other clusters. The signature distribution into different clusters over the time series is illustrated in Figure 4.

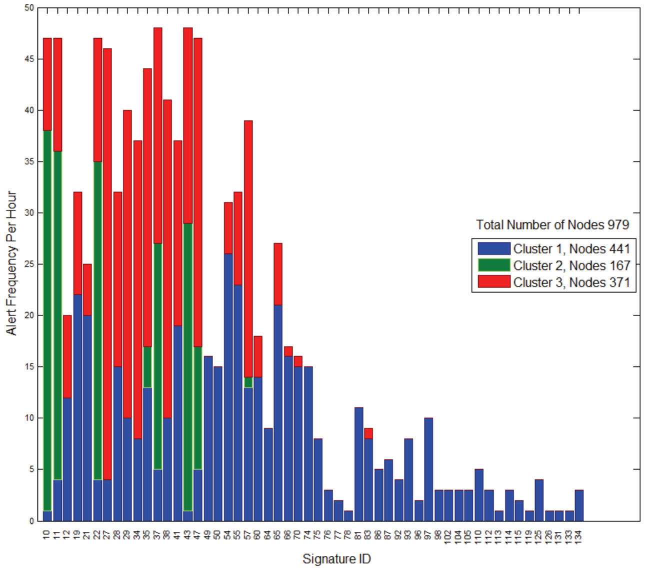

Fig. 4. Alert distribution into clusters, home network initiated alerts, round 2

If we examine the distribution of the signatures into different clusters, we observe that there are no more signatures appearing in every hour of the time series and belonging to the same cluster over the two-day time period. In this case, it can be concluded that signatures cannot be removed after this round, so there is no need to perform the third iteration round of clustering. Signatures 37 and 43 appear in every hour over the two days, but their behavior is inconsistent, which can be concluded from their distribution into different clusters. In contrast, signature 27 belongs mostly to cluster 3 but does not appear in every hour during the two days.

At this point the reader might wonder why clustering the time series data is used to rationalize the amount of alerts. Why not take the five biggest alerting signatures and simply ignore them? We will defend our approach using the following example. In some cases there can be a sudden event that causes an enormous amount of erroneous traffic, which the IDS system interprets as alerts. If in this case the five biggest alerting signatures are simply ignored, we could draw the wrong conclusions about the network behavior. The time series format and the clustering methods ensure that this misinterpretation cannot be made.

D. Reduction of external network related alerts

During the two-day period in question, IP traffic from the external to home network caused about 3 million alerts. In the time series format, the data set consisted of 1783 rows. As in the case of the home network, the data set consisted of eight arguments, but only two variables were used in mathematical analysis. The values of the arguments were scaled using the same MATLAB functions as in the case of the home network. The cluster estimation and analysis were carried out using the Davies-Bouldin index and the K-Means algorithm, respectively.

There were three different iteration rounds related to the external network. After those iteration rounds, a total of 13 alerting signatures could be ignored from further analysis and furthermore, removed from the Snort rule set. The signatures that were removed from future analysis were related to the ICMP, IPv6, and BitTorrent protocols. Generally it can be said that the results from the first iteration round were the most significant. In the first iteration round there were 79 alerting signatures under analysis and 69 in the second round analysis. Finally, in the third round, three more alerting signatures could be labeled as uninteresting.

During the two days, the 13 signatures formed a total of 2986773 alerts, covering 99 percent of the total amount of alerts, which is a very large proportion. When we analyzed the alerting signatures related to the external network, we noticed that 10 of the signatures were related to ICMP traffic. The share was so high that we suspected it to be port scanning related. Table I illustrates how the amount of significant alerts was reduced after each iteration round in the case of alerts originating in the home and external networks.

TABLE I. SUMMARY OF DIFFERENT ITERATION ROUNDS

	Home Network			**External Network**		
Round	***Signatures***	***Alerts***	***%***	***Signatures***	***Alerts***	***%***
1	57	3907585	100	79	3011376	100
2	54	3877369	99	69	2692465	89
3	-	-	-	66	294308	10
Result	**54**	**30216**	**1**	**66**	**24603**	**1**

VI. Conclusion And Future Work

The goal of this study was to obtain a method for rationalizing the amount of intrusion detection alerts by identifying the alerts related to the normal traffic in the target network. Removing the uninteresting alerts related to normal traffic resulted in a set of alerts that potentially held all the interesting alerts related to aberrant traffic. Our experience from the Snort IDS reveals that the number of alerts can easily exceed the limit of one million per day in a large-scale IP network. A manual examination of millions of alarms would drive a system administrator or network operator to desperation! We defined two arguments for mathematical analysis by post-processing the raw alert data. In total there were eight arguments. The other six variables offered valuable information during the analysis phase. Instead of using millions of rows in the mathematical analysis, we formed time series data from the original data set. To simplify the analysis phase, the alerts were divided into home and external data sets, based on the source IP address of the alert.

When the two data sets had been obtained, the reduction of the alerts was started. The amount of clusters was estimated with the aid of the Davies-Bouldin index, and the cluster analysis was carried out with the K-means algorithm. We analyzed signature behavior over the whole of the 48-hour time period. If a signature belonged to the same cluster over the whole time period and it appeared in every hour over the time period, it was associated with the normal traffic in the network, and thus ignored from further analysis and removed from the Snort rule set. When the inappropriate signatures had been removed from the data set, cluster estimation and analysis were processed again, until there were no signatures suitable for our definitions.

The method fulfilled our requirements surprisingly well, as seen in Table I. By using well-known mathematical functions and appropriate arguments, the number of alerts can be significantly reduced. This simplifies the further analysis of the alerts and optimization of the Snort rule set. When the optimized rule set is activated, a network operator can react faster to critical threats. On the other hand, the number of alerts is still fairly high after the iteration rounds, so alternative methods have to be investigated in order to further reduce the amount of alerts. In addition, we are planning to extend our study in the future to compare our method with other existing methods. Unfortunately, this could not be carried out in time for this paper because the time window of the study was coming to an end.

Another question is how well does the signature-based IDS system apply to the monitoring of large-scale IP networks, where the number of alerts might exceed the limit of one million per day. In addition to this, many malware applications use strong encryption algorithms to protect the control traffic, which, in turn, is difficult to detect from the normal network traffic. Consequently, there are still many open questions related to intrusion detection left to be addressed.

References

[1] J. Macqueen, “Some methods for classification and analysis of multivariate observations,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, 1967, pp. 281-297.

[2] M. Roesch, “Snort - Lightweight intrusion detection for networks,” Proceedings of the 13th Conference on Systems Administration (LISA-99), USENIX Association, Nov. 1999, pp. 229-238.

[3] A. Alharby and H. Imai, “IDS False Alarm Reduction Using Continuous and Discontinuous Patterns,” Applied Cryptography and Network Security, vol. 3531, 2005, pp. 423-442, doi:10.1007/11496137_14.

[4] R. Perdisci, G. Giacinto, and F. Roli, “Alarm clustering for intrusion detection systems in computer networks,” Engineering Applications of Artificial Intelligence, vol. 19 (4), Jun. 2006, pp. 429-438, doi:10.1016/j.engappai.2006.01.003.

[5] Libtrace, version 3.0.8, http://research.wand.net.nz/software/libtrace.php, retrieved Dec. 2011.

[6] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1 (2), Apr. 1979, pp. 224-227, doi:10.1109/TPAMI.1979.4766909.

[7] O. Knuuti, T. Seppälä, T. Alapaholuoma, J. Ylinen, P. Loula, P. Kumpulainen, and K. Hätönen, “Constructing communication profiles by clustering selected network traffic attributes,” Proceedings of the Fifth International Conference on Internet Monitoring and Protection ICIMP 2010, May 2010, pp. 105-109, doi:10.1109/ICIMP.2010.21.