International Journal of Cyber-Security and Digital Forensics (IJCSDF) 1(2): 130-139
The Society of Digital Information and Wireless Communications, 2012 (ISSN: 2305-0012)

An AIS Inspired Alert Reduction Model

Mohammad Mahboubian, Nur Izura Udzir, Shamala Subramaniam, Nor Asila Wati Abdul Hamid
mahboubian.uni@gmail.com, {izura, shamala, asila}@fsktm.upm.edu.my
Faculty of Computer Science and Information Technology
University Putra Malaysia, Serdang, Selangor, Malaysia

***Abstract*- One of the most important topics in the field of intrusion detection systems is to find a solution to reduce the overwhelming alerts generated by IDSs in the network. Inspired by danger theory which is one of the most important theories in artificial immune system (AIS) we proposed a complementary subsystem for IDS which can be integrated into any existing IDS models to aggregate the alerts in order to reduce them, and subsequently reduce false alarms among the alerts. After evaluation using different datasets and attack scenarios and also different set of rules, in best case our model managed to aggregate the alerts by the average rate of 97.5 percent.**

Keywords—Intrusion detection system; Alert fusion; Alert correlation, Artificial Immune system; Danger theory;

1.0 Introduction

In recent years intrusion detection systems (IDS) have been widely adopted in computer networks as a must-have appliances to monitor the network and look for malicious activities. It is possible to use and implement them either in the network level to monitor the activities in the network or to use them in host level to monitor activities on a particular machine in the system. In both cases after detecting a malicious activity they will send an alert to the network administrator.

Each alert contains information about this malicious activity such as source IP address, source port number, destination IP address, etc. Thus, for a single attack on a network or any of its hosts, there will be thousands of alerts generated and sent to the network administrator. Also some of these alerts may not be valid and are generated because of the wrong detection of an IDS (false positive) in the network. This is crucial as every day a significant number of alerts is generated and processing these alerts for network administrators can be a tedious task, especially if all of these alerts are not valid and can be result of false positive detection. Therefore, in the last few years one of the most focused topics in the field of network security and more specifically intrusion detection systems was to find solutions for this problem.

To reduce the overwhelming amount of generated alerts some researchers have suggested to aggregate alerts into clusters, which is also called alert fusion. The final objective of aggregation is to group all similar alerts together. During aggregation, alerts are put into groups based on the similarity of their corresponding features [25] such as Source IP, Destination IP, Source Port, Destination Port, Attack Class and Timestamp. On the other hand some of the researchers investigated different approaches to correlate the attack scenarios based on the alerts. Alert correlation provides the network administrator with a higher view of a multi staged attack.

Three main approaches have been used in the literature for correlating alerts in attack scenarios.
In the first approach the relationship between alerts are hardcoded in the system. These methods are limited to the

predefined rules available in the knowledge base of the system. In the second approach and to overcome the problem in the first approach, other approaches have been suggested such as machine learning and data mining techniques to extract relationships between alerts, but these approaches require a lengthy initial period of training. In these approaches, co-occurrence of alerts within a predefined time window is used as an important feature for the statistical analysis of alerts. This involves pair-wise comparison between alerts since every two alerts might be similar and therefore can be correlated [25]. But these repeated comparisons between alerts leads to a very huge computational overload, especially when they are going to be used in large-scale networks, in which we may expect thousands of alerts per minute.

Finally, in the third approach, some of the recent works focused on filtering and omitting false positive alerts.

In this paper we proposed a new aggregating method inspired by artificial immune system and more specifically danger theory which attempt to aggregate the generated alerts based on the prediction of attack scenarios. The proposed algorithm is able to reduce alerts before passing them to the network administrator and also to remove false positives from the generated alerts.

The remainder of this paper is organized as follows: in Section 2 we present a brief review of previous works in the literature. In Section 3 we describe the proposed model and discuss some of the aspects related to alert aggregation. Section 4 presents experimental results and finally we conclude this paper in Section 5.

Artificial Immune System:

Artificial Immune system is a mathematical model based on the human body defence system. Natural immune system is a remarkable and complex defence mechanism, and it protects the organism from foreign invaders, such as viruses. Therefore, it is vital for the defence system to distinguish between self-cells and other cells, as well as ensuring that lymph cells does not show any reaction against human body cells. To achieve this, the human body will go through a "Negative Selection" process [16] in which T-cells that react against self-proteins are destroyed therefore only those cells that do not have any similarity to self-proteins survive. These survived cells which are now called matured T-cells are ready to protect the body against foreign antigens.

Danger Theory:

This theory was first proposed in 1994 [17] by Matzinger. According to this theory not all foreign cells in our body should be considered an antigen. For instance the food which we eat is also a foreign ‘invader’ to our body but the human body does not react to this foreign invader.

Danger theory suggests that foreign invaders, which are dangerous, will induce the generation of danger signals by initiating cellular stress or cell death [19].

Then these molecules are detected by APCs, critical cells in the initiation of an immune response, this leading to protective immune defence system. In general there are two types of danger signals; in the first category the danger signals are generated by the body itself, and in the second category, the danger signals are derived from invading organisms, e.g. bacteria [20].

2.0 Related Works

Recently, there have been several proposals on alert fusion. Generally, each method is to combine duplicated alerts (alert which are very similar to each other) from the same or different sensors to reduce a large part of alerts. Here we overview some of the works which have been done in the last few years.

To measure similarities between alerts, the pioneers in the field of alert aggregation, Valdes and Skinner [1] proposed a method in which alerts are grouped into different clusters based on their overall similarity, determined based on their similarities on the corresponding features. Unfortunately, this method relies on expert knowledge to determine the similarity degree between attack classes.

In [2], the authors presented an algorithm to fuse multiple heterogeneous alerts to create scenarios, building scenarios by adding the alert to the most likely scenario. To do so it computes the probability that a new alert belongs to one of the existing scenarios.

Ning *et al*. [3] constructed a series of prerequisites and consequences of the intrusions. Then by developing a formal model they managed to correlate related alerts by matching the outcome of some previously seen alerts and the precondition of some later alerts. Julisch [4] used root causes to solve the problem of the alert attribute similarity. Although this approach was effective but finding root causes of the alert attributes is very difficult and in large networks seems to be impractical. Chung *et al*. [5] uses Correlated Attack Modelling Language (CAML) for modelling multistep attack scenarios and then to recognize attack scenarios he allowed the correlation engines to process these models. However, it is not easy for this algorithm to model new variant attacks. Valeur *et al*. [6] introduced a 10-step Comprehensive IDS Alert-Correlation (CIAC) system that uses exact feature similarity in two out of ten steps in their alert correlation system. Qin and Lee [7] proposed a statistical-based correlation algorithm to predict novel attack strategies. This approach combines the correlation based on Bayesian inference with a broad range of indicators of attack impacts and the correlation based on the Granger Causality Test. However, this algorithm cannot be used to predict complex multi staged attacks because of high false positive results.

In another work Qin and Lee [8] proposed an approach which applies Bayesian networks to IDS alerts in order to conduct probabilistic inference of attack sequences and to predict possible potential upcoming attacks. In [9] authors introduced a bi-directional and multi-host causality to correlate distinct network and host IDS alerts. But if the number of false positive alerts increases mistakes in recognition may occur. Zhu and Ghorbani [10] use the probabilistic output from two different neural network approaches, namely Multilayer Perception (MLP) and Support Vector Machine (SVM), to determine the correlation between the current alert and previous alerts. They used Alert Correlation Matrix (ACM) to store correlation level of any given two types of alerts. Wang *et al*. [11] proposed a new data mining algorithm to construct attack scenarios. This algorithm allows multi-stage attack behaviours to be recognized, and it also predicts the potential attack

steps of the attacker. However, to calculate the threshold used in this approach sufficient training is required. To detect DDoS attacks Lee [12] proposed clustering analysis using the concept of entropy .He then calculated the similarity value of attack attributes between two alerts using Euclidian distance. Fava *et al*. [13] proposed a new approach based on Variable Length Markov Models (VLMM), which is a framework for the characterization and prediction of cyber attack behaviour. VLMM can predict the occurrence of a new attack; however it does not know what kind of attack it is. Zhang *et al*. [14] uses the Forward and Viterbi algorithm based on HMM to recognize the attacker's attack intention and forecasts the next possible attack for the multi-step attack. By the design of Finite State Machine (FSM) for forecasting attacks, the Forward algorithm is used to determine the most possible attacking scenario, and the Viterbi algorithm is used to identify the attacker intension. Du *et al*. [15] proposed two ensemble approaches to project the likely future targets of ongoing multi-stage attacks instead of future attack stages.

3.0 Proposed Model

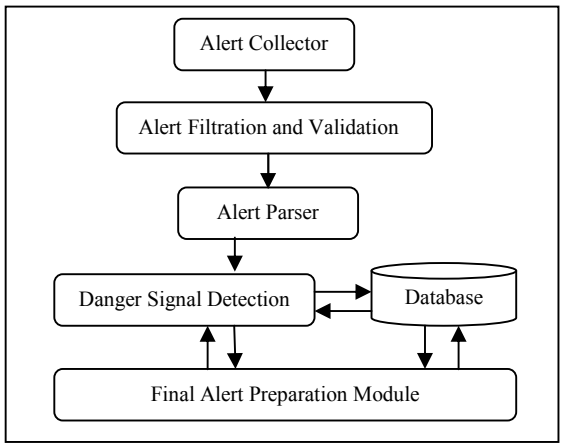

Figure 1 – The proposed model

We assumed that all types of computer attacks can be categorized into the following general groups:

a) One-to-One: in which the hacker attacks one of the machines on the network. This can be a Probe or a Dos attack or exploitation of services in that host.
b) Many-to-One: in which many machines (zombies) attack one of the machines on the network. Most probably this is a form of DDos attack.
c) One-to-Many: in which the hacker attacks many machines on the network such as probe attack.

According to danger theory only an alert or group of alerts can be considered valid (dangerous) if they initiate the danger signal.

To raise the danger signal some conditions must be satisfied and this conditions are defined prior to implementation of this system. Therefore we have a list of condition in which if any of these conditions is satisfied by a group of alarms, that group of alarm is considered dangerous and will be reported to the network administrator immediately.

Not only the proposed model tries to aggregate the alerts based on their common features but it correlates the attacks internally for better aggregating the alerts.

Figure 1 shows our proposed model. This model consists of six components and the design of this model is in such a way that any of these components can be replaced with a new implementation of that component depending on different network situation.

The following is an illustration of the main components of this model:

a) *Alert Collector module (CM)*: This module is responsible to collect all the alerts from all the IDS sensors in the network. Therefore after generation of an alert by an IDS sensor, instead of sending the alert directly to the administrator, it should be sent to this module. Once this module receives an alert, that alert will be registered into model to be processed. Another objective of this module is to standardize the alerts because IDS sensors might generate the alerts in different format. Therefore in order to process and compare the received alerts they should be in a same format. Another point about this module is that as this module receives enormous volume of alerts it must be implemented using a very robust multi-threaded software technology.

b) *Alert Filtration and Validation (FVM)*: One of the prerequisites of using this model is to keep the list of IP address and services running on all of the machines in the network under our administrative territory. By utilizing this information, this module filters out those alerts which do not make sense such as an alert of attack on a web server on a machine without a web server. Also this module aggregates those alerts which are exactly similar feature wise, helping to reduce redundant alerts.

c) *Alert Parser module (PM)*: The main objective of this module is to categorize and classify all validated alerts into one of the groups we mentioned earlier: one-to-one, many-to-one and one-to-many.

d) Danger Signal Detection Module (DSDM): This is the most important module in this model. This module is the implementation of the one of the most famous theories in the field of artificial immune system namely known as Danger Theory. Its main function is to analyze all received alerts in a specific time window in an attempt to correlate a multi-steps attack and aggregating all related alerts into a group of alerts which later will be represented to the administrator as a single alert. In order to achieve to this objective, a series of generalized rules are hardcoded into this module. Based on these rules and the actual characteristic of the available alerts this module dynamically decides if a group of alerts are related to an multi-steps attack and can be aggregated to a single alert.

e) Final Alert Preparation module (FAPM): The results of previous module are sent to this last module in order to make them presentable before passing to the administrator.

3.1 Model Implementation

The proposed model has a module namely Danger Signal Detection Module (DSDM) which decides if a group of alerts are likely to raise the danger signal or not, and will report a dangerous group of alerts to the network administrator.

The steps to implement this model are:

a) First we provide this model with information about the machines on the network, such as their IP address, list of services running on each machine and, in case of host IDS the id of the IDS, on that machine. This step should be repeated periodically in order to prevent concept drift.
b) Next, all alerts are grouped into one of the groups explained earlier. The priority is with the alerts which are exactly similar in terms of their features and after grouping these alerts the priority is with the second type of alerts namely "one-to-many". This is because before attacking a network the attacker needs to know about the machines on the network, so he/she initiates a probe to the network, which results in generating these types of alerts. After this group the least priorities belong to "one- to-one" and "many-to-one" types of alerts. The grouping is done within an adjustable time window value and based on the source IP address, destination IP address, destination Port number, timestamp and in case of host-based IDS, the id of the IDS.
c) Then each group is checked to find out if that group is capable of raising the danger alarm (Danger Theory).
d) For each group which satisfies the checking a record is registered in a database for the purpose of keeping track of the status of the attack as this is one of the sources which can indicate the existence of danger signal. Finally an alert will be sent to the network administrator containing the information about the attack, as well as all the machines IP addresses (source and destination) or port numbers which contribute to this alarm.
e) Alerts generated from network-based IDSs and host-based IDs are grouped separately but host-based IDSs' alerts are important in determining the severity of network-based IDS alerts.

3.2 Danger Signal Detection Module

This module indicates either a group of alerts are capable of raising the danger alarm or not and this is done by defining a list of rules. The following are some of the most important rules in this model:

a) In general an existence of one-to-many alert group (generated by network-based IDS) in database followed by one-to-one alert group type (generated by host-based IDS) will raise the danger alarm. This is because a hacker first scans the machines on a network and after he/she found a machine with a particular service running on it, he/she tries to exploit that service to gain access to that machine.
b) If in the alert group the source IPs are external and port number(s) are not matched with actual services running on the internal machines, this is an indication of danger signal and will be reported.
c) If in the alert group the source IPs are internal and port number(s) are matched with the actual services running on the destination machine(s), and the number of alerts in this group are not more than a predefined value, then this group is ignored.
d) If in the alert group there are more than one source IPs and a single destination IP, this will raise the danger alarm.
e) If in the alert group there are one single source IP and more than one IPs in destination IP this will raise the danger alarm.

f) If in the alert group there are more one source IP and one destination IP and we have recently a record in the database related to this source and IP address (probe), then this will raise the danger signal.

The similarity function *S* between two given alerts and b is calculated as follow:

$$S_{(a,b)}=\sum_{k=1}^{n} \propto_k . \partial(a_k, b_k) \quad (1)$$

Whereby n is total number of features and $\partial(a_k, b_k)$ is the similarity of feature k between these alerts which can be between 0 and 1, $\propto_k$ is the weight of that particular feature such that

$$\sum_{k=1}^{n} \propto_k = 1 \quad (2)$$

Having different weight for each feature leads to more precise grouping of the alerts. Among our features set source IP address and timestamp have the highest weights.

Therefore to calculate the similarity of two signals we need to calculate the following:

$\partial(a_{Src\,IP}, b_{Src\,IP})$, $\partial(a_{Dst\,IP}, b_{Dst\,IP})$,
$\partial(a_{Dst\,port}, b_{Dst\,port})$, $\partial(a_{time}, b_{time})$,

and in term of host based IDS:
$\partial(a_{sensor\,ID}, b_{sensor\,ID})$

After normalizing the formula in (1) the similarity value between two alerts can be between 0 and 1: 0 when two alerts are completely different and 1 when two alerts are identical.

4.0 Experimental Results

In order to evaluate our model first we setup a network of seven computers in which two computers play the role of attackers and with a different class of IP addresses so that they are considered as external machines (Figure 2). Next, for each machines inside the network we configured different services such as file server, web service, and remote desktop service and so on. As for the IDS we used our own proposed IDS in [21].

Table 1- Shows the services running on each workstation

Workstation	Service(s)
10.8.0.2	ftp (port 21)
10.8.0.3	Web server (port 80)
10.8.0.4	smtp (25) and imap (143)
10.8.0.5	RDP (port 3389)
10.8.0.6	SSH (port 22)

Next we simulated different kinds of attacks in order to generate alerts and starting with probe (including vertical and horizontal port scans) and Dos attacks and finally exploiting different services on the workstations to gain access to the machine and elevating the access level. Table 1 shows the services running on each of the workstations.

The first attacker (10.8.1.100) starts with scanning the whole range of network and finding the running services on each of the discovered workstations. Then he tries to exploit different services on different workstation one by one. At the same time the second attacker (10.8.1.200) scans the whole network and initiates a Dos attack against one of the discovered machines. These activities caused the IDS to generate more than 3000 alerts. These alerts was processed by this model and the final number of alerts was 31 therefore the proposed model showed a very good performance of 98.95% alerts reduction..

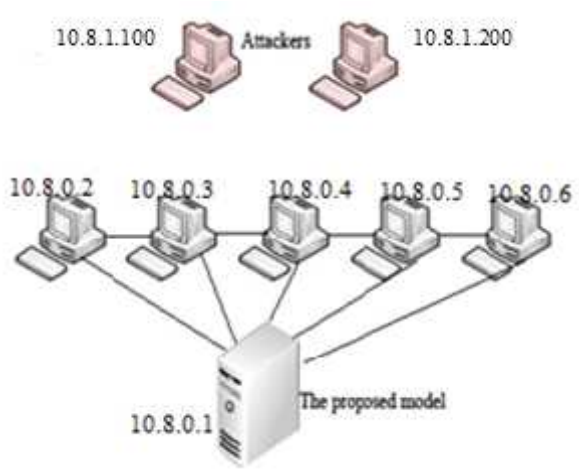

Figure 2 – The network setup for the first experiment

To better evaluate our proposed model we considered LLDOS1.0 and LLDOS2.0 attack scenarios of DARPA 2000 [16] as test datasets. These datasets contain a large number of normal data and attack data, well-known among IDS researchers [22, 23]. For this experiment we used a partial of these datasets. In order to simulate the networks we used NetPoke from DARPA to replay datasets and once again we used our own developed IDS for attack detection and also for generating alerts. Total number of 12068 alerts was generated by our IDS. Then we updated the model with the services running on each of the machines in these networks. Finally we run these experiments multiple times and each time with a different set of rules in "Danger Signal Detection Module". In all cases we make sure that these rules are enough generalized so that they can be utilized in other networks also therefore they are not crafted only for these experiments.

The following tables show the reduction percentage of each level of our model for the worse and best cases that we achieved. These results show that it is possible for this model to achieve the alerts reduction rate of 98.5% for LLDOS1.0, and 97.02% for LLDOS2.0 if we use the correct rules set in this model. Some of the modules are not meant for alert reduction and they mostly handle other issues such as parsing the incoming alerts or rearranging of alerts to make them more presentable for end user which in this case it is network admin.

Table 2- LLDOS1.0 worse case result

	FVM	PM	DSDM	FAPM	SUM
Input	7054	4901	4893	1951	7054
Output	4901	4893	1951	1945	1945
%			60.13		72.4

Table 3- LLDOS1.0 best case result after updating the rules

	FVM	PM	DSDM	FAPM	SUM
Input	7054	4901	4893	112	7054
Output	4901	4893	112	106	106
%			97.71		98.5

Table 4- LLDOS2.0 worse case result

	FVM	PM	DSDM	FAPM	SUM
Input	5014	3818	3812	1909	5014
Output	3818	3812	1909	1915	1915
%			49.92		61.8

Table 5- LLDOS2.0 best case result after updating the rules

	FVM	PM	DSDM	FAPM	SUM
In	5014	3818	3812	153	5014
Out	3818	3812	153	149	149
%			95.98		97.02

As one of our immediate future work we intend to experiment this model with "Capture the Flag 2010 dataset" [24].

5.0 Conclusion

In this paper we proposed a model to fuse the generated alerts by the IDSs in a computer network. Inspired by the human defence system, this model utilizes one of the most important theories in Artificial Immune System (AIS), danger theory, and attempts to aggregate alerts based on a general set of predefined rules, and also reduce the false alarms. In contrast with existing rule based alert correlation models which are limited to their set of predefined rules, this model does not have any limitation in terms of alert aggregation and this is because the predefined rules in this model are very general. After experimenting this model in a real network environment and also using existing datasets in literature, the proposed model managed to aggregate alerts with an average rate of 97.5 percent.

References

[1] A. Valdes and K Skinner, "Probabilistic Alert Correlation", In Proceedings of the 4th International Symposium on Recent Advances in Intrusion Detection, 2001, pp.54-68.

[2] O. M. Dain and R. K. Cunningham. Fusing a heterogeneous alert stream into scenarios. In Proceedings of the 2001 ACM Workshop on Data Mining for Security Applications, pages 1–13, 2001.

[3] P. Ning, Y. Cui, and D. S. Reeves. Constructing attack scenarios through correlation of intrusion alerts. In Proceedings of the 9th ACM Conference on Computer and Communications Security, pages 245–254, 2002.

[4] K.Julisch, "Using root cause analysis to handle intrusion detection alarms", PhD Thesis, University of Dortmund, Germany, 2003.

[5] S. Cheung, U. Lindqvist and M. W. Fong, Modelling multistep cyber attacks for scenario recognition, In Proceeding of Third DARPA Information Survivability Conference and Exposition (DISCEX III), Washington,D.C., April 2003.

[6] F. Valeur, G. Vigna, C. Kruegel and R. A. Kemmerer, A comprehensive approach to intrusion detection alert correlation, In Proceeding of IEEE Trans. Dependable Secure Computing., vol. 1, no. 3, pp. 146169, Jul.Sep. 2004.

[7] X. Qin and W. Lee, Discovering novel attack strategies from INFOSEC alerts, In Proceeding of 9th European Symposium on Research in Computer Security (ESORICS 2004), pp. 439-456, 2004.

[8] X. Qin and W. Lee, Attack plan recognition and prediction using causal networks, In Proceeding of 20th Annual Computer Security Applications Conference 2004.

[9] S. King, M. Mao, D. Lucchetti, and P. Chen, Enriching intrusionalerts through multi-host causality, In proceeding of the Network and Distributed Systems Security Symposium., San Diego, CA, 2005.

[10] B. Zhu and A. A. Ghorbani, Alert correlation for extracting attack strategies, International Journal of Network Security, Vol.3, No.3, pp.244-258, November 2006.

[11] L. Wang, Z. T. Li and Q. H. Wang, A novel technique of recognizing multi-stage attack behaviour, In Proceeding of IEEE International Workshop on Networking, Architecture and Storages, pp. 188, 2006.

[12] Keunsoo Lee, Juhyun Kim, Ki Hoon Kwon, Younggoo Han and Sehun Kim, "DDoS attack detection method using cluster analysis", *Expert Systems with Applications*, vol.34,no.3, 2007, pp.1659-1665 .

[13] D. Fava, S. R. Byers, S. J. Yang, Projecting Cyber Attacks through Variable Length Markov Models, IEEE Transactions on Information Forensics and Security, Vol.3, Issue 3, September 2008.

[14] S. H. Zhang, Y. D. Wang and J. H. Han, Approach to forecasting multistep attack based on HMM, Computer Engineering, Vol.34, No.6, pp. 131-133, Mar 2008.

[15] H. Du, D. Liu, J. Holsopple, and S. J. Yang, Toward Ensemble Characterization and Projection of Multistage Cyber attacks, In Proceeding of IEEE ICCCN10, Zurich, Switzerland, August 2-5, 2010.

[16] Duan Shan-Rong, LiXin The anomaly intrusion detection based on immune negative selection algorithm Granular Computing, 2009, GRC '09.In Proceeding of IEEE International Conference, 978-1-4244-4830-2, 2009.

[17] Matzinger P, Tolerance Danger and the Extended Family, Annual reviews of Immunology 12, 1994.

[18] U.Aickelin, P.Bentley, S.Cayzer, J.Kim,J. McLeod "Danger Theory: The Link between AIS and IDS second International Conference on Artificial Immune Systems, Edinburgh, U.K. September, 2003.

[19] Matzinger P, The Danger Model: A Renewed Sense of Self, Science 296: 2002.

[20] Gallucci S, Matzinger P, Danger signals: SOS to the immune system, Current Opinions in Immunology 13, pp 114-119. 2001

[21] M. Mahboubian, N. A. W. A Hamid "A Machine Learning based AIS IDS" In Proceeding of GCSE 2011 Dubai.

[22] G.Xiang, X.Dong, G.Yu "Gorrelating Alerts with a data mining based approach" In Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service.

[23] B.Cheng, G.Liao, C.Huang "A novel probabilistic matching algorithm for multi stage attack forecasts" IEEE Journal on Selected Areas in Communications, Vol. 29, No. 7, August 2011.

[24] "Capture the flag traffic dump, http://www.defcon.org/html/links/dc-ctf.html."

[25] Reza Sadoddin, Ali A. Ghorbani, An incremental frequent structure mining framework for real-time alert correlation, Computers & Security, Volume 28, Issues 3–4, May–June 2009, pp. 153-173, ISSN 0167-4048, 10.1016/j.cose.2008.11.010.