行政院國家科學委員會專題研究計畫期末報告

設計與實作具資訊融合能力之入侵偵測系統

計畫編號：NSC 92-2213-E-009-093
執行期間：92 年 08 月 01 日至　93 年 07 月 31 日
主持人：謝續平
執行單位：國立交通大學資訊工程學系

中文摘要

隨著網際網路的盛行，網路安全的議題也更為受到重視。入侵偵測系統可以偵測對系統的不當攻擊，是網路安全中不可或缺的一環。然而隨著網路環境以及攻擊方法日益複雜，入侵偵測系統需要更多方面的稽核資訊，以作更完整而詳實的分析。在本計畫中，我們提出一個可融合多方面的資訊的侵偵測系統，可整合以時序性分析為考量的入侵偵測模組，以及非時序性分析的入侵偵測模組。其中資訊融合的模組中，我們採用以模糊理論為基礎的專家系統，以期能分析更完整的稽核資料，達到提高偵測率，降低誤測率的效果提升。

Abstract

Intrusion detection systems intend to detect malicious attacks against computer systems. With network environment getting more and more complicated nowadays, intrusion tends to use combination of several types of attacks to increase its attacking power. Hence, an intrusion detection system that applies more than one detection model is required. In this project, a new intrusion detection model that fuses the result of both sequential sequence analysis model and evidence based analysis model is proposed to detect intrusion more intelligently. Rule-based fuzzy expert system is applied in the information fusion model to achieve higher detection rate and lower false alarm rate for intrusion detection.

1 Introduction[1]

Intrusion detection has been an important issue for computer security in resent years. Most common way to classify these different works is based on whether it may detect unknown attack. The first type of detection model is **misuse intrusion detection** [Kumar94] [Kumar95]. It uses signature for known attacks and compare it with audit data to determine if there is an attack. It has the advantage of high efficiency. However, it cannot detect unknown attacks. Beside, slight changes of intrusion behavior may not be detected. The second type of detection model is **anonymous intrusion detection** [Kumar95][**錯誤! 找不到參照來源。**]. It collects profile of system in normal state and compares audit data against it. If there is abnormal behavior, intrusion detection system generates the alarm. The advantage of anonymous detection is that it may detect unknown attacks. However, it usually suffers from lower detection rate and higher false alarm rate.

Intrusion detection systems (IDS) intend to detect attacks against computer systems. It analyzes audit data and raises the alarm if there is a

malicious attack. There are many types of attacks and different data sets should be selected properly for analysis according to characteristics of attacks. MIT Intrusion Detection System evaluation project [MITLL] has selected several intrusion detection systems and evaluated their performance. It categorizes intrusion detection into four types. These four types of intrusion intrusions are User-to-Root attack, Remote-to-User attack, DoS and DDoS attacks, and Probe attack. Different data sets should be chosen to detect different kinds of attacks. We may classify these data sets into four levels, including user-level data set, process-level data set, system-level data set and packet level data set.

In this project, a flexible intrusion detection model is proposed to amend the drawbacks of traditional misuse intrusion detection. The proposed model may endure disguised intrusion which conceals attack by camouflaging its intrusion behavior. In the proposed intrusion detection model, audit data are analyzed with two models, one is sequential sequence analysis model and the other is evidence based analysis model. Each of them has their advantages and characteristics to detect specific kind of attack. With combinations of these two models, it may detect more complex intrusion attacks. Fuzzy controller is applied to make a connection between these two models and fuse the result of them to determine if there is a malicious attack more flexibly and precisely.

2. Related Work

There has been much research on misuse-based intrusion detection systems. One of the criteria used to classify these different works is based on whether the order of sequence is taken into consideration or not. Sequential sequence analysis model analyzes the order of element in sequence and process these audit data serially. On the other hand, evidence based analysis collects audit data and analyzes it without taking order of element in audit data into consideration.

2.1 Sequential sequence analysis

The first type of intrusion detection model is **sequential sequence analysis**. It takes the orders of audit data into consideration. But sequential sequence analysis suffers from drawback of small deviation of orders of known attack behavior may not be detected. There has been much research on detecting intrusion with sequential sequence analysis model:

State transition analysis keeps a transition table to monitor user behavior step-by-step [Ilgun95]. Pattern-oriented model formalizes attack scenarios with formal definition [Shieh97]. Probability technique is also used to detect intrusion detection system. In intrusion detection system, the Morkov chain model of system's norm profile is learn from historic data of system's normal behavior [**錯誤! 找不到參照來源。**].

2.2 Evidence based analysis

The second type of intrusion detection model is **evidence based analysis**. In evidence-based analysis, audit data is analyzed according to intrusion occurrence without considering its ordering. Sometimes, attack can still be accomplished although order of attack pattern changes to escape from step-by-step monitor of intrusion detection system. Hence, evidence based analysis may tolerate slight changes of attack orders.

Data mining is also used to detect intrusion in evidence based analysis model [Lee01]. Neural network is also capable of analyzing data with

parallel processing of input data, which is the same as evidence base analysis. eXpert-BSM is an event-driven intrusion detection system that use audit log generated by Solaris Basic Security Module as data set for analyzing process [Lindqvist01] [BSM].

2.3 Information fusion for intrusion detection

Information fusion refers to the acquisition, processing and synergistic combination of information gathers by various knowledge sources. Fusion of various data source provides significant advantage over single data source, because it provides more complete information to enhance accuracy and preciseness. The fundamental architecture involves a hierarchical model that transforms multiple sources as input and processes the data with inference scheme.

Fundamental issues to be addressed in building a data fusion system have been discussed in [Hall97]. MADAM ID, Mining Audit Data for Automated Models for Intrusion Detection [Lee01], is an intrusion detection system that applies data mining model to detect intrusion. Multi-agent based intrusion detection architecture for LAN is proposed in [**錯誤! 找不到參照來源。**].

3 Modeling and methodology

In this section, an intrusion detection model is designed that fuses different detection models. In this way, sequential sequence analysis and evidence-based analysis are fused to complement each other in functionality.

3.1 Model Architecture

There are three components in the proposed architecture, as illustrated in Figure 3-1. In phase 1 analysis, system call traces and audit data logged by Solaris Basic Security Module are the analysis data source for sequential sequence analysis and evidence based analysis respectively. In phase 2 analysis, the results of two analysis models are inputted into information fusion model to evaluate overall malicious level. The advantage of this design of two-phases detection is that it takes the advantage of two models so as to increase detection rate and lower false alarm rate. There are detailed discussions in next sections.

3.2 Sequential sequence analysis

According to categorization of data sets described in introduction, process-level data can better represent ordering characteristic of process because process has it's own processing flow and execution routine. Hence system call traces of process are selected to be audit data sets for sequential sequence analysis.

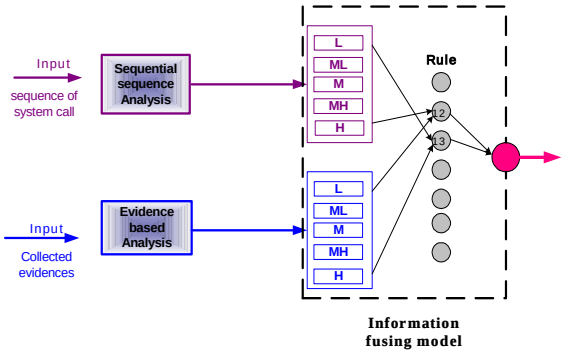

Figure 3-1 Model architecture

In this project, we focus on detecting intrusion toward programs such as sendmail、ftp、named、lpr services provided by system. We capture system calls for only specific programs running on system, so the volume of audit data is acceptable although system call traces may generate larger amount of audit log compared to other audit methodology.

Several sequence analysis methodologies that are applied to analyze system call traces have been

selected. In the proposed sequential sequence analysis model, sequence time-delay embedding is adopted to analyze system calls [Steven98].

3.2.1 Logging of system calls

Most of the system call traces utility provides functionality of selective log. For example, *struss* in Solaris provide the functionality to define filter to include or exclude dedicated system call when logging. Selective logging of system call is important and necessary issue because it may reduce the volume of audit data and improve the speed of analysis process enormously.

3.2.2 Signature database creation

Signatures are created by traces of system call sequence when intrusion is processing. After system call traces of attack pattern have be collected, each of them is sliced into subsequence with length window size. When window sliding through signature sequence one system call a time, system call sequences covered in the window are extracted and subsequence is created. This subsequence is inserted into signature database.

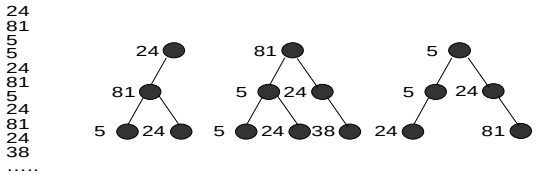

Figure 3-2 Data structure for insertion of sequential sequence

Figure 3-2 shows an example of inserting subsequence of system call trace into signature database.

3.2.3 Measure of malice

Sliding window mechanism is applied when analyzing audited system call traces and Hamming distance algorithm is used to compare difference between audit log sequence and signature sequence. Audit sequence with length of window size is compared with sequences in signature database, and Hamming distance between audit sequence and signature sequence is calculated. The smaller the Hamming distance is, the more similar it is between audit log and signature, the more possible it may be to indicate an intrusion. When subsequence of audit sequence is compared with sequences in signature database, the hamming distance between them is calculated.

Evaluation for malicious level of audit trace is modified in our sequential sequence model compared to the methodology proposed in [Steven98]. The reason for modification is because the type of analysis model proposed in this project is designed for misuse detection instead of anomaly detection. The design goal for original anomaly detection is to lower false alarm rate, so it calculate signal of anomaly S_A as:

$$S_A = \{d_{\max}(i) \forall sequences \quad i \}$$

where $d_{\max}$ is the maximal hamming distance of all the comparison between audit log and profile database. $d_{\max}$ refers to the most difference between audit log and profile, and the higher the $d_{\max}$ is, the more possible it is an attack.

However, misuse detection has consideration different from anomaly detection because it compares audit log against known signature instead uncertain profile. The original mechanism designed for anomaly can lower false alarm rate but suffers from the drawback of insertion or swap of a portion of code may lower detection rate dramatically. So

we modify signal of anomaly S_A into signal of malicious S_M as:

$$S_M = \frac{\sum W - d_{\max}}{N - W + 1}$$

,where N is number of system calls in audit sequence , W is window size, and N-W+1 is number of comparison.

3.3 Evidence based analysis

In this project, we use BSM audit log utility to log audit data [BSM]. BSM is an efficient audit log module provided by Solaris and is also used in EMERALD [Lindqvist01] as audit log data. In the proposed evidence based analysis model, one BSM record is regarded as evidence, and more than one evidence construct an attack signature. When comparing audit data with signature database, the matched evidence is recorded on the list, and if number of collected evidences belong to one specific signature is large enough to indicate an attack, analysis model generates an alarm. Figure 3-3 shows the architecture of evidence base analysis.

Our evidence based analysis model is composed of five modules: interface module, monitor module, pre-processing module, signature-loader module, and analysis module.

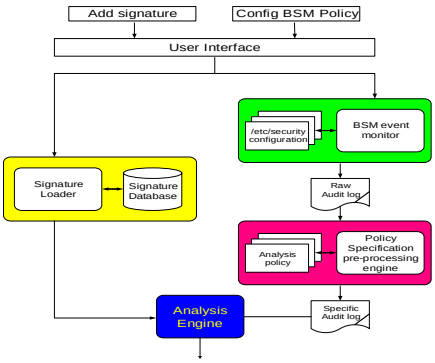

Figure 3-3 Evidence based analysis model

3.3.1 Measure of malice

Malicious level is measured by percentage of required evidences in one signature is collected. Filtered audit data generated by pre-processing module are inputted into analysis module and then compared against signature. One signature is composed of more than one evidences and the percentage of collected evidences higher than threshold may indicate an attack.

3.4 Information fusing

Information fusing technique is used to analyze two detection models of sequential sequence analysis and evidence-based analysis. It evaluates level of malicious behavior according to two analysis models. In this project, Fuzzy rule-based expert system is utilized to fuse two models.

Fuzzy controller is selected as information fusion model for three reasons [Dickerson01]. Defuzzy procedure of fuzzy controller provides a proper model to classify attack into different malicious level, which may help to strengthen response action of intrusion detection system to react correctly. There are several basic components in fuzzy controller [Jamshidi93][Meunier95].

3.4.1 Rule-based fuzzy expert system

Rule-based fuzzy expert system can be used to model complex or ill-defined system. A set of IF-THEN production rules are defined to fuse the analyzing results of different fuzzy sets.

Consider an n-input and m-output system shown in Figure 3-4. Let $\times$ be a Cartesian product of n universes x_i , for i=1,2,…,n, i.e. $X = x_1 \times x_2 \times x_3 \times ....x_n$; and $Y = y_1 \times y_2 \times y_3 \times ....y_m$.

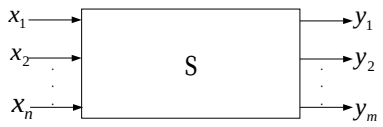

Figure 3-4 Block diagram for a n-input, m-output system

$X = (x_1, x_2, x_3, ...x_n)'$ is the input vector to the system defined on real space R^n and $y = (y_1, y_2, y_3, ...y_m)'$ is the output vector of the system define on real space R^m. The system *S* could represent any general static nonlinear mapping from *X* to *Y*, and a set of fuzzy rules are defined to describe the mapping relationship between input X and output Y.

3.4.2 Rule-based fuzzy expert system in IDS

Rule-based fuzzy expert system also inherits the advantage of fuzzy logic. It is suitable for handling varying data sources compared to traditional rule-based system. It is also more reliable and adaptive so as to meet the requirement of intrusion detection.

Fuzzy set of sequential sequence analysis for rule-based fuzzy expert system is defined as $\tilde{S}_x = \{(\tau_{\tilde{S}}(x), \mu_{\tau_x}(x)) \mid \tau_{\tilde{S}}(x) \in X\}$, where $0 \le x \le 1$, μ_{τ_x} is membership function which is normalized to the interval of [0,1], $\tau_{\tilde{S}}$ is mapping function denoted as $\tau_{\tilde{S}} : x \to X$ from numeric measurement *x* to symbolic measurement *X*, where $X = \{LOW, MED-LOW, MEDIUM, MED-HIGH, HIGH\}$. The linguistic variables are used to represent malicious level of analysis result for sequential sequence analysis.

Fuzzy set of evidence based analysis is defined as $\tilde{E} = \{(\tau_{\tilde{E}}(y), \mu_{\tau_y}(y)) \mid \tau_{\tilde{E}}(y) \in Y\}$, where $0 \le y \le 1$, μ_{τ_y} is membership function which is normalized to the interval of [0,1], $\tau_{\tilde{E}}$ is mapping function denoted as $\tau_{\tilde{E}} : y \to E$ from numeric measurement *y* to symbolic measurement *Y*,
where $Y = \{LOW, MED-LOW, MEDIUM, MED-HIGH, HIGH\}$. The linguistic variables are used to represent malicious level of analysis result for evidence based analysis model.

Rule set R^r of fuzzy rule-based expert system is defined to fuse two fuzzy sets, that is, the result of sequential sequence analysis and evidence based analysis in the form of

R^i: IF *x* is *X* and y is *Y* THEN m is M,

where *i* is *i*-th rules in rule set, *x* is malicious level of sequential sequence analysis, *y* is measure of evidence based analysis, m is final malicious level of fusion result, and $M = \{LOW, MED-LOW, MEDIUM, MED-HIGH, HIGH\}$.

Aggregation Rules are used to obtain overall output of individual outputs of fuzzy sets contributed by individual rules. In our proposed model, disjunction connection is applied because our model is designed for intrusion detection, and misuse system use signature comparison against analysis audit data instead of profile for anomalous detection. Hence, for misuse detection, detection rate is taken into considered more than anomaly detection.

Defuzzification function is used to find a value for the aggregated output. Several techniques

can be employed for defuzzifacation function including centroid of area method, mean of maximum method, smallest of maximum method, and largest of maximum method. In our proposed model, smallest of maximum method is applied.

3.4.3 Policies considered to define rule

The rules of fuzzy rule-based expert system are defined according to the characteristics of analysis model of sequential sequence and evidence based analysis model. The purpose of our design goal is to detect intrusion even if there are slight changes of intrusion pattern compared against signatures in signature database, so the changeability of attack pattern should be discussed to model the modification of attack behavior.

Three kinds of operations should be taken into considered when analyzing the effect of changes of intrusion behavior, including insertion of sequences, swap of sequences and changes of parameters. The analytical result of algorithm shows that sequential sequence analysis model may suffers from degrading of detection rate when insertion and swap operations is done, but it may endure changes of parameters because parameters is not taken into consideration in sequential sequence analysis model. On the other hand, the analytical result of evidence based analysis model shows that it may endure insertion and swap of sequences because it does not take orders into consideration. However, it suffers from degrading of detection rate when parameters are modified.

4. Evaluation

In this section, we evaluate the proposed model and present comparison of our approach and other research results. Data sets for sequential sequence analysis model are traces of system calls of processes (from UNIX program). Data sets for evidence based analysis model are traces of audit record of Basic Solaris Module provided by Solaris.

4.1 Selective log of system call traces

In order to save time and space for system call trace when sequential sequence analysis model is auditing and analyzing, it is necessary to log system calls selectively. We analyze the characteristic of data sets of system call traces obtained from MIT lab running sendmail program on SunOS 4.1.1.

Table 4-1 shows the top 5 system calls occur most frequently for normal and intrusion system call traces respectively. According to result from experiment, types of system calls that are suitable to be filtered without being logged are the ones that are functioning as kernel operation.

	Normal	Attack
First	lseek	sigvec
Second	sigstack	write
Third	rsvmsg	sigblk
Fourth	sendmsg	Fstatfs
Fifth	Readv	getpid

Table 4-1 Top 5 frequent occurrence of system calls

4.2 Evaluation for sequential sequence analysis model

In order to define rules for fuzzy rule-based expert system, the characteristic of detection model should be analyzed to evaluate the degree of degrading of detection rate when attack pattern is modified to escape from being detected. The detection rate goes down linearly with the size of insertion increase. But the detection date does not degrade dramatically with the length of swapping sequence increase, because it is bounded to the

length of window size.

According to difference level of tolerance toward changes of attack pattern, we may define the degree of changes allowed in information fusing model. With the tuning process of parameters and rules in fuzzy controller, our proposed misuse detection model can achieve flexibility and reliability because intrusion can still be detected properly even if attack pattern has been changed slightly.

4.3 Evaluation for evidence based analysis model

Data sets for BSM audit data are obtained from MIT intrusion detection evaluation project funded by DARPA. There are several types of attacks contained in audit data. Our evidence based analysis can detect host-based intrusion including Buffer overflow(eject, ffbconfig, fdformat, xterm attack contained in experiment data set)、Poor Environment Sanitation(loadable module attack contained in experiment data set)、poor temp file management(ps attack contained in experiment data set)、mis-configuration of program and bug and hole of program(such as sendmail, named, etc).

4.4 Information fusion with rule-based fuzzy controller

Level of tolerance toward modification of attack pattern has been discussed in former sections. With overall evaluation, evidence based analysis model suffer from less degrading of detection rate under modification of attack pattern. Different weights are given for these two analysis model according to the derived regulations.

Figure 4-1 is the result of information fusion model that fuses sequential sequence and evidence based analysis models, and it shows that evidence based analysis has more influence on overall malicious level than sequential sequence analysis does. The weight of evidence base analysis is heavier than sequential sequence analysis, and to what degree former analysis result is more important than latter one depends on how fuzzy rule sets are defined.

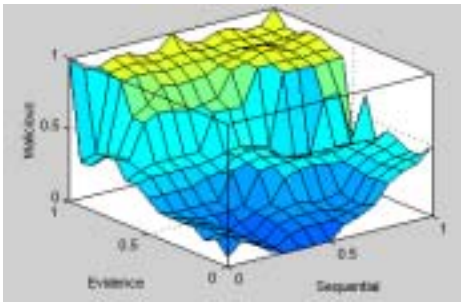

Figure 4-1 Information fusion for two intrusion detection analysis model

4.5 Comparison

Forrest proposed an algorithm that analyzes system call traces bases on Hamming distance calculation [Steven98]. Our sequential sequence analysis model also bases on Hamming distance calculation expect that our sequential sequence analysis model is misuse detection instead of anomaly detection. The advantage is that it is applicable to log system calls selectively in misuse detection than in anomaly detection. In anomaly detection, false alarm rate depends on whether profile of normal behavior is complete enough that does not miss any normal sequence and pure enough that does not contain any attack behavior. In order to keep false alarm rate low, large amount of system call traces for profile data sets should be created, hence it is not suitable to log system calls

selectively. However, misuse detection focuses on finding only misuse pattern so that selective logging can be applied to save storage and enhance analysis performance enormously.

5 Conclusion and Future work

In this project, information fusion is applied to fuse result of sequential sequences analysis and evidence based analysis to achieve higher detection rate and lower false alarm rate. It is endurable to slight changes of attack pattern so that it is more flexible compared to traditional misuse detection.

Sequential sequence analysis model analyzes system call traces with consideration of orders of system calls. It monitors step-by-step attack pattern so as to catch the intrusion more accurately. Evidence base analysis model does not take orders into consideration, so that modification of orders for attack pattern can still be detected. It monitors key parameters such as particular processes running on system or sensitive files that need to be protected.

Sequential sequence and evidence based analysis model are complement to each other in functionality, and each of them has it's characteristic for detecting specific kinds of attack. In this project, information fusing technique is applied to fuse two analysis results and take the advantages from each of them to achieve higher detection rate. Fuzzy controller is applied in our information fusion model to collect various data sources and calculate final malicious level through intelligent and reasonable inference process.

The issue of how to decide fuzzy rules for information fusion is an interesting topic. In the proposed intrusion detection model, the task of defining fuzzy rules require expert. With data analysis technique of neural-fuzzy model proposed in [Nurnberger99], rules can be obtained through automatic learning procedure, however, there has been little search on applying neural-fuzzy model in intrusion detection. The key point of automatic learning procedure relies on how to design proper input-output pair training data. It can be further studied to strengthen our intrusion detection model.

Reference

[Dickerson01] Dickerson, J.E., Juslin, J., Koukousoula, O., Dickerson, J.A, " Fuzzy intrusion detection," IFSA World Congress and 20th NAFIPS International Conference, July 2002.

[MITLL] http://www.ll.mit.edu.

[Hall97] David L. Hall, James Llinas, "An Introduction to Multisensor Data Fusion," Proceedings of the IEEE, Vol. 85, No. 1, January 1997.

[Heinbuch01] Susan C. Lee, David V. Heinbuch, "Training a neural-network based intrusion detector to recognize novel attacks", IEEE Transactions on System, Man, and Cybernetic-Partt A: System and Humans, VOL. 31, No 4, July 2001, pp. 294-299.

[Ilgun95] Koral Ilgun, Richard A. Kemmerer, Fellow, IEEE and Philip A. Porras, "State Transition analysis: A rule-based intrusion detection approach," IEEE Transactions on software engineering, vol 21, No 3, March 1995, pp. 1-22.

[Jamshidi93] Nohammad Jamshidi, Nader

Vadiee, Timothy J. Ross, ” Fuzzy Logic and Control, software and hardware applications,” Prentice Hall International Edition, 1993.

[Kumar94] Sandeep Kumar, Eugene H. Spafford, “A Pattern Matching Model for Misuse Intrusion Detection”, Proceedings of the 17th National Computer Security Conference,1994.

[Lee01] Wenke Lee, Salvatore J. Stolfo, “A Framework for Constructing Features and Models for Intrusion Detection Systems,” ACM Transactions on Information and System Security, 2001, pp.227-261.

[Lippmann98] Richard P. Lippmann, Robert K. Cunningham, “Improving Intrusion Detection Performance using Keyword Selection and Neural Networks,” First International Workshop on the Recent Advances in Intrusion Detection, Sept. 1998.

[Lindqvist01] Ulf Lindqvist, Phillip A. Porras, ” EMERALD eXpert-BSM: A Host-based Intrusion Detection Solution for Sun Solaris,” 17th Annual Computer Security Applications Conference, Dec. 2001, pp. 240-251.

[Meunier95] Bouchon-Meunier, Yager, Zadeh, “Fuzzy Logic and Soft Computing”,World Scientific, 1995.

[Nurnberger99] Nurnberger, D. Nauck, R. Kruse, “Neuro-fuzzy control based on NEFCON-model: recent developments,” Soft Computing, 1999.

[Pinheiro01] Robert Pinheiro, Alex Poylisher, Hamish Caldwell, “Mobile security agents for network traffic analysis,” IEEE Information Survivability Conference, 2001.

[Ryan98] Jake Ryan, Meng-Jang Lin, “Intrusion Detection with Neural Networks,” Advances in neural information processing systems 10, 1998 , pp. 1-7.

[Shieh97] Shiuh—Pyng Shieh, Virgil D Gligor, “On a pattern-oriented model for intrusion detection,” IEEE transactions on knowledge and data engineering, NOL. 9, No. 4, July/August 1997, pp. 661-667.

[Steven98] Steven A. Hofmeyr, Stephanie Forrest, Anil Somayaji, ” Intrusion detection using sequences of system calls,” Journal of Computer Security, August 1998, pp. 1-25.

[BSM] http://docs.sun.com/.

[Zhang01] Ran Zhang, Depei Qian, Chongming Bao, Weiguo Wu, Xiaobing Guo, ”Multi-agent based intrusion detection architecture,” IEEE Computer Networks and Mobile Computing Conference, 2001, pp. 494-501.