

PhD Stage 2 Application

Rare Event Data Mining for Forensic Purposes

Malcolm Corney

(n00213667)

October 2005

**School of Software Engineering and Data Communications
Information Security Institute
Faculty of Information Technology
Queensland University of Technology**

Table of Contents

PhD Stage 2 Application.....	1
Rare Event Data Mining for Forensic Purposes	1
Malcolm Corney	1
(n00213667).....	1
October 2005.....	1
School of Software Engineering and Data Communications	1
Information Security Institute	1
Faculty of Information Technology	1
Queensland University of Technology	1
Table of Contents.....	i
1. The Proposed Title.....	1
2. The Proposed Supervisors and their Credentials	1
3. Introduction.....	1
4. Program of Research and Investigation	2
4.1 Research Question	2
4.2 Individual Contribution to the Research Team.....	3
5. Design of the Proposed Research	3
5.1 Methodology and Research Plan	3
5.1.1 Initial Investigations.....	3
5.1.2 Events.....	4
5.1.3 Data Mining	4
5.1.4 Data Sources	7
5.1.5 Investigation of Scenarios Attacking Specific Operating Systems.....	8
5.1.6 Investigations of Scenarios Across Multiple Computers.....	9
5.1.7 Incorporation of Other Log Data	9
5.1.8 Resources	9
5.2 Collaborative Arrangement Evidence.....	9
5.3 Timeline for Completion of PhD Program	10
5.4 Preliminary Literature Review.....	11

5.4.1	Attack Recognition	11
5.4.2	Scenario Analysis and Discovery	12
5.4.3	Rare Events	13
5.4.4	Discussion	14
5.5	Coursework	14
6.	Research Ethics Statement	14
7.	Intellectual Property Statement.....	14
8.	Health and Safety Statement.....	15
9.	References.....	16

1. The Proposed Title

The proposed title for the thesis is “Rare Event Data Mining for Forensic Purposes”

2. The Proposed Supervisors and their Credentials

The proposed supervisors for this body of research are as follows:

Principal Supervisor: Adjunct Professor George Mohay

Associate Supervisor: Dr Andrew Clark

Adjunct Professor Mohay is the leader of the Computer Intrusion, Forensics and Evidence research domain in the Information Security Institute (ISI). Prior to this he was Head of the School of Computing Science and Software Engineering from 1992 to 2002. He supervises and has supervised PhD and Masters students in the above areas. He is currently involved as chief investigator in a number of related funded research projects: in the area of computer forensics with Australia’s DSTO (Defence Science and Technology Organization), a DEST FAST sponsored project on Internet security, and an ARC sponsored project on intrusion detection.

Andrew Clark is a Senior Research Fellow in the ISI. His research interests include network security, cryptography, and computer and intrusion forensics. He supervises and has supervised PhD and Masters students in these areas.

3. Introduction

Computer operating systems, services and software contain numerous vulnerabilities that are able to be exploited using specific attacks aimed at those vulnerabilities. Locard’s exchange principle suggests that some evidence of such exploits will be left behind on the attacked system and that evidence can be detected in a forensic investigation. In these cases, this evidence should be able to be found in the various event log files that are recorded on host and server systems.

The research problem to be investigated is the automated detection of attacks on computer systems by utilizing data mining techniques on computer and other event logs from heterogeneous sources. There is to be an emphasis on the use of rare events for the detection of attacks. The detection of attacks using information from event logs is currently done in a manual fashion and requires in depth knowledge of the systems that have been attacked.

Current research in attack scenario detection is centred on the use of intrusion detection systems and the correlation of results from different intrusion detection systems. The aim of this research is to use different information i.e. the information from the event logs, to detect those attack scenarios. The approach of correlating information from a number of event logs for detecting attack scenarios is a novel approach in this field of research.

4. Program of Research and Investigation

4.1 Research Question

The research question to be answered involves the detection of attack scenarios and exploits of vulnerabilities that are carried out on computers or across computer networks. This will be accomplished by using data mining techniques to discover attack scenarios on computer related events stored in log files from a variety of sources. The detection of attack scenarios is to be carried out with a particular emphasis on events which occur rarely in the event log population.

Computer events can come from a wide variety of sources. These event sources include events from different operating systems, e.g. different versions of Windows, Linux and UNIX. The events can also be generated by a range of services and applications running on computers using these operating systems. Events can be generated by web servers, mail servers, firewalls and intrusion detection systems or devices. Physical pieces of infrastructure such as elevators and doors can also generate loggable data.

It is now common for these events to be logged by the computers on which they are generated. Analysis of these log files to detect attacks is a difficult process as the logs are very populous, are sometimes difficult to understand and are usually stored in heterogeneous formats. In order to be able to compare these data sources with each other, a means for storing them in some common format is required.

An important objective of the research is to investigate the feasibility of identifying rare cases of events and activities that will lead to the detection of attacks or exploits against security vulnerabilities in computer systems. The intent of this research is to use data mining techniques for the discovery of rare cases of events and to produce clusters of events [3] with similarities in an effort to detect such attacks.

The importance of feature selection in data mining in general and also in rare event detection in particular is noted [3] and a significant aspect of this project will focus on feature selection and the level of activity to which it is applied. An important step in feature selection is the production of data vectors, which relate to each available event, and contain values for features which describe that event.

The data is to be stored in a database produced by the Event Correlation for Forensics (ECF) software [6]. This software allows event hierarchies to be formed so that activities can be detected that are not system specific. The software constructs higher order activities by correlating data from the raw event logs. The current system is capable of identifying activities such as file access, login sessions, web server sessions and can detect some known attack types.

The optimal level in the event/scenario hierarchy at which to apply data mining techniques will be investigated. This will to a large extent be dependent on the features that are recorded for the events at different levels in the hierarchy.

While the initial research will concentrate on clustering techniques to achieve the prime purpose of attack identification, the investigation of other possible approaches will not be ruled out as the project progresses. Other data mining techniques for rare case identification that may be considered are as follows:

- Temporal analysis of rare events [5]
- Statistical approaches to identifying outliers [17, 25] also use of ‘emerging patterns’ in AR for rare class identification
- The MINDS project [14, 16] has made good progress with data mining techniques in the Intrusion Detection System domain, and have identified a number that may be used for anomaly detection, in particular distance based techniques such as k -nearest neighbour and the Mahalanobis distance

4.2 Individual Contribution to the Research Team

The initial work to be carried out as part of this research is based on a research project being carried out in conjunction with the Defence Science Technology Organization (DSTO) of the Australian Department of Defence, entitled “**Event Abstraction and Data Mining for Forensic Purposes**”.

The research team is working on two main areas. The first area is that of recognition of common scenarios and classes of scenarios. This work follows on from a previous project with DSTO where software for event correlation was produced. The second area of the current project is to use data mining for the detection of rare events and classes of events for the discovery of attack scenarios.

My contribution to the research team is the investigation into data mining techniques for the identification of rare events and rare event classes that will aid in the detection of attacks and exploits of vulnerabilities in computer systems. It is my responsibility to investigate the usefulness of pre-existing data mining tools for use in this project. I will also be responsible for the definition of attributes or features that will be used in the data mining processes and testing the efficacy of those features.

5. Design of the Proposed Research

5.1 Methodology and Research Plan

5.1.1 Initial Investigations

The initial stages of the research have been and will continue to be centred on the investigation of event logs from Windows operating systems. The various versions of Windows operating system since Windows NT, record events into three separate logs: the application log, the security log and the system log. The most important of these logs is the security log and information relating to the contents of this log and the data recorded in the events has been located [21, 22].

To date work has been carried out to ensure that events from the Windows security log can be parsed and added to the ECF database. Various data

sources have been used for this including my local computer and saved log files from DSTO and from the 1999 DARPA Intrusion Detection data set [36].

5.1.2 Events

This familiarization stage with event logs and event data will allow the formation of feature sets which can be used for data mining purposes.

More and more frequently, operations that are carried out on or with computers are recorded in some manner in an event log. Operating systems such as Microsoft Windows and the many UNIX or Linux variants provide means for recording the details of events that occur. An important set of events for Microsoft Windows systems is reported by Smith [29].

A series of events that are recorded due to carrying out some exploit or attack can be defined as a scenario. The events in that scenario have both temporal and causal relationships. In the field of computer security and computer forensics, the post hoc detection of security exploits and vulnerabilities relies heavily on the events that have been logged by the computer or computers involved in the exploit.

If the series of events that define a scenario can be determined, a signature for the scenario can be prepared. This signature can then be used to evaluate whether or not a system has been exploited. It is impossible to know or to define all possible scenarios. A mechanism for automated scenario discovery would therefore be advantageous.

Scenarios could be discovered by concentrating on finding events which occur rarely and then correlating them with other events that have occurred at a similar time or within a time window. The correlation would have to take into account other attributes recorded in the events such as the IP address, the login identity, the name of a file, the process identification number and/or the computer name.

5.1.3 Data Mining

Data mining is an activity that can be carried out on large bodies of data in an attempt to discover new knowledge from the original data [12]. Data mining is typically performed on data sets from a variety of fields – weather data [9], medical data [30], astronomical data [34], financial data [17] and has been used in attempts to improve the performance of intrusion detection systems [4].

5.1.3.1. Data Mining Tools

It is not the purpose of this research to develop new data mining tools. The intent is to utilize existing data mining tool kits. The most promising set of tools that could be integrated with the existing ECF software is that provided by Weka [33].

This resource contains Java libraries for a multitude of classifiers and clusterers. Each of the tools has been made available in a suite of GUI based applications. The user has the ability to explore data, perform experiments between different learning schemes and a combination of the previous

activities in a drag and drop environment. Functionality has been provided to preprocess data, classify, cluster, associate, trim features and visualize data.

5.1.3.2. Feature Selection

From the investigations carried out so far a number of features should be able to be used for clustering raw event data. This is the data produced by ELDump [15] that is entered into ECF. If ECF is to be used to produce input for data mining, it will be necessary to reconstitute each event into a standard vector style output with specific features in specific columns, each having their associated value for the particular event or data point. Clustering of higher order events should also be possible. For each higher order event sequence, there should still be a series of features for the series that can be used for calculating distance measures in the clustering algorithms.

Examples of features that are present in the Windows Security event log include:

- time
- username
- computer
- domain
- workstation
- handle ID
- process ID
- logon type
- logon id
- primary user name
- primary domain
- primary logon id
- client user name
- client domain
- client logon id
- logon ID

Some of these attributes, such as the logon IDs (plain, primary and client) have values such as (0x0,0x2B1A1), and some user data is represented with values such as S-1-5-21-742865521-1025978620-313593124-500. While these values are unique, it may be necessary to massage the feature values into some form of numerical format if distance measures are to be used for clustering approaches. The use of a hash function of some sort may be one way of converting data from a variety of sources into a numerical format where needed.

There are a number of dimensionality reduction approaches which might be used to overcome the problems of a high dimensionality in the feature space [12]. These include: attribute transformations, such as Principal Components Analysis and Singular Value Decomposition; subspace clustering, for which a

number of algorithms exist; and co-clustering, where attributes are clustered into groups and proxy attributes are derived from these clusters.

5.1.3.3. Clustering

Clustering is a technique used to separate data points from a sample into distinct groups that are similar in some manner so that the data points in a cluster are more similar to each other than to the data points in other clusters [12]. To be able to generate the clusters, a metric is required to measure the similarity or dissimilarity between them. This is usually done by using some type of distance measure. The most common distance measures are the City Block distance measure:

$$d_1(x_i, x_j) = \sum_{k=1}^m |x_{ik} - x_{jk}|$$

and the Euclidean distance measure which is calculated as follows:

$$d_2(x_i, x_j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$$

where:

x_i and x_j are the i^{th} and j^{th} data points in the population

k is the attribute number

m is the number of attributes or the dimension of the feature space

Another metric that can be used is the Minkowski metric which is a more general formula:

$$d_p(x_i, x_j) = \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^p \right)^{1/p}$$

A measure known as the cosine correlation can be used to calculate the similarity between two data points rather than the distance between two data points:

$$s_{\cos}(x_i, x_j) = \frac{\left[\sum_{k=1}^m (x_{ik} \cdot x_{jk}) \right]}{\left[\sum_{k=1}^m x_{ik}^2 \cdot \sum_{k=1}^m x_{jk}^2 \right]^{1/2}}$$

There are two main approaches used for creating clusters. These are hierarchical clustering and partitional clustering.

With hierarchical clustering, data is organized in a nested sequence of groups and can be typically displayed in a dendrogram or a tree like structure. Every cluster node can contain child clusters. Sibling clusters contain data points that are covered by their parent cluster but are partitioned from each other.

Hierarchical clustering can be further subdivided into agglomerative and divisive algorithms. Algorithms include SLINK, COBWEB, CURE and CAMELEON.

Agglomerative algorithms build the clusters in a bottom up fashion by starting with the first data point in the population and then recursively merging two or more clusters. The agglomeration stops when the specified number of clusters has been built.

Divisive algorithms build the clusters in a top down nature by starting with the entire population in one cluster. This cluster is recursively split and the process ends when the specified number of clusters has been built.

In partitional clustering algorithms, data is organized by minimizing the within cluster squarer error between data points in the same cluster and by maximizing the square error between different clusters.

5.1.4 Data Sources

Data will be sourced from a number of places. DSTO have provided a proprietary data set which includes a particular scenario. This data set is quite small and will not contain enough data to be useful for base lining purposes.

5.1.4.1. DARPA Data Set

The initial research has focussed on determining the event types in the 1999 DARPA Intrusion Detection data set [36]. This data set contains over 10,000,000 separate events and consists of logs collected over a five week period on: days where no attacks were made; days where attacks were made and the type of attack and time of the attack are known; and days where attacks were made but no record of the details of the attacks have been released to the research community.

This data set is now considered old, as the Windows operating system has been updated twice since (Windows 2000 and Windows XP) and the attacks are aimed at vulnerabilities of the Windows NT operating system. This data set, however, is still in use by the intrusion detection research community and should still be able to be used for the proof of the concepts that are being researched.

5.1.4.2. Local Data Set

As part of the DSTO research project discussed in Section 4.2 it has been decided that a test network for capturing network traffic, the results of intrusion detection systems and event logs from computers running on both Windows and Linux operating systems will be built.

Furthermore, a series of attacks will be aimed at the vulnerabilities in these systems and all possible data will be recorded for later analysis. Attacks that can be carried out can be classified into probes, denial of service (DOS) attacks, Remote to Local (R2L) attacks, User to Root (U2R) attacks and data attacks [20].

The data generated from these attacks can then be used for attack scenario discovery. With this data set, it will be possible to validate the results of any

techniques used for scenario discovery. While a large amount of data will be generated by running these exploits, this will be a controlled data set.

5.1.4.3. ITS Data Set

An agreement has been reached with Information Technology Services (ITS) at QUT to gather data from their routine monitoring of the networks and systems at QUT. This data includes logs from numerous sources, including intrusion detection systems, syslogs from UNIX servers, proxy logs, logs from numerous Windows host machines and web server logs. ITS is currently logging approximately 3GB of uncompressed data daily, and maintains 45 days worth of compressed log data. This is an enormous amount of data to consider, and only portions of it will be experimented with. As this is real data there is no guarantee that attack scenarios will be discovered. Furthermore, there is no way to validate the data set to determine if the techniques used for attack scenario discovery actually work. This data set may be left until a discovery technique is near completion, when the data can be used as a production scale test of the technique.

The use of such data may pose an ethical question as the data is generated by the actions of people using computers. An ethics application will be made in the near future. More details of this are given in Section 6.

5.1.5 Investigation of Scenarios Attacking Specific Operating Systems

To investigate scenarios capable of exploiting vulnerabilities in the Windows operating system, data from the local data set will be investigated. This data set will consist of well controlled data containing known attacks. The event logs will be imported to the ECF database and manipulated to create data sets for data mining.

This will include the extraction of feature:value pairs for inclusion in data vectors. Each data vector relates to one event. Correlation of feature values will be carried out to create clusters of closely related events. As the timing of the attacks will be known from the details of the generation of the data set, it should be possible to identify event clusters that relate to the attacks, in an effort to determine attack signatures.

When various attack signatures that are known have been shown to be detectable, data from the DARPA data set or the ITS data set will also be analysed to determine if attacks can be identified in this data as well.

Investigations for exploits attacking computers using the Linux operating system will be carried out in a similar manner as for computers using the Windows operating system.

This step will allow further reasoning about attacks or exploits that are similar in nature but aimed at different operating systems. User to Root (U2R) exploits, for example, are possible on both operating systems.

It is possible that similar patterns of events can be detected in the logs from both operating systems. The ECF software allows the detection of higher order event sequences e.g. login events and login sessions. Login events can

be detected from the host machine's raw event log and designated as such in the database for use by higher level events that require the detection of such a login event. Login sessions consist of a login event and a corresponding logout event.

A prerequisite of a U2R exploit, for example, would have to be a login session. It should be possible, therefore, to build scenarios from these higher level events as detected by ECF rather than relying on the raw event logs.

5.1.6 Investigations of Scenarios Across Multiple Computers

Some attack exploits might involve more than one computer on a network. If the traffic and event logs from multiple computers can be incorporated into one data set, using the attacks identified in Section 5.1.5, it should be possible to build up a higher level picture of the cause and effect of various scenarios.

This will require reasoning about the temporal and causal nature of the identified event sequences.

5.1.7 Incorporation of Other Log Data

ECF is not intended as a tool just for storing event data from various computers on a network. Its purpose is to correlate events that come from a wide variety of sources. To date, a number of logs from heterogeneous sources can be incorporated into the database. Sources include Windows operating system event logs, UNIX or Linux Syslog, Apache web server logs, web browser logs, door logs and e-mail server logs.

Once a methodology has been determined for identification of attack scenarios using data mining techniques on logs from one or two sources, it is the intent of this research to incorporate logs from the variety of sources mentioned above. Inclusion of data from a wide variety of sources will be attempted to detect correlations across a wider range of sources.

5.1.8 Resources

The research is being supported by my employer, the School of Software Engineering and Data Communications by the provision of teaching relief and initially also, by funding from DSTO.

The main resource required will be suitable data. As discussed in Section 5.1.4, some data will be sourced from QUT ITS and some data will be generated. This latter data set will require a small network to be built. This will be undertaken by the technicians in the ISI.

5.2 Collaborative Arrangement Evidence

There is no collaborative arrangement in place for this research project.

5.3 Timeline for Completion of PhD Program

PhD Timeline

Rare Event Data Mining for Forensic Purposes

Activity	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1
	2005				2006				2007				2008				2009				2010			
Milestones																								
Progress Reports		■				■				■				■				■					■	
Ethics Application			■																					
Stage 2			■																					
Confirmation								■																
Final Seminar																								■
Submission																								■
Coursework																								
AIRS		■																						
Research Tasks																								
Data Collection and Generation		■	■																					
Investigation of Windows Events																								
data collection	■	■	■																					
data processing		■		■	■																			
data mining			■	■	■	■	■																	
Investigation of Unix Events																								
data collection			■	■	■																			
data processing							■	■																
data mining								■	■	■	■													
Investigation of Other Events																								
data collection												■												
data processing												■	■											
data mining												■	■	■	■									
Data Mining																								
discovery of common events											■	■	■	■	■									
discovery of rare events													■	■	■	■	■							
Scenario Discovery																■	■	■	■	■	■	■		
Thesis Preparation																								
Literature Review	■	■	■	■																				
Thesis Body																					■	■	■	■
Conclusions, Introduction & Abstract																						■	■	■

5.4 Preliminary Literature Review

In recent years the strengthening of computer security systems has not been sufficient to prevent attacks by malicious agents such as computer viruses and worms targeting the ever lengthening list of vulnerabilities present in current operating systems, services and application software.

5.4.1 Attack Recognition

Research has been carried out in a number of areas in attempts to recognize attacks. These areas include intrusion detection, computer process analysis and computer event analysis.

5.4.1.1 Intrusion Detection Systems for Attack Recognition

In an attempt to detect such attacks, research into Intrusion Detection Systems developed but IDSs have faults of their own. Anomaly based IDSs can produce many false positive alerts depending on their training and signature based IDSs cannot detect novel attacks with unknown signatures. Much research has been carried out in efforts to improve the results of IDS with correlation of alerts being one focus.

Lee et al [19] used a framework which allowed multiple intrusion detection sensors to exchange information to detect new and distributed intrusions. This system used a data mining approach to build intrusion detection models and distribute the newly discovered attacks to other sensors in the system.

Brugger [4] provides a review of how data mining has been applied to network intrusion detection including the use of statistical techniques, machine learning and correlation techniques.

Haines et al [10] measured the ability of correlation systems to detect attacks in a simulated military network using a large number of intrusion detection sensors. This work was aimed at testing the best IDSs at the time of the research against a variety of attack categories. A set of metrics was defined for the comparisons.

More recent research conducted by Xu and Ning [35], Noel et al [24], and Qin and Lee [26, 27] have all used correlation of intrusion detection alerts for the recognition and discovery of attack scenarios. This work recognizes the causal relationships between alerts in attack scenarios and highlights that certain preconditions must be satisfied if scenarios are to proceed along specific paths.

Aslam [2] presents the Kerf toolkit for intrusion analysis as part of a system where a language, SawQL, has been developed for querying recorded data to correlate records by time and features. This language allows the system administrator to query the data based on sequences of events and to apply temporal relationships between events in the sequence.

5.4.1.2 Analysis of System Calls for Attack Recognition

Ju and Vardi [11] studied sequences of shell commands and system calls in a UNIX system in an attempt to detect computer intrusions. A statistical

approach was taken to compare audit data with estimated signatures of normal behaviour. This study did not attempt to analyze the events recorded in the Syslog, instead focussing on the rarity of short command sequences.

Lee et al [18] used machine learning techniques to identify misuse and intrusion in UNIX systems. Sendmail data was used as the vehicle for this study and the results indicate that when properly trained, abnormal executions can be readily detected.

5.4.1.3. Analysis of Events for Attack Recognition

Chuvakin [8] discusses the correlation of events from computer logs, firewalls, and intrusion detection systems to use in a rule based Security Information Management System. The problem of heterogeneous logs and the need for data to be normalized to a common format is discussed. This system is not capable of detecting novel attacks as it relies on existing rules to detect possible attacks.

Allison [1] and Kramer [13] have used Perl scripts for log data reduction i.e. removal of irrelevant data, and specific event detection in their work. The approach taken in this work has been based mainly on TCP data and the detection of suspicious or anomalous activities.

Romig [28] has correlated data from a variety of UNIX sources in computer forensic investigations. He notes the difficulties with correlating times between different event sources and discusses the impact of missing log entries on the investigations.

5.4.2 Scenario Analysis and Discovery

Systems such as SAP and Management Information Systems can record work flows where the activities that take place are logged, but the process by which the work is done is not recorded. The operations in such an event series have both temporal and causal relationships. van der Aalst and de Medeiros [31] have used process modeling to rediscover work flows from the events recorded in the logs of such systems. This paper discusses how the event logs can be used for process discovery and delta analysis i.e. how a process differs from a prescribed process (anomaly detection). Each event in the log contains a timestamp, information about the performer or originator of the event, and information about the activity itself. The researchers investigated the information in the logs from three different perspectives – process (how it was done), organizational (who was involved) and case (what was done) perspectives. The process itself is visualized using a Petri Net.

Ning et al [23] have developed techniques to detect high level attack scenarios from a large collection of low-level intrusion detection alerts from complementary intrusion alert systems. They correlate these alerts based on measures of similarity using clustering and on prerequisites and consequences to describe causality to construct the attacks. After correlating with both techniques, the results are combined to build a better representation of the attack. The basic requirement for the discovery of scenarios is the reliance on the underlying intrusion detection systems generating reliable alerts. This is of course a limitation, as alerts can be easily missed and many generated alerts

are false positives. When alerts are correlated based on common attribute values, further reasoning can be conducted to infer that certain steps are missing due to alerts missed by the IDS. This technique relies on at least some of the attack steps in the sequence being known so that attack steps can be correlated and missing steps possibly unknown can be hypothesized about.

Security alerts can be produced by intrusion detection sensors, firewalls and file integrity checkers. Each alert is the result of some step in an attack scenario being possibly an exploit, probe or other event. Cheung et al [7] discuss how it would be desirable to have the recognition process automated in some way but recognize that there are many challenges to this process. Obstacles include: the representation of knowledge; the heterogeneous nature of alerts from different sources; steps in the attack being temporally distributed and spatially distributed among various sources; likely to be present in high numbers including the possibility of a high number of false positive alerts from IDS; the scenario may be carried out in numerous ways meaning that the order of the steps is temporally different; and steps might be missing from the attack scenario. They have developed a modeling language named Correlation Alert Modelling Language (CAML) and have developed a scenario recognition engine that takes as input, the low level alerts to create a model of the scenario. The language must be able to describe the scenario and express relationships between the steps such as the temporal relationships and the relationships between feature – value pairs, and prerequisite relationships.

CAML has been designed so that multistage scenarios can be specified in a modular manner. Attack steps are written in modules, each containing an activity section which outlines the matching steps in the attack for this module, a precondition section, which must be met before this module can hold, and a post-condition section which describes what will be satisfied once the activity and precondition sections are met. The post-conditions can be used as preconditions for other modules in the attack scenario.

5.4.3 Rare Events

Rare events when performing data mining are normally of special interest [32]. These rare events can be problematic for the data mining system as they can be outliers or disjuncts from the normal data and can be easily overlooked by the system. When data mining systems are trained with these rare events, generalization can be poor as there is normally a low number of these events in the training set. It is possible to improve the results by using different techniques including: obtaining additional training data; using different metrics that do not discount the importance of rare events – the precision metric is more important in these cases; by employing boosting algorithms, where different weights are placed on the training data in separate training iterations; and by placing rare events in their own training classes.

The temporal sequence of rare events can be another important factor in data mining. While some single events may occur rarely in a data set, temporal sequences including that event may be even less common. Chen et al [5] combine association and sequential pattern discovery to identify important rare events.

5.4.4 Discussion

While some work has been done in the area of computer event log analysis, this work has not been aimed at the detection of unknown attacks or attacks without signatures. Correlation of intrusion detection alerts has been a popular field of study in recent years for the detection of attack scenarios. Statistical approaches and data mining approaches have been used for this purpose. These studies should provide valuable insights for the proposed research where correlation of heterogeneous event log data will be used for the discovery of attacks.

With a vast array of data from computer event logs, it will most likely be necessary to use techniques to identify rare events and possibly rare temporal sequences of events to base the data mining investigations upon. If rare events can be located, clustering techniques can be used based on those events in the scenario discovery process.

5.5 Coursework

The only coursework planned for this research is the unit IFN001 Advanced Information Retrieval Skills. I have already attended the course and will submit the assessment item for this unit before the end of Semester 2, 2005.

6. Research Ethics Statement

The majority of the data to be used in this project has been generated from specific scripts that people have followed to generate various event logs. This will also be the case from the proposed experimental computer network, where various attack scenarios will be run and event logs recorded and analysed.

If real system data is requested from QUT ITS from their recorded event logs, there will be event data which has been generated by people. A letter of intent has been sent to Barry Lynam in ITS, outlining that the data we have requested will be used for investigative purposes. While the data will not contain personal details of the people whose actions have been recorded in the event logs, it could be possible to identify people from IP addresses and computer and network names. It is believed that ethical clearance will have to be sought before such data can be used.

If the research results in the production of any publications that need to contain raw data, such data will be deidentified before publication.

An application for ethical clearance is yet to be completed but an application is in progress and should be submitted before the end of 2005.

7. Intellectual Property Statement

The initial stages of this research are aligned with a research project being carried out in conjunction with the Defence Science and Technology Organization (DSTO) of the Department of Defence entitled **Event Abstraction and Data Mining for Forensic Purposes**.

Section 7, Paragraph 5 of the research contract states:

“The Research Institution shall cause all Specified Personnel prior to undertaking work in respect of the Research Project to assign any Developed Intellectual Property to the Commonwealth and to enter into confidentiality obligations in relation to all information arising from, or related to the Research Project as the Commonwealth shall require.”

8. Health and Safety Statement

My topic of research does not involve the use of any high risk materials and there are no health and safety implications arising from the project.

9. References

1. Allison, J., *Automated Log Processing*, in; *login: The Magazine of Usenix and Sage*. 2002. p. 17-20.
2. Aslam, J., et al., *The Kerf Toolkit for Intrusion Analysis*. IEEE Security & Privacy, 2004: p. 42-52.
3. Berkhin, P., *Survey of Clustering Data Mining Techniques*. 2002, Accrue Software.
4. Brugger, S.T., *Data Mining Methods for Network Intrusion Detection*, www.bruggerink.com/~zow/papers/brugger_dmnid_survey.pdf (last accessed 23rd November, 2004), 2004.
5. Chen, J., et al. *Temporal Sequence Associations for Rare Events*. in *The Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-04)*. 2004. Sydney, Australia.
6. Chen, K., et al. *ECF - Event Correlation for Forensics*. in *1st Australian Computer, Network & Information Forensics Conference*. 2003. Perth, WA, Australia.
7. Cheung, S., U. Lindqvist, and M.W. Fong. *Modeling Multistep Cyber Attacks for Scenario Recognition*. in *The Third DARPA Information Survivability Conference and Exposition (DISCEX III)*. 2003. Washington, D.C., U.S.A.
8. Chuvakin, A., *Security Event Analysis through Correlation*. Information Systems Security, 2004. **13**(2): p. 13-.
9. Ehrenman, G., *Mining What Others Miss*, in *Mechanical Engineering*. 2005. p. 26-31.
10. Haines, J., et al., *Validation of Sensor Alert Correlators*, in *IEEE Security & Privacy*. 2003. p. 46-56.
11. Ju, W.-H. and Y. Vardi, *Profiling UNIX Users And Processes Based on Rarity of Occurrence Statistics with Applications to Computer Intrusion Detection*. 2001, Avaya Labs Research. p. 1-22.
12. Kantardzic, M., *Data Mining Concepts. Models, Methods, and Algorithms*. First ed. 2003, Piscataway, NJ, USA: IEEE Press. 345.
13. Kramer, T., *Effective Log Reduction and Analysis Using Linux and Open Source Tools*. 2003.
14. Kumar, V., *Data Mining for Rare Class Analysis*, WWW Document, http://www-user.cs.umn.edu/~aleks/rare_class/ (last accessed 21st April, 2005), 2003.
15. Lauritsen, J., *ELDump* ver 0.13, <http://www.ibt.ku.dk/jesper/ELDump/default.htm>, 1998.
16. Lazarevic, A., et al. *A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection*. in *Third SIAM Conference on Data Mining*. 2003. San Francisco, CA.

17. Lazarevic, A., J. Srivastava, and V. Kumar, *PAKDD 2004 Tutorial: Data Mining for Analysis of Rare Events - A Case Study in Security, Financial and Medical Applications*, WWW, (last accessed 17/2/2005, 2005), 2004.
18. Lee, W., S.J. Stolfo, and P.K. Chan. *Learning Patterns from Unix Process Execution Traces for Intrusion Detection*. in *Proceedings of the AAAI Workshop on AI Methods in Fraud and Risk Management*. 1997.
19. Lee, W., et al. *A Data Mining and CIDF Based Approach for Detecting Novel and Distributed Intrusions*. in *Proceedings of the Third International Workshop on Recent Advances in Intrusion Detection*. 2000: Springer-Verlag.
20. Lippman, R., et al. *Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation*. in *RAID 2000*. 2000: Springer-Verlag.
21. Microsoft Corporation, *Windows 2000 Security Event Descriptions (Part 1)*, WWW Document, <http://support.microsoft.com/kb/299475/EN-US/> (last accessed 15th August, 2005), 2003.
22. Microsoft Corporation, *Windows 2000 Security Event Descriptions (Part 2)*, WWW Document, <http://support.microsoft.com/kb/301677/EN-US/> (last accessed 15th August, 2005), 2003.
23. Ning, P., et al. *Building Attack Scenarios through Integration of Complementary Alert Correlation Methods*. in *The 11th Annual Network and Distributed System Security Symposium*. 2004. San Diego, CA, USA.
24. Noel, S., E. Robertson, and S. Jajodia. *Correlating Intrusion Events and Building Attack Scenarios Through Attack Graph Distances*. in *20th Annual Computer Security Applications Conference (ACSAC 2004)*. 2004. Tucson, AZ, USA: IEEE Computer Society.
25. Pelleg, D. and A. Moore. *Active Learning for Anomaly and Rare-Category Detection*. in *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*. 2004. Whistler, BC.
26. Qin, X. and W. Lee. *Attack Plan Recognition and Prediction Using Causal Networks*. in *20th Annual Computer Security Applications Conference (ACSAC 2004)*. 2004. Tucson, AZ, USA: IEEE Computer Society.
27. Qin, X. and W. Lee. *Discovering Novel Attack Strategies from INFOSEC Alerts*. in *9th European Symposium on Research in Computer Security (ESORICS 2004)*. 2004. Sophia Antipolis, France.
28. Romig, S., *Correlating Log File Entries*, in; *login: The Magazine of Usenix and Sage*. 2000. p. 38-44.
29. Smith, R.F., *Monitoring Important Security Events*. *Windows & .NET Magazine*, 2003: p. 57 - 62.
30. Tsumoto, S., *Chance Discovery in Medicine - Detection of Rare Risky Events in Chronic Diseases*. *New Generation Computing*, 2003. **21**(2): p. 135-.
31. van der Aalst, W.M.P. and A.K.A. de Medeiros. *Process Mining and Security: Detecting Anomalous Process Executions and Checking Process Conformance*. in *2nd International Workshop in Security Issues with Petri Nets and other Computational Models*. 2004. Bologna, Italy.

32. Weiss, G.M., *Mining with Rare Cases*, in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Editors. 2005, Springer.
33. Witten, I.H. and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. 2005, San Francisco: Morgan Kaufmann.
34. Woodard, M., *Digital Dig - Data Mining in Astronomy*, WWW Document, (last accessed 25th January, 2005), 2002.
35. Xu, D. and P. Ning. *Alert Correlation through Triggering Events and Common Resources*. in *20th Annual Computer Security Applications Conference (ACSAC 2004)*. 2004. Tucson, AZ, USA: IEEE Computer Society.
36. Zissman, M., *1999 DARPA Intrusion Detection Evaluation Data Set*, WWW Document, http://www.ll.mit.edu/IST/ideval/data/1999/1999_data_index.html (last accessed 20th September, 2005), 1999.