Data Mining Tool for Intrusion Detection

GRADUATE PROJECT

Submitted to the Faculty of
The Department of Computing and Mathematical Sciences
Texas A&M University-Corpus Christi
Corpus Christi, Texas

In Partial Fulfillment of the Requirements for the Degree of
Master of Science in computer Science

by

Anitha Nalluri
Summer 2004

Committee Members

Dr. Dulal C. Kar
Committee Chairperson

Dr. Long-zhuang Li
Committee Member

Ms. Charlotte Busch
Committee Member
Data Mining Tool for Intrusion Detection

GRADUATE PROJECT

Submitted to the Faculty of
The Department of Computing and Mathematical Sciences
Texas A&M University-Corpus Christi
Corpus Christi, Texas

In Partial Fulfillment of the Requirements for the Degree of
Master of Science in computer Science

by

Anitha Nalluri
Summer 2004

Committee Members

Dr. Dulal C. Kar
Committee Chairperson

Dr. Long-zhuang Li
Committee Member

Ms. Charlotte Busch
Committee Member
# TABLE OF CONTENTS

Abstract......................................................................................................................ii

Table of Contents......................................................................................................iii

List of figures.............................................................................................................vi

1. Background and Rationale.....................................................................................1
   1.1. Introduction.....................................................................................................1
   1.2. Intrusion Detection.........................................................................................1
   1.3. Intrusion Detection Techniques......................................................................2
   1.4. Data Mining....................................................................................................3
   1.5. Background and Rationale ............................................................................4

2. Data Mining Tool for Intrusion Detection............................................................6
   2.1. Data Mining.....................................................................................................6
   2.2. Data Collection................................................................................................6
       2.2.1. Snort.................................................................................................6
       2.2.2. Data storage.......................................................................................7
   2.3. The Data Mining Tool....................................................................................7
   2.4. User Interface Design Description.................................................................8
       2.4.1. Login Page..........................................................................................9
       2.4.2. Main Page..........................................................................................10
       2.4.3. Clustering Page..................................................................................12
       2.4.4. Sequence Analysis Page......................................................................13
       2.4.5. Visualization Page..............................................................................15
       2.4.6. Drill down Page..................................................................................17
2.4.7. Sensors Page. .................................................................................. 19
2.4.8. Archives Page. .................................................................................. 20
2.4.9. Backup Page. .................................................................................... 22
2.4.10. Logout Page. .................................................................................... 22

3. Project Environment and Proposed System Design. .............................. 23
   3.1. Main Components. ........................................................................... 23
   3.2. System Requirements. ..................................................................... 23
      3.2.1. MySQL. .................................................................................... 24
      3.2.2. PHP. ........................................................................................ 24
   3.3. The Design of the Tool. ................................................................... 25
      3.3.1. Level 0 Data Flow Diagram. ....................................................... 25
      3.3.2. Level One Data Flow Diagrams. .............................................. 26
   3.4. Database Design. ............................................................................. 30
      3.4.1. Description of Database Tables. .................................................. 30
      3.4.2. Design and Entity Relationship. ................................................. 34
   3.5. Querying. ........................................................................................ 35

4. Implementation. .................................................................................... 37
   4.1. Login Module. .................................................................................. 37
   4.2. Sensors Module. .............................................................................. 38
   4.3. Alerts Module. .................................................................................. 38
   4.4. Clustering Module. .......................................................................... 39
   4.5. Sequence Analysis Module. ............................................................. 39
   4.6. Drill down Module. ......................................................................... 40
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7. Visualization Module</td>
<td>40</td>
</tr>
<tr>
<td>4.8. Backup Module</td>
<td>41</td>
</tr>
<tr>
<td>4.9. Archives Module</td>
<td>41</td>
</tr>
<tr>
<td>4.10. Logout Module</td>
<td>41</td>
</tr>
<tr>
<td>5. Testing and Evaluation</td>
<td>42</td>
</tr>
<tr>
<td>5.1. Software Testing</td>
<td>42</td>
</tr>
<tr>
<td>5.2. Usability Testing</td>
<td>43</td>
</tr>
<tr>
<td>6. Results and Conclusion</td>
<td>45</td>
</tr>
<tr>
<td>6.1. Results</td>
<td>45</td>
</tr>
<tr>
<td>6.2. Conclusion</td>
<td>45</td>
</tr>
<tr>
<td>7. Future Work</td>
<td>46</td>
</tr>
<tr>
<td>8. Acknowledgements</td>
<td>47</td>
</tr>
<tr>
<td>9. Bibliography and References</td>
<td>48</td>
</tr>
<tr>
<td>10. Appendix</td>
<td>50</td>
</tr>
</tbody>
</table>
ABSTRACT

Intrusion detection is the art of detecting and responding to resource misuse. An Intrusion detection system (IDS) assists in managing threats and vulnerabilities in the changing environment of the network. Data mining methods have been used to build automatic intrusion detection systems based on anomaly detection. Data mining attempts to extract implicit, previously unknown and potentially useful information from data. A Data mining based IDS helps in alleviating the problem of automatically detecting anomalous patterns in large volumes of audit data.

A Web-based data mining tool to analyze intrusions is developed. This data mining tool describes the behavioral forensics in intrusion detection. This tool finds anomalous activity that uncovers a real attack process and identifies long and ongoing patterns. This tool analyzes host based traffic features, time based traffic features, protocol based traffic features, and the associated intrusions. With the help of this tool rules can be generated to capture the behavior of the intrusions and normal activity.
1. BACKGROUND AND RATIONALE

1.1 Introduction

With the evolution of the Internet over the last few years, the need for security has been rising with it mainly due to the openness and connectivity nature of the Web. People and organizations are faced with more challenges every day to secure their data and all other assets of value to them. Prevention, detection, and response are part of the Network Defense model. Intrusion detection systems are important components of defensive measures protecting computer networks from abuse [Proctor 2000]. In this project a data mining tool has been developed that describes the behavioral forensics in intrusion detection.

1.2 Intrusion Detection

An intrusion detection system (IDS) attempts to detect an intruder breaking into the system or a legitimate user misusing system resources. Network based intrusion detection system and host based intrusion detection system are two primary intrusion detection models. A network intrusion detection system (NIDS) monitors traffic on the network wire and attempts to discover if a hacker/cracker is attempting to break into a system or cause a denial of service attack [IDS 2003]. A host based intrusion detection system audits data from a single host to detect intrusion.

Functions of NIDS include monitoring and analysis of network traffic, recognition of activity patterns and statistical analysis for abnormal activity patterns and generating security alerts.
The structure and components of NIDS are illustrated in Figure 1.1.

![Diagram of NIDS components](image)

**Figure 1.1 Components of IDS**

The network provides the data to the rules engine and the activity normalizer. The rules engine searches the data for patterns from the known malicious activity database. The activity normalizer performs analysis (usually statistical) of the data. The sensor is a network packet capturer. [Proctor 2000]

### 1.3 Intrusion Detection Techniques

Intrusion detection techniques can be categorized into misuse detection and anomaly detection, based on the strategy for detection. Misuse detection finds intrusions
by looking for activity corresponding to known techniques for intrusions [IDS 2003]. This involves the monitoring of network traffic in search of direct matches to known patterns of attack called signatures. This is a rule based approach. The disadvantage of this approach is that it can only detect intrusions that follow pre-defined patterns [IDS 2003]. In anomaly detection, the system defines the expected behavior of the network or profile in advance. Any significant deviations from this expected behavior are then reported as possible attacks. Such deviations are not necessarily attacks.

Intrusion detection systems based on misuse detection lack the ability to detect attacks that do not fit a pre-defined signature. Those shortcomings can be overcome by using anomaly detection. It has potential ability to recognize unforeseen attacks. A critical issue for anomaly detection is the need to reduce false alarms, since any activity outside a known profile raises an alarm [IDS 2003]. Data mining is a major approach to anomaly detection. This helps in identifying the anomalous pattern from the audit data.

1.4 Data Mining

Data mining attempts to extract implicit, previously unknown and potentially useful information from data. Data mining of intrusion detection involves processing large quantities of collected data to look for patterns of interest.

Data mining normally excels at observing general regularities. This feature is used in anomaly detection in network data at a relatively high level and the delegation of suspected problems to dedicated applications. Data mining analysis determines the bounds for normal network activity [Bloedorn 2003]. Data mining techniques determine which characteristics of previously identified attack activity distinguish the pattern from
normal network usage [DSTO 2003]. Following are the data mining techniques used with this tool:

- Clustering
- Sequence Analysis
- Visualization
- Drill down

1.5 Background and Rationale

Network security is the need of the hour. Intrusion detection is a critical component of secure information systems. An intrusion detection system generates lots of alarms, out of which many are false positives and false negatives. False positive alarms are issued when normal behaviors are incorrectly identified as abnormal and false negative alarms are issued when abnormal behaviors are incorrectly identified as normal [Hossain 2003].

On a regular basis, security analysts sift through this data looking for the most serious attacks. There can be thousands of suspicious activity alarms and each requires further analysis to fully understand its purpose.

Network management requires analyzing data to identify true alerts and to respond to them. This can be made easy by having a data mining tool that analyzes audit trails to spot abnormal patterns of usage.

This tool offers a common platform to analyze host based traffic features, time based traffic features, protocol based traffic features, port based traffic features, and the associated intrusions. This tool also offers visual and static reports. The rules mined form audit data are merged and added into an aggregate rule set to detect intrusion. The tool is
a Web-based application, widely accessible across the Internet. This provides flexibility by allowing remote monitoring of intrusions.

This document is organized into seven chapters to describe the project. Chapter one gives a brief background and overview of intrusion detection and describes the need for the development of this tool. Chapter two describes the basic modules and the graphical user interface of this tool. Chapter three describes the software design with data flow diagrams. Chapter four describes the implementation details of the modules in this tool. Chapter five describes the usability testing and evaluation of the Web-based data mining tool. Chapter six brings out the results of this project. Chapter seven describes the scope of future work.
2. DATA MINING TOOL FOR INTRUSION DETECTION

2.1 Data Mining

As previously stated, data mining has been defined as the nontrivial extraction of implicit, previously unknown and potentially useful information from data [Frawley 1992]. In the context of intrusion detection, data mining is defined as processing large quantities of data collected at a central location to look for patterns of interest. The data exists in different forms and formats. The goal of mining is to derive multi-feature correlations from audit data [Bloedorn 2003].

2.2 Data Collection

Data collection is an important part of this tool. Without the audit data this tool is of no significance. Snort is used to collect the audit data [Snort 2003].

2.2.1 Snort

Snort is an intrusion detection system. Snort is a cross-platform, lightweight network intrusion detection tool that can be deployed to monitor small TCP/IP networks and detect a wide variety of suspicious network traffic as well as outright attacks [Snort 2003]. By lightweight, it means it is suitable for small TCP/IP networks. There are other intrusion detection systems for larger networks. Snort is appropriate for a network in a school or a small business. It can provide administrators with enough data to make informed decisions on the proper course of action in the face of suspicious activity. Snort can also be deployed rapidly to fill the holes of security in a network.

Snort uses a flexible rules language to describe traffic that it should collect or pass, as well as a detection engine that utilizes modular plug-in architecture. Snort provides with the facility to log the audit data directly into a database. With the help of
this feature and by running Snort in intrusion detection mode, required audit data is collected.

2.2.2 Data storage

Data collected through snort is stored in a database for further data mining. Snort is operated in network intrusion detection mode. In this mode, it generates alerts and captures network packets. These alerts and other network traffic form the audit data. This raw data is stored in the MySQL database for further mining. Figure 2.1 illustrates the collection of the audit data.

Figure 2.1 Flow of Data Collection

2.3 The data mining tool

This data mining tool is a Web-based tool that analyses the audit data for intrusion detection. The tool uses data mining techniques like clustering, sequence analysis, drill down, and visualization. The tool helps the network administrator in differentiating actual intrusions from false intrusions. The rules mined from audit data are merged and added into an aggregate rule set to detect intrusion.
Intrusions are clustered based on the host name, protocol type, and some known intrusion signatures. This tool also provides time based sequence analysis. The tool broadens the analysis by narrowing the search through the audit data.

2.4 User Interface Design Description

The user interface primarily consists of eleven Web pages, which are referred as page(s) in this document. A user authentication is required to use this tool. The user is provided with navigation information from the second page (Main Page) onwards to any page of ones choice at any given time. Figure 2.2 depicts the interface design of the system. The Login Page validates the user. Upon successful validation, the user is led to Main Page. If the user validation fails, the user stays in the Login Page only.

From the Main Page, the user has the choice to choose any of the other pages. After validation, the user has an option to navigate to other pages from any page. Once, the user exits the system, he or she is taken to the Logout Page. Details about the tool and the glossary of terms used are explained in the Help Page. This page is accessible from any other page.
2.4.1 Login Page

The first page authenticates the user to use the data mining tool. Two specific input boxes to enter the username and the password are provided. Open Secure Socket Layer (OpenSSL) is used to enforce security during the exchange of sensitive information. Figure 2.3 depicts a screen shot of the Login Page.
2.4.2 Main Page

After successful authentication, the user is led to the Main page. This page displays brief information about the sensors deployed in the system. A user can navigate to any page of his or her choice from the options provided on the left hand navigation bar of the page or from the options on the tool bar at the top. These options include Clustering, Sequence Analysis, Visualization, Drill down, Alerts, Archives, Sensors, Help, and Exit. Based on the choice, the user is redirected to one of the following pages:

- *Clustering Page*

- *Sequence Analysis Page*
• Visualization Page
• Drill down Page
• Alerts Page
• Sensors Page
• Backup Page
• Archives Page
• Help Page
• Exit Page

This page also provides the links to navigate through out the data mining tool.

Figure 2.4 depicts a screen shot of the Main Page.

Figure 2.4 A Screen Shot of Main Page
2.4.3 Clustering Page

When the user selects the **Clustering** option in the *Main page* or the *Clustering* link in any other pages, the user reaches this page. Here the user is asked to choose the type of clustering he or she is interested in. The user can select clustering based on source host or destination host or protocols or alerts. Figure 2.5 depicts a screen shot of the *Clustering Page*.

![Data mining Tool for Intrusion Detection](image)

**Figure 2.5 A Screen Shot of Clustering Page**

After the user makes a choice, results are displayed. Details of the results are displayed by clicking on the view link. The results page displays the clustered data based on the user’s choice. Figure 2.6 depicts a screen shot of the results of *Clustering Page*. 
2.4.4 Sequence Analysis Page

When the user selects the **Sequence Analysis** option in the *Main page* or the *Sequence Analysis* link in any other page, the user reaches this page. Here the user is asked to make a choice between time-span analysis or analysis particular to a day. Figure 2.7 depicts a screen shot of the *Sequence Analysis Page*. 
Figure 2.7 A Screen Shot of results of Sequence Analysis Page

The user is prompted to enter date or date and time based on the choice made.

Figure 2.8 depicts a screen shot of the Sequence Analysis Page.
2.4.5 Visualization Page

When the user selects the Visualization option in the Main page or the Visualization link in any other pages, the user reaches this page. This page provides the options to choose the choice of graphical representation such as bar graphs, pie charts. The user has to select the option of Source Host or Destination Host or Protocol or Alerts, based on which audit data is visualized. Figure 2.9 depicts a screen shot of the Visualization Page.
After the user makes the choice, results are displayed graphically. Figure 2.10 depicts a screen shot of the results of Visualization Page.
2.4.6 Drill down Page

When the user selects the **Drill down** option in the *Main page* or the *Drill down* link in any other page, the user reaches this page. With this option, deeper analysis is performed by narrowing down the search. Figure 2.11 depicts a screen shot of the *Drill down Page*. 

**Figure 2.10 A Screen Shot of results of Visualization Page**
Figure 2.11 A Screen Shot of Drill down Page

On selecting the option, results are displayed and the user can select a choice which narrows down the search or analysis by providing a search within a search option to the user. This takes the user through several pages similar to Figure 2.12.
Figure 2.12 A Screen Shot of results of Drill down Page

2.4.7 Sensors Page

When the user selects the Sensors option in the Main page or the sensor link in any other page, the user reaches this page. This page displays lists the sensors deployed in this system. It also displays sensor hostname, interface details, and other information about those sensors. Figure 2.13 depicts a screen shot of the Sensors Page.
2.4.8 Archives Page

When the user selects the Archives option in the Main page or the Archives link in any other page, the user reaches this page. This page provides with an option to archive the analyzed data and to view old data from the archives database. Figure 2.14 depicts a screen shot of the Archives Page.
Archiving shrinks the database and this increases the performance. It also saves time by avoiding the already analyzed data. Figure 2.15 depicts a screen shot of adding to archives.
2.4.9 Backup Page

When the user selects the Backup option in the Main page or the Backup link in any other page, the user reaches this page. This page provides with an option to take a backup of the entire database. For security purpose, the user is prompted for password. On authentication, the database is backed up. Figure 2.16 depicts a screen shot of the Backup Page.

![Backup Page Screenshot](image)

**Figure 2.16 A Screen Shot of Backup Page**

2.4.10 Logout Page

When the user selects the Exit option in the Main page or the Main link in any other page, the user reaches this page. This page logs out the user and the user is no longer permitted to access other pages. The user has to login again to use the system.
3. PROJECT ENVIRONMENT AND PROPOSED SYSTEM DESIGN

3.1 Main Components

The following are the main components of the proposed data mining tool:

- Data Collecting tool - snort
- Database – to store audit data
- Client (Web pages) – to display the analysis

Snort, an intrusion detection system for small networks is used to collect data. Snort is an open source network intrusion detection system. Snort can be set up in various modes. Snort is set up in network intrusion detection mode to log the required audit data. This data is stored to a database for further analysis.

The data collected using snort is stored in a MySQL database. Snort supports logging the audit data directly into a database. Through this support, audit data is stored in a MySQL database.

The tool is a Web-based tool. The results of the analysis are displayed in a Web browser. Any Web browser such as Internet Explorer or Netscape Navigator serves the purpose of the client. This also provides compatibility among various types of machines.

3.2 System Requirements

This system is designed to run on any web browser such as Internet Explorer and Netscape. MySQL database system running under the Linux operating system is used as the back end. The Hypertext Markup Language (HTML) is used to generate all front-end documents. JavaScript is used to validate user input on the Web browser at client’s end. PHP Hypertext preprocessor (PHP), a server side scripting language is used to access the MySQL database. For visualization, JpGraph, an object oriented library for PHP is used.
Structured Query Language (SQL) is used for querying the database. Snort, an open source and freely available intrusion detection system is used.

3.2.1 MySQL

MySQL is a very fast, robust, relational database management system (RDBMS). MySQL is the most popular Open Source SQL database. It is developed and provided by MySQL AB [MySQL 2002]. A database enables us to efficiently store, search, sort, and retrieve data. The MySQL server controls access to the data to ensure that multiple users can work with it concurrently, provides fast access to data, and ensures that only authorized users can obtain access. MySQL is a multi-user, multi-threaded server. In addition to supporting standard Structured Query Language (SQL), it compiles on a number of platforms and has multithreading abilities on UNIX, which is good for great performance.

3.2.2 PHP

PHP is a server-side scripting language [Merral 2003]. PHP is a general-purpose scripting language that is used for Web development and can be easily embedded into HTML. PHP is mainly used in server side scripting like collecting the information from the user, generating dynamic pages, and sending and receiving cookies.

PHP provides the luxury of procedural programming, object-oriented programming, and a mixture of both to the user. When using PHP the user is not limited to output only in HTML but also has capabilities like outputting images, PDF files, and even Flash movies and text in XHTML and XML [Weiling 2001].
3.3 The Design of the Tool

The tool has a client and a database serving as a data repository. This tool is secured with login and cookies.

A cookie is a small file that the server embeds on the user's computer. A cookie typically contains the name of the domain from which the cookie has come, the "lifetime" of the cookie, and a value, usually a randomly generated unique number. This Web-based data mining tool uses session cookies to make the system more secure and allows a user to carry information across pages of the system.

This tool queries the database to perform the data mining techniques. The functionality and design of the system is represented using data flow diagrams in this section. Data Flow Diagrams give a picture of flow of data among the processes in the system [Pressman 1996]. All the data techniques query the database and display those details on the Web browser. The following sections describe level 0 and level 1 Data Flow Diagrams (DFD).

3.3.1 Level 0 Data Flow Diagram

Level 0 DFD or context level diagram depicts the overall data flow in the system as a whole. Figure 3.1 shows the flow of data in the system at a higher level.

![Figure 3.1 Context Level Diagram of the System](image)
3.3.2 Level One Data Flow Diagrams

Level One Data Flow Diagram (DFD) explains the flow of data in the system in detail, compared to the context level DFD. Here all the major processes of the system, their inputs and outputs are depicted.

Login Page

The login page provides access to the tool. This page accepts the username and password. These are passed to Validate User process as an input. Here the username and password are validated against the Users table from the database. Once the user is authenticated, the user is redirected to the Main Page. If the user validation fails, the user is redirected to the Login Page. Figure 3.2 shows the level one Data Flow Diagram for this page.

Figure 3.2 Level One Data Flow Diagram for the Login Page

Main Page

In the Main Page, brief information about the sensors and alerts are displayed. User input is obtained and sent as an input to redirect option process. After processing the user request, this process takes the user to the requested page. Figure 3.3 shows the level one Data Flow Diagram for this page.
Clustering Page

In this page, the user provides the choice of clustering as the input. This is passed to Clustering Data module. This, in turn, interacts with the database to get the data and clusters the data. After clustering, the results are displayed on the output screen. Figure 3.4 shows the level one Data Flow Diagram for this page.

Sequence Analysis Page

In this page, the user provides the time during which the sequence analysis has to be performed as input. This is passed to the Analyze Data module. This, in turn, interacts with the database to get the data and analyzes that data. After the analysis, the results are displayed on the output screen. Figure 3.5 shows the level one Data Flow Diagram for this page.
Visualization Page

The Visualization Page gets the type of display from the user. This is passed as input to the Visualize Data module. This module in turn interacts with the database to get the data and visualize that data. After that, results are displayed on the output screen.

Figure 3.5 Level One Data Flow Diagram for the Sequence Analysis Page

Drill down Page

In this page, data analysis is made more specific by narrowing down the search. The user input is obtained by the Get Input module. This module gets the input from the user. Then this is passed to the Drill down module which performs the specific analysis. The user can further narrow the choice at any point. When the results are displayed, user
can input choice to further narrow down the analysis. The results are displayed on the output screen. Figure 3.7 shows the level one Data Flow Diagram for this page.

![Figure 3.7 Level One Data Flow Diagram for the Drill down Page](image)

**Backup Page**

In this page, the user is provided with an option to backup the database. For security purpose, the user is prompted for a password. On authentication, the database is backed up. Figure 3.8 shows the level one Data Flow Diagram for this page.

![Figure 3.8 Level One Data Flow Diagram for the Backup Page](image)
Logout Page

When the user selects to exit, user is logged out of the system. This user can no longer access any of the pages of the tool. Figure 3.9 shows the level one Data Flow Diagram for this page.

![Figure 3.9 Level One Data Flow Diagram for the Exit Page](image)

Alerts Page

In this page, the user can view the information of known intrusion signatures or alerts.

Sensors Page

In this page, the user can view the detailed information of the sensors deployed in the system.

3.4 Database Design

3.4.1 Description of Database Tables

The database for this tool is in MySQL. Database contains tables that store the audit data and the users’ information who use this system. This subsection gives a brief description of the tables and fields in the database and their relationships. All the tables are normalized to avoid redundancy. The following is the list of tables in the database.

- Data
- Details
- Encoding
- Event
- Flags
- Icmphdr
- Iphdr
- Opt
- Protocols
- Schema
- Sensor
- Services
- Signature
- Tcphdr
- Udphdr
- Users

Following is a brief description of the above tables and the prominent fields in them. For complete description of the fields, see Appendix A.

**Data:**

This table consists of **sid**, **cid**, and **data_payload** fields. Sid and cid identify the audit data and data_payload contains the payload part of that network packet.

**Detail:**

Detail is a lookup table. It consists of **detail_type**, and **detail_text** fields. They identify the type of logging of the network packets.
**Encoding:**

Encoding is a lookup table. It consists of `encoding_type`, and `encoding_text` fields. They identify the type of encoding of the network packets.

**Event:**

Event table consists of `sid`, `cid`, `signature`, and `timestamp` as fields. Sid and cid identify the network packet. Signature identifies the event and timestamp is the time when the event has occurred.

**Flags:**

This is a lookup table. This table consists of `urg`, `ack`, and other fields that indicate the TCP (Transmission Control Protocol) flag list [Protocols 2003].

**Icmp hdr:**

This table consists of `sid`, `cid`, `icmp_type`, `icmp_code`, and other ICMP (Internet Control Message Protocol) fields. Sid and cid identify the network packet. Other fields identify the icmp fields associated with that packet [Protocols 2003].

**Iphdr:**

This table consists of `sid`, `cid`, `ip_src`, `ip_dst`, `ip_proto`, `ip_ttl`, and other IP (Internet Protocol) fields [Protocols 2003]. Sid and cid identify the network packet. Other fields identify the source IP address, destination IP address, protocol, time to live, and other IP fields respectively that are associated with that packet.

**Protocols:**

This table is a lookup table that contains the protocol list of Layer-4. It consists of `protocol`, `name`, and `description`. 
**Schema:**

This table consists of self-documented information about the database.

**Sensor:**

This table contains sensor details with **sid**, **hostname**, **interface**, and **encoding** fields. Sid and hostname identify the sensor id and the hostname where the sensor is deployed respectively.

**Services:**

This table is a lookup table with details of the TCP and UDP (User Datagram Protocol) services list. It consists of **port**, **protocol**, **name**, and **description** fields.

**Signature:**

This table contains the normalized listing of alert or signature names, ids and priorities. It consists of **sig_id**, **sig_name**, and **sig_priority** fields.

**Tcp hdr:**

This table consists of **sid**, **cid**, **tcp_sport**, **tcp_dport** and other TCP fields. Sid and cid identify the network packet [Protocols 2003]. Other fields identify the source port number, destination port number, and other tcp fields respectively that are associated with that packet.

**Udphdr:**

This table consists of **sid**, **cid**, **udp_sport**, and **udp_dport**. Sid and cid identify the network packet. Other fields identify the source port number, and destination port number associated with that packet.
Users:

The table consists of fields like username and password. Users present in this table are only permitted to access the system.

3.4.2 Design and Entity Relationship

The database for this tool is in MySQL. Database contains tables that store the audit data and user information. All the tables in this database are normalized to avoid redundancy. Figure 3.10 shows the connections of database tables:
3.5 Querying

The major portion of this tool lies in querying the database and analyzing the obtained results. Querying is a major portion of data mining. Structured Query Language (SQL) is used to query the database. SQL comprises one of the fundamental building
blocks of modern database architecture. SQL defines the methods used to create and manipulate relational databases on all major platforms [Sql 2002]. Queries are written based on the data mining technique and the type of analysis chosen by the user.

Performance of the tool is controlled to a great extent by the efficiency of the queries. Optimized queries are written in SQL to increase the performance of the tool.
4. IMPLEMENTATION

This chapter narrates the project implementation details. Entire project is divided into ten main modules. Each module is sub-divided into several units. Following is the list of modules:

- Login Module
- Sensors Module
- Alerts Module
- Clustering Module
- Sequence Analysis Module
- Drill down Module
- Visualization Module
- Backup Module
- Archives Module
- Logout Module

4.1 Login Module

Login module is the first module which provides the access to this system. login.php is the file that gets the username and password as input from the user. These details are passed as input to login_action.php, a PHP script that connects to the database to verify the authenticity of the user. On successful authentication, the user is redirected to the Main Page. A session cookie is also set here. If authentication fails, a login failure message is displayed and the user is redirected to the Login Page. Following is the code snippet to verify user’s authenticity:
<?
include "connection.php";
$table = "users";
$username = $_POST["username"];
$password = $_POST["password"];
$query = "select * from $table where username = "$username";"
$result = mysql_query ($query);

if( mysql_num_rows($result) != 0 )
{
    $row = mysql_fetch_array($result);
    $temp = $row["password"];
    if($password == $temp)
    {
        setcookie ("user", md5($_POST[userid]),time()+6000);
        header("Location:main.php");
    }
    else
    {
        $flag = 1;
    }
}
else
{
    $flag = 2;
}
?>

4.2 Sensors Module

This module displays the detailed information of all the sensors deployed in the system. The PHP script sensors.php accesses the database and retrieves the information about all the sensors present in the system. This script obtains the list of sids (unique sensor id associated with every sensor) present in the system. Using the sid, detailed information about the sensor is retrieved from the database.

4.3 Alerts Module

This module displays the detailed information of all the alerts present in the system. This list includes the universal known intrusion signatures as well as new alerts
recognized by this tool. The script `alerts.php` displays the details. This script connects to
the database and queries the database for the list of alert signatures and their details.

4.4 Clustering Module

Clustering module is where clustering, one of the data mining techniques used in
this tool is implemented. The script `clustering.php` gets the choice of clustering the user is
interested in as the input. This input is passed to `clustering_action.php` script which gets
the information of the intrusions according to the given choice from the database. This
script displays the results. The script `view.php` displays the detailed information of the
records.

4.5 Sequence Analysis Module

In this module, the sequence analysis technique is implemented. The script `seq-anal.php`
script displays the choice of analysis to the user. The user has to select time span or day-
based analysis. Once the user selects one of them, that choice is sent as input to
`seq_anal_action.php` file. This file queries the database and displays the information
according to the user’s choice. Figure 4.1 is a flow chart describing the data flow in this
module.
4.6 Drill down Module

In this module, the drill-down data mining technique is implemented. The `drilldown.php` script displays the choices available to the user. The user has to select among the source host, destination host, protocol, and alert. Once the user selects one of them, that choice is sent as input to `drill1.php` file. This file queries the database and displays the information according to the user’s choice. Here the user can further narrow down the choice by selecting the links provided in this page. That choice is passed as input to `drill2.php` file. This file obtains the information from the database and displays it to the user.

4.7 Visualization Module

This module implements the visualization technique of data mining. The `visualization.php` file prompts the user to make choice of visual option. After this choice is made, the user is prompted to choose among the protocol, source host, destination host,
and alerts. Both these choices are passed as input to the `visualization_action.php` script. This file queries the database and obtains the information. `visualization_action.php` displays the results graphically using JpGraph, an object-oriented library for PHP.

### 4.8 Backup Module

This module helps in taking the backup of the system. The `backup.php` is the file that prompts the user for username and password as a security measure. This password is password as input to the `backup_action.php` file. This file matches the user’s password against the database and authenticates the user. On successful validation, this file takes a backup of the database with the last known stable information. Backup is done to a file that is named as the current date.

### 4.9 Archives Module

This module displays the list of archives and also provides with an option to create an archive. The `Archives.php` script provides the choice to view or create archive. This page leads to `add-archives.php` or `view-archives.php` based on the choice made. If the choice is to create a new archive, the `add-archives.php` script prompts the user to enter the date to archive the information pertaining to that date. If the choice is to view archives, the `view-archives.php` script lists the available archives.

### 4.10 Logout Module

This module logs out the user from the system. The `logout.php` file has the PHP script to logout the user from the system. Once the user selects exit option, all the cookies are unset and the user is forbidden to access the system. The user is permitted to use the system with a login again. This module redirects the user to the login page.
5. TESTING AND EVALUATION

The data mining tool has undergone both Software Testing and Usability Testing. Being a Web-based, the Web pages are tested for proper usability.

5.1 Software Testing

Software testing involved the functional performance check. It has been tested to check if the tool performs all the functions of classification, clustering, visualization, sequence analysis, and drill down. The system is tested while it is being developed. The project is divided into modules and each module, in turn, is divided into many units. Several distinct test cases are evolved for each unit as well as for each module and finally for the entire tool. Both black box and white box testing are adopted. A coding standard is adopted and proper documentation of the code is followed for ease of inspection and maintenance.

Entire system is divided into modules. Independent and integrated testing are performed at every stage of development. Each unit is developed and tested simultaneously. Once the units of a module are tested independently, the units are combined to form the module. Each module is tested independently for data integrity, data consistency, consistency of data input and output formats and for other features. Each module is tested individually to check if it has achieved the objectives of the set of requirements.

Once all the modules are developed, the modules are combined and tested. Progressive process is used to combine the modules. First two modules are combined and tested with several cases. The next module is added to the first two and tested, and so on.
After all the modules are integrated and produce the desired results and comply with the requirements, the entire tool is tested with several test cases.

5.2 Usability Testing

The Web pages are tested and evaluated. During the initial stages, the developer tested the Web pages. Later after the tool is completely developed, the primary user, a network manager is asked to evaluate the usability. Suggestions from the network manager have been incorporated for better usability of the system. Dr. John Fernandez, a usability expert, performed a heuristic evaluation, against the general guidelines [Rosson 2000] listed below:

a. Use simple and natural dialog
b. Speak the user’s language
c. Minimize the memory load
d. Be consistent
e. Provide feedback
f. Provide clearly marked exits
g. Provide short cuts
h. Provide good error messages
i. Prevent errors
j. Include good help and documentation

Dr. Fernandez suggested few changes to the user interface which were adopted to improve the usability of the interface.
The Web pages are tested on different platforms - IBM compatible PC and Macintosh machine. The Web pages were also tested using different kinds of browsers such as Netscape Navigator, Internet Explorer, and Mozilla.

The Web pages are tested in the CI Labs of Texas A&M University, Corpus Christi. The usability of the Web site is evaluated for the following:

a. Ease of finding information
b. Legibility or ease of reading the content.
c. Completeness with which site’s subject is treated
d. Appearance of the site
e. Consistency in Layout, Menus, and Visual Cues
f. Help on how to use the site
6. RESULTS AND CONCLUSION

6.1 Results

The primary objective of building a Web-based Data mining Tool is achieved. A tool that is robust and has sufficiently good performance is developed. The tool has been tested and found satisfactory by the Network Manager at the Department of Computing and Mathematical Sciences. The observations and rules mined from audit data with the help of this tool are merged and added into an aggregate rule set to detect intrusion.

6.2 Conclusion

A data mining tool for intrusion detection that is robust and has sufficiently good performance is developed. This tool helps the network administrator to identify anomalies in the network traffic, create rules to detect intrusion based on the analysis and to distinguish false alarms from positive alarms.

This tool provides a better insight into differentiating false intrusions from the real ones. It also provides with more rules to perform better intrusion detection. The observations and rules mined from audit data with the help of this tool are merged and added into an aggregate rule set to detect intrusion. This makes the existing intrusion detection system more robust. There is a lot of scope for future work; most importantly, other data mining techniques could be added to further extend and enhance the system.
7. FUTURE WORK

There is a scope for future work with this project. This tool can be upgraded with additional features. Some of them are listed here:

1. Include services while clustering the data. A source host is vulnerable to attacks because of the services running on it. So it will be a good idea to cluster the data based on the services, a host is running.

2. Another data mining technique, classification, can be used to classify the intrusion into one of the known intrusion categories.

3. This tool could be developed as a stand alone application by using Visual Basic to provide a better user interface.

Right now this tool collects audit data from a particular network and then analyzes that data. This tool can be generalized by providing an option to upload data collected over a different network. The uploaded data can be analyzed remotely.
ACKNOWLEDGEMENTS

I would like to render my sincere thanks to Dr. Dulal Chandra Kar, my Project Committee Chair for his constant support, valuable insights and guidance through out the project. I sincerely thank Dr. Long-zhuang Li and Ms. Charlotte Busch for being on my project committee.

I am also thankful to Dr. Fernandez for sparing time to evaluate this tool at a short notice and for his feedback about the usability of the tool. I thank Mr. Abhijeet Trivedi for providing access to the Linux server in Cl342. I would like to thank all the faculty and staff at CAMS for their support.

Above all, I thank the Almighty for giving me the strength and will to finish the undertaken task, successfully.
8. BIBLIOGRAPHY AND REFERENCES


APPENDIX

DATA DICTIONARY

Tables used in the database of this project are as follows:

Data

This table contains the content of packet payload of the audit data with the following fields:

- sid: Sensor ID, a unique number identifying a sensor
- cid: Event ID, a unique number identifying an event
- data_payload: Packet payload encoded according to sensor encoding

Detail

This table contains the level of detail with which a sensor is logging with the following fields:

- detail_type: a unique number associated with the type of logging
- detail_text: Description of the type of logging

Encoding

This is a lookup table that contains the details of type of encoding used for the packet payload. This table has the following fields:

- encoding_type: a unique number associated with the type of encoding
- encoding_text: Description of the type of encoding

Event

This table contains the details of meta-data about the detected alert or the audit data with the following fields:

- sid: Sensor ID, a unique number identifying a sensor
• cid: Event ID, a unique number identifying an event
• signature: Signature ID
• timestamp: Timestamp of when the event was logged

**Flags**

This is a lookup table of TCP flag list. This table has the following fields:

• number: Value of the number when the flags are set
• RES1: Reserved bits
• RES2: Reserved bits
• URG: Flag for urgent data
• ACK: Flag to acknowledge data
• PSH: Flag to deliver data to application
• RST: Flag to drop connection
• SYN: Flag for a new connection
• FIN: Flag to close connection
• valid: Flag to indicate the validity
• description: Description of the packet based on the flags

**IcmpHdr**

This table contains the ICMP protocol fields associated with a network packet.

This table has the following fields:

• sid: Sensor ID, a unique number identifying a sensor
• cid: Event ID, a unique number identifying an event
• icmp_type: ICMP type
• icmp_code: ICMP code
• icmp_csum: ICMP checksum
• icmp_id: ICMP ID
• icmp_seq: ICMP sequence number

Iphdr

This table contains the IP protocol fields associated with a network packet. This table has the following fields:
• sid: Sensor ID, a unique number identifying a sensor
• cid: Event ID, a unique number identifying an event
• ip_src: Source IP Address
• ip_dst: Destination IP Address
• ip_ver: IP version
• ip_hlen: IP header length
• ip_tos: IP type-of-service
• ip_len: IP datagram length
• ip_id: IP ID
• ip_flags: IP flags
• ip_off: IP fragment offset
• ip_ttl: IP time-to-live
• ip_proto: IP protocol
• ip_csum: IP checksum

Protocols

This is a lookup table with an IP encoded protocol list with the following fields:
• protocol: Unique number associated with a protocol
• name: Name of the protocol
• description: Description of the protocol

Schema

This table contains self-documented information about the database. This table has the following fields:

• vseq: Database schema ID number
• ctime: Timestamp of database creation time

Sensor

This table contains the details of the sensors placed on the network through which audit data is collected. This table has the following fields:

• sid: Sensor ID, a unique number identifying a sensor
• hostname: Hostname of the sensor
• interface: Name of the network interface (e.g. eth0)
• filter: BPF filter, used to narrow or refine the collection of audit data
• detail: Detail level of the logging
• encoding: Encoding format of the payload

Signature

This table contains the normalized listing of alert or signature names, priorities used while collecting audit data. This table has the following fields:

• sig_id: Signature ID, a unique number identifying a signature
• sig_name: Description of the signature
• sig_class_id: Classification ID, a unique number identifying the classification of the signature
• sig_priority: Number indicating the priority of the signature
• sig_rev: Revision number
• sig_sid: Number representing the internal signature ID

**Tcphdr**

This table contains the TCP protocol fields associated with a network packet. This table has the following fields:

• sid: Sensor ID, a unique number identifying a sensor
• cid: Event ID, a unique number identifying an event
• tcp_sport: TCP source port
• tcp_dport: TCP destination port
• tcp_seq: TCP sequence number
• tcp_ack: TCP ACK number
• tcp_off: TCP offset
• tcp_res: TCP reserved
• tcp_flags: TCP flags
• tcp_win: TCP window
• tcp_csum: TCP checksum
• tcp_urp: TCP urgent pointer

**Udphdr**

This table contains the UDP protocol fields associated with a network packet.

This table has the following fields:

• sid: Sensor ID, a unique number identifying a sensor
• cid: Event ID, a unique number identifying an event
• udp_sport: UDP source port
• udp_dport: UDP destination port
• udp_len: UDP length
• udp_csum: UDP checksum

Users

This table contains the login information of users authorized to use the system with the following fields:

• username: Name used to login
• Name: Full name of the user
• Password: password selected by the user to login
LIST OF FIGURES

1. Figure 1.1 Components of IDS ..................................................... 2
2. Figure 2.1 Flow of Data Collection ............................................... 7
3. Figure 2.2 User Interface Design of the System ............................. 9
4. Figure 2.3 A Screen Shot of Login page ...................................... 10
5. Figure 2.4 A Screen Shot of Main page ....................................... 11
6. Figure 2.5 A Screen Shot of Clustering page ................................. 12
7. Figure 2.6 A Screen Shot of results of Clustering page ................... 13
8. Figure 2.7 A Screen Shot of Sequence analysis page ..................... 14
9. Figure 2.8 A Screen Shot of results of Sequence analysis page ....... 15
10. Figure 2.9 A Screen Shot of Visualization page .......................... 16
11. Figure 2.10 A Screen Shot of results of Visualization page .......... 17
12. Figure 2.11 A Screen Shot of Drill down page ............................. 18
13. Figure 2.12 A Screen Shot of results of Drill down page ............... 19
14. Figure 2.13 A Screen Shot of Sensors page ................................. 20
15. Figure 2.14 A Screen Shot of Archives page ............................... 21
16. Figure 2.15 A Screen Shot of Add Archives page ......................... 21
17. Figure 2.16 A Screen Shot of Backup page .................................. 22
18. Figure 3.1 Context level Diagram of the System ......................... 25
19. Figure 3.2 Level One Data Flow Diagram for the Login page .......... 26
20. Figure 3.3 Level One Data Flow Diagram for the Main page .......... 27
21. Figure 3.4 Level One Data Flow Diagram for the Clustering page .... 27
22. Figure 3.5 Level One Data Flow Diagram for the Sequence Analysis page .... 28
23. Figure 3.6 Level One Data Flow Diagram for the Visualization page……….28
24. Figure 3.7 Level One Data Flow Diagram for the Drill down page………….29
25. Figure 3.8 Level One Data Flow Diagram for the Backup page……………..29
26. Figure 3.9 Level One Data Flow Diagram for the Exit page………………..30
27. Figure 3.10 Connections of Database Tables………………………………35
28. Figure 4.1 Flow chart of Sequence Analysis Module………………………40