Payload-based Anomaly Detection in HTTP Traffic

A Thesis submitted for the degree of

Doctor of Philosophy

By

Aruna Jamdagni

In

Faculty of Engineering and information Technology

School of Computing and Communications

UNIVERSITY OF TECHNOLOGY, SYDNEY AUSTRALIA

SUBMITTED NOVEMBER 2012
UNIVERSITY OF TECHNOLOGY, SYDNEY
SCHOOL OF COMPUTER AND COMMUNICATIONS

The undersigned hereby certify that they have read this thesis entitled “Payload-based Anomaly Detection in HTTP Traffic” by Aruna Jamdagni and that in his opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Principal Supervisor          Co-Supervisor

Prof. Xiangjian (Sean) He     Dr. Priyadarsi Nanda
CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

------------------------------------------------------
Signature of Author
Abstract

Payload-based Anomaly Detection in HTTP Traffic

Internet provides quality and convenience to human life but at the same time it provides a platform for network hackers and criminals. Intrusion Detection Systems (IDSs) have been proven to be powerful methods for detecting anomalies in the network. Traditional IDSs based on signatures are unable to detect new (zero days) attacks. Anomaly-based systems are alternative to signature based systems. However, present anomaly detection systems suffer from three major setbacks:

(a) Large number of false alarms,
(b) Very high volume of network traffic due to high data rates (Gbps), and
(c) Inefficiency in operation.

In this thesis, we address above issues and develop efficient intrusion detection frameworks and models which can be used in detecting a wide variety of attacks including web-based attacks. Our proposed methods are designed to have very few false alarms. We also address Intrusion Detection as a Pattern Recognition problem and discuss all aspects that are important in realizing an anomaly-based IDS.

We present three payload-based anomaly detectors, including Geometrical Structure Anomaly Detection (GSAD), Two-Tier Intrusion Detection system using Linear Discriminant Analysis (LDA), and Real-time Payload-based Intrusion Detection System (RePIDS), for intrusion detection. These detectors perform deep-packet analysis and examine payload content using n-gram text categorization and Mahalanobis Distance Map (MDM) techniques. An MDM extracts hidden correlations between the features within each payload and among packet payloads. GSAD generates model of normal network payload as geometrical structure using MDMs in a fully automatic and unsupervised manner. We have implemented the GSAD model in HTTP environment for web-based applications.
For efficient operation of IDSs, the detection speed is a key point. Current IDSs examine a large number of data features to detect intrusions and misuse patterns. Hence, for quickly and accurately identifying anomalies of Internet traffic, feature reduction becomes mandatory. We have proposed two models to address this issue, namely two-tier intrusion detection model and RePIDS.

Two-tier intrusion detection model uses Linear Discriminant Analysis approach for feature reduction and optimal feature selection. It uses MDM technique to create a model of normal network payload using an extracted feature set.

RePIDS uses a 3-tier Iterative Feature Selection Engine (IFSEng) to reduce dimensionality of the raw dataset using Principal Component Analysis (PCA) technique. IFSEng extracts the most significant features from the original feature set and uses mathematical and graphical methods for optimal feature subset selection. Like two-tier intrusion detection model, RePIDS then uses MDM technique to generate a model of normal network payload using extracted features.

We test the proposed IDSs on two publicly available datasets of attacks and normal traffic. Experimental results confirm the effectiveness and validation of our proposed solutions in terms of detection rate, false alarm rate and computational complexity.
Acknowledgement

This research would not have been possible without the guidance and the help of many people. First and foremost, my utmost gratitude to my supervisor, Prof. Xiangjian He, for his excellent guidance, support and steadfast encouragement that I will never forget. His comments and suggestions during preparation of this thesis have been extremely valuable. Without his support and supervision, I could not have come this far. I would thank him for his helpfulness and to have been by far more than a simple supervisor.

I would like to express my gratitude to my research co-supervisors, Dr. Priydarshi Nanda and Dr. Ren Liu, for their friendly guidance and unfailing support. Their encouragement has kept me moving ahead at a critical time. Without their help, I would not have been able to complete this thesis.

I appreciate the financial assistance of Australian Postgraduate Award (APA) provided by Australian government and Top-up scholarship provided by the Commonwealth Scientific and Industrial Research Organisation (CSIRO).

Many thanks to my Employer University of Western Sydney and Prof. Simeon Simoff, Dean, School of Computing and Engineering, who gave me time off from work. I will be always grateful for that.

I also appreciate Dr. Qiang Wu and Dr. Wenjing Jia for providing helpful suggestions. My special thanks to collaborator and my good friend Thomas Tan for brain storming discussions. My friends: Thomas Tan and Sheng Wang, they are always helpful whenever I have questions not only on research but also on other matters. It would have been a lonely lab without them.

Last but not the least, I would like to express my love and gratitude to my family members, especially my daughter Divya, my husband Rishi, my sister Meera and brother in-law Satya for their endless love, understanding and encouragement to work on this thesis.
“What we are is God's gift to us. What we become is our gift to God.”

Eleanor Powell

Dedicated to Dear God
Table of Contents

Table of Contents ........................................................................................................................................ viii
List of Tables ............................................................................................................................................... xii
List of Figures ............................................................................................................................................ xiii
List of Acronyms ......................................................................................................................................... xv
Chapter 1 Introduction ............................................................................................................................... 1
  1.1 Motivations: Need for Information Security .................................................................................... 2
    1.1.1 Reasons of Network Threats ........................................................................................................ 3
  1.2 Challenges for Payload Based Anomaly Detection .......................................................................... 6
  1.3 Research Objectives ......................................................................................................................... 7
  1.4 Research Approach ............................................................................................................................ 7
    1.4.1 Design Objectives ......................................................................................................................... 8
    1.4.2 Design Approach ........................................................................................................................... 9
  1.5 Contributions to Thesis ...................................................................................................................... 10
    1.5.1 Framework for Payload-based Anomaly Intrusion Detection .................................................... 10
    1.5.2 Implementation and Evaluation of proposed prototype ............................................................... 10
    1.5.3 Payload Feature Selection for Network Intrusion Detection Using Linear Discriminant Analysis
echnique .................................................................................................................................................... 11
    1.5.4 Cumulative Profile Generation ................................................................................................... 11
    1.5.5 Framework for Real-time Intrusion Detection Using Principal Component Analysis Technique.................................................................................................................... 11
  1.6 Thesis Organization ............................................................................................................................ 12
Chapter 2 Taxonomy of Intrusion Detection systems and Related work .............................................. 14
  Introduction ................................................................................................................................................. 14
  2.1 Strategies for Threat Mitigation ......................................................................................................... 15
  2.2 Taxonomy of Intrusion Detection Systems ....................................................................................... 18
    2.2.1 Intrusion Detection Systems Based on Data Sources ................................................................. 20
    2.2.2 Intrusion Detection System Based on Detection Method ......................................................... 21
Chapter 4 Feature Selection and Two Tier Based Intrusion Detection using LDA

Introduction

4.1 Feature Selection Algorithms

4.2 Linear Discriminant Analysis

4.3 LDA-based Intrusion Detection System

4.3.1 Framework of LDA-based Intrusion Detection System

4.3.2 Framework Modules

4.4 Experimental Results and Analysis

4.4.1 Experimental Results

4.4.2 Analysis of Results

4.5 Two-Tier Intrusion Detection System

4.5.1 Framework of Two-Tier System

4.6 Experimental Results and Analysis

4.6.1 Experimental Results

4.6.2 Analysis of Results

4.7 Common Profile (Signature) for Integrated Feature Set

4.8 Conclusion

Chapter 5 RePIDS: a Multi Tier Real Time Payload Based Intrusion Detection System

5.1 Introduction

5.2 State-of-Art Systems

5.3 RePIDS: Real-time Payload Based Network Intrusion Detection System

5.3.1 Framework of Real-Time Intrusion Detection System

5.3.2 Framework Modules

5.4 Experimental Results and Analysis
5.4.1 Experimental Setup..................................................................................140
5.4.2 Datasets ....................................................................................................140
5.4.3 Model Training and Testing Process............................................................141
5.4.4 Results and Analysis ................................................................................145
5.5 Comparison of RePIDS..............................................................................149
  5.5.1 Detection Performance ............................................................................150
  5.5.2 Complexity Analysis ..............................................................................150
5.6 Conclusions ..................................................................................................154

Chapter 6 Conclusion and Future work ..............................................................155
  6.1 Summary .....................................................................................................156
    6.1.1 Geometrical Structure Anomaly Detection Detector .........................157
    6.1.2 Two-tier LDA-Based Detector ..............................................................158
    6.1.3 Real-time Payload Based Intrusion Detection System .......................158
    6.1.4 Single Profile (Signature) for a Group of Similar Types of
      Attacks..........................................................................................................159
  6.2 Thesis Contributions....................................................................................160
  6.3 Future Work................................................................................................161

References..........................................................................................................163
List of Tables

2.1 Mitigation of attack strategies.................................................................16
2.2 Analogy between text categorization and intrusion detection .................31
2.3 Confusion matrix....................................................................................32

3.1 Performance comparison........................................................................88
3.2 Comparison of GSAD, McPAD and PAYL on GATECH attack dataset ....89
3.3 Summary of experimental results for Generic attacks on various dataset ....89

4.1 Performance of Phf attacks for various selected features .........................106
4.2 Confusion matrix for LDA-based IDS using integrated feature set ..........108
4.3 Performance of LDA-based IDS for four types of attacks .......................118
4.4 Performance of two-tier system using features from 3-types of attacks ......119
4.5 Comparison of IDSs ..............................................................................120

5.1 Principal Component (PC) selection .......................................................144
5.2 Performance Scores corresponding to number of principal components ....146
5.3 Performance score .................................................................................149
5.4 Performance comparison ......................................................................150
5.5 Computational complexity of RePIDS, PAYL and McPAD .......................152
# Table of Figures

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Taxonomy of intrusion detection system</td>
<td>19</td>
</tr>
<tr>
<td>2.2</td>
<td>Generic pattern recognition process</td>
<td>28</td>
</tr>
<tr>
<td>2.3</td>
<td>Pattern recognition process for intrusion detection</td>
<td>29</td>
</tr>
<tr>
<td>3.1</td>
<td>Framework of Geometrical Structure Anomaly Detection System</td>
<td>56</td>
</tr>
<tr>
<td>3.2</td>
<td>Average relative frequency of each byte, (a) Normal Http payload,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(b) Crashiiis attack payload, (c) Back attack payload</td>
<td>65-66</td>
</tr>
<tr>
<td>3.3</td>
<td>Average MDM Images, (a) Normal Http payload, (b) Crashiiis attack payload,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(c) Back attack payload</td>
<td>67</td>
</tr>
<tr>
<td>3.4</td>
<td>Weight factor scores, (a) Normal Http request, (b) Back attack packets</td>
<td>68</td>
</tr>
<tr>
<td>3.5</td>
<td>ROC Curve for accuracy of the GSAD model</td>
<td>69</td>
</tr>
<tr>
<td>3.6</td>
<td>A Typical HTTP (GET) request with parameters</td>
<td>71</td>
</tr>
<tr>
<td>3.7</td>
<td>Nimda attack</td>
<td>72</td>
</tr>
<tr>
<td>3.8</td>
<td>Back attack, 790 /s,</td>
<td>73</td>
</tr>
<tr>
<td>3.9</td>
<td>Average relative frequency of characters for normal HTTP GET request</td>
<td></td>
</tr>
<tr>
<td></td>
<td>payloads, (a) marx, (b) hume</td>
<td>79</td>
</tr>
<tr>
<td>3.10</td>
<td>Average MDM images of normal HTTP GET request, (a) marx, (b) hume</td>
<td>80</td>
</tr>
<tr>
<td>3.11</td>
<td>MDM images of attack packets, (a) Apache2 attack, (b) Phf attack</td>
<td>82</td>
</tr>
<tr>
<td>3.12</td>
<td>Weight factor scores of attack, (a) Apache2, (b) Phf</td>
<td>83-84</td>
</tr>
<tr>
<td>3.13</td>
<td>MDMs of generic attacks</td>
<td>85-86</td>
</tr>
<tr>
<td>3.14</td>
<td>MDM of shell-code attacks</td>
<td>86-87</td>
</tr>
<tr>
<td>3.15</td>
<td>MDM of polymorphic attack</td>
<td>87</td>
</tr>
<tr>
<td>4.1</td>
<td>Framework of LDA-based intrusion detection system</td>
<td>98</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.2</td>
<td>Flow model for feature selection process</td>
<td>101</td>
</tr>
<tr>
<td>4.3</td>
<td>Average MDMs, (a) normal HTTP request, (b) Phf attack packets</td>
<td>107</td>
</tr>
<tr>
<td>4.4</td>
<td>Difference distance map between normal HTTP and Phf attack packets</td>
<td>107</td>
</tr>
<tr>
<td>4.5</td>
<td>Framework of LDM based two-tier intrusion detection system</td>
<td>111</td>
</tr>
<tr>
<td>4.6</td>
<td>Character relative frequencies of Crashiis attack</td>
<td>115</td>
</tr>
<tr>
<td>4.7</td>
<td>Average MDM image of normal HTTP request packets</td>
<td>115</td>
</tr>
<tr>
<td>4.8</td>
<td>Average MDM (a) Phf attack packets, (b) difference distance map</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>between normal HTTP and Phf attack packets</td>
<td></td>
</tr>
<tr>
<td>4.9</td>
<td>Average MDM (a) Apache2 attack packets, (b) difference distance map</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>between normal HTTP and Apache2 attack packets</td>
<td></td>
</tr>
<tr>
<td>4.10</td>
<td>ROC curve of LDA-basedIDS</td>
<td>121</td>
</tr>
<tr>
<td>4.11</td>
<td>ROC curve of a two-tier IDS</td>
<td>122</td>
</tr>
<tr>
<td>5.1</td>
<td>Framework for real-time payload based intrusion detection system</td>
<td>131</td>
</tr>
<tr>
<td>5.2</td>
<td>Scree test plot, (a) Full screen plot, (b) Enlarged scree plot with</td>
<td>143</td>
</tr>
<tr>
<td></td>
<td>first 25 eigenvectors</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Trends of F-Value</td>
<td>146</td>
</tr>
<tr>
<td>5.4</td>
<td>MDM of normal HTTP payload</td>
<td>147</td>
</tr>
<tr>
<td>5.5</td>
<td>MDMs of (a) Apache2 attack, (b) Phf attack payloads</td>
<td>147-48</td>
</tr>
</tbody>
</table>
Acronyms and Abbreviations

ABS       Anomaly Based System
DARPA    Defense Advanced Research Projects Agency
DDoS       Distributed Denial of Service
DoS        Denial of Service
IDES       Intrusion Detection Expert System
IDS        Intrusion Detection System
GATECH    Georgia Institute of Technology
GSAD      Geometrical Structure Anomaly Detection System
GSPM      Geometrical Structure Payload Model
HIDS      Host-based Intrusion Detection System
HTTP      Hyper Text Transport Protocol
IFSEng    Iterative Feature Selection Engine
IDPS      Intrusion Detection and Prevention System
KDD       Knowledge Discovery in Databases
LDA       Linear Discriminant Analysis
LDM       Linear Discriminant Module
McPAD     Multi classifier Payload Based Anomaly Detection
MD        Mahalanobis Distance
MDM       Mahalanobis Distance Map
MIT       Massachusetts Institute of Technology
MS-SQL    MiscroSoft Structured Query Language
NIDS      Network Intrusion Detection System
PA        Parallel Analysis
PAYL      Payload Based Anomaly Detection System
PCA       Principal Component Analysis
PC         Principal Component
RePIDS    Real-time Payload-based Intrusion Detection System
R2L       Remote to Local
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBS</td>
<td>Signature Based System</td>
</tr>
<tr>
<td>SRI</td>
<td>Stanford Research International</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TC</td>
<td>Text Categorization</td>
</tr>
<tr>
<td>U2R</td>
<td>User to Root</td>
</tr>
</tbody>
</table>
Authors Publications for the Ph.D

Published papers

Journal Papers


Conference papers


