

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

2.2.3	Hybrid Intrusion Detection System .....	24
2.2.4	Data Audit Time .....	25
2.2.5	System Structure .....	26
2.2.6	Action after Intrusion Detection .....	27
2.3	Pattern Recognition Approach .....	27
2.4	Performance Evaluation of Intrusion Detection System .....	32
2.5	Related Research Works .....	34
2.5.1	Review from the Perspectives of Intrusion Detection Techniques .....	35
2.5.2	Review from the Perspective of Payload-based Intrusion Detection System .....	44
2.6	Conclusions .....	49
Chapter 3	GSAD: Geometrical Structure Anomaly Detection System.....	51
Introduction	.....	51
3.1	GSAD-Geometrical Structure Anomaly Detection System.....	54
3.1.1	Framework of the Proposed Intrusion Detection System.....	54
3.1.2	Framework Modules .....	56
3.1.3	Base-line Profile Generation .....	62
3.1.4	Model Testing .....	62
3.2	GSAD Evaluation .....	63
3.2.1	Experimental Setup.....	63
3.2.2	DARPA 1999 Dataset .....	63
3.2.3	Experimental Results and Analysis .....	64
3.3	HTTP and Examples on Attacks .....	70
3.3.1	HTTP .....	70
3.3.2	HTTP Attack Examples.....	72
3.4	Implementation of GSAD in HTTP Environment .....	73
3.5	Evaluation in HTTP Environment .....	74
3.5.1	Experimental Setup.....	74
3.5.2	Datasets .....	75

3.5.3	Experimental Results and Analysis.....	76
3.6	Analysis of eResults.....	88
3.7	Conclusion.....	90
<b>Chapter 4</b>	<b>Feature Selection and Two Tier Based Intrusion Detection using LDA.....</b>	<b>91</b>
	<b>Introduction.....</b>	<b>91</b>
	<b>4.1 Feature Selection Algorithms.....</b>	<b>93</b>
	<b>4.2 Linear Discriminant Analysis .....</b>	<b>95</b>
	<b>4.3 LDA-based Intrusion Detection System.....</b>	<b>96</b>
	<b>4.3.1 Framework of LDA-based Intrusion Detection System.....</b>	<b>97</b>
	<b>4.3.2 Framework Modules.....</b>	<b>98</b>
	<b>4.4 Experimental Results and Analysis .....</b>	<b>104</b>
	<b>4.4.1 Experimental Results.....</b>	<b>104</b>
	<b>4.4.2 Analysis of Results.....</b>	<b>1099</b>
	<b>4.5 Two-Tier Intrusion Detection System.....</b>	<b>109</b>
	<b>4.5.1 Framework of Two-Tier System.....</b>	<b>110</b>
	<b>4.6 Experimental Results and Analysis .....</b>	<b>114</b>
	<b>4.6.1 Experimental Results .....</b>	<b>114</b>
	<b>4.6.2 Analysis of Results.....</b>	<b>120</b>
	<b>4.7 Common Profile (Signature) for Integrated Feature Set .....</b>	<b>123</b>
	<b>4.8 Conclusion.....</b>	<b>123</b>
<b>Chapter 5</b>	<b>RePIDS: a Multi Tier Real Time Payload Based Intrusion Detection System.....</b>	<b>125</b>
	<b>5.1 Introduction .....</b>	<b>126</b>
	<b>5.2 State-of-Art Systems .....</b>	<b>129</b>
	<b>5.3 RePIDS: Real-time Payload Based Network Intrusion Detection System .....</b>	<b>130</b>
	<b>5.3.1 Framework of Real-Time Intrusion Detection System.....</b>	<b>131</b>
	<b>5.3.2 Framework Modules .....</b>	<b>133</b>
	<b>5.4 Experimental Results and Analysis .....</b>	<b>140</b>

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

2.1	Taxonomy of intrusion detection system .....	19
2.2	Generic pattern recognition process .....	28
2.3	Pattern recognition process for intrusion detection .....	29
3.1	Framework of Geometrical Structure Anomaly Detection System .....	56
3.2	Average relative frequency of each byte, (a) Normal Http payload, (b) Crashiis attack payload, (c) Back attack payload.....	65-66
3.3	Average MDM Images, (a) Normal Http payload, (b) Crashiis attack payload, (c) Back attack payload.....	67
3.4	Weight factor scores, (a) Normal Http request, (b) Back attack packets .....	68
3.5	ROC Curve for accuracy of the GSAD model .....	69
3.6	A Typical HTTP (GET) request with parameters .....	71
3.7	Nimda attack.....	72
3.8	Back attack, 790 /s, .....	73
3.9	Average relative frequency of characters for normal HTTP GET request payloads, (a) marx, (b) hume .....	79
3.10	Average MDM images of normal HTTP GET request, (a) marx, (b) hume .....	80
3.11	MDM images of attack packets, (a) Apache2 attack, (b) Phf attack .....	82
3.12	Weight factor scores of attack, (a) Apache2, (b) Phf .....	83-84
3.13	MDMs of generic attacks .....	85-86
3.14	MDM of shell-code attacks .....	86-87
3.15	MDM of polymorphic attack .....	87
4.1	Framework of LDA-based intrusion detection system .....	98

4.2	Flow model for feature selection process .....	101
4.3	Average MDMs, (a) normal HTTP request, (b) Phf attack packets .....	107
4.4	Difference distance map between normal HTTP and Phf attack packets .....	107
4.5	Framework of LDM based two-tier intrusion detection system .....	111
4.6	Character relative frequencies of Crashiis attack .....	115
4.7	Average MDM image of normal HTTP request packets .....	115
4.8	Average MDM (a) Phf attack packets, (b) difference distance map between normal HTTP and Phf attack packets .....	116
4.9	Average MDM (a) Apache2 attack packets, (b) difference distance map between normal HTTP and Apache2 attack packets .....	116
4.10	ROC curve of LDA-based IDS .....	121
4.11	ROC curve of a two-tier IDS .....	122
5.1	Framework for real-time payload based intrusion detection system .....	131
5.2	Scree test plot, (a) Full screen plot, (b) Enlarged scree plot with first 25- eigenvectors.....	143
5.3	Trends of $F$ -Value .....	146
5.4	MDM of normal HTTP payload .....	147
5.5	MDMs of (a) Apache2 attack, (b) Phf attack payloads .....	147-48

# 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

SBS	Signature Based System
SRI	Stanford Research International
SVM	Support Vector Machines
TC	Text Categorization
U2R	User to Root

# Authors Publications for the Ph.D

## Published papers

### Journal Papers

1. **A. Jamdagni**, Z. Tan, P. Nanda, X. He, R. Liu, “RePIDS: a Multi Tier Real-Time Payload-Based Intrusion Detection System,” *Computer Networks (ERA Tier A)*, Elsevier, accepted (Final) on 8 October 2012.
2. **A. Jamdagni**, Z. Tan, P. Nanda, X. He, R. Liu, “Mahalanobis Distance Map Approach for Anomaly Detection of Web-Based Attacks,” *Journal of Network Forensics*, 2(2), 25-39, 2011.

### Conference papers

3. Z. Tan., **A. Jamdagni**, P. Nanda, X. He, R. Liu, “Network Intrusion Detection based on LDA for payload feature selection,” IEEE Globecom 2010 Workshop on Web and Pervasive Security (WPS 2010), Miami, USA, 2010, pp.1590-1594.
4. Z. Tan, **A. Jamdagni**, X. He, P. Nanda, R. Liu, W. Jia, W. Yeh, “A Two-Tier System for Web Attack Detection Using Linear Discriminant Method,” *Information and Communications Security, LNCS*, Vol. 6476/2010, pp.459-471.
5. **A. Jamdagni**, Z. Tan, X. He, P. Nanda, R. Liu, “Mahalanobis Distance Map Approach for Anomaly Detection of Web-Based Attacks,” in 8th Australian Information Security Management Conference. November 2010. Perth.

6. **A. Jamdagni**, Z. Tan, P. Nanda, X. He, R. Liu, “Intrusion detection using GSAD model for HTTP traffic on web services,” in IWCMC’ 10 Proceedings of the 6th International Wireless Communications and Mobile Computing Conference 2010. France: ACM.
7. **A. Jamdagni**, Z. Tan, R. Liu, P. Nanda, X. He, “Pattern Recognition Approach for Anomaly Detection of Web-based Attacks,” in the Seventh Annual CSIRO ICT Centre Science and Engineering Conference, November 2010.
8. **A. Jamdagni**, Z. Tan, P. Nanda, X. He, R. Liu, “Intrusion Detection Using Geometrical Structure,” in 4th International Conference on Frontier of Computer Science and Technology (FCST 2009), Shanghai, China, December 17-19, 2009, pp. 327-333.
9. **A. Jamdagni**, Z. Tan, R. Liu, P. Nanda, X. He, “A Framework for Geometrical Anomaly Detection Model,” in the Sixth Annual CSIRO ICT Centre Science and Engineering Conference, November 2009.