

TABLE OF CONTENTS

APPROVAL PAGE	ii
DECLARATION.....	iii
FOREWORD.....	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF CODE	xi
ABSTRACT	xii
CHAPTER I INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	2
1.3 Research Scope	3
1.4 Research Objectives.....	3
1.5 Research Benefits.....	3
CHAPTER II LITERATURE REVIEW.....	5
CHAPTER III THEORETICAL BASIS.....	12
3.1 Binary Files	12
3.2 ELF Files.....	12
3.3 Instruction and Operand Extraction	13
3.4 Shannon Entropy.....	14
3.5 Kullback-Leibler Divergence (KL-Divergence).....	15
3.6 Agglomerative Hierarchical Clustering (AHC).....	16
3.7 Silhouette Coefficient as an Evaluation Metric	19
3.8 Block Feature Correlation Using Cross-Correlation Function	20
3.9 Correlated Data Flow Graph.....	21

CHAPTER IV RESEARCH METHODOLOGY	22
4.1 Research Description	22
4.2 Environment Set-Up and Data Acquisition	22
4.2.1 Building Custom Programs	22
4.2.2 Disassembling ELF Files	23
4.3 Feature Extraction	24
4.3.1 Entropy Based Significant Feature Extraction	24
4.4 Block Characterization.....	25
4.4.1 Implementing Agglomerative Hierarchical Clustering (AHC).....	25
4.4.2 Evaluating the Clusters Using Silhouette Coefficient.....	26
4.5 Constructing Correlated Data Flow Graphs.....	27
4.5.1 Calculating Block Feature Correlation	27
4.5.2 Generating Correlation Data Flow Graphs	27
CHAPTER V IMPLEMENTATION	28
5.1 Development Environment	28
5.2 Dataset Generation.....	28
5.2 Feature Extraction Program	29
5.3 Extracting Significant Features by Entropies Level	32
5.3.1 Noise Removal via Entropy Thresholding.....	32
5.4 Block Characterization Using KL-Divergence and AHC.....	35
5.4.1 Calculating KL-Divergence	35
5.4.2 AHC Implementation.....	36
5.5 Silhouette Coefficient as Evaluation Metric	38
5.6 Generating Correlated Data Flow Graphs (CDFGs).....	39
5.6.1 Calculating Block Correlation	40

5.6.2 Visualizing Highly Correlated Data Flow Graphs (CDFGs)	40
CHAPTER VI RESULTS & ANALYSIS.....	42
6.1 Block Characterization Visualized as Dendrograms	42
6.1.1 Functional Clustering of Simple Calculator.....	42
6.1.2 Functional Clustering of Dynamic Array Allocator	44
6.1.3 Functional Clustering of Dynamic CSV Parser	47
6.2 Silhouette Coefficient as an Evaluation Metric	50
6.3 Block Correlation Analysis Using Correlated Data Flow Graphs	52
CHAPTER VII CONCLUSION.....	56
REFERENCES.....	58
APPENDIX.....	62
Appendix A (Entropy Calculation)	62
Appendix B (Simple Calculator).....	64
Appendix C (Simple Calculator Assembly)	65
Appendix D (Dynamic Array Allocator).....	70
Appendix E (CSV Parser)	71
Appendix F (Dendrogram of CSV Parser Block Clusters)	72
Appendix G (Coefficient Value of Simple Calculator Blocks).....	73
Appendix H (Coefficient Value of Dynamic Array Allocator Blocks).....	74
Appendix I (Coefficient Value of CSV Parser Blocks).....	75
Appendix J (Block Similarity Heatmap of Simple Calculator).....	78
Appendix K (Block Similarity Heatmap of Dynamic Array Allocator)	79

LIST OF TABLES

Table 2.1 Table of Comparison.....	9
Table 5.1 Environment Specifications	28
Table 5.2 Data frame of disassembled code of simple calculator program	30
Table 5.3 First 5 assembly codes distribution probability and entropy	32
Table 5.4 First 5 assembly codes probability and entropy after noise removal....	35
Table 5.5 Top 5 highly correlated blocks (Simple Calculator).....	40
Table 6.1 Cluster Assignment Table (Simple Calculator).....	42
Table 6.2 Cluster Assignment Table (Dynamic Array Allocator).....	45
Table 6.3 Cluster Assignment Table (CSV Parser)	48

LIST OF FIGURES

Figure 3.1 ELF Format	13
Figure 3.2 Pseudocode for Calculating Entropy using Shannon Entropy.....	15
Figure 3.3 Agglomerative Hierarchical Clustering Algorithm Process.....	17
Figure 3.4 Pseudocode of Agglomerative Hierarchical Clustering	18
Figure 3.5 Example of Dendrogram Produced by AHC Algorithm	19
Figure 3.6 An example of highly correlated data flow graph	21
Figure 4.1 Example of a disassembled binary file	23
Figure 5.1 Content size of each block (CSV Parser)	31
Figure 5.2 Content size of each block (Dynamic Array Allocator).....	31
Figure 5.3 Content size of each block (Simple Calculator)	31
Figure 5.4 Distinct instruction (ϵ) Shannon entropies. (X-axis = Instructions)....	33
Figure 5.5 Distinct left operand (δ) Shannon entropies. (X-axis = Left operand)	33
Figure 5.6 Distinct right operand (γ) Shannon entropies (X-axis = Right operand)	34
Figure 5.7 Block similarity matrix heatmap (Simple Calculator).....	36
Figure 5.8 AHC dendrogram output (Dynamic Array Allocator)	38
Figure 5.9 Silhouette Coefficient values for each block (CSV Parser (Left) and Dynamic Array Allocator (Right)).....	39
Figure 5.10 Maximum correlation values for each block (Dynamic Array Allocator)	41

Figure 5.11 Highly correlated blocks (Dynamic Array Allocator).....	41
Figure 6.1 Dendrogram of Block Clusters (Simple Calculator)	43
Figure 6.2 Block 1 Content (Simple Calculator)	43
Figure 6.3 Average Entropy of Distinct Right Operand per Cluster (Simple Calculator).....	44
Figure 6.4 Dendrogram of Block Clusters (Dynamic Array Allocator)	45
Figure 6.5 Average Entropy of Distinct Right Operand per Cluster (Dynamic Array Allocator)	46
Figure 6.6 Average Entropy of Distinct Instruction per Cluster (CSV Parser)	49
Figure 6.7 Average Entropy of Distinct Right Operand per Cluster (CSV Parser)	49
Figure 6.8 Silhouette Coefficient for Each Block (CSV Parser)	50
Figure 6.9 Silhouette Coefficient for Each Block (Dynamic Array Allocator)....	51
Figure 6.10 Silhouette Coefficient for Each Block (Simple Calculator).....	51
Figure 6.11 Highly Correlated Blocks (CSV Parser).....	53
Figure 6.12 Highly Correlated Blocks (Dynamic Array Allocator)	53
Figure 6.13 Highly Correlated Blocks (Simple Calculator)	54

LIST OF CODE

Code 5.1 Parsing instructions function to format assembly code	29
Code 5.2 A function to convert the extracted code into CSV format	30
Code 5.3 A loop to write with entropy above the threshold into a new CSV	34
Code 5.4 A function to calculate KL-Divergence between two blocks.....	35
Code 5.5 A function to convert similarity matrix into distance matrix	37
Code 5.6 Implementation of AHC	37
Code 5.7 Calculating Silhouette Coefficient values for each block	38