Computational Intelligence in E-mail Traffic Analysis

by

Mark Jyn-Huey Lim, BEng (Hons)

Submitted in fulfilment of the
requirements for the Degree of
Doctor of Philosophy

University of Tasmania

May 2008

Statement of Originality

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Mark Jyn-Huey Lim
Date: Friday 30th May 2008.

Statement of Authority of Access

This thesis may be made available for loan and limited copying in accordance with the *Copyright Act 1968*.

Mark Jyn-Huey Lim
Date: Friday 30th May 2008.

Abstract

In law enforcement, tools and techniques are required that enable forensic analysts to uncover electronic evidence about the communication activities of possible criminal or terrorist suspects. This is needed in order to better understand the actions of criminal or terrorist groups and to also understand the communication patterns of suspected individuals. The extraction of useful information from electronic communication data is a difficult task, due to the large amounts of data and also due to the difficulty in making sense of unusual activities in the data. This thesis considers the problem of aiding the analyst to provide a better understanding about the communication behaviour of suspected individuals. The type of data considered for the thesis is e-mail traffic, which is based on information obtained from e-mail message headers but not the content of e-mails.

This thesis proposes a "computational intelligence" approach for analysing e-mail traffic, by using a set of computational techniques to provide different perspectives for examining the communication behaviour of suspect e-mail accounts. This is considered important, since a range of views on e-mail traffic behaviour can provide the user/analyst a more overall understanding about the behaviour of suspect e-mail accounts. The purpose of using a set of computational techniques is to utilise the capabilities of each technique, so that the combined effect of using those techniques present useful information to the user/analyst about a suspect e-mail account's traffic behaviour.

The computational techniques used for the research in this thesis are visualisation and feature extraction techniques, which each provide different ways of examining e-mail traffic behaviour. Visualisation is used to provide a visual method of interpreting, exploring, and understanding the communication patterns present in e-mail traffic data. The two visualisation techniques used for visualization are social network visualisation and time-series visualisation. Feature extraction techniques are another type of technique used to analyse e-mail traffic behaviour, by providing information that locate features in the data, indicating where unusual changes in communication activity are occurring. The two techniques used for

feature extraction in the research are decision tree classification and hierarchical fuzzy inference.

Two case studies are provided in this thesis. The first case study explores the detection of unusual variations in traffic behaviour from simulated e-mail traffic data, while the second case study explores the rating of abnormal communication changes from the Enron e-mail corpus dataset. Both case studies demonstrate that computational intelligence is a useful approach for providing the user/analyst a better understanding about the traffic behaviour of suspect e-mail accounts.

Acknowledgements

The work produced in this thesis has only been made possible through the support of a number of people, whom I would like to thank. Firstly, to Prof. Michael Negnevitsky and Mrs. Jacky Hartnett, both of whom I am grateful to have as my supervisors. Prof. Michael Negnevitsky has been a great mentor who has provided me insightful advice that has helped me to develop as a person and as a researcher. Mrs. Jacky Hartnett has been wonderful as my co-supervisor and is someone I have always enjoyed sharing lively discussions with during our meetings.

Secondly, I would like to thank all of my colleagues with whom I have shared room 337 during my PhD candidature: Warwick Gillespe, Shao Kwan Ng, Dao Thuan An Le, Yin Chin Choo, Jeffrey Poh, Jonathan Culberg, and Ahsan Latheef. All of these people have shared parts of the PhD journey with me and have been wonderful people who have made my postgraduate experience an enjoyable and interesting one. I would like to make special mention of my colleague Cameron Potter, who has been really great in helping me adjust to postgraduate life during the beginning of my PhD candidature. Cameron was someone I looked up to as a fellow PhD candidate and whose drive and motivation for his research has inspired me to do the same with mine.

Thirdly, I owe special thanks to my friend Andrew Durdin, whose in-depth knowledge of Python and Mac computers have been really essential in providing me the knowledge I need for doing my work. Andrew has been a great friend who has helped me troubleshoot technical problems that are just out of my reach. Also special thanks to Deborah Ploughman, who helped proof read Chapters 2 and 3 of my thesis.

Finally, none of this would have been possible without the love and support of my mum and dad, Kim and Yong Seng Lim, and my brother David Lim, who have been behind me through the whole of my PhD studies.

Contents

Abstract **ii**

Acknowledgements **iv**

Contents **ix**

List of Figures **xv**

List of Tables **xvii**

Preface **xix**

1 Introduction **1**

1.1 E-mail Communications 1

1.2 Involvement Of E-mail In Illegal Activities 3

1.3 Investigation of Suspected Individuals 4

1.3.1 Computer Forensics 5

1.4 E-mail Analysis Methods 7

1.4.1 E-mail Content Analysis 7

1.4.2 E-mail Traffic Analysis 9

1.5 Difficulties Associated With Analysing E-mail Data 11

1.6 Problem Statement 12

1.7 Thesis Contribution and Layout 13

2 Behaviour Analysis of E-mail Traffic **15**

2.1 Introduction 15

2.2 Behaviour Analysis Methods 16

2.2.1 Individual Behaviour Analysis 18

2.2.2 Behaviour Comparison Analysis 25

2.2.3 Clique Behaviour Analysis 31

2.2.4 Network Behaviour Analysis 34

2.3 Discussion 38

2.4 Summary 43

3 E-mail Traffic Analysis Using Computational Intelligence **45**

3.1 Introduction 45

3.2 Computational Intelligence 46

3.3 Visualisation Techniques 49

3.3.1 Social Network Visualisation 52

3.3.2 Time-Series Visualisation 56

3.4 Feature Extraction Techniques 61

3.4.1 Decision Tree Classification 63

3.4.2 Hierarchical Fuzzy Inference 70

3.5 Summary 82

4 Development of the E-mail Traffic Analysis System **85**

4.1 Introduction 85

4.2 System Architecture 86

4.2.1 Overview of the System 87

4.2.2 Implementation 92

4.3 E-mail Traffic Data 93

4.3.1 Simulated E-mail Traffic Data 94

4.3.1.1 Conceptual E-mail System Model 95

4.3.1.2 Personality Traits of the Behaviour Model . . 96

4.3.1.3 Sending and Replying Delay Distributions . . 99
4.3.1.4 Generation of Simulated Data 102
4.3.2 The Enron E-mail Dataset 106
4.4 Analysing The Data 110
4.5 Summary 115

5 The E-mail Traffic Analysis System Evaluation and Case Studies 118
5.1 Introduction 118
5.2 E-mail Traffic Data Simulation and Analysis 119
5.2.1 Case Study 1 119
5.2.2 Case Study 2 128
5.2.3 Discussion 133
5.3 Case Study: Enron E-mail Traffic Data 140
5.3.1 Analysis of Enron Employee's Traffic Behaviour 140
5.3.2 Discussion 147
5.4 Concluding Remarks 149
5.5 Summary 150

6 Summary and Further Studies 154
6.1 Investigations Involving E-mail As Evidence 154
6.2 Analysis of E-mail Traffic Behaviour 155
6.3 Major Contributions 157
6.3.1 Utilising A Combination Of Techniques To Examine E-mail Traffic Behaviour 157
6.3.2 Development Of A Personality Trait Based E-mail Traffic Simulation Tool 159
6.3.3 Development Of A Novel System Architecture For E-mail Traffic Analysis 159
6.4 Further Studies 161
6.4.1 Extending the Conceptual E-mail System Model 161
6.4.2 Further Evaluation Of Decision Tree Classification . . . 162

6.4.3 Investigation Of Other Hierarchical Fuzzy Inference Architectures . . . 162
6.4.4 Correlating E-mail Content Analysis And E-mail Traffic Analysis Results . . . 164
6.5 Conclusion . . . 165

References **166**

Appendices **178**

A E-mail Traffic Database Schema **178**
A.1 ‘emailSystems’ Table . . . 180
A.2 ‘emailClients’ Table . . . 180
A.3 ‘emailTraffic’ Table . . . 180
A.4 ‘sendingDelays’ Table . . . 181
A.5 ‘replyingDelays’ Table . . . 181
A.6 ‘suspiciousClients’ Table . . . 182
A.7 ‘MITDNormProfiles’ Table . . . 182
A.8 ‘TDNormProfiles’ Table (Obsolete) . . . 184
A.9 ‘ADAnalysisInfo’ Table . . . 184
A.10 ‘AnomalyDetectionTypes’ Table . . . 185

B Class Diagram For The E-mail System Simulation Model **186**

C Enron Events Timeline **188**

D Former Employees of Enron **190**

E Hierarchical Fuzzy Inference System Development **193**
E.1 Architecture . . . 193
E.2 Summary Of The Rule Bases . . . 193
E.3 Input And Output Variable Fuzzy Sets . . . 194
E.4 Input-Output Mappings Of The Fuzzy System Modules . . . 202

F Simulated E-mail Traffic Data Analysis Results **206**

F.1 Case Study 1 Decision Tree Results 206

F.2 Case Study 2 Decision Tree Results 210

G Enron E-mail Traffic Data Analysis Results **214**

List of Figures

1.1 The parts of an e-mail message that are used for content and traffic analysis. 7

2.1 Different levels of analysis for examining e-mail traffic behaviour. 17

2.2 Diagram illustrating the idea of individual behaviour analysis. . 18

2.3 Example of a user's 24-hour sending usage profile. 19

2.4 Example of the frequency table and histogram used for the Recipient Frequencies method [1], with kind permission of Springer Science and Business Media. 20

2.5 Comparison of the past and recent sending usage patterns for an individual. 21

2.6 Example of the rolling window approach used to detect variations in recipient frequencies [2] ©ACM, Inc. Reprinted by permission. 23

2.7 Observing abnormal behaviour from different perspectives. . . . 24

2.8 Example of how the abnormal behaviour of individual e-mail accounts may be summarised 25

2.9 The concept of behaviour comparison analysis. 26

2.10 Concept map of the 15 behavioural features defined by [3]. . . . 28

2.11 Concept of the comparison approaches used. 29

2.12 Diagram illustrating the idea of clique behaviour analysis. 32

2.13 Diagram of clique profiling, based on the clique diagrams originally drawn by [2]. 33

2.14 The concept of network behaviour analysis. 34

2.15 Diagram showing the hidden group's e-mail traffic communication and structure, based on the diagrams originally drawn by [4]. 36

2.16 Further investigation of abnormal network activity. 38

3.1 Extracting and comparing different types of information about the data. 47

3.2 Simple example of multi-dimensional data. 48

3.3 Computational techniques used for e-mail traffic analysis. 49

3.4 Steps of the visualisation process, based on a diagram drawn by [5]. 50

3.5 Simple diagram showing the components of a social network graph. 52

3.6 The social network visualisation process. 53

3.7 Social network visualisation of data from Table 3.2. 55

3.8 Example of a large network of 355 e-mail users. 56

3.9 Example of different notations used for representing time-series data. 57

3.10 Time-series visualisation process. 58

3.11 Example of e-mail traffic volume using a weekly time-scale. . . 59

3.12 Example of e-mail traffic volume using a daily time-scale. . . . 60

3.13 Steps of the feature extraction process. 61

3.14 The information obtained through each step of feature extraction. 61

3.15 Steps of the decision tree classification process. 63

3.16 The data produced after each step of the decision tree process. . 63

3.17 Example of how decision tree classification partitions the data. . 64

3.18 Decision tree representation used for describing the data. 65

3.19 Decision tree classification process used for e-mail traffic analysis. 66

3.20 Data produced for finding unusual interactions. 66

3.21 Two decision trees produced for incoming and outgoing e-mail traffic. 69

3.22 Comparison between Boolean logic and fuzzy logic. 71

3.23 Example of a linguistic variable and its associated fuzzy set. . . 71
3.24 Different architectures for a hierarchical fuzzy system of 4 variables. 72
3.25 Analysis of the suspect's communication links. 73
3.26 Mapping of e-mail traffic behaviour measurements. 76
3.27 Process used for finding abnormal changes in communication behaviour. 78
3.28 Data produced for finding abnormal communication links. . . . 79
3.29 Analysing each of the suspect's communication links. 79
3.30 Examples of unknown sending delay and replying delay cases. . 81
3.31 The architecture used for the hierarchical fuzzy inference system. 82

4.1 Overview of the e-mail traffic analysis system. 88
4.2 Implementation of the e-mail traffic analysis system. 93
4.3 The entities making up the e-mail system model. 95
4.4 Layout for the sending delay normal distribution. 100
4.5 Layout for the replying delay normal distribution. 101
4.6 Flow diagram of the simulation model set-up and simulation process. 104
4.7 The main graphical user interface for the e-mail system model generation program. 105
4.8 The behaviour editor part of the e-mail system model generation program. 105
4.9 Events diagram for events in the conceptual e-mail system model. 106
4.10 Overview of how simulated data is entered into the e-mail traffic analysis system. 106
4.11 Steps for exploring the data using visualisation techniques. . . . 111
4.12 Steps for analysing the data with decision tree classification. . . 113
4.13 Steps for analysing the data with hierarchical fuzzy inference. . 114

5.1 The setup configuration given for the social connections between e-mail clients. 120

5.2 Stacked column chart showing the behavioural profiles of each e-mail client and their behaviour model ID. 120

5.3 Column chart showing the personality trait degree values of all the e-mail clients. 121

5.4 Number of e-mail messages sent and received by the e-mail clients over 120 simulation days. 121

5.5 Overview of communications between the 10 e-mail clients. . . 122

5.6 Weekly time-series overview of the 10 e-mail clients, with *clientB* selected. 123

5.7 Unusual changes in interactions found from both the inboxes and outboxes of e-mail clients (i.e. changes found in both directions). 125

5.8 Unusual changes in interactions found from either the inbox or outbox of e-mail clients (i.e. changes found only in one direction). 125

5.9 Decision tree for *clientG* showing unusual changes in incoming interactions. 126

5.10 Decision tree for *clientG* showing unusual changes in outgoing interactions. 126

5.11 Weekly time-series data of e-mail traffic between *clientG* and *clientI*, showing a drop in traffic after week 6 or around day 47. . 127

5.12 Weekly time-series data of e-mail traffic between *clientG* and *clientD*, showing an increase in traffic after week 6. 127

5.13 Daily time-series data for the interaction between *clientG* and *clientF* from days 35 to 56 (i.e. weeks 5 to 8). 127

5.14 The behavioural profiles of each e-mail client in the simulation model. 129

5.15 The setup configuration given for the e-mail clients' social connections for case study 2. 129

5.16 Number of e-mail messages sent and received by e-mail clients over 182 simulation days. 130

5.17 Social network overview of the 9 e-mail clients. 130

5.18 Social network diagrams showing where the decision tree identified unusual changes in interaction from two directions. 132

5.19 Social network diagrams showing where the decision tree identified unusual changes in interaction in one direction. 132

5.20 Weekly time-series data for *clientG*, showing the general e-mail traffic activity. 134

5.21 Decision tree classification output for *clientG*'s unusual changes in incoming e-mail traffic interactions. 134

5.22 Decision tree classification output for *clientG*'s unusual changes in outgoing e-mail traffic interactions. 135

5.23 Weekly time-series of traffic between *clientG* and *clientA*, highlighting the week 15 ($\approx$ day 108) with a drop in number of outgoing e-mails on week 14. 135

5.24 Weekly time-series data for *clientG* and *clientC*, highlighting the week 15 ($\approx$ day 108) with a significant rise in number of e-mails from *clientG* to *clientC*. 135

5.25 Daily time-series data for *clientG* and *clientC*, showing an unusual increase in daily e-mail traffic activity after day 108. . . . 136

5.26 Social network diagrams highlighting the change in communications between *clientG* and *clientC* before and after day 108. . . . 136

5.27 Jeffrey Skilling's e-mail addresses (orange) and his circle of associates (blue) from 1st January 1999 to 1st August 2000 (17 months). 143

5.28 Jeffrey Skilling's e-mail addresses (orange) and his circle of associates (blue) from 1st February 2001 to 1st September 2001 (7 months). 143

5.29 Weekly time-series overview of traffic from Jeffrey Skilling's e-mail accounts from 1st January 1999 to 1st September 2001. . . 144

5.30 Social network diagrams of Jeffrey Skilling's top ten ranked communication links. 145

5.31 Weekly time-series of Jeffrey Skilling and Rosalee Fleming, focusing on the surveillance period. 146

5.32 Weekly time-series of Jeffrey Skilling and Steven Kean, focusing on the surveillance period. 146

6.1 A suggestion for a trimmed down version of the hierarchical architecture. 163

A.1 Diagram of the schema layout for the e-mail traffic database. . . 179

B.1 Legend of the symbols used in the class diagram. 186

B.2 Class diagram of the implementation of the conceptual e-mail system simulation model. 187

C.1 Events associated with Jeffrey Skilling's employment and the collapse of Enron. 189

E.1 Architecture of the hierarchical fuzzy inference system. 194

E.2 The fuzzy sets used for the input and output variables of L1RB1. 195

E.3 The fuzzy sets used for the input and output variables of L1RB2. 196

E.4 The fuzzy sets used for the input and output variables of L1RB3. 197

E.5 The fuzzy sets used for the input and output variables of L1RB4. 198

E.6 The fuzzy sets used for the input and output variables of L2RB1. 199

E.7 The fuzzy sets used for the input and output variables of L2RB2. 200

E.8 The fuzzy sets used for the input and output variables of L3RB1. 201

E.9 The input-output mappings for the first and second rule bases of layer 1. 203

E.10 The input-output mappings for the third and fourth rule bases of layer 1. 204

E.11 The input-output mappings for layer 2 of the hierarchy. 205

E.12 The input-output mapping for L3RB1 in layer 3 of the hierarchy. 205

List of Tables

3.1 Example of e-mail traffic log data. 51

3.2 Example of relational data extracted from e-mail traffic data. . . 54

3.3 Example of time-series e-mail traffic data. 59

3.4 Example of data attributes and the assigned class labels. 64

3.5 Example of attributes and class label used for unusual incoming e-mail traffic. 68

3.6 Example of attributes and class label used for unusual outgoing e-mail traffic. 68

4.1 Filtering options available through the use of the Data Parameter Selector component. 90

4.2 Description of the personality trait dimensions, through the use of trait pair examples. 97

4.3 The relationship between the effect of personality trait dimensions on e-mail communication behaviour, based on intuitive assumptions. 99

5.1 Unusual changes in interaction behaviour found by decision tree classification. 124

5.2 Unusual changes in interactions derived from the two decision tree outputs produced using WEKA. 131

5.3 A listing of e-mail addresses possibly belonging to Jeffrey Skilling. 141

5.4 A listing of the top ten abnormality ratings from the communication links analysed. 144

D.1 Information on Kenneth Lay, Jeffrey Skilling, and Andrew Fastow. 191

D.2 Information on Richard A. Causey and John M. Forney 192

E.1 Summary of the rule bases used for the hierarchical fuzzy inference system. 194

F.1 Decision tree results for *clientA* to *clientD* from simulation case study 1. 207

F.2 Decision tree results for *clientE* to *clientH* from simulation case study 1. 208

F.3 Decision tree results for *clientI* and *clientJ* from simulation case study 1. 209

F.4 Decision tree results for *clientA* to *clientC* from simulation case study 2. 211

F.5 Decision tree results for *clientD* to *clientF* from simulation case study 2. 212

F.6 Decision tree results for *clientG* to *clientI* from simulation case study 2. 213

G.1 Enron case study behaviour measurements, rows 1 to 40. 216

G.2 Enron case study behaviour measurements, rows 41 to 80. . . . 217

G.3 Enron case study behaviour measurements, rows 81 to 120. . . . 218

G.4 Enron case study behaviour measurements, rows 121 to 160. . . 219

G.5 Enron case study behaviour measurements, rows 161 to 200. . . 220

G.6 Enron case study behaviour measurements, rows 201 to 240. . . 221

G.7 Enron case study behaviour measurements, rows 241 to 280. . . 222

G.8 Enron case study behaviour measurements, rows 281 to 320. . . 223

G.9 Enron case study behaviour measurements, rows 321 to 360. . . 224

G.10 Enron case study behaviour measurements, rows 361 to 400. . . 225

G.11 Enron case study behaviour measurements, rows 401 to 440. . . 226

G.12 Enron case study behaviour measurements, rows 441 to 480. . . 227

G.13 Enron case study behaviour measurements, rows 481 to 520. . . 228

G.14 Enron case study behaviour measurements, rows 521 to 525. . . 229

Trademark Notice

The following are trademarks or registered trademarks of their respective companies:

MATLAB is a registered trademark of The Mathsworks, Inc.;

Preface

E-mail is an Internet application that has become a popular form of electronic communications, allowing people to quickly send messages to others and to distribute messages to large groups of people. However, there are certain individuals in society who use e-mail as a means for aiding with the conduct of illegal or criminal activities. Law enforcement agencies require ways of analysing a suspected individual's e-mail data, in order to extract and obtain important evidence linking the suspect with the occurrence of a crime.

The work in this thesis considers the problem of examining the e-mail data of a known suspected individual by analysing the traffic component of their communications, but not the content of their messages. This thesis proposes an approach, termed "computational intelligence", which utilises a combination of computational techniques to provide different perspectives on a suspected individual's communication behaviour. This thesis describes how computational intelligence can be used to investigate the overall changes in an individual's e-mail traffic behaviour patterns and also provide useful information that aids in understanding those changes in behaviour.

The work described in this thesis was conducted from February 2004 to January 2008 at the School of Engineering, University of Tasmania. The candidate's research supervision was provided by Prof. Michael Negnevitsky from the School of Engineering, University of Tasmania, and Mrs. Jacky Hartnett from the School of Computing, University of Tasmania.

Thesis Organisation

This thesis is organised into six chapters. The following is a brief description of each chapter:

- Chapter 1 provides an introduction to e-mail communications and overviews the range of problems associated with the use of e-mail by society. It then

describes the involvement of e-mail in illegal or criminal activities and how law enforcement agencies investigate electronic data for evidence of a suspected individual's involvement with a crime. The chapter then provides an overview of e-mail analysis methods that may be used to extract useful information from e-mail data and finally provides a statement of the problem considered for the research.

- Chapter 2 examines the different methods used for performing behaviour analysis of e-mail traffic and describes how these methods detect unusual or abnormal e-mail traffic behaviour. The chapter then highlights the main limitations associated with the existing e-mail traffic behaviour analysis methods.

- Chapter 3 describes the computational intelligence approach for analysing e-mail traffic behaviour. The chapter defines the meaning of "computational intelligence" and describes how the approach utilises a combination of computational techniques to analyse e-mail traffic behaviour. The chapter then describes the visualisation and feature extraction techniques used in the research for obtaining information about an individual's e-mail traffic behaviour patterns. The visualisation techniques described are time-series and social network visualisation, and the feature extraction techniques described are decision tree classification and hierarchical fuzzy inference.

- Chapter 4 describes the e-mail traffic analysis system developed for the research, which integrates each of the visualisation and feature extraction techniques described in Chapter 3. The chapter provides an overview of the e-mail traffic analysis system's architecture and how the system can be used to analyse the e-mail traffic behaviour patterns of suspected individuals. The chapter also describes the two types of e-mail traffic data used to evaluate the e-mail traffic analysis system, which are: simulated e-mail traffic data and data from the Enron e-mail corpus.

- Chapter 5 presents two sets of case studies that evaluate the e-mail traffic analysis system and demonstrate the use of computational intelligence. The first set of case studies evaluate the e-mail traffic analysis system using simulated e-mail traffic data, while the second set of case studies evaluate the system using the Enron e-mail corpus. The chapter then discusses and compares the results of both sets of case studies.

- Chapter 6 finally summarises the major contributions of the thesis and provides suggestions for further studies to extend the research work presented in this thesis.

Supporting Publications

There have been a number of journal and conference papers published during the course of the candidate's study. These publications are listed below in chronological order:

1. M. J.-H. Lim, M. Negnevitsky, and J. Hartnett, "Artificial Intelligence Applications for Analysis of E-mail Communication Activities," in *The 2nd International Conference on Artificial Intelligence in Science and Technology (AISAT 2004)*, M. Negnevitsky, Ed. Hobart, Tasmania, Australia: University of Tasmania, 2004, pp. 109 - 113.

2. M. Negnevitsky, M. J.-H. Lim, J. Hartnett, and L. Reznik, "Email Communications Analysis: How to Use Computational Intelligence Methods and Tools?," in *IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety*. Orlando, FL, USA: IEEE, 2005, pp. 16 - 23.

3. M. J. Lim, M. Negnevitsky, and J. Hartnett, "Tracking and Monitoring E-mail Traffic Activities of Criminal and Terrorist Organisations Using Visualisation Tools," in *6th Australian Information Warfare & Security Conference*, G. Pye and M. Warren, Eds. Geelong, Victoria, Australia: School of Information Systems, Deakin University, 2005, pp. 112 - 124.

4. M. J.-H. Lim, M. Negnevitsky, and J. Hartnett, "Personality Trait Based Simulation Model of the E-mail System," *International Journal of Network Security*, vol. 3, no. 2, pp. 164-182, 2006.

5. M. J.-H. Lim, M. Negnevitsky, and J. Hartnett, "E-mail Traffic Analysis Using Visualisation and Decision Trees," in *IEEE International Conference on Intelligence and Security Informatics*, ISI 2006, vol. 3975 / 2006, S. Mehrotra, D. D. Zeng, H. Chen, B. Thuraisingham, and F.-Y. Wang, Eds. San Diego, CA, USA: Springer Berlin / Heidelberg, 2006, pp. 680-681.

6. M. J. Lim, M. Negnevitsky, and J. Hartnett, “Tracking and Monitoring E-mail Traffic Activities of Criminal and Terrorist Organisations Using Visualisation Tools,” *Journal of Information Warfare*, vol. 5, no. 2, pp. 46 - 60, 2006.

7. M. J.-H. Lim, M. Negnevitsky, and J. Hartnett, “A Fuzzy Approach For Detecting Anomalous Behaviour in E-mail Traffic,” in *4th Australian Digital Forensics Conference*, C. Valli and A. Woodward, Eds. Perth, Western Australia: School of Computer and Information Science, Edith Cowan University, 2006, pp. 36 - 49.

8. M. J.-H. Lim, M. Negnevitsky, and J. Hartnett, “Detecting Abnormal Changes in E-mail Traffic Using Hierarchical Fuzzy Systems,” in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2007)*. Imperial College, London, UK: IEEE, 2007, pp. 1309-1314.

Glossary

Acronyms

CEO	Chief Executive Officer
COO	Chief Operating Officer
FERC	Federal Energy Regulatory Commission
FK	Foreign Key
ISP	Internet Service Provider
MIME	Multipurpose Internet Mail Extensions
NLP	Natural Language Processing
PK	Primary Key
RFC	Request for Comments
SEC	Securities and Exchange Commission
SMTP	Simple Mail Transfer Protocol
UML	Unified Modeling Language

Terms

Abnormal Behaviour	Behaviour that significantly deviates from typical or historically observed behaviour.
Analyst	Refers to an individual whose role is to analyse and to obtain useful information from data.
Anomaly Detection	A method commonly used in computer network intrusion detection, where it is used for detecting new types of intrusion attacks previously unknown to the computer system or computer network.

Artificial Intelligence	Computational techniques that perform tasks that would require “intelligence” if it were performed by humans.
Behaviour	Refers to the overt actions of an individual that can be directly observed by others.
Behaviour Comparison Analysis	Defined in this thesis as the comparison of two or more e-mail accounts, to determine the similarity or differences in communication behaviour.
Chapter 11 Bankruptcy Protection	A part of the United States Bankruptcy Code that allows companies to reorganise their business in order to recover from crippling debt.
Clique	Generally defined as a term referring to small clusters of individuals that frequently communicate with each other.
Clique Behaviour Analysis	Defined in this thesis as the analysis of small clusters of e-mail accounts to examine their group interaction behaviour.
Computational Intelligence	An approach for using a set of computational techniques, to extract information from data and present the information to the user/analyst in a useful and intelligent manner.
Computational Techniques	Defined in this thesis as techniques of extracting information from data in regard to the data properties.
Computer Forensics	A process that applies an array of tools and techniques to obtain forensic evidence from digital or electronic data.
Digital Forensics	Same meaning as “Computer Forensics”
Discrete-Event Simulation	A type simulation methodology whereby a system is modelled with state variables that change instantaneously at separate points in time. These points in time are ones where an “event” occurs.
E-mail Client	See “E-mail User”.

E-mail Content Analysis	Defined in this thesis as a type of e-mail analysis method that extracts information related to the textual content of messages sent by e-mail users.
E-mail Message Header	The initial section of an e-mail message that contains information such as the delivery route of the message, the time/date when the message was sent and received, the type of formatting used for the main body of the e-mail message.
E-mail Traffic	Information derived from the header section of e-mail messages, but not the content of e-mails (e.g. sender, receiver, and date/time information)
E-mail Traffic Analysis	Defined in this thesis as a type of e-mail analysis method that extracts information related to the transit and delivery of e-mail messages sent by e-mail users.
E-mail User	Refers to an individual who uses e-mail communications to communicate with others.
Feature Extraction Techniques	Computational techniques that extract and process information from data, to locate data records that possess particular features or patterns.
Individual Behaviour Analysis	Defined in this thesis as the analysis of an individual e-mail account's traffic communication behaviour.
Law Enforcement Agency	A department or organisation that performs the service of enforcing the law, to ensure that individuals are not conducting illegal activities.
Natural Language Processing	A field of artificial intelligence that focuses on analysing and understanding spoken or written language.
Network	A term often used to describe complex systems that consist of a number of interconnected and interacting components.

Network Behaviour Analysis	Defined in this thesis as the analysis of the relationships and connections between a large group of e-mail users, to extract information about their overall “network” behaviour.
Personality Trait Dimensions	A set of trait dimensions that describe particular aspects of an individual’s personality.
Replying Delay	The time delay between the sending of an e-mail message to a recipient and the recipient sending a reply message back to the sender.
Sending Delay	The time delay between the sending of one e-mail message and the sending of the next e-mail message.
Social Network	A system of agents that interact with each other and have some pattern of contact between them.
Text mining	An area where techniques are applied for extracting information from text documents in order to discover critical patterns or features describing the document’s contents.
Time-Series	A series of values where the quantity varies over time.
Traits	Words that describe the manner in which someone acts, thinks or feels.
Unusual Behaviour	Behaviour that seems different from what is normally expected.
User	Refers to the individual who uses the software application or software system.
Visualisation	A process that involves the transformation of data into graphical images.

Chapter 1

Introduction

1.1 E-mail Communications

Electronic mail or e-mail, is an Internet communications application that has transformed the way people can conduct person-to-person communications. E-mail originated during the 1970s, where the idea of computer-based messaging systems were being developed as experimental systems [6]. The current e-mail system originated from the US Ministry of Defense's ARPANET project and the ARPANET's e-mail proposals were published as the SMTP protocol RFC 821 [7] and RFC 822 [8] in 1982 [6, 9]. These were later updated by RFC 2821 [10] and RFC 2822 [11].

Following the early development of e-mail, there has been a rapid growth in the use and popularity of e-mail as an Internet communications application since the early 1990's [9]. The popular use of e-mail is due to a number of features that make e-mail a convenient method of communication over the Internet. In terms of composing a message and sending it, e-mail is easy to write, quick to send, and allows a single message to be sent to large groups of people. A feature that makes e-mail unique in comparison to other forms of person-to-person Internet communication is the asynchronous nature of using e-mail. Synchronous communication applications like instant messaging and Internet chat rooms for example, allow people to communicate while requiring both parties to be on-line simultaneously. E-mail on the other hand, is asynchronous in that it does not require both parties to be on-line at the same time and provides more time for individuals to compose their message before sending it. These features make e-mail a useful electronic communication medium in which individuals can easily send and read messages at their own convenience.

Apart from the features of e-mail just described, there are other factors that have been attributed to the rapid growth in e-mail usage by society. One of these is the introduction of graphical based e-mail client software, which enables the user to easily compose, manage, and read e-mail messages. Graphical based client software, such as Eudora or Microsoft Outlook, provided a more interactive interface compared to the command based interfaces that were originally used by early e-mail clients on UNIX systems [12]. Another factor was the introduction of web-based e-mail clients [12], which means that an individual's e-mail messages can be checked from any computer with connection to the Internet. This helped to simplify access to e-mail accounts by requiring only a web browser to view and manage e-mail messages, thus making it another convenient method of accessing e-mail accounts. Other factors that have been attributed to the rapid growth of e-mail are: development of multimedia encoding systems such as MIME (Multi-purpose Internet Mail Extension), enabling the sending of formatted documents, images, sound, web pages, and small computer applications [12]; and the setup of mailing list servers, which enabled the formation of discussion lists, allowing users to conduct informal discussions about topics of interest [12]. Each of these factors have led to making e-mail more accessible to society, hence leading to its now popular and widespread use.

Problems Associated With E-mail Communications

While e-mail has transformed the way individuals can communicate with large numbers of people and over vast distances, there are problems that have emerged as a result of wider e-mail use by society. Firstly, spam e-mail (unsolicited e-mail) and e-mail based computer viruses have been a major problem, and have been the subject for research into methods of detecting or preventing the propagation of spam e-mails (e.g. [13, 14, 15, 16]) or virus e-mails (e.g. [3, 17, 18]). Likewise, there have also been problems with e-mails that that try to solicit confidential information (e.g. credit card details) from e-mail users by falsely posing as a trusted agent [19]. These types of e-mails are known as "phishing" e-mails and have also been the subject of research into methods of detecting these types of fake messages [19].

Another type of problem associated with e-mail is the management of large amounts of e-mail messages received by individuals. The sheer number of e-mail messages present in people's inboxes often causes a burden for e-mail users when trying to search through or sort their e-mail messages. Methods are currently being investigated for assisting the user with sorting and organising their

e-mails, given the increasing amount of e-mail messages received by individuals on a day-to-day basis [18, 20, 21]. The problem that is the subject for this research is the examination of e-mail data belonging to suspected individuals who use e-mail as a means for conducting illegal or criminal activities.

1.2 Involvement Of E-mail In Illegal Activities

While e-mail may be used in association with lawful activities, there are certain individuals who use e-mail communications as a means in which to help plan, coordinate, or conduct illegal activities. The possible use of e-mail by such individuals is due to a number of features of e-mail, which make it a convenient method of communication to aid with illegal activities. Firstly, e-mail has a global reach, meaning that it is easy for an individual to communicate with another person located anywhere around the world. Due to this, there are generally no restrictions that would prevent an individual from using e-mail to communicate across different countries. Secondly, using e-mail is inexpensive since there are no costs for the user to send any number of e-mail messages to other people. This makes it a cheap and cost effective method for individuals involved in illegal activities, since there are no added financial costs for using the technology in comparison to other electronic communication technologies such as mobile phones, where there is usually a charge associated with using the service. Thirdly, e-mail is ideal for direct group communication since it allows the addresses of the intended recipients to be specified as part of the e-mail message header or placed as part of a mailing list. This makes it very quick to send a message directly to a particular group of people. Finally, e-mail allows an individual to remain anonymous since a false name or false identity can be used when subscribing to new e-mail accounts. This allows the individual's true identity to remain hidden while using e-mail. Thus, the combinations of these features make e-mail a convenient communication medium that could be used to aid certain individuals with conducting illegal activities.

Given the possible use of e-mail to aid with illegal activities, there are various types of illegal or criminal activities in which an individual may use computers or information technology, such as e-mail, to aid in performing unlawful acts. These types of activities may be grouped into two categories. The first category is where computers are used as a communications tool to help conduct the crime [22]. The crimes that usually fall into this category are "traditional" crimes, for example: stalking, child sex exploitation, identity theft, financial fraud, and

corporate espionage, which can also be conducted off-line without involving the use of a computer. The crimes that fall into this category may involve the use of e-mail, for example in a case where e-mail has been used to distribute child pornography [23].

The second category is where computers are used as targets for the illegal activity [22]. In this particular category, computer systems are targeted for the purpose of either acquiring information stored on a system without authorization, altering the data stored on the system, or to interfere with the availability of the computer or server. Again, e-mail may be involved in these types of crimes, for example where an e-mail server is attacked by being overwhelmed by an unusually large number of e-mail messages, causing it to stop functioning properly [24].

In either situation where computers or information technology is involved with the conduct of illegal or criminal activity, there are usually traces of data left on the computer system that leaves a record of the activity that has occurred. These traces of data can be used as important evidence to prove that a suspected individual has been involved with the conduct of unlawful activity. The use of such evidence is required by law enforcement agencies when investigating cases of illegal or criminal activity.

1.3 Investigation of Suspected Individuals

In the criminal justice process, law enforcement agencies are involved in the role of investigating cases of illegal or criminal activities [25, 26]. As part of investigations, physical or electronic evidence is usually obtained for analysis to determine whether particular suspects have been associated with committing criminal offences. After investigation, law enforcement agencies then decide whether a case should be followed on by further actions. These actions may involve either an arrest of the suspect or summoning the suspect before a magistrate or judge, and then laying criminal charges against the suspect [25].

During investigations by law enforcement agencies, a process called “forensics” is used to collect, extract, and analyse evidence associated with a case of illegal or criminal activity. Forensics applies various scientific methods or techniques to help reconstruct the series of events that have occurred in relation to a crime [22, 27]. The purpose of the process is to obtain a more complete understanding of the unlawful activity that has occurred, in order to link a suspect to a crime using the available trace evidence [22].

1.3.1 Computer Forensics

In investigations where electronic evidence is involved, law enforcement agencies rely on the processes of "forensic computing" or "computer forensics" to examine the activities of suspected individuals. This process draws upon methods and techniques from a range of disciplines (e.g. software engineering, cryptography, electronic engineering, and data communications), and involves the application of information technology to search for electronic evidence [28]. Computer forensics has been defined by [28] as: "*the process of identifying, preserving, analysing and presenting digital evidence in a manner that is legally acceptable*".

Computer forensics comprises of four key elements, as described by [28]:

- **Identification of digital evidence** - This is the first step in the forensic process, which involves identifying what type of evidence is present and knowing what type of information is stored on a computer system or electronic device.
- **Preservation of digital evidence** - This is a crucial part of the forensic process, whereby any examination of the electronic data is to be carried out in the least intrusive manner (i.e. minimising changes to the original data). This is important given the likelihood of judicial scrutiny of the evidence in a court of law. Any alteration of the data that is of evidential value must be accounted for and justified.
- **Analysis of digital evidence** - This step involves the extraction, processing and interpretation of digital data. The product of the analysis step is information interpreted from the data that is understandable by humans.
- **Presentation of digital evidence** - The final step of the forensic process is the actual presentation of the evidence in a court of law.

Computer forensics is becoming an increasingly important part of forensic examinations carried out by law enforcement agencies. This is because of the increasing pervasiveness of computers and information technology into people's daily lives and also people's increasing reliance on the Internet for various services [29]. Due to this, law enforcement agencies have to be able to deal with the increasing body of electronic evidence associated with investigations.

E-mail Investigations

For investigation of crimes involving e-mail, it firstly needs to be determined by computer forensic investigators whether e-mail use was involved in the illegal or criminal activity concerned [30]. For example, in a possible case of child pornography distribution, it needs to be determined whether the suspected individual used e-mail to send and share child pornography images. This is an important step to consider, given there needs to be a clear reason for examining e-mail data to use it as evidence to prove a suspect's involvement in a crime. Once it has been determined that e-mail use was involved, the next step in the investigation is for investigators to obtain access to the computers or electronic devices that store the e-mail data [30].

There are a variety of places where e-mail data may be stored. E-mail data may reside on the computers or electronic devices used by the suspect, for example: their desktop computer, laptop computer, or mobile phone. Other places where e-mail data may be stored is on the Internet Service Provider's e-mail servers. These are possible sources of e-mail data storage that law enforcement agencies may need to search in order to access, obtain, and seize e-mail data as evidence of the crime. After seizure and preservation of the original e-mail data, the next step in the investigation is analysis of the e-mail data.

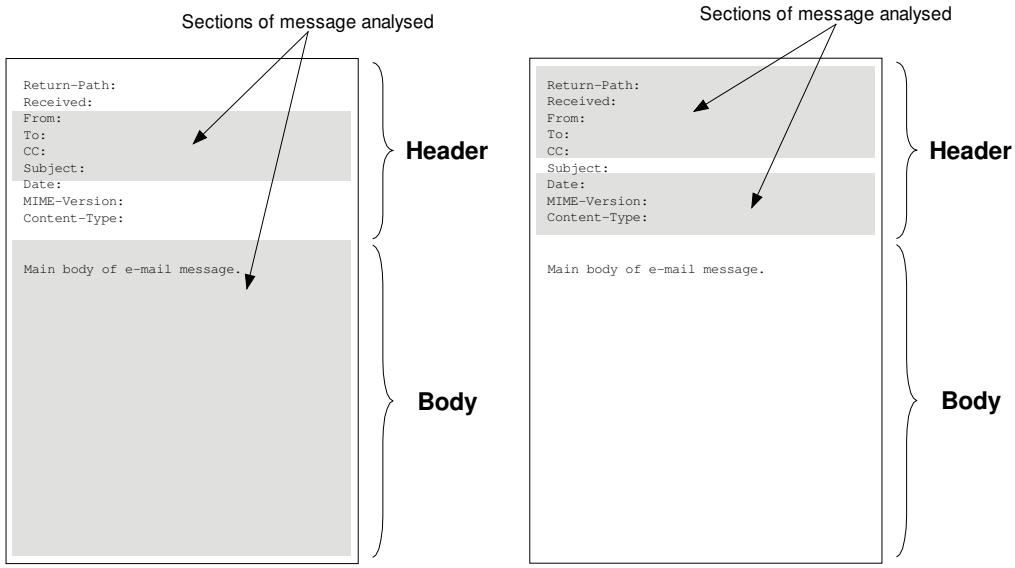
There have been only few publications that describe the exact tools or techniques used by law enforcement agencies to analyse e-mail data. This is due to the fact that law enforcement agencies avoid the disclosure of methods used for computer forensic analysis [31]. What is known however, is that there are a wide variety of commercial "off-the-shelf" software products that may be used to perform forensic analysis of e-mail data [30].

Despite this, it is still important to consider the tools and techniques used to analyse e-mail data, given that there is still a need by law enforcement to develop better tools and techniques to help analyse electronic data [31]. Also, there is a need for methods that enable law enforcement investigators to search through massive amounts of data to locate vital evidence associated with a crime [31]. Thus, this is where one can consider the different types of e-mail analysis methods available that may be utilised for examining a suspected individual's e-mail data.

1.4 E-mail Analysis Methods

E-mail data in its original text format contains a great deal of information that can be examined to provide useful knowledge about patterns of communication occurring between e-mail users. However, human analysis of e-mail data in its original text format is an extremely difficult and time consuming task, given the large amounts of data that may be involved. This is where e-mail analysis methods can be utilised for the examination of e-mail data, in order to aid in extracting information on meaningful patterns or features hidden in the data.

There are two types of e-mail analysis methods that may be employed to examine e-mail data for computer forensic purposes: e-mail content analysis and e-mail traffic analysis. Both of these methods provide different ways of analysing the data, and extract different types of information about the communication activities of individuals using e-mail. These methods also use different parts of an e-mail message for analysis, as shown in Figure 1.1.



(a) Parts of message used for e-mail content analysis.

(b) Parts of message used for e-mail traffic analysis.

Figure 1.1: The parts of an e-mail message that are used for content and traffic analysis.

1.4.1 E-mail Content Analysis

E-mail content analysis is defined here as a type of analysis method that extracts information related to the textual content of messages sent by e-mail users. The purpose of this form of analysis is to obtain useful information about the patterns

or features associated with the use of words entered into e-mail messages. There are various parts of an e-mail message that can be examined for content analysis, such as the *Subject*, *To*, *From*, and *Body* fields [32, 33, 34]. Each of the fields used for e-mail content analysis are highlighted in Figure 1.1(a).

Based on the text information contained in each e-mail message, e-mail content analysis techniques are able to analyse for particular patterns or features related to the linguistic characteristics of the text used in the e-mail messages. The techniques used to extract this information come from areas such as text mining [35], natural language processing (NLP) [36], or statistics. Techniques in each of these areas provide different ways of examining the content-based information contained in e-mail. Some brief examples of approaches that have been proposed for analysing e-mail content are: analysis of user topic interests [37], analysis of an individual's word usage patterns [38], message thread analysis [39], and authorship analysis [40].

The use of e-mail content analysis for computer forensic analysis of e-mail data offers a number of advantages compared to human analysis of individual e-mail messages. One of these is that the techniques used for e-mail content analysis enable the forensic analyst to identify patterns of word usage or particular linguistic features in e-mail messages, which would otherwise be unnoticed by the analyst. Another advantage is that content analysis offers a method of summarising or overviewing the textual content of e-mail messages without requiring the analyst to examine each individual e-mail message.

However, there are certain drawbacks that limit the use of e-mail content analysis. Firstly, if the body of certain e-mail messages is encrypted, then content analysis techniques will be unable to extract any information about what is contained in the main body of those messages. This defeats the purpose of e-mail content analysis and provides another problem for the analyst in terms of needing to break the encryption if the text content of the encrypted message is to be analysed. Another drawback is that content analysis techniques mainly focus on the textual content of e-mails, but provide little information about other characteristics such as time-based information or the presence of attachments. This limits the use of e-mail content analysis from analysing other types characteristics that may be present in e-mail data. While these are drawbacks limit the use of e-mail content analysis in these situations, e-mail traffic analysis is another type of method that can be considered for analysing e-mail data.

1.4.2 E-mail Traffic Analysis

E-mail traffic analysis is defined here as a type of analysis method that extracts information related to the transit and delivery of e-mail messages sent by e-mail users. The main purpose of e-mail traffic analysis is to obtain useful information about the exchange of messages between e-mail users and the types of messages sent, without examining the content of the e-mail messages. The information examined for traffic analysis is derived from the header section of e-mail messages, given that the header contains fields related to the transport of e-mail [9].

The header fields that can generally be used for traffic analysis are those related to the sender, recipient, and date/time information [4] (e.g. *From*, *To*, *CC*, *BCC*, and *Date* fields). Other header fields such as *Content-Type*, may be used for traffic analysis to provide other types of information about the e-mail messages (e.g. the presence of attachments or the types of attachments being sent) [3]. A header field that is generally not used for traffic analysis the *Subject* field, since it provides user specified information related the content of e-mail messages. Examples of header fields that can be used for traffic analysis are illustrated and highlighted in Figure 1.1(b).

Using the information contained in e-mail message headers, traffic analysis techniques are able to analyse for patterns or features associated with the communication interactions occurring between e-mail users. The techniques that can be used for e-mail traffic analysis come from a variety of areas, including: statistics, visualisation, and artificial intelligence. Techniques from each of these areas enable particular types of traffic information to be obtained about the individuals using e-mail. Some brief examples of types of information that can be extracted from e-mail using traffic analysis methods are: the usage patterns of individual e-mail users [41, 42], the interaction behaviour between an e-mail user and their contacts [2], and the connections emerging from interactions between groups of e-mail users [43, 44].

The application of e-mail traffic analysis for computer forensic purposes offers a number of benefits in comparison to human analysis of individual e-mail messages and e-mail content analysis. One of these benefits is that e-mail traffic analysis provides a way of overviewing the interactions occurring between e-mail users and enable the observation of events that occurred as a result of e-mail exchanges between e-mail users. This may provide useful information for forensic analysts, given that they need to obtain information from the data that allows the reconstruction of events associated with a crime. The second benefit

of e-mail traffic analysis is that it enables the identification of patterns of interaction occurring between e-mail users, which would otherwise be unnoticed by the analyst when examining individual e-mail messages. Finally, in comparison to e-mail content analysis, e-mail traffic analysis techniques are not hindered by the presence of encrypted e-mails, since traffic analysis focuses on the transit and delivery of e-mail messages rather than its contents.

Although there are benefits of using e-mail traffic analysis techniques, there are also some problems that make it difficult to apply traffic analysis techniques in certain situations. The first of these is the use of fake or falsely generated e-mail addresses by certain individuals, in order to hide their true identity and the origin of where their e-mails are sent from. This can be done by the individual sending e-mails through anonymous remailers, which are able to conceal the sender's true identity from the recipient [45]. The use of such tactics can cause problems for e-mail traffic analysis techniques, since it is dependent on knowing the identity of the sender and recipient addresses, and also how e-mail messages are being transported through the Internet. The second problem is knowing whether certain e-mail accounts belong to a particular individual. While traffic analysis techniques are able to examine the patterns of communications associated with interactions between e-mail users, it is still difficult to determine whether certain e-mail accounts are associated with a particular individual. This problem is due to the fact that an individual may use multiple e-mail accounts and also by the fact that the e-mail addresses used by individuals do not necessarily represent the individual's real identity. It is still yet to be determined whether e-mail traffic analysis techniques are able to accurately identify a particular individual from their e-mail traffic communication activity.

Overall, while e-mail content analysis and e-mail traffic analysis provide different methods of analysing e-mail data, the use of these methods may not necessarily have to be mutually exclusive. E-mail content analysis techniques may be used alongside e-mail traffic analysis techniques, in order to provide wider coverage about the communication activities of the individuals using e-mail. This approach may prove to be useful, given that the combination of both analysis methods could be used to overcome limitations that exist when using these methods individually. However, to determine the practical application of e-mail content and e-mail traffic analysis methods for computer forensic purposes, one needs to consider the current difficulties faced by law enforcement agencies in analysing e-mail data.

1.5 Difficulties Associated With Analysing E-mail Data

Law enforcement agencies currently face a number of challenges associated with the investigation of electronic evidence [46]. The reports by [29, 31] highlight a range of problems related to the investigation of crimes involving the use of computers and information technology. Some of the problems identified by [29, 31] are particularly relevant to the analysis of e-mail communications data. Each of these problems are described below.

Knowing Where To Look For Evidence

Firstly, there is the problem of sifting through large amounts of data to search for evidence indicating the occurrence of illegal or criminal activities, hence the proverbial term: "finding the needle in the haystack". The large amounts of electronic data collected for investigation makes it a difficult and time consuming task to analyse the data for evidence of unlawful activities. Knowing where to begin searching for vital evidence of such activities is a task which may become insurmountable for investigators [31].

Anonymity and Traceability

Another problem is the ability of individuals to remain anonymous and difficult to trace when using the on-line environment [29]. On the Internet, individuals can use false names or provide false identification to sign up for new e-mail accounts (e.g. signing up for free web-based e-mail accounts such as Hotmail or Yahoo Mail). Along with the use of false information for e-mail accounts, individuals may also use multiple e-mail accounts, which makes it difficult to determine how many e-mail accounts are used by a particular individual and to confirm the e-mail accounts that belong to the same individual. In terms of traceability, it is also possible that some individuals may use anonymous remailer servers to hide the source and destination addresses of their e-mail messages, making it difficult to trace the path travelled by those messages [45, 47]. Thus, these problems associated with anonymity and traceability makes it difficult to accurately identify and locate possible suspects in the data.

Encryption of Data

One of the major difficulties faced by law enforcement is the possibility of electronic data and communications being encrypted, meaning that offenders can keep information secure from outside perusal [29, 31]. This makes it difficult for law enforcement agencies to examine electronic evidence, given that this can hinder investigation efforts due to the additional effort and cost associated with cracking the encryption [29]. Encryption is a problem relevant to e-mail communications, given that e-mail messages can be sent as encrypted messages [48]. In the case of e-mail encryption, the message content is scrambled to prevent people other than the sender and intended recipient from reading the e-mail message.

Development of Better Tools to Analyse The Data

The final problem relevant to the analysis of e-mail data is the availability of tools that aid with analysing electronic evidence. It has been noted by [31] that law enforcement require the need to develop or obtain their own tools and techniques in order to better deal with the difficulties associated with investigating electronic evidence. The use of cutting-edge technology by law enforcement for investigations may prove to be valuable for uncovering useful information about suspected individuals.

1.6 Problem Statement

This research considers the use of e-mail traffic analysis to examine the e-mail communication activities of a suspected individual. The purpose of the research is to determine how to aid the user or analyst in examining a suspect's overall changes in e-mail traffic communication behaviour, given the situation where the identity and e-mail addresses of the suspect is already known. The objective of the research is to focus on investigating analysis techniques that can provide useful information about a suspect's changes in communication behaviour, based on the examination of their e-mail traffic data. The scope of this research is to consider how traffic analysis techniques can be applied during the analysis stage of the computer forensic process. Other stages of the computer forensic process: *Identification of digital evidence*, *Preservation of digital evidence*, and *Presentation of digital evidence*, while still important, are considered to be outside the scope of this research.

In addition to objective set out for the research, this research is to also consider the following problems:

- *Analysing large data sets:* Given the large amounts of e-mail data usually required to be analysed, how does one determine where to start the analysis? What would be the best method of approaching this?
- *Making the analysis tractable:* From the large amounts of e-mail data available, what is the best way of narrowing down the search? What approach can be used to narrow down the amount of data analysed?
- *Finding particular patterns or features in the data:* What type of patterns or features could be used for finding and locating a suspect's change in e-mail traffic communication behaviour? What will these patterns or features tell the user or analyst about the individual being investigated?
- *Making sense of patterns, relationships, or features present in the data:* What type of analysis techniques can be used to help the user/analyst make sense of hidden patterns, relationships, or features present in the e-mail data?
- *Helping the user/analyst do their investigation:* What is the best approach for allowing the user to carry out their investigation?

1.7 Thesis Contribution and Layout

This thesis proposes an approach, termed "computational intelligence", which utilises a combination of computational techniques to provide different viewpoints on a suspected individual's communication behaviour. This thesis describes how computational intelligence can be used in e-mail traffic analysis to provide useful information and aid with understanding a suspected individual's overall changes in communication behaviour. The remainder of this thesis is set out as follows:

Chapter 2 describes the existing methods used for analysing e-mail traffic behaviour and describes how these methods detect occurrences of unusual or abnormal communication behaviour.

Chapter 3 describes the computational intelligence approach and explains the purpose of utilising a combination of computational techniques for analysing e-

mail traffic behaviour. The chapter then describes the visualisation and feature extraction techniques used in the research for examining e-mail traffic behaviour.

Chapter 4 describes the e-mail traffic analysis system developed for the research and how it integrates the visualisation and feature extraction techniques to provide different perspectives on a suspect's e-mail traffic behaviour. The chapter then describes the e-mail traffic data used for evaluating the e-mail traffic analysis system and how the e-mail traffic analysis system can be used to analyse the data.

Chapter 5 presents two sets of case studies that evaluate the e-mail traffic analysis system. The first set of case studies uses simulated e-mail traffic data to evaluate the e-mail traffic analysis system, while the second set of case studies uses data from the Enron e-mail corpus.

Finally, Chapter 6 summarises the major contributions of the thesis and discusses suggestions for further studies to extend the research work presented in this thesis.

Chapter 2

Behaviour Analysis of E-mail Traffic

2.1 Introduction

Behaviour is a term that is used to refer to the overt actions of an individual that can be directly observed by others [49]. Actions such as driving a car or using the Internet, are examples of an individual's behaviour that be observed by other people. In the analysis of e-mail traffic communications, behaviour relates to a wide variety of observable actions made by individuals or groups of individuals communicating via e-mail. These actions performed by individuals or groups of individuals can reveal information about their patterns of communication as well as information about the level of communication activity occurring between e-mail users.

The purpose of this chapter is to review the behaviour analysis methods that have been used for studying e-mail traffic communication behaviour. The behaviour analysis methods reviewed in this chapter generally relate to methods that detect unusual or abnormal communication behaviour, since such methods may have useful applications for law enforcement work. The chapter firstly describes behaviour analysis and how e-mail traffic behaviour has been studied from different levels of analysis. This is then followed by a description of the behaviour analysis methods that have been used for analysing e-mail traffic behaviour, which are grouped into four categories: *Individual Behaviour Analysis*, *Behaviour Comparison Analysis*, *Clique Behaviour Analysis*, and *Network Behaviour Analysis*. The final part of the chapter then discusses the current limitations with the behaviour analysis methods used for analysing e-mail traffic behaviour.

2.2 Behaviour Analysis Methods

The purpose of performing behaviour analysis of e-mail traffic is to extract and reveal useful information about the communication behaviour of individuals, through the examination of their e-mail traffic data. The type of behaviour information that can be extracted from the data varies, depending on the purpose of the analysis and also on the type of method used to extract the information. Examples of behaviour information that may be extracted from e-mail traffic are: how an individual generally behaves through the use of their e-mail account, how an individual communicates with particular associates, or how a set of individuals communicate as a group. The extraction of such information may be useful for law enforcement purposes, since it may aid with understanding the communication habits of suspected individuals, the relationships between different suspects, and how the suspects' communication activities change over time.

There has been a variety of methods proposed for performing behaviour analysis of e-mail traffic, each of which examine e-mail traffic behaviour from different perspectives. While some of the methods proposed have only been demonstrated for detecting abnormal e-mail traffic behaviour caused by e-mail based computer viruses [2, 3] or spam (i.e. unsolicited e-mail messages) [18], certain information provided by these methods may be useful for law enforcement purposes. The behaviour analysis methods covered in this chapter generally relate to those methods that are able to detect unusual or abnormal e-mail traffic behaviour. The detection of such types of behaviour may be useful for law enforcement applications, since it may aid in signifying the presence of possible criminal or terrorist activities. Such information may also help in predicting the possibility of an upcoming criminal or terrorist event and identify potential suspects.

It has been noted that the current methods used for behaviour analysis of e-mail traffic can be divided into categories according to the "level of analysis" performed by each method. The idea behind this categorisation is that the analysis of e-mail traffic behaviour can be considered based on the level of detail provided on the e-mail accounts being examined. Some methods such as those mentioned in Section 2.2.1 provide a great level of detail about the behaviour of specific e-mail accounts, while other methods such as those mentioned in Section 2.2.4 provide a much broader overview of the behaviour of several e-mail accounts. To conceptualise the types of analyses performed by each of the current behaviour analysis methods, the diagram in Figure 2.1 was drawn to determine the different levels of analysis provided by current behaviour analysis methods.

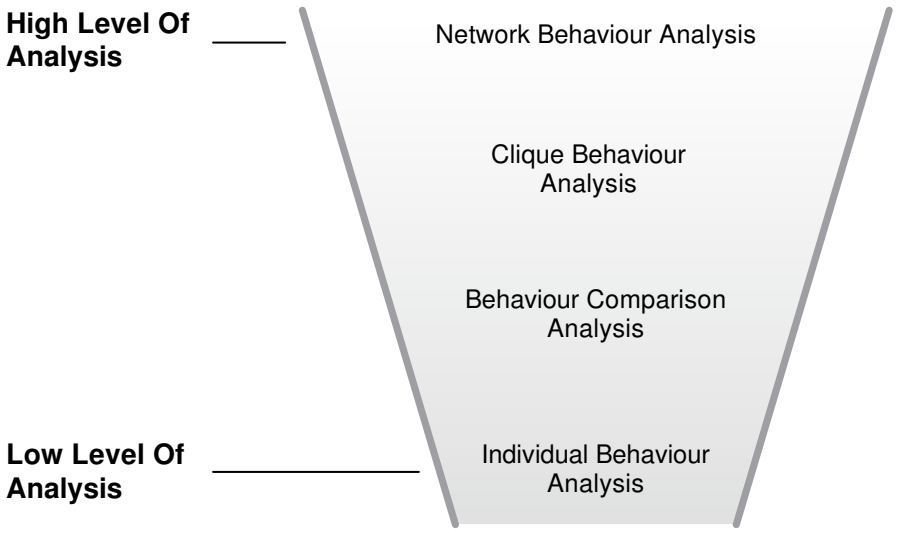


Figure 2.1: Different levels of analysis for examining e-mail traffic behaviour.

Each of the categories shown in Figure 2.1 relate to the level of detail provided by the current behaviour analysis methods. At the lowest level, *Individual Behaviour Analysis* considers how individual e-mail accounts behave in terms of providing details on how particular e-mail accounts are being used by their owner, and how those e-mail accounts are used to communicate with certain associates. At the highest level, *Network Behaviour Analysis* considers a much higher level overview in terms of how multiple e-mail accounts are connected to each other, and the significance of the connections between particular e-mail accounts. In between these two levels are *Behaviour Comparison Analysis* and *Clique Behaviour Analysis*, where *Behaviour Comparison Analysis* considers the similarities or differences in behaviour of multiple e-mail accounts and *Clique Behaviour Analysis* considers the behaviour of small groups or clusters of e-mail users that frequently communicate with each other.

Although this is not the only way of reviewing the current behaviour analysis methods, this perspective may allow one to consider whether there is a relationship between the information provided at different levels of analysis. The best analogy for understanding the "levels of abstraction" perspective would be details revealed by satellite images viewing the Earth from space. At the highest level one can see a layout of all of the continents and countries on the Earth. As one begins to zoom in, details are gradually revealed about the cities, roads, and buildings. The "levels of analysis" perspective used here for categorising behaviour analysis methods follows this analogy.

2.2.1 Individual Behaviour Analysis

Individual behaviour analysis is defined here as the analysis of an individual e-mail account's traffic communication behaviour. At this level of analysis, the focus is on the extraction of information about the usage behaviour of a particular e-mail account, by considering an enclosed view of the behaviour involving that e-mail account. The diagram in Figure 2.2 illustrates the idea of performing individual behaviour analysis. This approach of analysing e-mail traffic allows behaviour information to be extracted that reveals details on how an e-mail account is used by their owner and how the owner interacts with their associates.

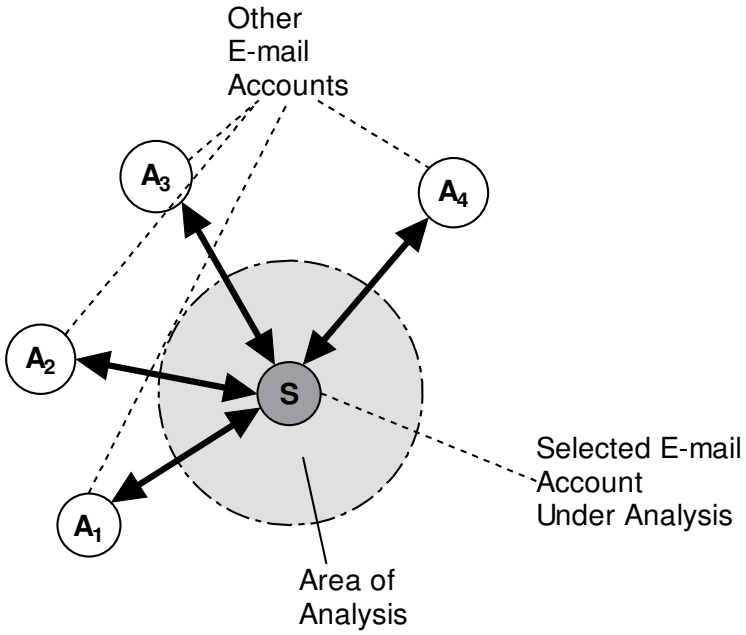


Figure 2.2: Diagram illustrating the idea of individual behaviour analysis.

There has been two methods proposed for individual behaviour analysis, each of which considers different types of communication patterns for detecting abnormal behaviour. The first method by [18, 41, 42], called the "Sending Usage Model" [18], considers the behaviour of an individual by examining the times of the day that the individual is most likely to send e-mail messages. The basis for this method is the assumption that an individual develops a particular habit for sending e-mail messages at particular times of the day, which can be measured by computing statistics from the individual's e-mail traffic data. The computation of these statistics involves examining how many e-mails are sent during each hour of the day and then computing a 24-bin histogram of the average number of messages sent during each hour. The resulting histogram from the statistical analysis of the individual's e-mail traffic behaviour then presents a profile of the

individual's sending usage patterns, like the example shown in Figure 2.3.

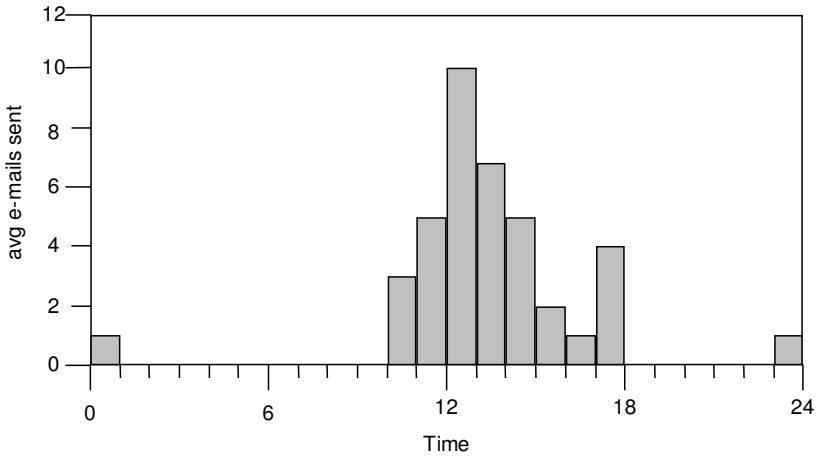


Figure 2.3: Example of a user's 24-hour sending usage profile.

The second method by [2, 18, 42], called "Recipient Frequencies", considers the frequency of communication between an individual and each of their recipients. The idea behind this method is that an individual has certain recipients or associates that they will communicate with on a frequent basis, while there are other associates that the individual will communicate with on a less frequent basis. Through this process, the individual is assumed to develop a particular pattern of behaviour based on how frequently they communicate with each of their associates. Based on this idea, an individual's pattern of communication can be measured by examining their e-mail traffic data for the frequency of communication with each associate. To measure the individual's pattern of communication, [2, 18, 42] computes a table of frequencies by examining the frequency of e-mail messages sent by the individual to particular associates. Each of the frequencies calculated is represented as a percentage, reflecting how often the individual communicates with a particular associate [2, 18, 42]. After computing the table of frequencies, the information can be sorted in descending order and displayed as a histogram. An example of the frequency table and recipient frequency histogram is shown in Figure 2.4, which is based on those shown by [2, 18, 42]. The information presented by the frequency table and recipient frequency histogram represents an individual's pattern of communication with their associates.

Detection of Abnormal Behaviour for Individual E-mail Accounts

When detecting abnormal e-mail traffic behaviour associated with individual e-mail accounts, both the "Sending Usage Model" and "Recipient Frequencies"

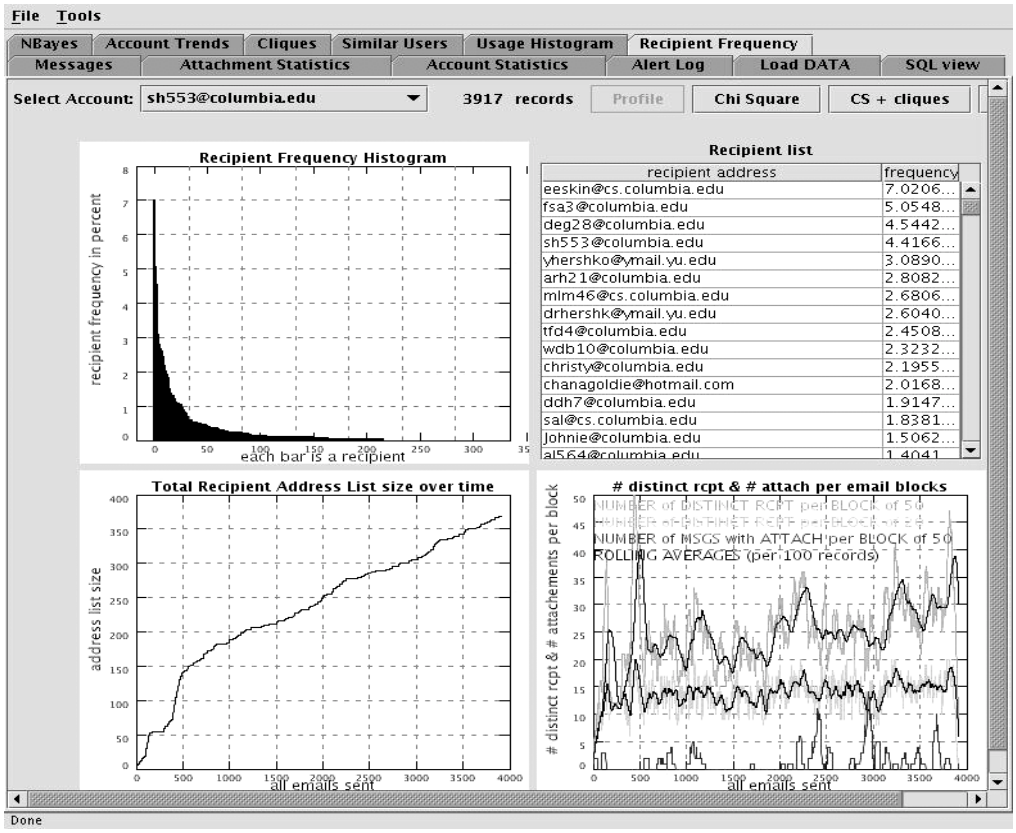


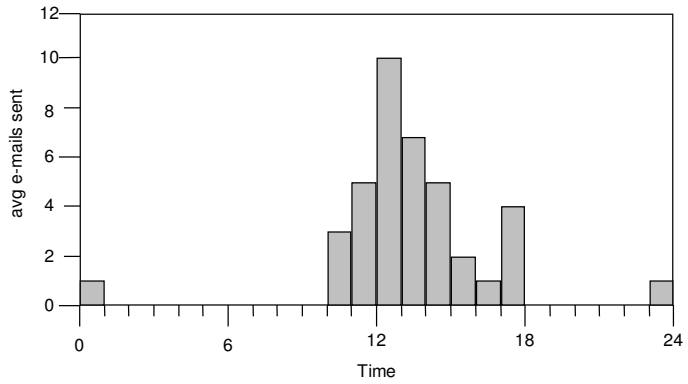
Figure 2.4: Example of the frequency table and histogram used for the Recipient Frequencies method [1], with kind permission of Springer Science and Business Media.

methods use approaches that involve examining how an individual's e-mail traffic behaviour changes over time. This is done in order to measure the amount of change in behaviour that has occurred and to also determine whether the amount of change detected is significant enough to indicate the presence of abnormal communication behaviour. Both methods however, use different approaches for notifying the user about the presence of abnormal behaviour.

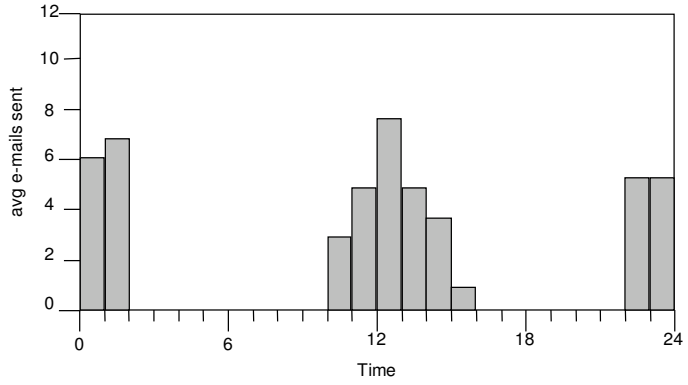
In the "Sending Usage Model" method [18, 41, 42] the approach used is to look at two different periods from an individual's e-mail archive and compare an individual's past e-mail usage with their recent e-mail usage. This is done to determine if the difference in the individual's sending usage patterns indicates the presence of an abnormal change in communication behaviour. Any significant difference in sending usage patterns found then indicates the presence of an abnormal change in communication behaviour.

To determine whether the individual has been exhibiting abnormal communication behaviour, [18, 41, 42] initially profiles the individual's past e-mail account usage by building a 24-bin histogram of the individual's typical sending usage

patterns. After profiling the individual's typical sending usage patterns, another 24-bin histogram is built from the individual's recent e-mail account usage, reflecting the individual's current or most recent sending usage patterns. Examples of an individual's past and recent usage histograms that is presented to the user is illustrated in Figure 2.5, which is based on the histogram profiles presented by [18, 41, 42].



(a) User's past sending usage patterns).



(b) User's recent sending usage patterns.

Figure 2.5: Comparison of the past and recent sending usage patterns for an individual.

Once the two histograms have been created, these are then compared by computing a distance measure to determine the differences between the histograms. The distance is computed by [18, 42] using the weighted Mahalanobis distance function:

$$D_4\left(h_1, h\right) = \sum_{i=0}^{n-1} \frac{w_i \left(h_1[i] - h[i]\right)}{\sigma_i} \tag{2.1}$$

$$\text{Weight } w_i = \frac{h_1[i]}{\sum_{j=0}^{n-1} h_1[j]} \tag{2.2}$$

where n is the number of bins in the histogram, vector h represents the histogram of the individual's profiled period, h_1 represents the histogram of the individual's

recent usage period, and σ_i is the variance around the arithmetic mean. If the distance between the histograms is found to be large, then an alert is generated to inform the user that the individual's sending usage patterns have exhibited an abnormal change in communication behaviour.

In the "Recipient Frequencies" method [2, 18, 42], a different approach is used for detecting abnormal behaviour by using a "rolling window" to find significant variations in an individual's pattern of communication with their associates. The basis for this rolling window approach is to scan through blocks of e-mail messages to look for points in time where there are fluctuations in the frequency of communication between an individual and their associates. Such fluctuations are assumed by [2, 18, 42] to be signs of where an individual changes their rate of e-mail message transmissions to their associates.

To detect abnormal behaviour, an initial block of e-mail messages is selected from the individual's e-mail archive by [2, 18, 42] to establish and profile the individual's "normal" frequency of communication with their associates. While the number of e-mail messages required for this initial profiling period is not strictly specified by [2, 18, 42], the example provided by [2] suggests that for a user consisting of 2500 outbound e-mail messages, the first 400 outbound messages are used for profiling the individual's recipient frequencies. After establishing the size and location of the profiling period, a scanning window is then specified (e.g. a scanning window of size 100 e-mail messages), which is used as the test period. The test and profiling periods are then compared with each other by examining the table of frequencies calculated from each period. The frequencies of are compared by [2, 18, 42] using the Hellinger distance function:

$$HD\left(f_p[], f_t[]\right) = \sum_{i=0}^{n-1} \left(\sqrt{f_p[i]} - \sqrt{f_t[i]} \right)^2 \tag{2.3}$$

where $f_p[]$ is the array of normalised frequencies for the profiling period, $f_t[]$ is the array of normalised frequencies for the testing period, and n is the total number of distinct recipients observed during both periods. The array of frequencies $f_p[]$ and $f_t[]$ are further defined by [2] as:

$$f_p[i] = N(i)_p / ws_p \tag{2.4}$$

$$f_t[i] = N(i)_t / ws_t \tag{2.5}$$

where ws_p is the Hellinger training window size, and ws_t is the Hellinger testing window size, $N(i)_p$ is the number of times that the current recipient of the e-mails appears during the profiling period window, and $N(i)_t$ is the number of times that the current recipient of the e-mails appears during the testing period window. This comparison of the frequencies between the profiling and test period windows is rolled forward along the individual's entire e-mail archive, shifting along by steps of one record at a time. During each step, the Hellinger Distance is calculated and is output onto a graph for the user to visualise the fluctuations in recipient frequencies. An example of the Hellinger Distance graph output to the user is shown in Figure 2.6. After the Hellinger Distance graph has been presented to the user, the user then uses the graph to search for significant fluctuations in the plot for indications of abnormal changes in behaviour. Points in the plot, such as peaks or troughs, are considered as possible indications of abnormal behaviour, since they indicate that a significant change in recipient frequencies has occurred.

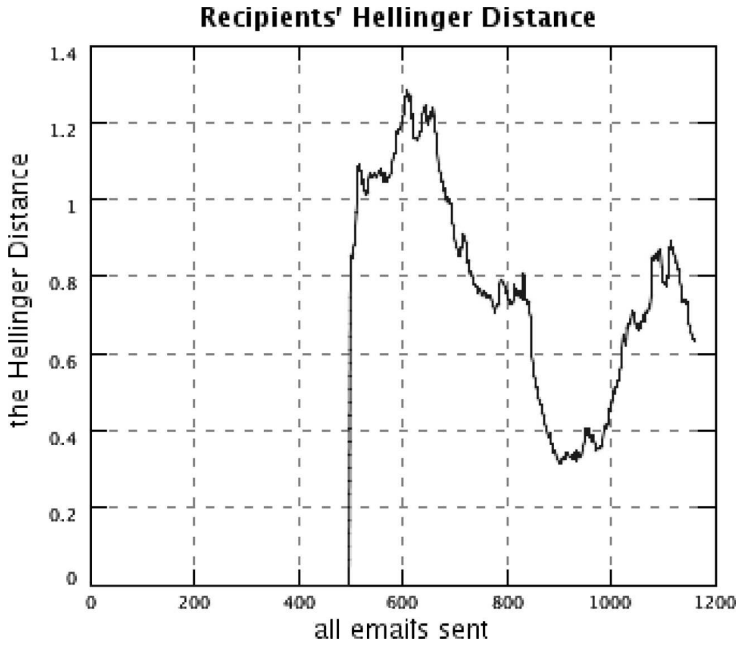


Figure 2.6: Example of the rolling window approach used to detect variations in recipient frequencies [2] ©ACM, Inc. Reprinted by permission.

Current Limitations Of Individual Behaviour Analysis

There is however, several drawbacks with the methods currently used for individual behaviour analysis. Firstly, there has been little consideration about allowing the user/analyst to view an individual e-mail account's abnormal behaviour

from different perspectives. In the “Sending Usage Model” method [18, 41, 42], the user is only allowed to search for abnormal behaviour based on viewing an individual e-mail account’s 24-hour sending patterns. There has been no consideration by [18, 41, 42] on viewing the same individual e-mail account using a different behaviour pattern, for example by comparing it with the “Recipient Frequencies” method. Viewing an individual e-mail account’s traffic behaviour from different perspectives, as illustrated in Figure 2.7, could provide the user/analyst a broader overview of a suspect’s e-mail account behaviour. This is because it may aid the user/analyst to verify the abnormal behaviour found by comparing two types of behaviour patterns to determine if both behaviour patterns detect abnormal changes in communication behaviour. This may be useful in confirming whether the abnormal behaviour patterns observed belongs to a particular type of individual that is required to be investigated.

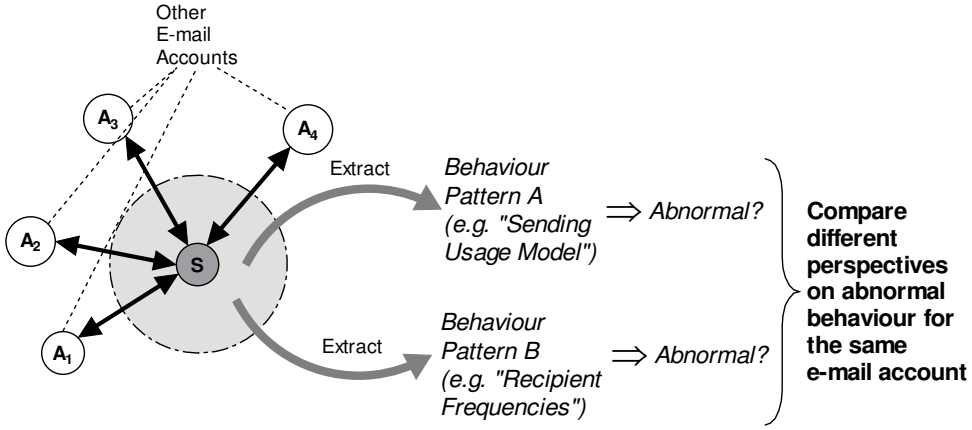


Figure 2.7: Observing abnormal behaviour from different perspectives.

Another drawback is that both the “Sending Usage Model” and the “Recipient Frequencies” methods lack the capability to present the user/analyst a summary of the abnormal behaviour observed for several e-mail accounts. The approach used by both of these methods only allows for the user/analyst to study a single e-mail account at a time to determine the presence of abnormal behaviour. For example, the “Sending Usage Model” method [18, 42] only allows the user to pick a single e-mail account to examine whether the past and recent 24-hour sending patterns have exhibited significant differences in behaviour . The same is also evident for the “Recipient Frequencies” method [2, 18, 42] where only a single e-mail account at a time can be analysed for abnormal behaviour. The disadvantage of analysing each e-mail account individually is that it may take the user/analyst a significant amount of time to search through a large collection of e-mail accounts (e.g. 100 e-mail accounts) in order to find something that

is “interesting” to the user/analyst. If a summary of the abnormal behaviour observed from a selection of e-mail accounts is listed to the user/analyst, then it would allow the user/analyst to more quickly select a particular e-mail account that is of interest for the investigation. An example of this summary concept is illustrated in Figure 2.8.

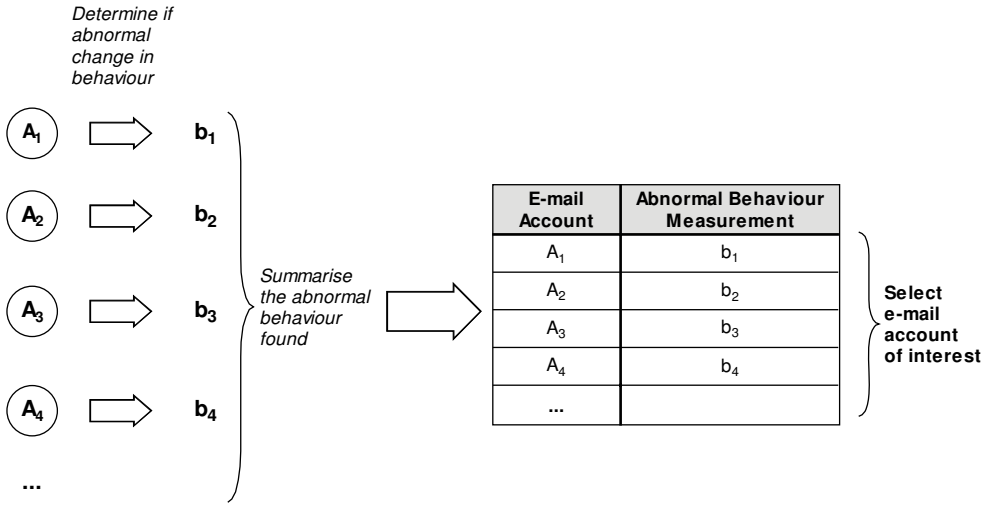


Figure 2.8: Example of how the abnormal behaviour of individual e-mail accounts may be summarised

2.2.2 Behaviour Comparison Analysis

In contrast to individual behaviour analysis, behaviour comparison analysis is defined here as the comparison of two or more e-mail accounts, to determine the similarity or differences in communication behaviour. The general concept for this type of analysis is that a number of behaviour measurements are extracted from each e-mail account, to establish and identify their typical communication behaviour patterns. The behaviour measurements obtained from each e-mail account are then compared with each other, to determine the similarity or differences in behaviour between particular e-mail accounts. E-mail accounts that are similar in behaviour share very similar communication patterns, while e-mail accounts that are different in behaviour exhibit communication patterns that can be distinguishable from other e-mail accounts. This type of analysis approach allows for the identification of e-mail accounts based on the comparison of each account’s e-mail traffic behaviour patterns. The diagram in Figure 2.9 illustrates the concept of behaviour comparison analysis.

For behaviour comparison analysis, there has been two methods proposed that each use different approaches for comparing e-mail traffic behaviour. The first

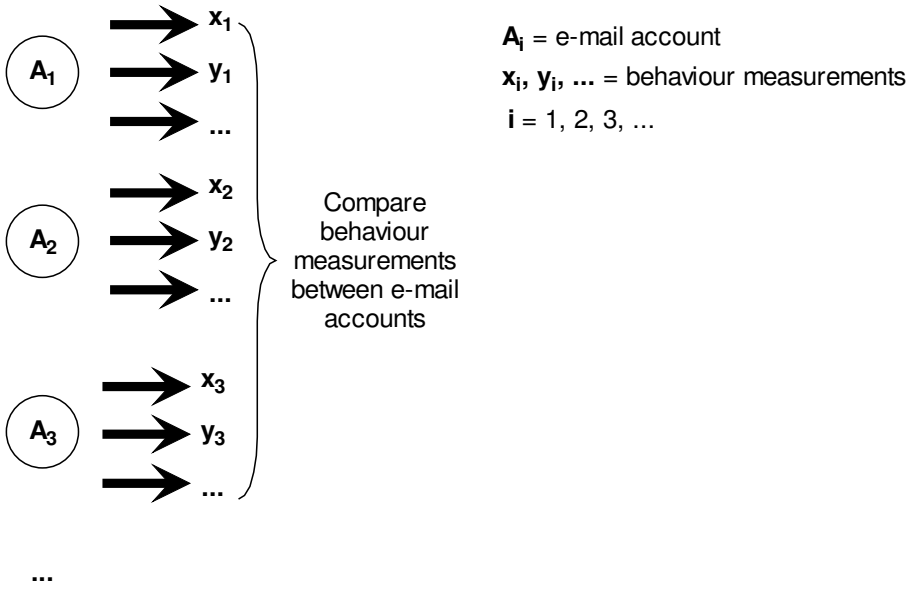


Figure 2.9: The concept of behaviour comparison analysis.

method by [18, 42], called the "Similar User Model", compares the hourly usage profiles for a group of e-mail accounts to a selected e-mail account. This comparison is made in order to determine how closely other e-mail accounts match the behaviour patterns of the selected e-mail account. The behaviour patterns used for comparing each e-mail account is based on examining the times of the day that each e-mail user is likely to send e-mail messages. The statistical approach used for obtaining this information is similar to the 24-bin sending usage histogram used for the "Sending Usage Model" method [18, 41, 42] described previously in Section 2.2.1. The difference however, is the addition of other metrics that can be used to build the 24-bin histogram of each e-mail account's usage patterns. In [18], the user is given the choice of selecting one of four types of behaviour patterns for comparing the hourly sending usage behaviour of e-mail accounts:

- The average number of e-mails sent out during each hour.
- The average size of e-mails sent out during each hour.
- The average number of attachments sent out during each hour.
- The average number of recipients of e-mails sent out during each hour.

Each of the above behaviour patterns can be used to profile each e-mail account's sending usage patterns, by sampling each e-mail account's traffic data using one

of the four measurements mentioned above. The result from the profiling of each e-mail account is similar to the histogram shown for the "Sending Usage Model" method in Figure 2.3.

The second method by [3], employs a different approach whereby a number of e-mail accounts are compared with each other by extracting several behaviour measurements from each e-mail account. The behaviour measurements used by [3] are defined as "behavioural features", where each behavioural feature is a statistical representation of some aspect of each e-mail user's activity or behaviour. Through the use of these behavioural features, [3] seeks to identify particular e-mail accounts share similar communication behaviour or exhibit abnormally different communication behaviour to other e-mail accounts.

Altogether there are 15 behavioural features defined by [3], which provide a broad spectrum of characteristics for comparing the communication behaviour of multiple e-mail accounts. The behavioural features used are sorted into two categories based on how the features are calculated from the archive of each e-mail account. The grouping of these 15 behavioural features is shown as a concept map in Figure 2.10, which is based on the groupings originally described by [3]. The first category of features, called "Per-Email Features", are numerical values that are calculated on a per e-mail message basis. The "Per-Email Features" category consists of two subgroups, where the "Single Email Multinomial-Valued Features" subgroup have values represented as bit strings consisting of one or more bits, and the "Per-Email Continuous Features" subgroup have values represented by positive integers (i.e. counting values). In the "Features Calculated Over a Sending Window" category, the features are calculated over a sending window consisting of x number of e-mail messages. The typical number of messages used for the sending window is the e-mail user's last twenty e-mail messages [3]. All of these behavioural features are used by [3] to establish the communication behaviour characteristics of each e-mail account being analysed.

Finding E-mail Accounts with Similar or Abnormal Behaviour

The comparison methodology used by the "Similar User Model" [18, 42] method and the "behavioural features" method [3], each use different approaches for comparing behaviour and finding particular e-mail accounts based on the compared behaviour. The difference between the approaches used is due to the number of e-mail accounts that are compared against each other. In the "Similar User Model" method [18, 42], a many-to-one approach is used whereby the behaviour

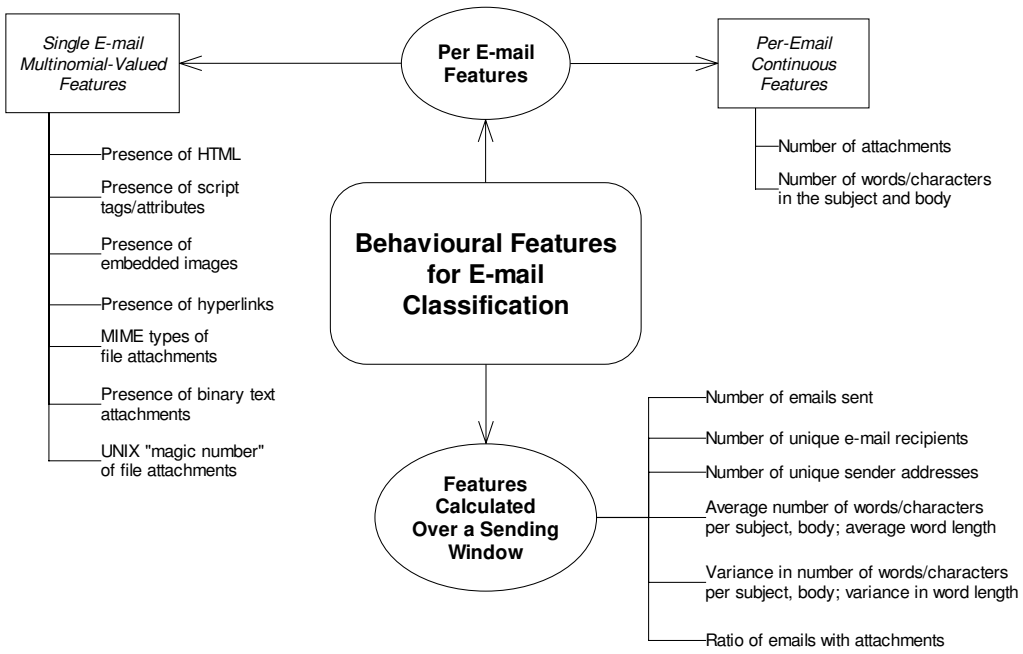
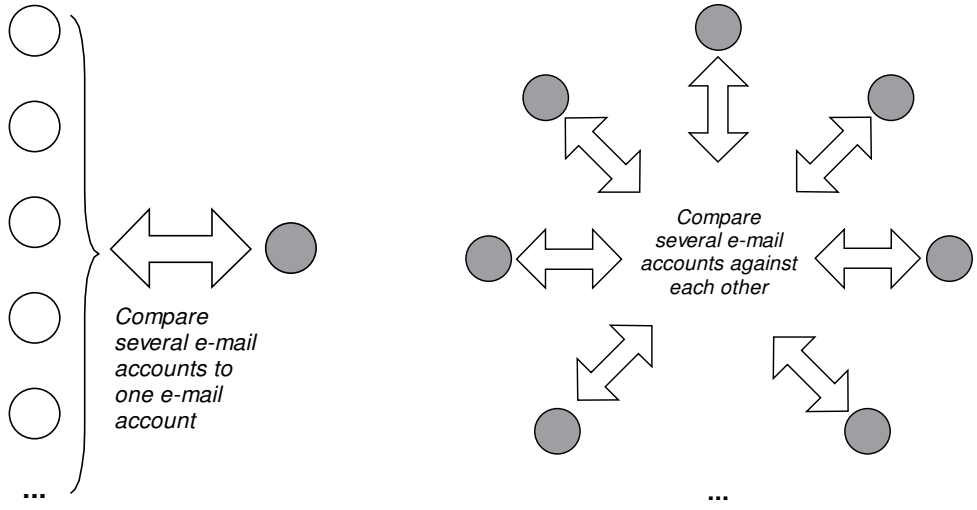


Figure 2.10: Concept map of the 15 behavioural features defined by [3].

patterns for a number of e-mail accounts are compared to that of a selected e-mail account. A diagram of this many-to-one approach is illustrated in Figure 2.11(a). The "behavioural features" method by [3] instead uses a many-to-many approach whereby the behavioural features for several e-mail accounts are compared against each other. This many-to-many comparison approach is illustrated in Figure 2.11(b), showing how it compares to the many-to-one approach in Figure 2.11(a).

To locate particular e-mail accounts based on the comparison of their behaviour, each of the current behaviour comparison analysis methods use comparison techniques that are related to the comparison methodology used. In the "Similar User Model" method [18, 42], the hourly usage patterns of several e-mail accounts are compared to that of a selected e-mail account through the choice of one of four histogram comparison functions. The four histogram comparison functions used by [18] are: simplified histogram intersection L1-form, histogram Euclidean distance L2-form [50], histogram quadratic distance [50], and the Kolmogorov-Smirnov test [50]. The formulas used for these histogram comparison functions are shown in Eqs. (2.6) - (2.10), where: h_1, h_2 are a pair of histograms, n is the number of bins in the histogram, i is the bin number, $D_1\left(h_1, h_2\right)$ is the simplified histogram intersection L1-form, $D_2\left(h_1, h_2\right)$ is the histogram Euclidean distance L2-form, $D_3\left(h_1, h_2\right)$ is the histogram quadratic distance, A is a matrix, $D_{KS}\left(h_1, h_2\right)$ is the result used for the Kolmogorov-Smirnov test, N is the total



(a) Many-to-one comparison.

(b) Many-to-many comparison.

Figure 2.11: Concept of the comparison approaches used.

number of samples, and $n(x)$ signifies the number of points less than x. The variable x is ordered from smallest to largest.

$$D_1(h_1, h_2) = \sum_{i=0}^{n-1} |h_1[i] - h_2[i]| \tag{2.6}$$

$$D_2(h_1, h_2) = \sum_{i=0}^{n-1} (h_1[i] - h_2[i])^2 \tag{2.7}$$

$$D_3(h_1, h_2) = (h_1[i] - h_2[i])^T A (h_1[i] - h_2[i]) \tag{2.8}$$

$$D_{KS}(h_1, h_2) = \max_{x \subseteq X} (|F_{h_1}(x) - F_{h_2}(x)|) \tag{2.9}$$

$$F_h(x) = \frac{n(x)}{N} \tag{2.10}$$

The user is given the choice by [18] to select one of Eqs. (2.6) - (2.9) to compute the behavioural distance between the group of e-mail accounts and the selected e-mail account. The chosen equation is then used to compute the histogram distance between each e-mail account and the selected account, and a number is output signifying the differences in behaviour. Numbers close to zero indicate very little difference in behaviour and numbers greater than zero indicate a larger difference in behaviour. Based on the use of the usage histogram and histogram distance functions, the “Similar User Model” method is able to identify similarly

behaving e-mail accounts.

The "behavioural features" method [3] on the other hand, uses two types of techniques for comparing e-mail traffic behaviour. These techniques relate to finding similar behaviour and finding abnormally behaving e-mail accounts. The first technique used by [3], is to use maximal covariance analysis [51] to find e-mail accounts that share similar behavioural features. The output of the maximal covariance analysis was represented as a gray scale cell chart where each row represents a behaviour feature and each column represents individual e-mail accounts. This grayscale cell chart was then used to provide an overview of all e-mail accounts and to search for e-mail accounts that shared similar behaviour feature values based on the colour of matching cells.

The second set of techniques used by [3] are Support Vector Machines (SVM) and Naïve Bayes classification, which were used to train models that classify the difference between worm behaviour and "normal" e-mail account behaviour. The SVM and Naïve Bayes were trained using a combination of the 15 behavioural features defined by [3] to identify worm e-mail traffic behaviour as "abnormal" behaviour. The performance accuracy of the trained models were adjusted by changing the number of behavioural features used by SVM and Naïve Bayes. It was found from their results that a low number of behavioural features caused the SVM and Naïve Bayes models to produce a high rate of false positives (false alarms) due to lack of generality caused by less features, while a high number of behavioural features caused the SVM and Naïve Bayes models to produce a high rate of false negatives (missed alarms) due to addition of extraneous features.

Current Limitations of Behaviour Comparison Analysis

While the existing methods proposed for behaviour comparison analysis [3, 18, 41, 42] provide different ways of comparing the traffic behaviour of multiple e-mail accounts, these methods have certain limitations. Firstly, there seems to be a lack of in-depth understanding about the relationship between different behaviour measurement variables and the type of e-mail traffic behaviour they are useful for identifying. In the work by [3], a feature selection process was applied to rank the behavioural features or behaviour measurement variables most appropriate for identifying the difference between a normal e-mail account and virus-infected e-mail accounts. It was found through the process that the top three ranking behavioural features: ratio of e-mails with attachments, binary attachment, and MIME type application/octet-stream, were related to the presence

of attachments. These were noted to be redundant features since those three features did not provide any additional information about the e-mail traffic behaviour of computer viruses. This shows that although one can define a set of behavioural features to examine e-mail traffic behaviour, further understanding is required for knowing how particular behavioural features are related to the type of traffic behaviour to be identified. This may be important in helping to determine what behavioural features are the most useful for detecting certain types of e-mail traffic behaviour, such as computer virus behaviour or a known type of criminal behaviour.

Another limitation of the current work is that it is not well understood how many behaviour measurement variables may be required for comparing behaviour and accurately identifying e-mail accounts that exhibit a particular type of behaviour pattern. Both of the methods proposed by [18, 42] and [3], either use a single behaviour measurement [18] or up to 15 behaviour measurements [3] for comparing behaviour. It is not well understood whether one behaviour measurement variable provides enough information to compare and identify the behaviour of e-mail accounts. Also, it is not well understood whether 15 behaviour measurement variables is over-redundant in the information provided. Careful consideration is needed for deciding whether a particular behaviour measurement variable is useful for the investigation task and how many behaviour measurement variables are required to provide sufficient information about the behaviour of suspicious or abnormal e-mail accounts.

2.2.3 Clique Behaviour Analysis

While there are a number of definitions regarding the term “clique” [52], it can generally be defined as a term referring to small clusters of individuals that frequently communicate with each other [18]. In clique behaviour analysis, it is defined here as the analysis of small clusters of e-mail accounts, to examine their group interaction behaviour. At this level of analysis, the focus is on understanding the common group communication behaviour exhibited amongst several e-mail accounts, rather than the behaviour of individual e-mail accounts. The diagram in Figure 2.12 illustrates the idea of clique behaviour analysis.

In the work by [2, 18, 42], a method called “User Cliques” is proposed for examining an individual’s e-mail archive for unusual clique behaviour. The basis for the “User Cliques” method is that an individual may often send e-mail messages to different groups of recipients, based on how those recipients relate to

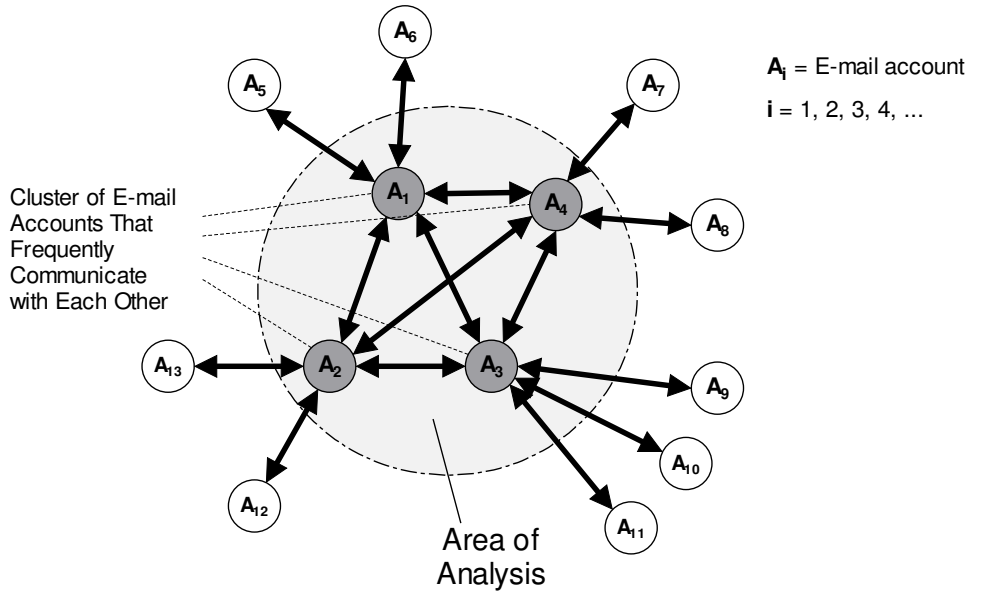


Figure 2.12: Diagram illustrating the idea of clique behaviour analysis.

the individual. For example, the individual may send a certain set of messages to friends, another set of messages to work colleagues, and another set of messages to relatives. In most cases, it may be unlikely that the individual will distribute the same set of e-mail messages to different groups of recipients (e.g. the individual does not send the same messages to work colleagues and relatives). Based on this, it is assumed by [2, 18, 42] that there will be different social cliques that can be found by examining how an individual sends e-mail messages to certain sets of recipients.

To detect unusual clique behaviour, [2, 18, 42] searches for e-mail traffic behaviour that violates an individual's typical clique e-mailing behaviour. This is performed by establishing an individual's normal clique e-mailing behaviour by examining the recipient list of each e-mail message sent from the individual's e-mail account (i.e. taken from the TO, CC, and BCC fields). The list of recipients found from each e-mail message is then used to build models of the different cliques that the individual sends messages to. The individual's cliques are profiled by using a rule defined by [2], which specifies that a set of recipients cannot be subsumed by another set of recipients, when determining whether the recipients form a separate clique. For example, for an e-mail account U consisting of recipient sets: {A,B}, {A,B,C,D}, and {D,E,F}, the first set will be subsumed by the second set, meaning that there will be two cliques where: $Clique_1 = \{A, B, C, D\}$ and $Clique_2 = \{D, E, F\}$. Figure 2.13 shows a visual representation of how these cliques are profiled, which is based on the clique

diagrams originally drawn by [2].

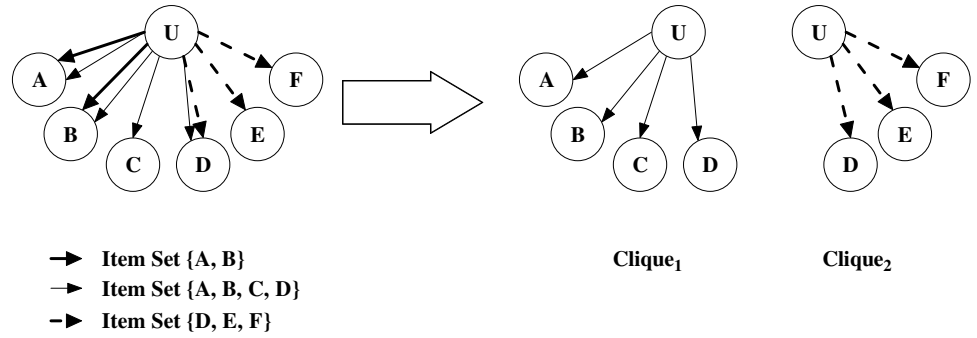


Figure 2.13: Diagram of clique profiling, based on the clique diagrams originally drawn by [2].

After establishing the individual's normal cliques, the "User Cliques" method detects unusual e-mail traffic behaviour by scanning new or recent outgoing e-mail messages for recipient lists that violates any of the cliques previously profiled. If an outgoing e-mail message has a recipient list that is not a subset of any previously profiled clique, then that e-mail message is deemed to have caused a violation. This information can then be used to alarm the user/analyst that the individual has started communicating with new sets of recipients.

Although the "User Cliques" method proposed by [2, 18, 42] is able to analyse the cliques present in an individual's e-mail archive, it suffers from drawbacks related to what it is able to analyse. A major problem is that the "User Cliques" method only works well for cases where the individual under analysis rarely or occasionally communicates with new recipients. For cases where the individual under analysis is a new e-mail account user or the individual constantly communicates with new recipients, it was noted by [2] that it may be possible that too many alarms would be generated during these situations. Such cases may render the user cliques method useless since a large number of alarms may not necessarily be useful for the user/analyst.

Another limitation of the "User Cliques" method is that it does not examine the variations in behaviour of the existing cliques. In [2], the user cliques method has only been shown to be useful for detecting the appearance of new cliques. However, the problem with the method proposed by [2] is that it does not provide any information about the dynamics of the existing cliques or whether the cliques originally profiled still continue to communicate with each other. Information on the variations in traffic behaviour of the existing cliques may be useful for understanding the level of activity exhibited by particular clusters of e-mail accounts.

2.2.4 Network Behaviour Analysis

The term "network" is often used to describe complex systems that consist of a number of interconnected and interacting components [53]. Examples of complex systems that can be described as "networks" are [54]: the World Wide Web, an electrical power grid, or a river system. In relation to analysing e-mail traffic behaviour, a large group of e-mail users (e.g. more than 20 e-mail users) may be treated as a network since the connections and relationships between the e-mail users are often complex to analyse. For network behaviour analysis, it is defined here as the analysis of the relationships and connections between a large group of e-mail users, to extract information about their overall "network" behaviour. The purpose of this type of analysis is to obtain information about the behaviour for a large group of e-mail users, by taking into account the communication ties or connections between particular e-mail users and also the type of interaction that occurs between each member of the group. At this level of analysis, the focus is on understanding the "network" properties for the large group of e-mail users, rather than identifying the small clustering of e-mail users as previously described for clique behaviour analysis in Section 2.2.3. The diagram in Figure 2.14 illustrates the idea of network behaviour analysis.

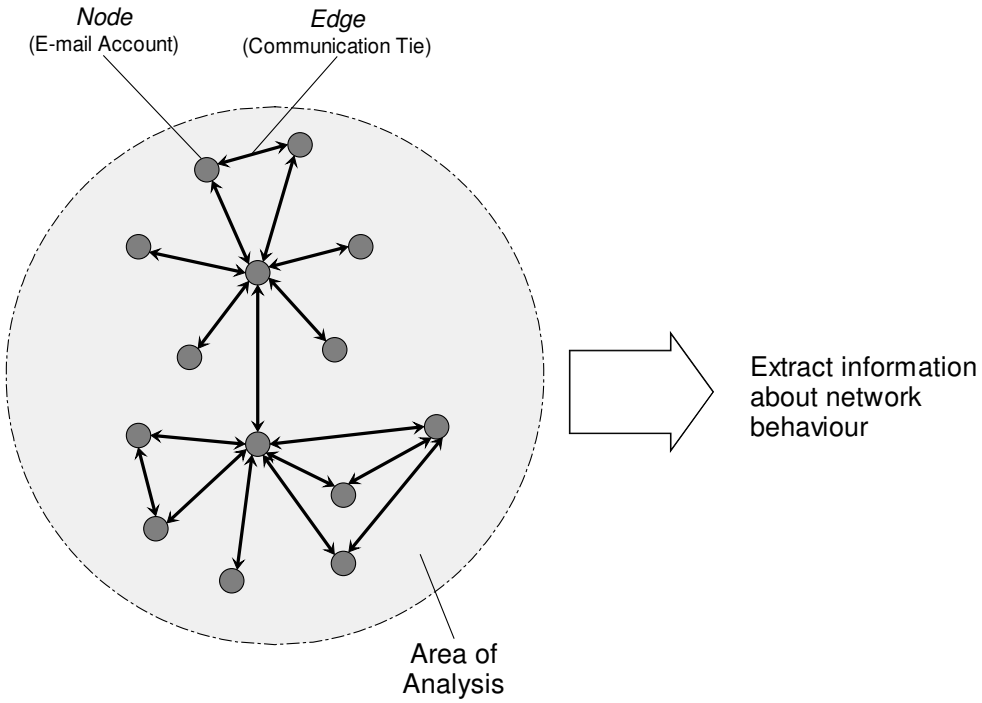


Figure 2.14: The concept of network behaviour analysis.

There has been a number of methods proposed and applied to analysing networks

of e-mail users (also referred to as “e-mail social networks”). These network behaviour analysis methods can be subdivided into two categories of analyses: methods that examine the static properties of e-mail social networks, and methods that examine the dynamic properties of e-mail social networks. For methods that analyse the static properties of e-mail social networks, these methods work from a “snapshot” of the network of e-mail users, by capturing the connections and interactions occurring between e-mail users over a particular period of time. This network “snapshot” is then used for the analysis and is used to obtain network information about the connections and relationships between e-mail users. Examples of work using the static method of analysis are: [44, 55, 56, 57, 58].

Although the static method of analysis is useful for providing a glimpse of the properties of an e-mail social network, what it lacks is the ability to provide information on how the network changes over time. This is important to consider given that:

- The communication between e-mail users may vary in intensity over time.
- The communication ties or connections between certain e-mail users may not always be persistent.
- New connections may form between previously unconnected e-mail users.

To take such factors into account, dynamic methods of network behaviour analysis are used to capture additional information about the temporal aspects of e-mail social networks, as well as the connection and relationship information. There are two dynamic methods of network behaviour analysis that may be considered useful for law enforcement applications.

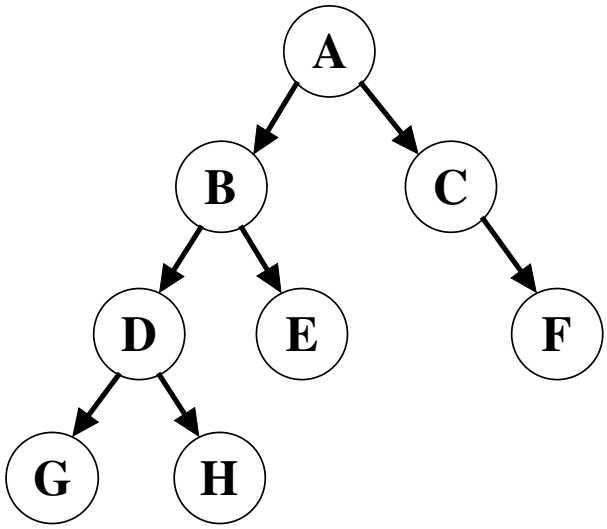
Detecting Hidden Group Structure

The first of these is the work by [4], which searches a network of e-mail users for hidden group structure relating to a “chain of communication” between the e-mail users. The basis of the method proposed by [4] is that a hidden group structure can be found from a network of e-mail users, by examining how messages are passed on from one group member to another. The connections between each group member involved in the message passing forms a “chain of communication” that shows how the members of the hidden group are connected. Based on this “chain of communication”, the members of the hidden group can be identified from a network of e-mail users.

To illustrate the method proposed by [4], Figure 2.15 shows how messages may be propagated through each member of the hidden group. In Figure 2.15(a), the list shown illustrates the timestamp and address information recorded from the group members' e-mail traffic logs. This list of timestamps and addresses shows how messages may be passed from one individual to another with some delay between each message. Figure 2.15(a) also displays two waves of communication, indicated in bold and italic, showing how messages may be passed along as overlapping waves through the chain of communication links. Using the available timestamp and address information from the e-mail traffic data, it is assumed by [4] that a hidden group structure can be found from the chain of communications. An example of the hidden group structure found from Figure 2.15(a) is shown in Figure 2.15(b).

00 **A→B**
00 **A→C**
04 **B→D**
05 **B→E**
08 *A→B*
09 **C→F**
10 *A→C*
12 **D→G**
12 **D→H**
13 *B→D*
13 *B→E*
13 *C→F*
15 *D→G*
15 *D→H*

(a) Hidden group communication showing timestamp and addresses.



(b) Hidden group structure.

Figure 2.15: Diagram showing the hidden group's e-mail traffic communication and structure, based on the diagrams originally drawn by [4].

To detect the hidden group structure, [4] proposed an algorithm that examines the e-mail traffic data for evidence of message passing. The algorithm is dependent on the threshold parameters specified for the minimum and maximum propagation delay for messages to be passed on between e-mail users. The propagation delay defines the amount of delay that occurs before a message is passed down the chain (e.g. $A \rightarrow C$ causes $C \rightarrow F$ to occur within a specified time frame). The use of the propagation delay parameters determine which individuals will be added to the communication chain structure, identifying the members of the

hidden group.

Detection of Abnormal Hot Spots of Activity

The second method proposed for detecting unusual network behaviour is the work by [59]. This method scans a network of e-mail users for "hot spots" of abnormal bursts of activity, based on the measurements calculated from three statistical measures. The statistical measures used by [59] each compute statistics from each node of the network, to determine whether there are spikes of abnormal network activity.

The detection technique used by [59] is called "scan statistics". It is based upon the concept of a moving window of analysis where local statistical metrics are computed for each window of time. In the case of the network of e-mail users examined by [59], scan statistics were used to compute statistical metrics based on the neighbourhood of connections surrounding each individual in the network. The result from the use of scan statistics was that [59] was able to perform a temporal analysis of the activity surrounding each individual in the network and used a threshold to determine whether there were spikes of activity indicating abnormal behaviour. If the scan statistics for an individual were found to exceed the threshold, the individual was identified as causing abnormal activity in the network.

The scan statistics approach used by [59] was applied to a case study involving the Enron e-mail corpus (more information about the Enron e-mail corpus will be provided in Section 4.3.2). It was found by [59] that the use of scan statistics were able to locate individuals in the network of e-mail users whom were causing abnormal spikes of activity. Further investigation of the suspected individuals by [59] found very unusual behaviour, such as the case of aliasing where an individual switched to using a different e-mail account, causing a spike of unusual e-mail traffic activity in the network. The work by [59] demonstrate the usefulness of a network level view of analysis for detecting abnormal e-mail traffic behaviour caused by suspect e-mail accounts.

Other Considerations For Network Behaviour Analysis

Both of the network behaviour analysis methods proposed by [4] and [59] demonstrate that they are useful for finding unusual connections between e-mail users and locating suspect e-mail accounts exhibiting abnormal e-mail traffic activity.

However, even though these methods are able to locate suspiciously behaving e-mail accounts, other methods of behaviour analysis may still be required to provide further information for a user or analyst. It was shown by [59] that scan statistics was useful for locating abnormal activity in the network, but it was also noted by [59] that scan statistics can only assist with identifying potentially interesting activity in the network. This means that further investigation would be required to understand the cause of the observed abnormal network activity, as illustrated in Figure 2.16. Based on such information from the network level of analysis, this indicates that other behaviour analysis methods, such as those from clique behaviour analysis, behaviour comparison analysis, or individual behaviour analysis, may be useful for providing further details about the abnormal behaviour observed.

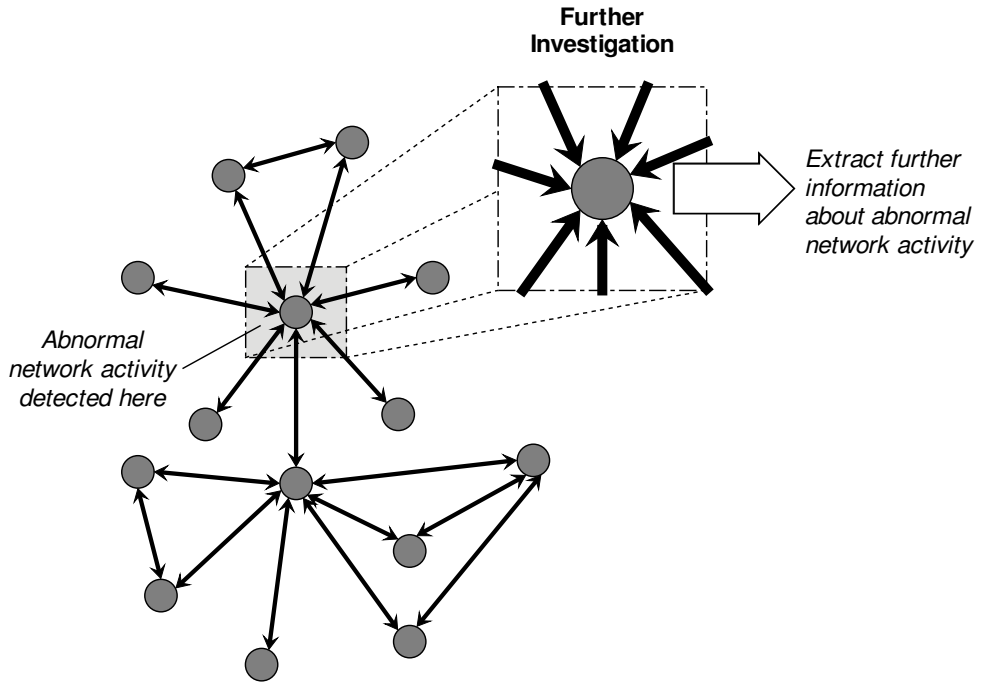


Figure 2.16: Further investigation of abnormal network activity.

2.3 Discussion

After reviewing the relevant work related to behaviour analysis of e-mail traffic, it has been observed that there are certain limitations with the existing work that require important consideration. This section discusses some of the current limitations for behaviour analysis of e-mail traffic, which will be addressed in various parts of this thesis.

Integrating Different Levels of Analysis

In Section 2.2, an overview was provided on the various behaviour analysis methods used for examining e-mail traffic behaviour. Although each of the methods described covered different levels of analysis, there has been little consideration about the integration of different levels of analysis. This is important to consider since there are various aspects of e-mail traffic behaviour that may be examined when searching for unusual or suspicious communication behaviour.

The main benefit for integrating different levels of analysis is that it may provide a way for the user/analyst to conveniently zoom in and out of details in the data. This sort of approach may enable a user/analyst to utilise information available at particular levels of analysis in order to gain a quick overview of the data, for example: using network behaviour analysis or behaviour comparison analysis methods. After gaining an overview of the data, a user/analyst may then want to zoom into details about particular e-mail accounts by utilising a lower level analysis method (e.g. using individual behaviour analysis methods). Such an approach like this may prove to be useful for a user/analyst when investigating abnormal e-mail traffic activity.

The approach of zooming in and out of the data has not yet been properly demonstrated for behaviour analysis of e-mail traffic. In the work by [60], a location-based analysis system (i.e. using IP address and geospatial information) was developed that examines for abnormal e-mail communication activity related to SPAM e-mail messages. The work by [60] demonstrated how different location-based analysis methods were integrated for examining e-mail messages. It described how the user can utilise different located-based methods to zoom in on details related to unusual or abnormal activities of interest (e.g. using an map overview of geographical locations, then zooming in on specific details relating to volume of traffic originating from different countries). The work by [60] shows how integrating different levels of analysis works when analysing e-mail traffic from a location-based point of view.

However, for behaviour-based analysis of e-mail traffic the existing work does not clearly demonstrate the use of integrating different levels of analysis. In the work by [2, 18, 41, 42] a behaviour-based analysis system was developed for examining e-mail traffic, which uses a set of different behaviour-based analysis methods. Although [2, 18, 41, 42] describes each of the methods used for their system, it has not been clearly demonstrated how the user can zoom in and out of details through different levels of analysis. For example: whether the user

can view e-mail traffic behaviour from a clique behaviour analysis level using the "User Cliques Model" (described previously in Section 2.2.3) and then zoom in on a partcular e-mail account at an individual behaviour analysis level using the "Sending Usage Model". This shows that there is a current lack of consideration about the integration of different levels of analysis in order to allow the user/analyst to zoom in and out of details on e-mail traffic behaviour.

The approach of integrating different levels of analysis was considered for the research in this thesis when selecting the computational techniques to use for analysing e-mail traffic (covered in Chapter 3) and developing the system for analysing e-mail traffic (covered in Chapter 4). The use of the different analysis viewpoints, as well as levels of analysis, will be demonstrated in Section 5.2 for the evaluation of simulated e-mail traffic data and also in Section 5.3 for the Enron case study.

Understanding The Effect of Behaviour Measurement Variables

In Section 2.2.2, it was noted that the "behavioural features" work by [3] defined several types of behavioural features or behaviour measurement variables for examining the traffic behaviour of multiple e-mail accounts. What the work by [3] demonstrates is that there are a wide variety of behaviour measurements that can be defined to examine e-mail traffic behaviour. However, what is currently lacking is an in-depth understanding of how certain behaviour measurement variables are related to each other. The work by [3] covered a wide range of behaviour measurement variables, but it was not well understood what the exact relationships between each of the variables were, other than how the variables were calculated from the data. Another problem with the current methods is knowing what type of traffic behaviour each behaviour measurement variable is useful for identifying (e.g. is the behaviour measurement variable better for detecting e-mail viruses or the behaviour of a known type of criminal?). As noted previously in Section 2.2.2, the work by [3] shows that further work is required for understanding the appropriate types of behaviour measurement variables to use for accurately identifying particular types of traffic behaviour.

There are two approaches used for the research in this thesis to obtain a better understanding about the effect of behaviour measurement variables. The first approach, covered in Sections 4.3.1 and 5.2, involves creating a behaviour-based simulation model of e-mail clients and observing the effects of the clients' behaviour on the generated e-mail traffic. This simulation approach considers prob-

lem of understanding measurable behaviour from the reverse point of view by attempting to model behaviour and then determining how to measure the effects of particular behaviour parameters from the observed e-mail traffic. The second approach, covered in Sections 3.4.2 and 5.3, involves the use of nine behaviour measurement variables to measure the changes in e-mail traffic behaviour that are occurring for suspect e-mail accounts. These nine behaviour measurements are fused together, to determine the amount of overall change in e-mail traffic behaviour. This fusion approach is being used to better understand the relationships between particular behaviour measurement variables and the types of traffic behaviour they are useful for identifying.

Summarising Abnormal Behaviour For Individual Accounts

Previously in section 2.2.1, it was noted that the one of the current limitations for individual behaviour analysis was the lack of consideration about providing a summary of abnormal behaviour observed for each of the e-mail accounts under analysis. This is considered as important since the current individual behaviour analysis methods (i.e. "Sending Usage Model" [18, 41, 42] and "Recipient Frequencies" [2, 18, 42]) only enable the user/analyst to examine the details of individual e-mail accounts separately. These existing methods do not provide a slightly higher level summary of the observed abnormal behaviour for each of the individual e-mail accounts.

A summary of abnormal behaviour may be useful for the user/analyst, since it summarises information about the abnormal changes in behaviour that have occurred for each of the e-mail accounts. Such an approach may allow the user/analyst to select the e-mail account of interest, based on the abnormal behaviour information provided by the summary. The selected e-mail account can then be investigated by the user/analyst for further detail about its communication behaviour. The approach of summarising the abnormal behaviour observed from a number of individual e-mail accounts is covered in Section 3.4.2. The use of the summarised abnormal behaviour information is further described in Section 4.4 and then later demonstrated in the Enron case study, covered in Section 5.3.

Placing the User/Analyst as Part of the Analysis Process

An important part of analysing e-mail traffic data is to provide useful information to the user or analyst about the behaviour of suspect e-mail accounts. Although the existing behaviour analysis methods are able to provide information about unusual or abnormal e-mail traffic behaviour, these methods are still dependent on the user to decide how to proceed with the investigation of particular e-mail accounts. For example, the scan statistics method by [59] (previously described in Section 2.2.4) only locates areas of abnormal activity and requires the user to decide whether to investigate an e-mail account for further detail. It is therefore important to consider the user or analyst as part of the analysis process when developing a system that incorporates different behaviour analysis methods for analysing e-mail traffic.

One of the advantages for considering the user/analyst as part of the analysis process is that the user/analyst knows the purpose of the investigation and what type of suspect they need to search for. Such knowledge about the purpose of the investigation and type of suspect would be difficult to build into an analysis system. By using the user's or analyst's knowledge about the purpose of the investigation, the user can decide which e-mail accounts should be included as part of the investigation or which e-mail accounts should be excluded from the investigation. For example, the user/analyst may only have the rights to examine the evidence for a particular suspect e-mail account, where the owner is known to have formally performed illegal activities. But the user/analyst may also not be allowed to view the communication evidence of certain e-mail accounts because the owners have not been formally suspected of any illegal activities.

Another reason for considering the user/analyst as part of the process is that they are able to use their judgement and experience to determine whether there is enough useful information provided to understand the behaviour of suspect e-mail accounts. Current behaviour analysis methods perform well at finding unusual or abnormal traffic behaviour, but lack the decision making capabilities required to determine whether further investigation is required to provide more information for the user. By utilising the user/analyst's judgement, the user controls how much information is presented to them and whether they require more information about particular suspect e-mail accounts.

Given that there are certain tasks that the user/analyst is able to do well, it may be useful to utilise the user/analyst's knowledge of investigation's purpose and allow them to determine the direction of the analysis process. This could ensure

that the correct set of e-mail accounts is being analysed for the investigation and the correct type of information is being extracted from the e-mail traffic data. Along with the inputs supplied by the user/analyst, one may also want to utilise the capabilities of certain behaviour analysis methods. This may be approached by leveraging each method's ability to perform certain tasks and utilise those to take advantage of tasks that would otherwise be difficult and time consuming for the user/analyst to perform (e.g. searching for particular types of abnormal traffic patterns). These two approaches of utilising the inputs from the user and the capabilities of different behaviour analysis methods are considered in the development of the e-mail traffic analysis system, which is covered in Chapter 4.

2.4 Summary

This chapter described behaviour analysis of e-mail traffic as a way of extracting and revealing useful information about the communication behaviour of individuals, through the examination of their e-mail traffic data. A number of behaviour analysis methods were described, each of which were grouped into four categories according to the level of detail provided on e-mail traffic behaviour. The four categories used for reviewing the current behaviour analysis methods were: individual behaviour analysis, behaviour comparison analysis, clique behaviour analysis, and network behaviour analysis. The behaviour analysis methods described in each of these categories provided different approaches for locating unusual or abnormal communication behaviour in e-mail traffic data.

A number of limitations and drawbacks were identified with the current methods used for behaviour analysis of e-mail traffic. One of the major limitations was the lack of consideration for integrating different levels of analysis to enable the user to interactively zoom in and out of the data. The second limitation identified was a need for having a more in-depth understanding about behaviour measurement variables in terms of the relationships between particular behaviour measurements and also the type of traffic behaviour that particular behaviour measurement variables would be useful for identifying. A third limitation discussed was the need to summarise the abnormal behaviour exhibited by individual e-mail accounts and allowing the user/analyst to choose the e-mail account of interest from the summary. The final limitation identified was the importance of the role of the user/analyst as part of the analysis process and allowing for the user/analyst to provide inputs on the direction of the analysis process.

In the next chapter, the use of the "computational intelligence" approach will be

described and an overview will be provided of the computational techniques used for guiding the user to find abnormal e-mail traffic behaviour and enabling the user to examine the behaviour of suspect e-mail accounts from different levels of analysis.

Chapter 3

E-mail Traffic Analysis Using Computational Intelligence

3.1 Introduction

Previously in Chapter 2, an overview was provided of the current methods used for behaviour analysis of e-mail traffic. The chapter highlighted limitations with the approaches used to present unusual or abnormal behaviour information to the user, and also limitations with the understanding of the behaviour measurements used to extract e-mail traffic behaviour. This chapter follows on by proposing an approach for combining sets of computational techniques to present useful information about unusual or abnormal e-mail traffic behaviour. The main purpose of the chapter will be to describe the computational techniques used in this research to extract and present information to the user about the traffic behaviour of suspect e-mail accounts.

The first section of this chapter describes the "computational intelligence" approach, by defining what it means and explains how it is used for analysing e-mail traffic behaviour. The second section of this chapter, provides a description of the visualisation techniques used in this research for e-mail traffic analysis and how these present useful information to the user about the patterns and relationships hidden in the data. The final section describes feature extraction techniques and how these are used to assist the user with finding unusual/abnormal traffic behaviour exhibited by suspect e-mail accounts.

3.2 Computational Intelligence

While there are a number of definitions used to describe computational intelligence [61, 62], the definition used in this thesis is to describe a type of approach for analysing e-mail traffic behaviour. Computational intelligence is defined here as an approach for using a set of computational techniques, to extract information from data and present the information to the user/analyst in a useful and intelligent manner. The purpose of defining this approach is to provide the user/analyst with a more overall understanding of e-mail traffic behaviour. This is considered important, given the variety of behaviours that can be analysed from e-mail traffic and the different levels of abstraction that may be used for analysing e-mail traffic (as previously mentioned in Chapter 2). Because of the complexity of information available for the user to analyse, the aim of using computational intelligence is to reduce the effort required by the user to understand the e-mail traffic data being studied. However, before the user can understand the data, one needs to consider what computational techniques may be used to extract information from the data.

Computational Techniques

The term "computational techniques" is defined here as techniques of extracting information from data in regard to the data properties. The purpose of using computational techniques is to supply the user with useful knowledge about the data. The computational techniques that may be used for computational intelligence, include any type of data analysis technique from the areas of:

- **Statistics** - mathematical techniques that are used for summarising and interpreting large amounts of data [63].
- **Visualisation** - computational techniques that provide information about the data by transforming the data into visual images, and allowing the user to explore and understand the data visually [5, 64].
- **Artificial Intelligence** - computational techniques that perform tasks that would require "intelligence" if it were performed by humans [65, 66]. The types of computational techniques that would be considered "intelligent", would be those that are able to learn and understand data, or make decisions based on information contained in the data [66]. Examples of "intel-

ligent" tasks performed are: classifying objects, learning and recognising patterns, prediction, or clustering objects into groups.

The computational techniques from each of the above areas provide different ways of analysing the data, with each presenting a certain perspective of the data to the user. For instance, some statistical techniques are able to describe the general characteristics of the data to the user (e.g. mean, standard deviation), while some artificial intelligence techniques inform the user about the presence of patterns in the data (e.g. artificial neural networks [66, 67]). However, each type of computational technique can only provide a limited perspective of the data to the user, meaning that the user cannot obtain an overall understanding of the data by using one particular type of technique. Therefore, consideration needs to be given towards using a set of computational techniques to provide the user a more overall understanding of the data.

Using Sets of Computational Techniques

The purpose of using sets of computational techniques for computational intelligence is to provide the user a variety of perspectives about the data and allow the user to compare those perspectives. This is important given that each computational technique used can only provide limited information about the data in relation to the type of analysis it can perform. By using a set of computational techniques, each technique provides information about a particular aspect of the data. The user can then use the information provided by each of the computational techniques to gain a better overall understanding of the data, as illustrated in Figure 3.1.

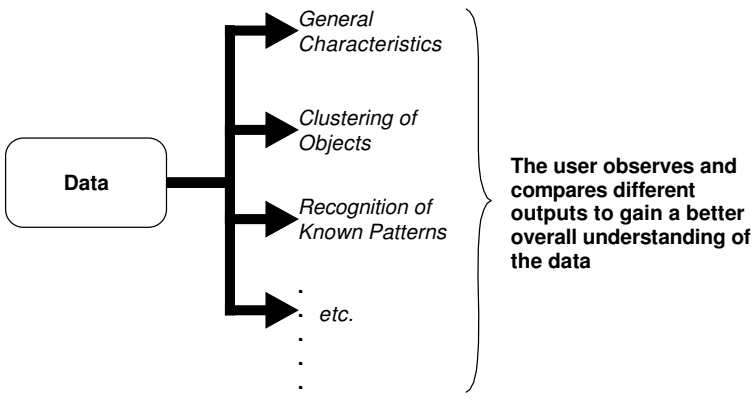


Figure 3.1: Extracting and comparing different types of information about the data.

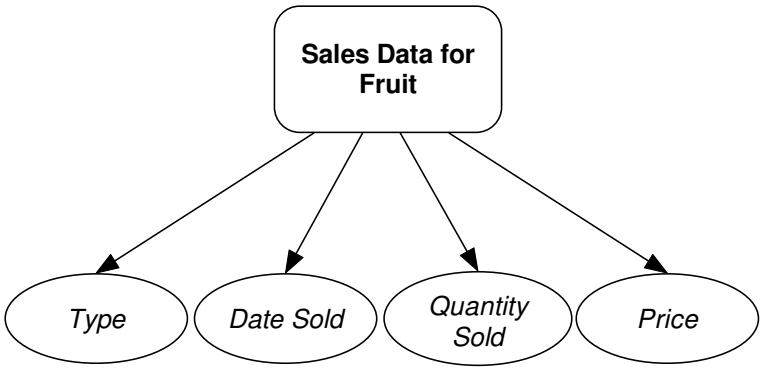


Figure 3.2: Simple example of multi-dimensional data.

Another important point for using sets of computational techniques is for dealing with multi-dimensional data. Much of the data analysed is often multi-dimensional and may contain a large number of variables (e.g. hundreds of variables) that describe particular characteristics of the data [68]. As a simple example of multi-dimensional data, Figure 3.2 shows a concept map of the dimensions associated with the sales data for a fruit store. Each of the dimensions in Figure 3.2 (*Type, Date Sold, Quantity Sold,* and *Price*) indicate particular attributes of the data. The benefit of using a set of computational techniques is that each technique may be used to analyse a certain number of dimensions, so that the whole set covers a range of data dimensions.

Analysis of E-mail Traffic Behaviour

The purpose of using computational intelligence for e-mail traffic analysis is to provide the user/analyst a set of perspectives for analysing e-mail traffic behaviour. This can be approached by assigning a number of computational techniques to analyse e-mail traffic behaviour at different levels of analysis, for example: overviewing the behaviour of a selection of e-mail accounts or examining the behaviour interactions between pairs of e-mail accounts. This can also include using particular techniques to pinpoint unusual or abnormal behaviour, in order for the user/analyst to find these from large amounts of data. The overall effect of using a set of computational techniques is that it provides a variety of ways for the user/analyst to analyse and understand the behaviour of suspect e-mail accounts.

For this research, two types of computational techniques are utilised. The first type, visualisation techniques, is used to enable the user to explore and understand the behaviour of selected e-mail accounts. The second type, feature extrac-

tion techniques, is used to aid the user with locating unusual or abnormal changes in traffic behaviour exhibited by suspect e-mail accounts. Both visualisation and feature extraction techniques are used as a set, to provide the user/analyst different perspectives on the behaviour of suspect e-mail accounts. The diagram in Figure 3.3 illustrates each of the perspectives presented by the computational techniques used. The visualisation techniques used provide different levels of analysis for analysing e-mail traffic behaviour, while the feature extraction techniques provide ways of pinpointing unusual or abnormal changes in behaviour. The remainder of this chapter describes how each of these techniques is used to analyse e-mail traffic behaviour.

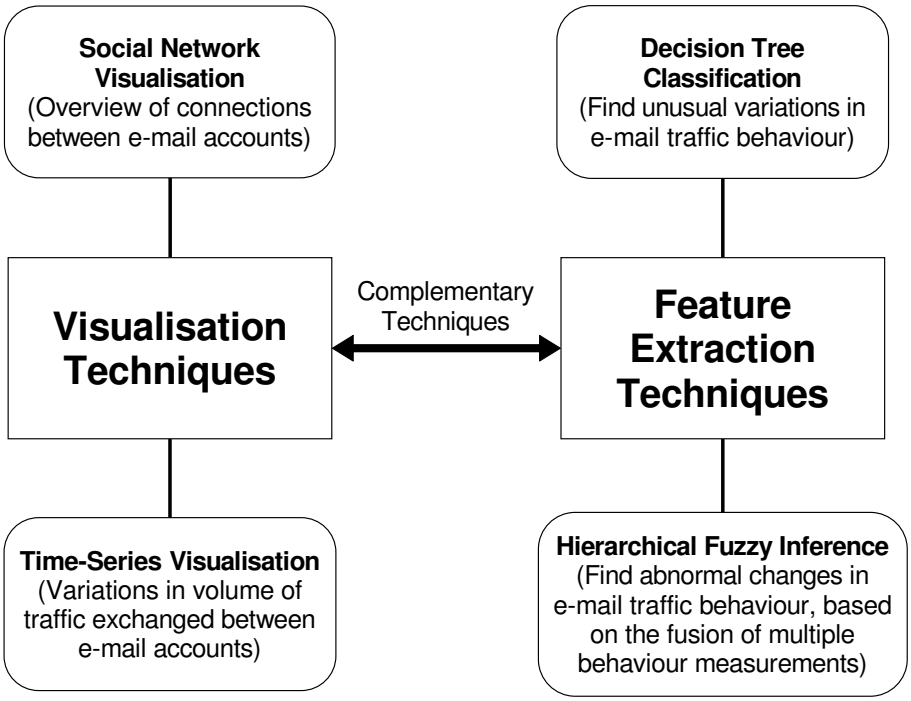


Figure 3.3: Computational techniques used for e-mail traffic analysis.

3.3 Visualisation Techniques

Visualisation techniques are part of a process called “visualisation”, which involves the transformation of data into graphical images. The purpose of this process is to enable data to be visually interpreted by the user, allowing them to make sense of large amounts of information [5, 64]. Visualisation is useful because it takes advantage of the human vision system’s natural abilities to recognise structures and patterns from visual images [5]. It is also useful because it allows the user to explore the data and gain a better understanding of the data

[5]. Without visualisation, there would be certain structures and patterns that would otherwise remain hidden in the data.

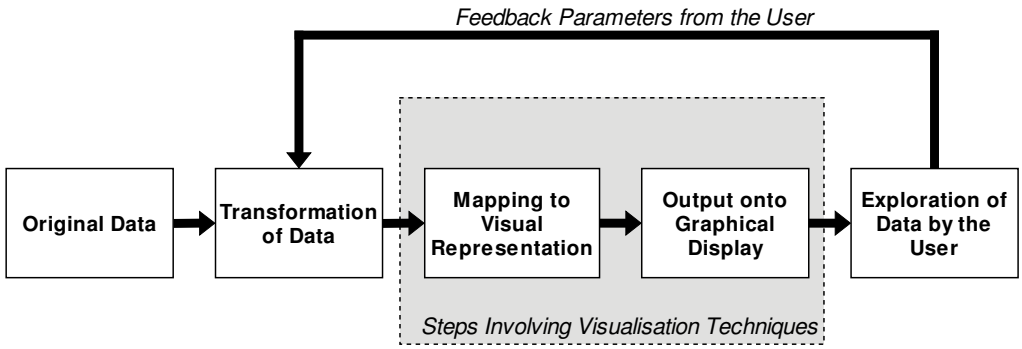


Figure 3.4: Steps of the visualisation process, based on a diagram drawn by [5].

A diagram of the visualisation process is shown in Figure 3.4, which is based on a visualisation process diagram drawn by [5]. The diagram in Figure 3.4 has been modified to identify the steps of the visualisation process and also indicate where visualisation techniques fit into the process. Here, visualisation techniques are defined as computational techniques that map the transformed data into a suitable visual format and output the mapped image onto a graphical display. Visualisation techniques are being defined in this manner to identify a set of common steps that is not directly dependent on the type of data used. This is to help separate visualisation techniques from the step of data transformation, which is directly dependent on the type of data being used. One other point to note is the role of the user as part of the visualisation process. Figure 3.4 shows the user interacting with the data by specifying parameters that control how data is transformed for visualisation. This interaction with the user allows the user to control the analysis process, so that the user determines what will be visually analysed [5].

For e-mail traffic analysis, visualisation techniques are being used to overcome a number of difficulties associated with analysing e-mail traffic data. Firstly, traffic logs like the example shown in Table 3.1, are difficult to examine because the data is presented in a manner that does not allow the user/analyst to easily recognise patterns and relationships in the data. Since visualisation techniques are designed to aid with the recognition of patterns and relationships in the data, this is being used to enable the user/analyst to gain an understanding of e-mail traffic behaviour. A second reason for using visualisation techniques is to allow the user to quickly overview large amounts of data. Given the large amounts of e-mail traffic data that may be analysed, visualisation techniques can be used to

aid the user to quickly overview the information contained in the data. Finally, visualisation techniques are being used to help facilitate interaction between the user and the data, by enabling the user to explore the data and investigate any unusual or abnormal e-mail traffic behaviour.

Table 3.1: Example of e-mail traffic log data.

From	To	Date and Time
clientG@utas.edu.au	clientC@utas.edu.au	4/01/2000 1:09:25 AM
clientC@utas.edu.au	clientG@utas.edu.au	5/01/2000 6:31:50 PM
clientB@utas.edu.au	clientG@utas.edu.au	6/01/2000 12:12:35 PM
clientB@utas.edu.au	clientG@utas.edu.au	6/01/2000 3:36:36 PM
clientG@utas.edu.au	clientC@utas.edu.au	8/01/2000 8:13:38 AM
clientG@utas.edu.au	clientC@utas.edu.au	8/01/2000 4:42:05 PM
clientG@utas.edu.au	clientB@utas.edu.au	9/01/2000 2:29:54 AM
clientG@utas.edu.au	clientD@utas.edu.au	9/01/2000 3:27:43 PM
clientG@utas.edu.au	clientB@utas.edu.au	10/01/2000 6:07:58 AM
clientB@utas.edu.au	clientG@utas.edu.au	11/01/2000 9:37:38 AM
clientG@utas.edu.au	clientA@utas.edu.au	11/01/2000 4:36:38 PM
clientG@utas.edu.au	clientB@utas.edu.au	12/01/2000 9:49:26 AM
clientB@utas.edu.au	clientG@utas.edu.au	12/01/2000 8:18:43 PM
clientA@utas.edu.au	clientG@utas.edu.au	13/01/2000 4:37:39 AM
clientB@utas.edu.au	clientG@utas.edu.au	13/01/2000 6:35:03 AM
clientC@utas.edu.au	clientG@utas.edu.au	14/01/2000 3:21:26 AM
clientG@utas.edu.au	clientC@utas.edu.au	14/01/2000 4:59:20 AM
clientB@utas.edu.au	clientG@utas.edu.au	14/01/2000 6:20:24 PM
clientG@utas.edu.au	clientB@utas.edu.au	15/01/2000 6:54:48 AM
clientG@utas.edu.au	clientC@utas.edu.au	15/01/2000 3:12:10 PM
clientG@utas.edu.au	clientB@utas.edu.au	16/01/2000 8:51:00 PM
clientG@utas.edu.au	clientA@utas.edu.au	17/01/2000 1:02:02 AM
clientC@utas.edu.au	clientG@utas.edu.au	17/01/2000 7:10:58 PM

Although there are a number of visualisation techniques that may be used for analysing e-mail traffic, this research only used two types of visualisation techniques. Social network visualisation and time-series visualisation are used in this research to provide different perspectives of analysis for examining e-mail traffic behaviour. Social network visualisation is being used to provide a high-level visual overview of connections between particular e-mail accounts, while time-series visualisation is being used to provide a lower-level visual analysis of the volume of e-mail traffic sent and received by particular e-mail accounts. Each of these techniques are described below in Sections 3.3.1 and 3.3.2.

3.3.1 Social Network Visualisation

A “social network” is defined as a system of agents that interact with each other and have some pattern of contact between them [52, 54]. Social network visualisation is a computational technique that generates images of these social networks, based on the connections and relationships that exist between individual agents. The technique of social network visualisation originated through sociology [69] and mathematical graph theory [52], where the idea was developed for using points and lines (called “graphs” or “social network graphs”) to represent systems of agents and their connections. The purpose for using the points and lines representation was to aid with visualising relational data, which describe how individual agents relate to each other in the social network.

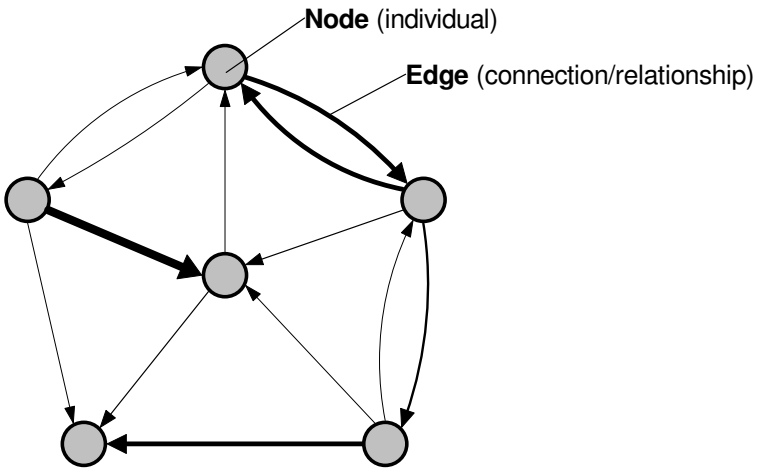


Figure 3.5: Simple diagram showing the components of a social network graph.

Social network visualisation is often used for analysing relational information, due to its easy-to-understand graph representation. The diagram shown in Figure 3.5 indicates the basic components of a social network graph. Each “Node” represents an individual and each “Edge” represents a connection or relationship between two particular individuals. The edges can be assigned arrows to indicate a directional relationship that is one-way, and also be assigned different line thicknesses to indicate the strength of a relationship. For example, thin lines indicate a weak relationship, while thick lines indicate a strong relationship. This type of visual representation enables the user to immediately perceive and understand the network of connected individuals, which is one of the useful features of using social network visualisation.

Analysing The Social Networks In E-mail Traffic

The purpose of using social network visualisation for e-mail traffic analysis is to present the user with an overview of the communication ties between particular e-mail accounts. E-mail traffic log data like the one in Table 3.1 is difficult for the user/analyst to understand when examining the data for connections between different e-mail accounts. Social network visualisation aids the user/analyst with gaining an overview of the social network of e-mail users, so that the user/analyst can understand the communication ties and relationships between particular e-mail accounts. This also enables the user/analyst to spot areas of interest in the e-mail social network, such as the clustering of e-mail users into distinct social groups or communities [43, 44, 70].

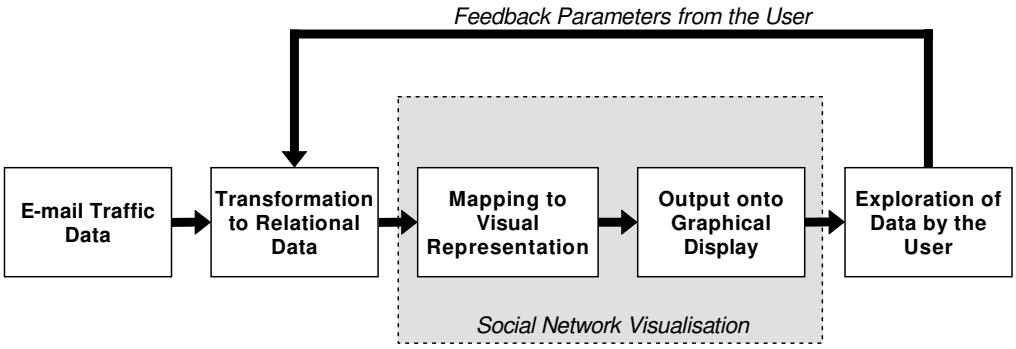


Figure 3.6: The social network visualisation process.

To convert e-mail traffic data into social network graphs, the data is processed using the steps shown in Figure 3.6. Firstly, the e-mail traffic data is transformed into relational data, like the one shown in Table 3.2. This relational data is extracted from the traffic data, producing information about connections between e-mail accounts. The relational data may also contain information about the number of e-mails sent between e-mail accounts, which can be used to indicate the connection strength of the relationship. The relational data is then mapped into its graph image through social network visualisation, which computationally maps the relational data into its visual representation and outputs it onto a graphical display. When creating the visual representation, social network visualisation also applies a type of layout algorithm (e.g. multidimensional scaling [71, 72], spring embedder algorithm [73, 74]) that organises the layout of the points and lines in the image. This is so that the graph image is automatically arranged in a convenient manner for the user to look at and analyse. In the final output, a social network graph of the e-mail traffic data is produced, like the one shown in Figure 3.7, which was created using a program called GUESS [75].

The process of social network visualisation shows how it is useful for providing the user/analyst a visual representation of e-mail traffic data and also an overview of the connections between e-mail accounts.

Table 3.2: Example of relational data extracted from e-mail traffic data.

E-mail Account Address	Associate Address	Number of E-mail Messages Sent
clientA@utas.edu.au	clientG@utas.edu.au	170
clientA@utas.edu.au	clientI@utas.edu.au	186
clientA@utas.edu.au	clientB@utas.edu.au	105
clientA@utas.edu.au	clientC@utas.edu.au	104
clientB@utas.edu.au	clientG@utas.edu.au	95
clientB@utas.edu.au	clientF@utas.edu.au	36
clientB@utas.edu.au	clientA@utas.edu.au	93
clientB@utas.edu.au	clientD@utas.edu.au	63
clientC@utas.edu.au	clientA@utas.edu.au	104
clientC@utas.edu.au	clientG@utas.edu.au	213
clientC@utas.edu.au	clientF@utas.edu.au	26
clientC@utas.edu.au	clientI@utas.edu.au	145
clientD@utas.edu.au	clientG@utas.edu.au	126
clientD@utas.edu.au	clientF@utas.edu.au	37
clientD@utas.edu.au	clientI@utas.edu.au	92
clientD@utas.edu.au	clientB@utas.edu.au	62
clientE@utas.edu.au	clientF@utas.edu.au	33
clientE@utas.edu.au	clientH@utas.edu.au	62
clientF@utas.edu.au	clientC@utas.edu.au	16
clientF@utas.edu.au	clientB@utas.edu.au	30
clientF@utas.edu.au	clientE@utas.edu.au	23
clientF@utas.edu.au	clientD@utas.edu.au	32
clientG@utas.edu.au	clientC@utas.edu.au	216
clientG@utas.edu.au	clientB@utas.edu.au	108
clientG@utas.edu.au	clientA@utas.edu.au	175
clientG@utas.edu.au	clientD@utas.edu.au	138
clientH@utas.edu.au	clientI@utas.edu.au	128
clientH@utas.edu.au	clientE@utas.edu.au	64
clientI@utas.edu.au	clientC@utas.edu.au	148
clientI@utas.edu.au	clientA@utas.edu.au	188
clientI@utas.edu.au	clientD@utas.edu.au	103
clientI@utas.edu.au	clientH@utas.edu.au	136

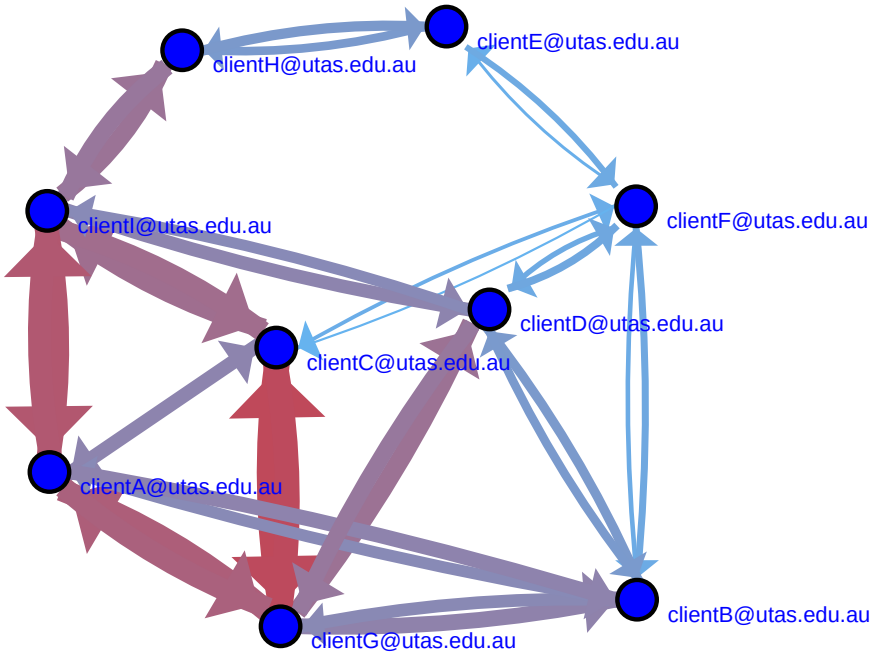


Figure 3.7: Social network visualisation of data from Table 3.2.

Limitations of Social Network Visualisation

While social network visualisation is useful for visualising the communication ties and relationships between e-mail accounts, it does have a number of limitations. Firstly, when the number of individuals visualised becomes extremely large (e.g. more than 200 individuals), it becomes increasingly difficult for the user to perceive the distinct communication ties and relationships between particular individuals. This is because as the number of individuals increases, the number of visualised connections also increases, resulting in a social network graph that is crowded with communication ties between the large number of individuals. An example of this is shown in Figure 3.8. Due to the crowding of large numbers of individuals in the social network graph, this makes it difficult for the user to locate and investigate e-mail accounts that may be exhibiting unusual communication behaviour.

Another limitation with social network visualisation is that the social network graphs produced are not ideal for representing the temporal information contained in e-mail traffic data. The temporal aspect of e-mail traffic data is important, given that the relationship between e-mail accounts does not always remain the same and may fluctuate due to the occurrence of particular events. The information displayed in social network graph images (e.g. the name/address of the

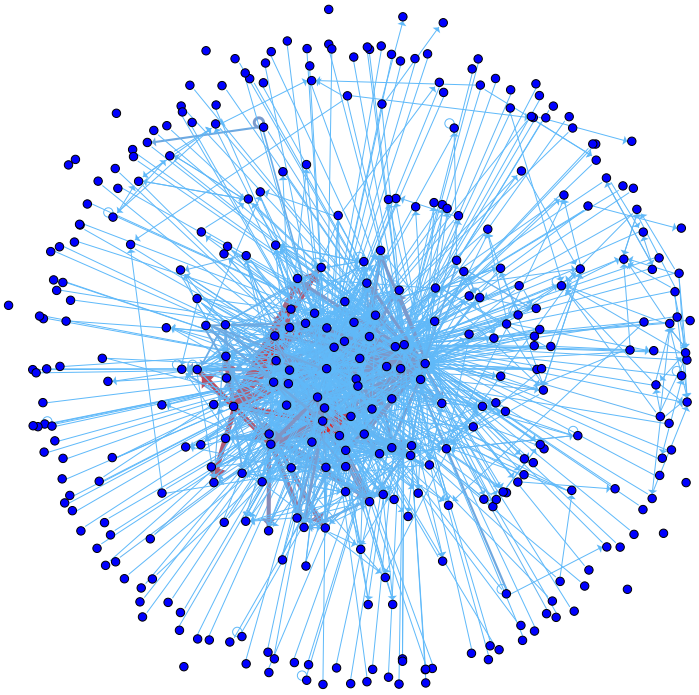


Figure 3.8: Example of a large network of 355 e-mail users.

individual, the number of connections, the strength of the connections) cannot show temporal changes using the points and lines representation. To overcome the limitations of displaying temporal information with social network visualisation, time-series visualisation was used in this research as another type of visualisation technique for analysing e-mail traffic.

3.3.2 Time-Series Visualisation

Time-series visualisation is a type of computational technique that provides a one-dimensional view of data and presents a visual image of how the data changes over time. This type of visualisation is often used for statistical analysis of time-series data [76], whereby the plotted image provides the user information about the temporal variations of the data. The simple graph plot presented by time-series visualisation makes it easy for the user to interpret, since the data is simply represented as a sequence of plotted values that displays important features of the temporal data (e.g. trends, outliers) [76].

The information presented by time-series visualisation is usually drawn as a series of values plotted between two axes. The horizontal axis is used to represent the time variable, while the vertical axis is used to represent the variable or dimension being analysed. The plotted values for the sampled data may be rep-

resented using notations such as: dots, connected lines, or columns. Figure 3.9 shows an example of different notations used for representing time-series data.

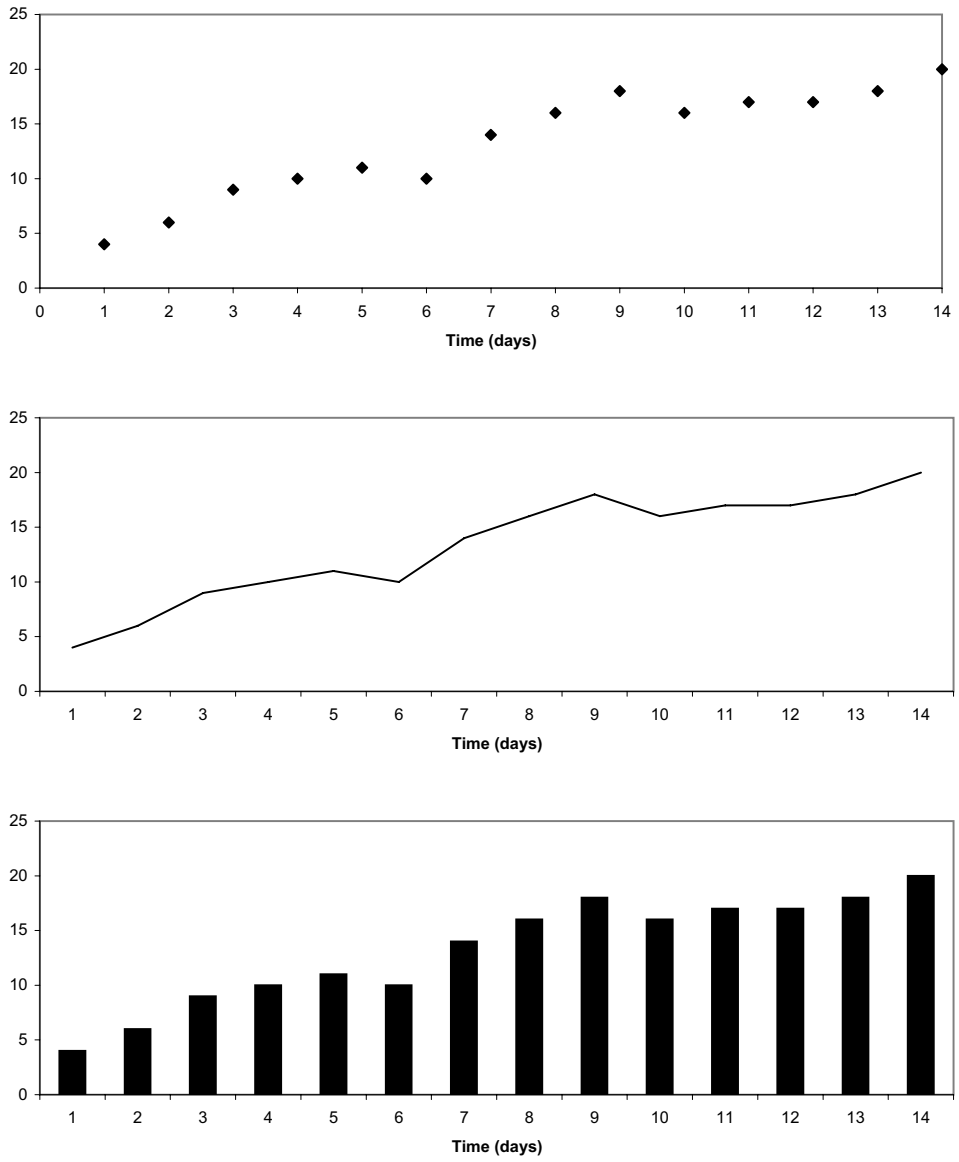


Figure 3.9: Example of different notations used for representing time-series data.

Analysing Time-Series Information in E-mail Traffic

The purpose of applying time-series visualisation to analyse e-mail traffic data is to examine the volume of e-mail messages exchanged between particular e-mail accounts over time. This is to provide the user/analyst information about the variations in the number of e-mail messages sent or received by e-mail accounts, as well as information about particular trends (e.g. rise in traffic volume or drop

in traffic volume). Another reason for using time-series visualisation is for examining the relationship between particular e-mail accounts. This can enable the user to investigate time periods of intense or low traffic activities, as well as find unusual interactions between those e-mail accounts. As a result, this provides another way of examining relationship information, in addition to social network visualisation (described previously in Section 3.3.1).

It should be noted that different time-scales (e.g. minutes, hours, days, weeks, or months) can be applied to the sampling of e-mail traffic data used for time-series visualisation. The advantage of using different time-scales is that it provides different levels of granularity for analysing the data [77]. This provides the user/analyst with a variety of options for analysing the data, since each time-scale used may reveal different types of patterns through the resulting time-series plot. For example, the pattern of a person who sends e-mail messages only on certain days of the week, may be better noticed when the time-scale is adjusted to e-mail messages sent per day. Thus the use of different time-scales can aid the user in visually analysing the e-mail traffic data for a variety of temporal behaviours.

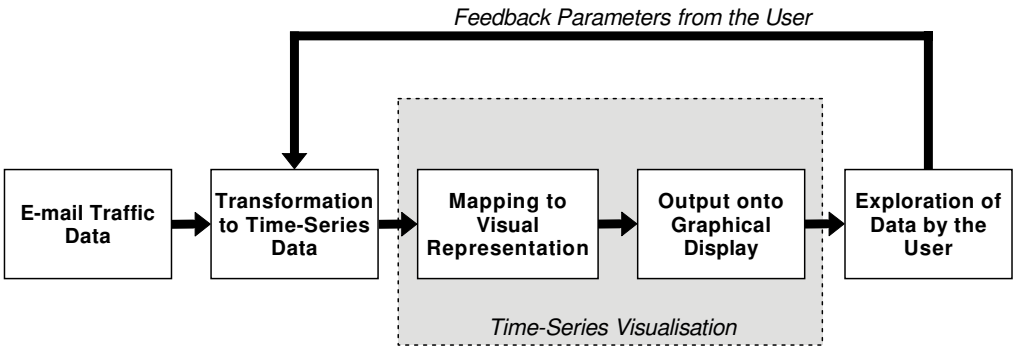


Figure 3.10: Time-series visualisation process.

To create the time-series graph for time-series visualisation, the e-mail traffic data is processed using the steps shown in Figure 3.10. Firstly, the e-mail traffic data is transformed to extract temporal and traffic volume information, and sampled at a particular time-scale to produce the time-series data. An example of the resulting time-series e-mail traffic data is shown in Table 3.3. The time-series data is then processed using time-series visualisation and plotted as a time-series graph. The graph is then output onto graphical display, showing an image like those in 3.3 and 3.12. Both of these Figures were visualised using TimeSearcher 2 [78] under different time-scales. The resulting time-series graph shows that time-series visualisation is a useful technique for aiding the user/analyst to understand the temporal aspects of e-mail traffic data.

Table 3.3: Example of time-series e-mail traffic data.

Week Number	Number of Incoming Messages	Number of Outgoing Messages
0	4	6
1	9	10
2	11	10
3	14	16
4	18	16
5	17	17
6	18	20
7	20	18
8	20	22
9	22	21
10	24	26
11	35	31
12	22	20
13	20	17
14	15	20

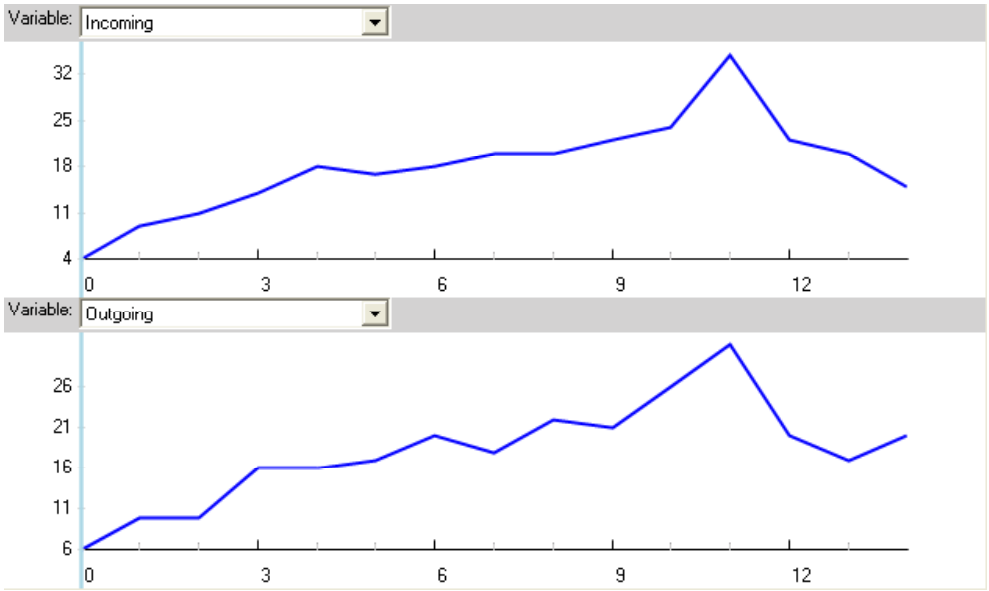


Figure 3.11: Example of e-mail traffic volume using a weekly time-scale.

Limitations of Time-Series Visualisation

While time-series visualisation is useful for displaying the temporal information associated with particular e-mail accounts, there are a number of limitations associated with using time-series visualisation. Firstly, it is limited in its ability to provide an overview of connections between multiple e-mail accounts. This is because time-series visualisation only provides information about aspects of

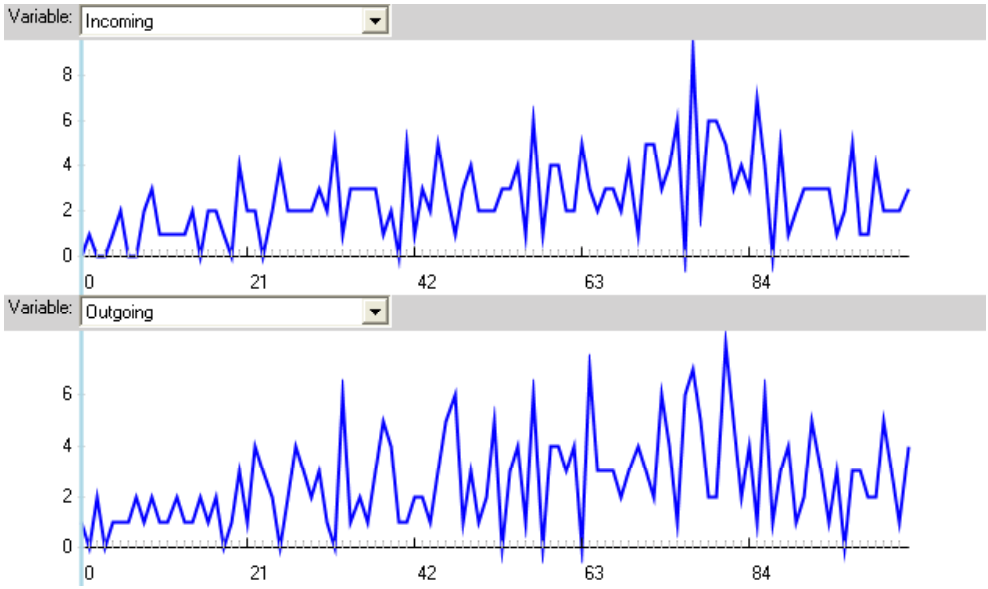


Figure 3.12: Example of e-mail traffic volume using a daily time-scale.

the data that vary with time, but is unable to display more complex relational information such as those shown by social network visualisation.

Another limitation of time-series visualisation is that it is difficult to display information for a large selection of e-mail accounts (e.g. more than 10 e-mail accounts). This is due to the fact that while multiple plots for several e-mail accounts may be displayed on the same time-series graph, the graph will gradually become overcrowded and incomprehensible when the number of time-series plots is increased. This makes it difficult for the user to visually locate in the e-mail traffic data which e-mail accounts may be exhibiting unusual or abnormal traffic behaviour. It also makes it difficult for the user to search and determine which e-mail account may be of interest for detailed investigation.

Overall, what the social network and time-series visualisation techniques show is that these are useful methods for aiding the user/analyst to visually explore e-mail traffic data. While both of these techniques do have certain limitations, these limitations can be overcome by utilising both techniques to complement each other, hence using them as a set for computational intelligence. The approach of utilising both visualisation techniques will be described in Section 4.4, and demonstrated in Sections 5.2 and 5.3. However, when searching for signs of unusual or abnormal behaviour, the exploration approach used by the visualisation techniques would take the user a great deal of time to look for those behaviours. This is because of the large amount of e-mail traffic data that may have to be examined and also the time and effort required for determining the

presence of unusual or abnormal behaviour. To aid the user with finding unusual or abnormal e-mail traffic behaviour, feature extraction techniques are considered, to provide methods for quickly locating these types of behaviour.

3.4 Feature Extraction Techniques

Feature extraction techniques are defined here as computational techniques that extract and process information from data, to locate data records that possess particular features or patterns. These computational techniques are defined here as being part of the feature extraction process, which is shown as a process diagram in Figure 3.13 and a data flow diagram in Figure 3.14. The diagram in Figure 3.13 illustrates the steps involved in extracting features from the data while the diagram in Figure 3.14 illustrates the information obtained through each step in feature extraction .

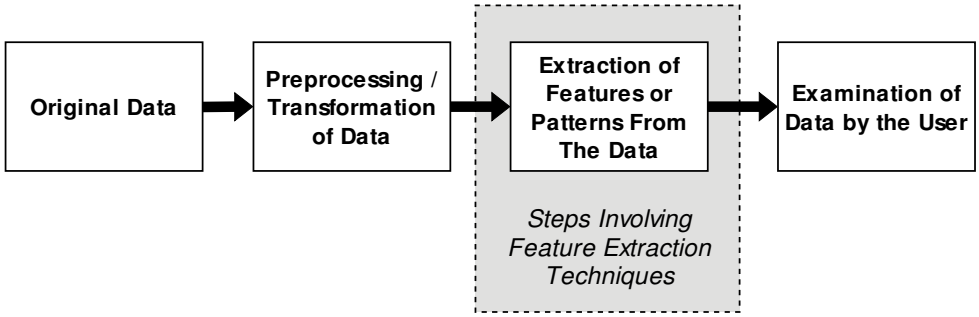


Figure 3.13: Steps of the feature extraction process.

Figure 3.14: The information obtained through each step of feature extraction.

In the first step of the feature extraction process, the original data is preprocessed to extract data records with selected attributes and measurements. The attributes of the data are characteristics or variables that describe particular aspects of each record in the original data [68] (e.g. the data for student marks contains the attributes “student id”, “first name”, “last name”, “assignment 1 mark”, “assignment 2 mark”). Measurements are the result of additional processing of the original data to compute numerical values from particular data attributes. These

numerical values provide quantitative information about a particular aspect of the data records (e.g. the student's average assignment marks). After the original data has been preprocessed, feature extraction is performed to locate features or patterns present in the data. These features provide information about the distinctive properties possessed by certain data records and also inform the user or analyst where these records can be found in the data.

There are a variety of computational techniques that may be considered for performing feature extraction. The computational techniques that may be suitable for feature extraction are:

- **Classification techniques** - techniques that assign data records to one of several predefined categories [68].
- **Clustering techniques** - techniques that divide the data into groups or clusters [68].

Each of the techniques listed above are able to extract certain types of information from the data and also aid the user in understanding the features and patterns hidden in the data.

For e-mail traffic analysis, feature extraction techniques are being used to primarily overcome the difficulties associated with using visualisation techniques for finding unusual or abnormal traffic behaviour. The visualisation techniques described in Section 3.3 were noted to be useful for visual exploration of data, but have limited capabilities for allowing the user to pinpoint specific features or patterns from the data. Feature extraction techniques on the other hand, present information to the user that enables them to locate specific data records associated with particular features in the data. Thus feature extraction techniques are considered to be useful for finding unusual or abnormal e-mail traffic behaviour.

While there are a number of feature extraction techniques that may be used to locate unusual or abnormal e-mail traffic behaviour, this research only investigated two types of feature extraction techniques. The first technique, decision tree classification, is used for analysing suspect e-mail accounts to find unusual variations in traffic behaviour. The second technique, hierarchical fuzzy inference, is used for analysing a selection of suspect e-mail accounts to determine the degree of change in behaviour that has occurred. Both of these techniques are described in Sections 3.4.1 and 3.4.2.

3.4.1 Decision Tree Classification

Decision tree classification is a type of computational technique that builds a model of the data to describe where to locate particular types of data records. These data records are located by defining discrete boundaries that separate areas of the data where certain types of features or data records can be found. The first algorithms developed for decision tree classification were ID3 (Iterative Dichotomiser) [79] and CART (Classification and Regression Trees) [80], which were invented during the 1980s [81]. These were then followed later by algorithms such as C4.5 [82] and C5.0 (a commercial version of C4.5) [83]. The popular use of decision tree classification is due to the fact that it does not require the input of domain knowledge and can be used for exploratory analysis of the data [81]. It is also popular due the visual tree representation that is used for describing the data, making it intuitive and easy to be interpreted by the user [81].

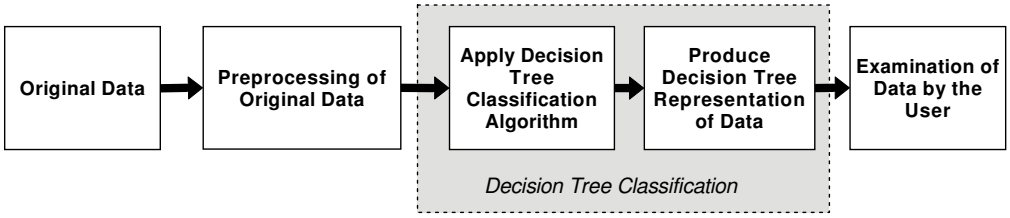


Figure 3.15: Steps of the decision tree classification process.

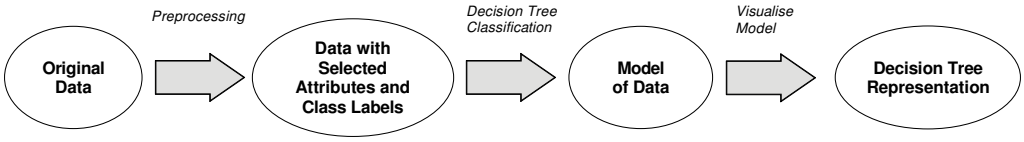


Figure 3.16: The data produced after each step of the decision tree process.

Data is analysed by decision tree classification by processing it through a series of steps, as shown in Figures 3.15 and 3.16. These steps are based on those described by the ID3 algorithm [79] and from [84]. Firstly, the original data is preprocessed to select data attributes and assign class labels that will be used by the decision tree classification algorithm. The data attributes selected are variables that will be tested by the decision tree classification algorithm to determine where to place the boundary for separating data records. The class labels assigned to each data record identifies what category each record belongs to and will be used by the decision tree classification algorithm to group similar classes of data records. An example of selected data attributes and class labels produced through preprocessing is shown in Table 3.4.

Table 3.4: Example of data attributes and the assigned class labels.

Attribute1 x	Attribute2 y	Class Label
0.5	0.9	A
0.3	0.3	A
0.8	0.9	B
0.9	0.6	A
0.9	0.9	B
...		

Once the original data has been preprocessed, the decision tree classification algorithm is applied to the preprocessed data. The method generally used for analysing the data is to recursively process the data to determine the placing of boundaries between different classes of data records [84]. At each step of the recursive process, the algorithm determines whether a particular attribute (e.g. x or y) is able to optimally subdivide the data into different classes. If an area of the data can be subdivided using that attribute, then the attribute is used to define a new boundary in the data. This division process is repeated until each area of the data consists of a single class of data records. An example of this recursive process is illustrated in Figure 3.17.

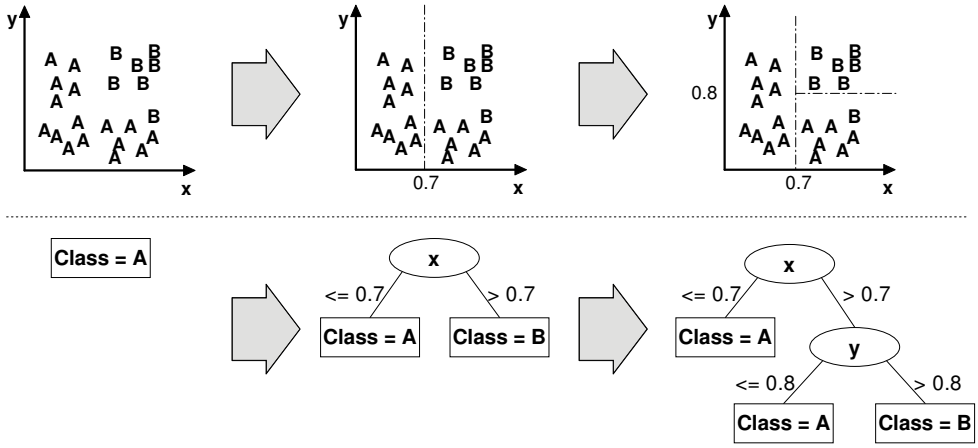


Figure 3.17: Example of how decision tree classification partitions the data.

At the end of the recursive process, a model of the data is produced that contains structured information identifying where particular classes of data records are located. This model of the data can be visualised as a tree representation, as shown in Figures 3.17 and 3.18. The tree representation presented by decision tree classification consists of three types of tree nodes [68, 81]:

- **Root Node** - a node with no incoming edges and only outgoing edges.
- **Internal Node** - a node with incoming edges and outgoing edges.

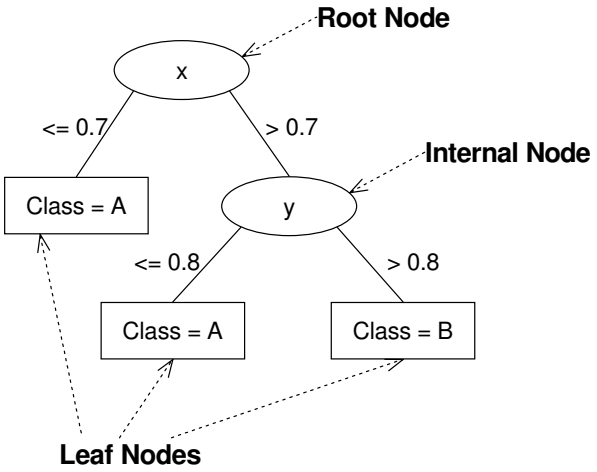


Figure 3.18: Decision tree representation used for describing the data.

- **Leaf Node** - a terminal node that has only one incoming edge.

The non-terminal nodes (root node and internal nodes) display tests performed on certain data attributes to show how the data has been split. The edges outgoing from each non-terminal node display the outcome of the test on an attribute (e.g. $x <= 0.7$ or $x > 0.7$), indicating where the boundary was placed in the data. The terminal nodes or leaf nodes display the class of data records found in a particular area of the data (e.g. class "A" or class "B") and usually also the number of data records belonging to a particular class (e.g. class A = 30 records, other classes = 2 records). To determine where to find a certain class of data records, the user follows the path from the root node to a particular leaf node (e.g. $x > 0.7$ AND $y <= 0.8$) to understand how to locate those data records. The overall result of using the tree representation is an intuitive and easy-to-understand diagram that presents information to the user on where to locate particular types of data records.

E-mail Traffic Analysis with Decision Tree Classification

Decision tree classification is being used in this research to identify occurrences of "unusual" interactions between a suspect e-mail account and its associates. As an example, an unusual interaction may be a case where the suspect starts sending large volumes of e-mail messages to a particular associate after a long period of little activity. Another example of this may be where the suspect stops communicating with a particular associate and begins communicating with another associate. Both of these examples illustrate that there may be certain interactions where there is an unusual change in the suspect e-mail account's traffic

behaviour. The aim of using decision tree classification is to aid the user/analyst with locating these unusual interactions in the suspect e-mail account's traffic data.

To find unusual interactions, the approach used here is to apply decision tree classification to perform two separate analyses:

- Analysis of the suspect's incoming interactions, based on messages received by the suspect.
- Analysis of the suspect's outgoing interactions, based on messages sent by the suspect.

The purpose of this approach is to produce two decision trees in order to simplify the information provided by the decision tree classification output and to reduce the size and amount of decision tree information viewed by the user/analyst. This is considered as a way of making the decision tree information easier for the user/analyst to interpret, rather than using a much larger decision tree to present information on both unusual incoming and outgoing interactions.

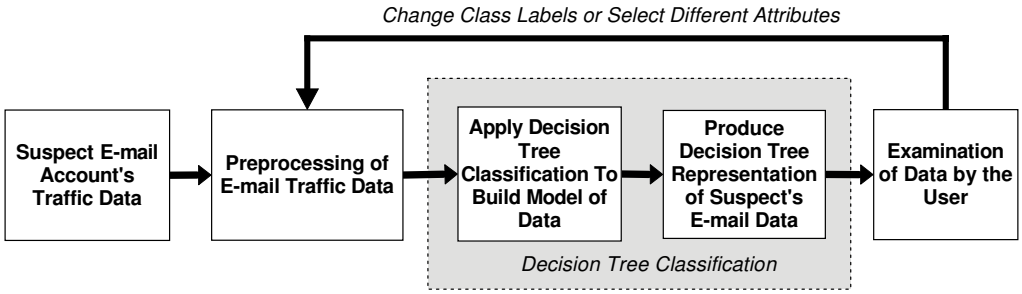


Figure 3.19: Decision tree classification process used for e-mail traffic analysis.

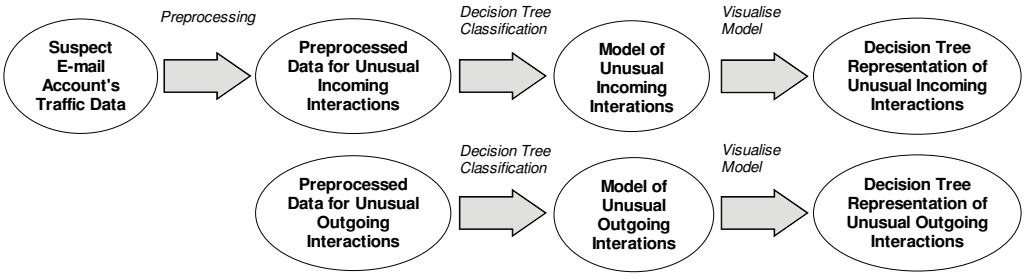


Figure 3.20: Data produced for finding unusual interactions.

The process used for analysing a suspect's e-mail traffic data for unusual incoming and outgoing interactions is shown in Figures 3.19 and 3.20. During the preprocessing step, the suspect's e-mail traffic data is formatted so that messages

with multiple recipients are treated as separate messages sent to many recipients at the same time. The e-mail messages are treated in this way to enable the decision tree classification algorithm to analyse the individual interactions between the suspect and each of its associates.

After the messages have been formatted, particular attributes and class labels are selected from the data to determine how the data records are categorised and grouped to find the unusual interactions. The attributes and class labels selected can be one of two sets: one for finding unusual incoming interactions (shown in Table 3.5) and the other for finding unusual outgoing interactions (shown in Table 3.6). Each of the attributes and class labels used for these two sets are described as follows:

- **Direction** - a nominal or name-based attribute that indicates whether the e-mail message was sent or received by the suspect e-mail account. This attribute is created during preprocessing stage.
- **MessageDate** - a numerical attribute that indicates the date/time the e-mail message was sent or received. The date/time information is represented as a number to enable the decision tree classification algorithm to numerically split the date/time information. This attribute is computed during the preprocessing stage and is based on the relative difference between the original timestamp and a reference date/time (e.g. 23.4 days).
- **'From' Class Label** - a nominal value where the values of the attribute are based on the e-mail addresses found in the 'From' field of each e-mail message's header. This class label is used for finding unusual incoming interactions by specifying that e-mail messages originating from particular sender addresses be grouped together during classification.
- **'To' Class Label** - a nominal value where the values of the attribute are based on the e-mail addresses found in the 'To','CC', or 'BCC' field of each e-mail message's header. This class label is used for finding unusual outgoing interactions by specifying that e-mail messages sent to particular recipient addresses be grouped together during classification.

After the preprocessing stage has been completed, the preprocessed data is processed by the decision tree classification algorithm to locate the unusual interactions present in the suspect's e-mail traffic data. The decision tree classification algorithm is applied to the two sets of preprocessed data, as shown in Figure 3.20,

Table 3.5: Example of attributes and class label used for unusual incoming e-mail traffic.

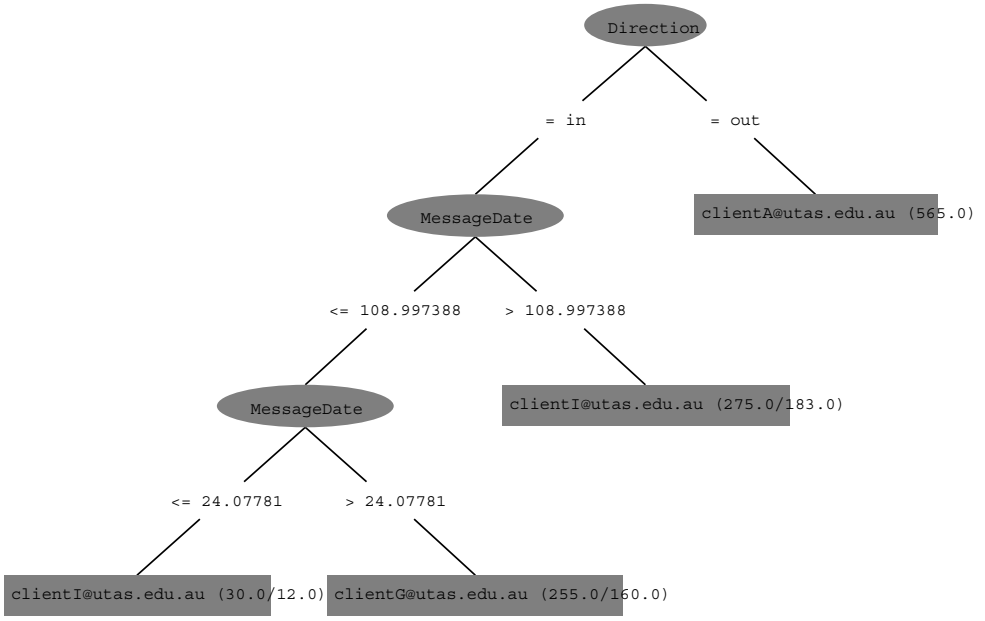
Attribute1 *Direction* (nominal value)	**Attribute2** *MessageDate* (numerical value)	**Class Label** *From* (nominal value)
'out'	23.4	clientA@utas.edu.au
'out'	24.0	clientA@utas.edu.au
'in'	25.0	clientB@utas.edu.au
'in'	30.0	clientC@utas.edu.au
...		

Table 3.6: Example of attributes and class label used for unusual outgoing e-mail traffic.

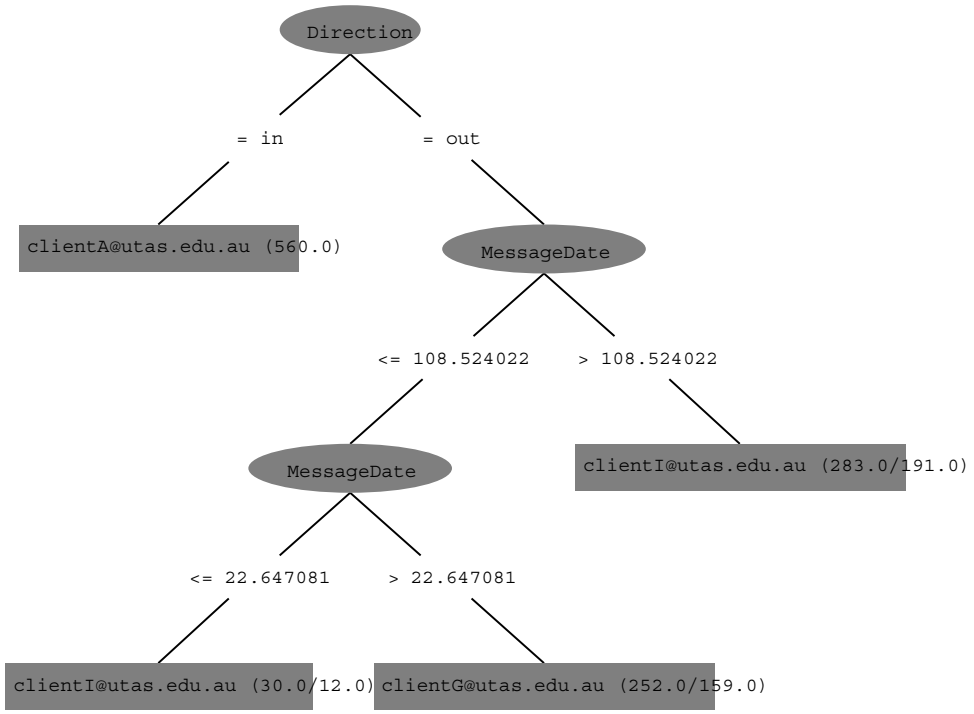
Attribute1 *Direction* (nominal value)	**Attribute2** *MessageDate* (numerical value)	**Class Label** *To* (nominal value)
'out'	23.4	clientB@utas.edu.au
'out'	24.0	clientC@utas.edu.au
'in'	25.0	clientA@utas.edu.au
'in'	30.0	clientA@utas.edu.au
...		

to produce a model of unusual incoming interactions and a model of unusual outgoing interactions. The models created are developed from the categorising and grouping of e-mail messages, which are determined by the sender addresses (for unusual incoming interactions) or by the recipient addresses (for unusual outgoing interactions). The models produced by the decision tree classification can then be used to identify locations in the data where there has been "unusual" changes in interaction behaviour. Both of the models can be visualised as tree representations, like the example shown in Figure 3.21, which was visualised using WEKA (Waikito Environment for Knowledge Analysis) [84].

To interpret the interaction information presented in the visualised decision tree, the user/analyst follows a path from the root node to one of the leaf nodes on the decision tree. The path taken along the decision tree tells the user/analyst about the direction of the e-mail messages (incoming or outgoing) and then the period of time when the interaction occurred. For example, in Figure 3.21(a) a path from the root node to the first leaf node on the left is: Direction = 'in', MessageDate <= day 24. After following the path, the information contained in the leaf node tells the user/analyst about the associate that the suspect mostly communicated with during a particular period of time and also the number of e-mail messages exchanged with that associate. For example: in Figure 3.21(a), the suspect *clientA@utas.edu.au* mostly communicated with *clientI@utas.edu.au* prior to day 24, sending $(30 - 12) = 18$ messages. Comparison of one leaf node to



(a) Unusual incoming interactions, classifying the 'From' addresses.



(b) Unusual outgoing interactions, classifying the 'To' addresses.

Figure 3.21: Two decision trees produced for incoming and outgoing e-mail traffic.

a neighbouring leaf node tells the user/analyst that the suspect's interaction with their associates changed, indicating that something "unusual" happened. For example, in Figure 3.21(a), comparison of the first left leaf node to the second left leaf node reveals that the suspect *clientA@utas.edu.au* started communicating

more with associate *clientG@utas.edu.au* after day 24 and before day 109. This shows a change in behaviour from how the suspect *clientA@utas.edu.au* mostly communicated with *clientI@utas.edu.au* prior to day 24. These simple examples show how easy it is for the user to read information from the visualised decision tree and to interpret the information to reveal where unusual changes in behaviour are occurring.

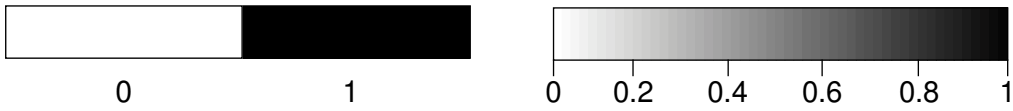
Although decision tree classification describes where unusual interactions can be found in the e-mail traffic data, the user/analyst still requires additional information to verify what was found. This can be done by using decision tree classification in combination with visualisation techniques, so that the user/analyst is able to use visualisation to further explore and understand the interactions found. This computational intelligence approach of using both decision tree classification and visualisation techniques is described in more detail in Section 4.4 and is evaluated in Section 5.2. The next feature extraction technique, hierarchical fuzzy inference, is another computational technique used in this research for analysing the communication behaviour exhibited by suspect e-mail accounts.

3.4.2 Hierarchical Fuzzy Inference

Fuzzy inference is a type of computational technique that models the uncertain and imprecise reasoning process used by humans to interpret information [66]. As humans, we often use vague terms to describe things that we observe in the world around us. The vague terms used are often uncertain or imprecise, for example: "the weather is *hot*", "that man is *tall*", "the danger risk is *high*". Computers normally cannot understand vague terms and must compute observations using crisp numbers, for example: "the weather is 37.5°C", "that man is 182 cm", "the danger risk is 89 %". Fuzzy inference allows computers to use vague terms to interpret observations or information, enabling computers to use an uncertain and imprecise reasoning process that is similar to that used by humans.

The computational technique of fuzzy inference is built upon the foundations of fuzzy logic and fuzzy set theory [85]. The basic concept behind fuzzy logic [85] is the representation of knowledge through *degrees of membership*. This means that knowledge can be expressed as having a degree of truth within a spectrum of values, rather than having a strict true or false value as represented through Boolean logic [66]. This comparison between fuzzy logic and Boolean logic is illustrated in Figure 3.22, where Boolean logic is represented by two discrete

values and fuzzy logic is represented through a spectrum of values between 0 to 1.



(a) Boolean logic (b) Fuzzy logic

Figure 3.22: Comparison between Boolean logic and fuzzy logic.

Fuzzy set theory extends the idea of fuzzy logic through the concept of *fuzzy sets*, which are like the classical crisp set used in mathematical theory but have fuzzy boundaries [66]. Fuzzy sets are used in fuzzy set theory to represent the fuzzy values of a *linguistic variable*, which is a variable that is labelled by a word to describe a particular term or concept (e.g. temperature). A linguistic variable contains a number of fuzzy subsets, each of which is labelled by a word that vaguely describes the range of values covered by the subset (e.g. “Cold”, “Warm”, or “Hot”). These fuzzy subsets can be used to map a crisp input value from the linguistic variable’s universe of discourse, to determine its “degree of membership” to each of the fuzzy subsets. An example of this mapping is shown in Figure 3.23. This mapping of crisp input values into its degree of membership of a fuzzy set, allows input information to be represented in a vague and imprecise manner.

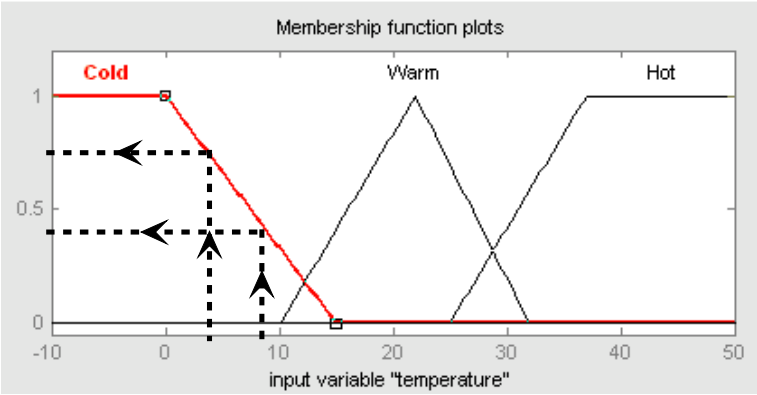


Figure 3.23: Example of a linguistic variable and its associated fuzzy set.

Fuzzy inference builds upon fuzzy logic and fuzzy set theory, by processing information through a series of inputs, which are then mapped into one or more outputs. The mapping of inputs into outputs is done by using fuzzy rules, which encode knowledge using vague or uncertain terms. For example: “IF temperature is hot, THEN air_conditioner_output is high”, “IF temperature is warm, THEN air_conditioner_output is medium”. Fuzzy inference systems operate by

processing input data that is crisp (e.g. 37.5°C), interpreting that value by "fuzzifying" it (e.g. 37.5°C is a member of the term "hot"), applying the fuzzy rules to determine the output (e.g. air conditioner output is high), then "defuzzifying" the output to produce a crisp number (e.g. air condition output level = 90%) [66]. One of the advantages of fuzzy inference is that it is able to process data that contains uncertain information and also has the ability to process input from several measurement sensors in parallel. Fuzzy inference is often used in decision support systems [86] to provide advice on things that contain a level of uncertainty or risk, such as, for example, real estate evaluation [87].

However, in complex domains where there are a large number of input variables, using a single fuzzy rule base is not ideal. This is because when the number of input variables for a fuzzy inference system increases, the number of fuzzy rules required exponentially increases. Assuming that n is the number of variables and m is the number of membership functions for each variable, a complete set of rules will require m^n rules [88, 89]. This exponential rule explosion causes problems with computational complexity, real-time performance, and system definition [89]. One of the solutions proposed by [88] to overcome the exponential rule explosion problem is the use of hierarchical fuzzy inference systems.

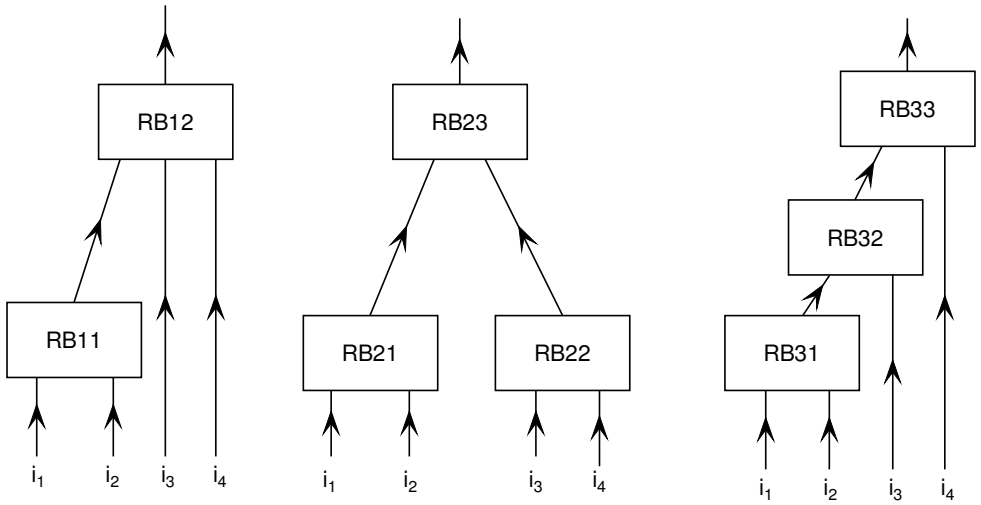


Figure 3.24: Different architectures for a hierarchical fuzzy system of 4 variables.

A hierarchical fuzzy inference system consists of a number of low-dimensional fuzzy system modules, linked together in a hierarchical structure where the outputs of the lower levels become the inputs of the higher levels [66]. Generally there are several levels to a hierarchical fuzzy inference system, but the architecture used to construct the connections varies, depending on how the inputs

from each level is to be grouped. Figure 3.24 shows an example of some of the architectures that can be used, which is based on those drawn by [89].

The architecture for a hierarchical fuzzy inference system can be defined in one of two ways. The first method is the use of an expert, whereby the expert supplies all of the required knowledge for building the system [89].

E-mail Traffic Analysis Using Hierarchical Fuzzy Inference

Hierarchical fuzzy inference is used in this research to analyse a suspect e-mail account's traffic data to find abnormal changes in communication behaviour between the suspect and particular associates. The purpose of this is to aid the user/analyst to locate communication links that have behaved "abnormally", so that the user/analyst can be informed of when and where the abnormal change in behaviour occurred. This search for abnormally behaving communication links is illustrated in Figure 3.25.

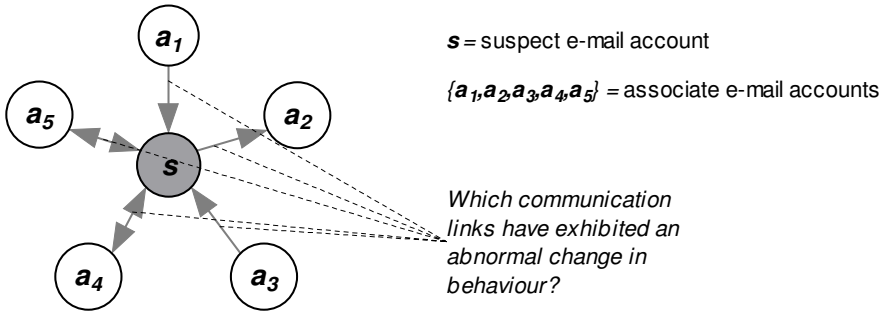


Figure 3.25: Analysis of the suspect's communication links.

In order to find "abnormal" behaviour with hierarchical fuzzy inference, a method called *anomaly detection* is used. Anomaly detection is a method that is commonly used in an area of computer network security called intrusion detection [90], where it is used for detecting new types of intrusion attacks previously unknown to the computer system or computer network. The concept of anomaly detection is based on the idea of establishing a baseline profile of what is considered "normal" system behaviour, then analysing the current or recent system behaviour for significant deviations from the baseline behaviour. If there are significant deviations from the baseline behaviour, then the user or administrator is alerted of the "abnormal" change in behaviour. Although anomaly detection is often used in computer network security [22], the same principles may also be applied for computer forensic applications when investigating suspected individuals for changes in communication behaviour.

Hierarchical fuzzy inference is used here for anomaly detection, to analyse the difference between a suspect e-mail account's profiled and recent communication behaviour. This difference is to be determined through the fusion of several behaviour measurements. The reason for using hierarchical fuzzy inference to do this comparison is that fuzzy techniques offer a number of advantages compared to the statistical and thresholding techniques currently used by [2, 42]. First, fuzzy techniques have been shown to be useful for overcoming the rigid boundary or "sharp boundary problem" found in anomaly detection [91]. Fuzzy techniques help to provide a softer distinction between "normal" and "abnormal" behaviour and can be used to rate changes in behaviour by degrees of abnormality. Second, fuzzy techniques have also been proven to be useful for dealing with uncertain or imprecise information [66], which is ideal for working with data extracted from observations of human behaviour. Thirdly, fuzzy techniques have also been shown to be useful for fusing together information from a number of inputs and applying heuristics to determine the overall status of the inputs [92]. This is ideal for improving the detection of "abnormal" changes in behaviour, based on the information fused from several behaviour measurements and summarising the overall abnormality status to the user/analyst. These advantages of fuzzy techniques make it useful to apply hierarchical fuzzy inference to compare the difference between a suspect e-mail account's recent and profiled e-mail traffic behaviour.

E-mail Traffic Communication Behaviour Measurements

Before abnormal communication behaviour can be detected with hierarchical fuzzy inference, a number of communication behaviour measurements need to be defined in order to determine what will be used to record a change in e-mail traffic behaviour. Thus, it is necessary to define communication behaviour measurements in order to describe particular aspects of an individual's e-mail traffic behaviour and to describe how that individual's communication behaviour may have changed over different periods of time. In this work, communication behaviour measurements are defined based on a basic set of e-mail traffic dimensions, which can be computed from the sender, recipient, and date/time information obtained from a collection of e-mail messages. These basic set of e-mail traffic dimensions are:

- **E-mail Traffic Volume** - based on a count of the number of e-mail messages sent/received by an individual per hour, per day, per week, or per

month, through interactions with a particular associate. This provides information on the traffic flow of e-mail messages sent/received by an individual and the rate at which messages are being exchanged with particular associates.

- **Delays Between E-mails Sent (or "Sending Delays")** - based on a measure of the time delays between each e-mail message sent/received by an individual. This provides information on expected delay times between each message sent/received by an individual through interaction with particular associates.

- **Replying Response Time (or "Replying Delays")** - based on a measure of the time it takes for an individual to write a response e-mail message in reply to messages received from particular associates. Also considers the reverse action for the time it takes for particular associates to reply back to an individual. This provides information on how quickly an individual is expected to reply or receive a reply through interaction with particular associates.

The purpose for defining these three e-mail traffic dimensions is to establish a set of behaviours from which information can be obtained about the level of traffic activity occurring between an individual and their associates. The type of behaviours being examined is the way in which individuals act and respond to others through e-mail communications. The idea behind this is to use e-mail traffic volume to capture information about the amount and variation of e-mail messages sent/received, and to use sending and replying delays to capture information about the waiting times between each e-mail sent. By using each of these aspects of an individual's e-mail traffic communication behaviour, these dimensions provide information about the level of traffic activity occurring between an individual and particular associates.

After defining the basic set of e-mail traffic dimensions, a set of communication behaviour measurements can be defined to compute measures for each of the e-mail traffic dimensions. Each of the measures computed extracts information that describes and summarises a particular aspect of an individual's e-mail traffic behaviour. To establish the type of communication behaviour measurement that can be used, a concept map was developed to determine the behaviour measurements used for each e-mail traffic dimension. This concept map is shown in Figure 3.26, illustrating the linkage between the communication behaviour measurements and the three e-mail traffic dimensions. It should be noted that

these communication behaviour measurements take into account the *direction* of e-mail traffic, meaning that the measurements for incoming and outgoing e-mail messages are considered as different aspects of an individual's communication behaviour.

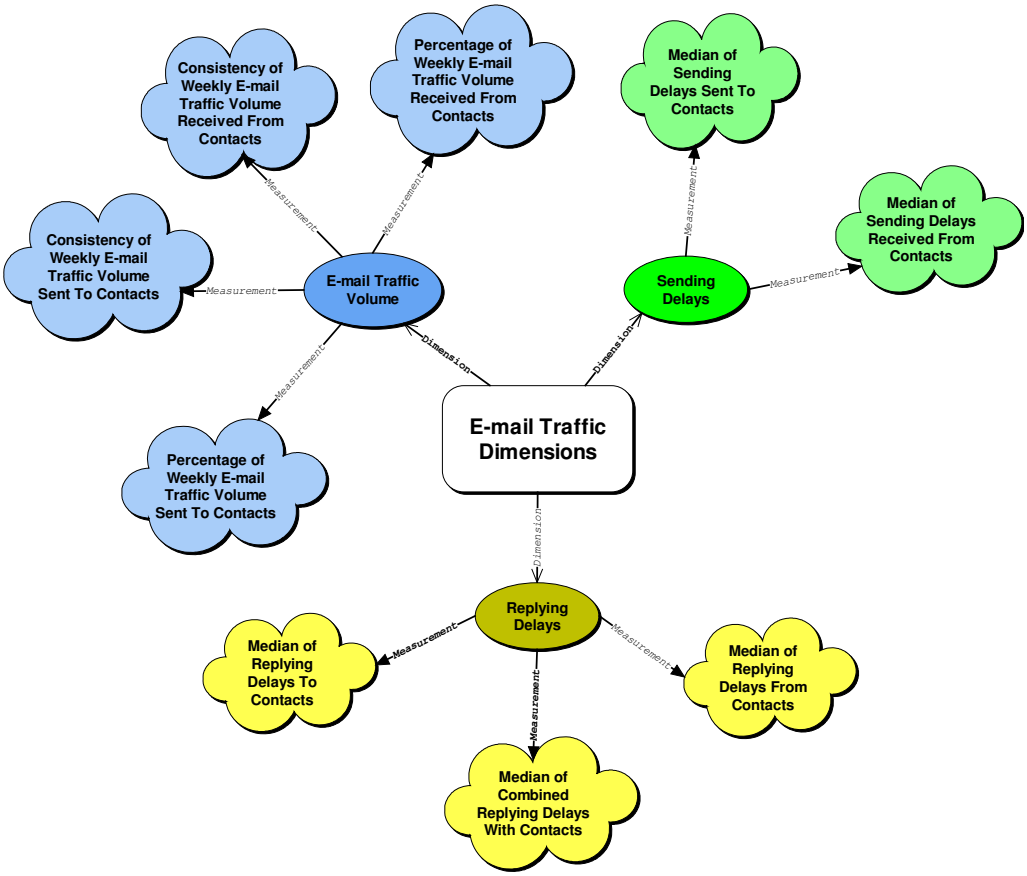


Figure 3.26: Mapping of e-mail traffic behaviour measurements.

The communication behaviour measurements shown in Figure 3.26 each use statistical methods [63, 76, 93] to compute measures for each e-mail traffic dimension. The method used to compute the measures for incoming and outgoing e-mail traffic behaviour are described as follows:

- **Consistency of Weekly E-mail Traffic Volume** - considers how "consistent" or "reliable" an individual is with the weekly number of e-mail messages sent to particular associates. Also considers how consistent particular associates are with the weekly number of e-mail messages sent to the individual. This behaviour measurement is measured by computing an autocorrelation of the time-series of the weekly number of e-mail messages sent (or received) by an individual. The purpose of using autocorrelation is to examine the variability or randomness of the number of e-mails sent

or received by an individual through interaction with particular associates. The autocorrelation formula used is Eq. (3.1) from [76], where: $x_1, ..., x_N$ are a set of N observations, $\bar{x}_{(1)}$ is the mean of the first $N-1$ observations, and $\bar{x}_{(2)}$ is the mean of the last $N-1$ observations.

- **Percentage of Weekly E-mail Traffic Volume** - considers the percentage e-mail messages sent by an individual to each of their associates and also the percentage of e-mail messages received by an individual from each associate. This behaviour measurement is measured by computing the average percentage of e-mails sent to or received from particular associates each week (e.g. 10% of e-mails per week to contact A, 40% per week to contact B, 50% per week to contact C).

- **Median of Sending Delays** - considers the most likely expected time delay between e-mail messages sent to particular associates, or the expected time delay between messages received from particular associates. This behaviour measurement is measured by computing the statistical median of the sending delays for messages sent or received by an individual.

- **Median of Replying Delays** - considers the most likely expected response delay between e-mail messages replied to particular associates, or the expected response delay for messages received back from particular associates. This behaviour measurement is measured by computing the statistical median of the replying delays for messages sent or received by an individual.

$$r = \frac{\sum_{t=1}^{N-1}\left(x_t - \bar{x}_{(1)}\right)\left(x_{t+1} - \bar{x}_{(2)}\right)}{\sqrt{\left[\sum_{t=1}^{N-1}\left(x_t - \bar{x}_{(1)}\right)^2 \sum_{t=1}^{N-1}\left(x_{t+1} - \bar{x}_{(2)}\right)^2\right]}} \tag{3.1}$$

$$\bar{x}_{(1)} = \sum_{t=1}^{N-1} x_t / \left(N-1\right) \tag{3.2}$$

$$\bar{x}_{(2)} = \sum_{t=2}^{N} x_t / \left(N-1\right) \tag{3.3}$$

Each of the nine communication behaviour measurements described above all extract information about different aspects of an individual's e-mail traffic behaviour. However, it should be noted that the e-mail traffic dimensions and communication behaviour measurements defined is not an exhaustive list of all

possible behaviours that can be extracted from e-mail traffic data. These e-mail traffic dimensions and communication behaviour measurements are the set that have been chosen as the focus for this work. Given that a set of communication behaviour measurements have been defined, these can be used to provide input information for the hierarchical fuzzy inference system to analyse for abnormal changes in e-mail traffic behaviour.

Processing of E-mail Traffic For Anomaly Detection

To analyse the suspect e-mail account's traffic data for abnormal changes in communication behaviour, the data is processed through a number of steps as shown in Figures 3.27 and 3.28. In the preprocessing step, the suspect's e-mail traffic data is formatted so that messages with multiple recipients are treated as separate messages sent to many recipients at the same time. This is to allow the communication links between the suspect and each associate to be analysed separately to find abnormal changes in communication link behaviour. After the e-mail traffic data has been formatted, two portions of the suspect's e-mail traffic data are selected for the "profiling" and "surveillance" periods. The profiling period is a historical period of time that is specified by the user/analyst to establish the "normal" behaviour of the suspect, while the surveillance period is a more recent period of time specified by the user/analyst to establish the "current" behaviour of the suspect.

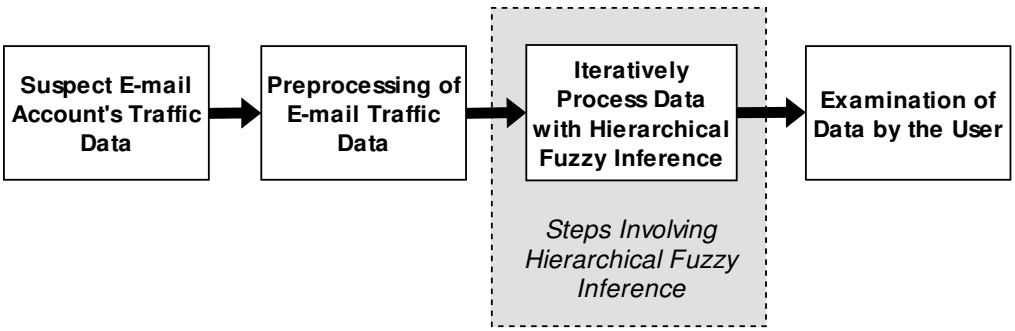


Figure 3.27: Process used for finding abnormal changes in communication behaviour.

Once the profiling and surveillance period portions of the data have been selected, nine communication behaviour measurements are computed for each communication link. These nine communication behaviour measurements are computed for both the profiling and surveillance periods, as illustrated in Figure 3.29. For this analysis, it is assumed that all of the suspect's communication links that

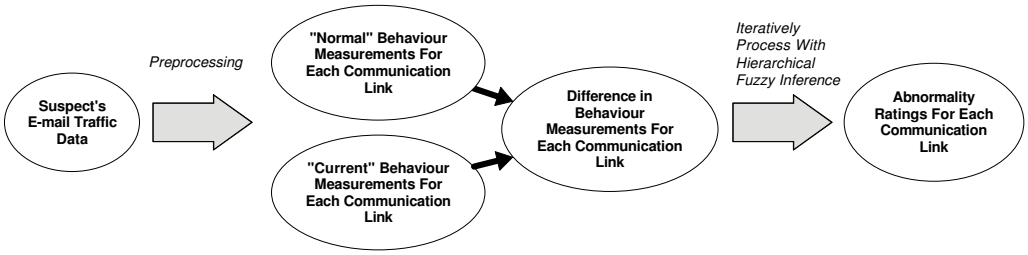


Figure 3.28: Data produced for finding abnormal communication links.

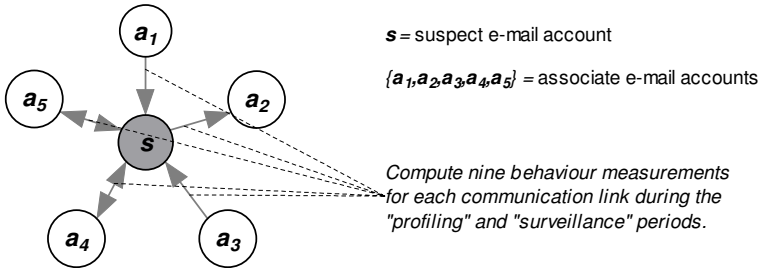


Figure 3.29: Analysing each of the suspect's communication links.

are analysed must initially exist during the profiling period. Any new communication links that are formed between the suspect and new associates during the surveillance period are not considered for this analysis. Based on this assumption, the nine communication behaviour measurements are computed for the suspect's profiling period in the following manner:

$$x_{ij} = f_j\left(\mu_i\right), \text{for } i = 1, ..., K; j = 1, ..., 9 \tag{3.4}$$

where i is the iteration number of each associate, K is the total number of associates during the profiling period, j is the iteration number of each communication behaviour measurement, μ_i is a set of observations of interaction with associate i during the profiling period, f_j is a function used to compute behaviour measurement j during the profiling period, and x_{ij} is the "normal" behaviour measurement for the interaction with associate i measured using behaviour measurement j. Similarly, the nine communication behaviour measurements are computed for the suspect's surveillance period as follows:

$$y_{ij} = f_j\left(\nu_i\right), \text{for } i = 1, ..., K; j = 1, ..., 9 \tag{3.5}$$

where ν_i is a set of observations of interaction with associate i during the surveillance period, f_j is a function used to compute communication behaviour measurement j during the surveillance period, and y_{ij} is the current behaviour measurement for the interaction with associate i measured using behaviour measure-

ment j.

After computing the profiling and surveillance period communication behaviour measurements, these measurements are further preprocessed to produce the behaviour change measurements (difference measurements) specifying the change in communication behaviour for each of the suspect's communication links. The difference between the "current" and "normal" behaviour is computed as follows:

$$d_{ij} = |y_{ij} - x_{ij}| \text{, for } i = 1, ..., K; j = 1, ..., 9 \tag{3.6}$$

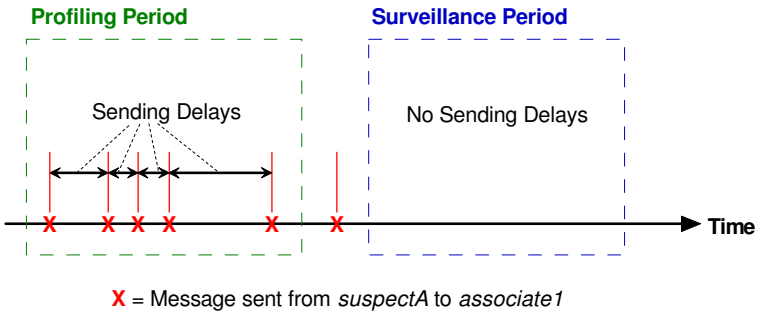
where d_{ij} is the difference between the "current" and "normal" behaviour measurements. It should be noted that the communication behaviour measurements related to "sending delays" and "replying delays" may have cases where the communication behaviour measurements cannot be computed during either the profiling or surveillance period. These cases may occur when no e-mail messages or only one e-mail message has been exchanged between the suspect and associate. Such cases mean that a sending or replying delay cannot be recorded, since at least two e-mail messages are required to record the time delay between messages. If such a case occurs while computing Equation 3.6, then these are flagged during preprocessing to signify the existence of unknown values for that particular difference measurement. Examples of these special cases are shown in Figure 3.30.

Once the behaviour change measurements have been computed for each of the suspect's communication links, the behaviour change measurements (including any flagged unknown value cases) are processed by the hierarchical fuzzy inference system. These behaviour change measurements are processed as follows:

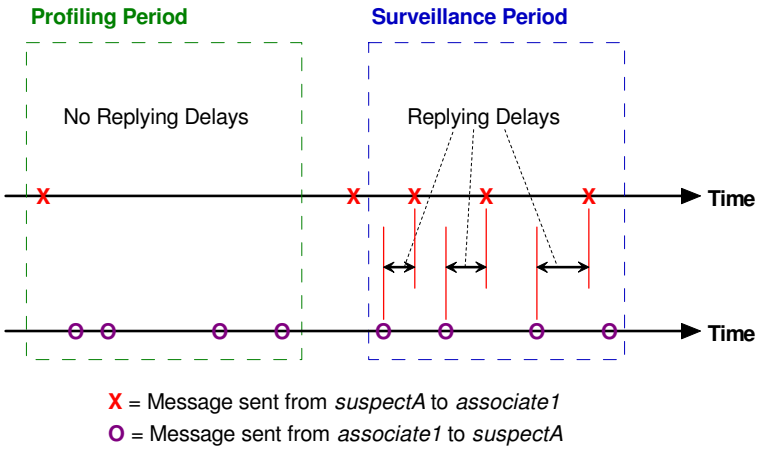
$$z_i = g\left(d_{i1}, d_{i2}, ..., d_{i9}\right) \text{, for } z_i \text{ in } [0, 1] \tag{3.7}$$

where $(d_{i1}, d_{i2}, ..., d_{i9})$ are the behaviour change measurements for communication with associate i, g is the function represented by the hierarchical fuzzy inference system, and z_i is the overall change in e-mail traffic behaviour for the communication link between the suspect and associate i. The final output from the hierarchical fuzzy inference system, the "abnormality rating" z_i, produces a number in the range $[0, 1]$ to summarise the degree of abnormality for communication link i's change in behaviour. Values where $z_i \rightarrow 0.0$ indicate little change in e-mail traffic communication behaviour, while values where $z_i \rightarrow 1.0$ indicate a large change in e-mail traffic communication behaviour.

The process just described shows how the hierarchical fuzzy inference system



(a) Example of an unknown sending delays case.



(b) Example of an unknown replying delays case.

Figure 3.30: Examples of unknown sending delay and replying delay cases.

can be used to compare differences in a suspect e-mail account's profiled and surveillance period communication behaviour. The final abnormality rating numbers given by the hierarchical fuzzy inference system, summarises all of the changes in communication behaviour observed for each communication link. The abnormality rating produced for each communication link can be used by the user/analyst to get a quick overview of any abnormal changes in communication behaviour observed from each of the suspect's communication links.

The architecture used for the hierarchical fuzzy inference system is shown in Figure 3.31. This shows one of the possible architectures that can be used for fusing the behaviour change measurements d_{i1} to d_{i9}. The behaviour change measurements are processed by the hierarchical fuzzy system through three layers. The low-level fuzzy system modules in the first layer of Figure 3.31 processes the behaviour change measurements by grouping the inputs according to: *changes in incoming traffic volume, changes in incoming traffic delays, changes in outgoing traffic volume*, and *changes in outgoing traffic delays*. The outputs from the first layer are then used as inputs into the second layer of the architecture, which groups the new inputs as: *changes in incoming traffic* and *changes in outgoing*

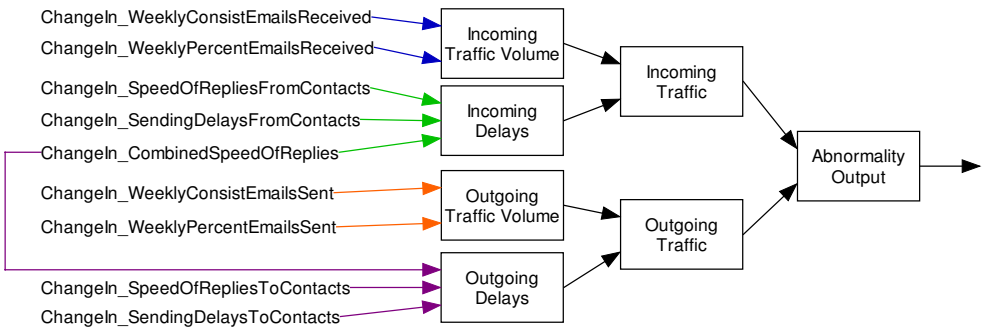


Figure 3.31: The architecture used for the hierarchical fuzzy inference system.

traffic. Finally, the outputs from the second layer are passed as inputs into the fuzzy system module in the third layer. This last fuzzy system module determines the overall change in e-mail traffic behaviour patterns, based on changes observed from either incoming e-mail traffic or outgoing e-mail traffic.

In the final output of the hierarchical fuzzy inference system, a number z_i is produced that indicates the degree of change occurring between the suspect and a particular associate. This number however, only provides a high level summary of the change in e-mail traffic behaviour observed from the suspect's data. If further details about the change in e-mail traffic behaviour is required, then the computational intelligence approach can be applied by utilising other computational techniques to provide additional perspectives on those changes found. A further description of the combined use of hierarchical fuzzy inference and visualisation techniques is given in Section 4.4 and the hierarchical fuzzy inference system shown in Figure 3.31 is evaluated in a case study in Section 5.3.

3.5 Summary

This chapter proposed the use of a "computational intelligence" approach for analysing e-mail traffic behaviour. The computational intelligence approach described was defined as a method for using a set of computational techniques, to extract information from data and present the information to the user/analyst in a useful and intelligent manner. Based on this definition, any combination of computational technique from areas such as statistics, artificial intelligence, or visualisation, may be used for highlighting particular patterns, relationships, or features present in the data.

Two types of computational techniques were described in this chapter, which are utilised for computational intelligence. The first type, visualisation techniques, is

used for presenting information about the relationships and patterns hidden in e-mail traffic data. The purpose of using this type of technique is to quickly convey useful information to the user about the traffic behaviour of suspect e-mail accounts. For this research, there were two types of visualisation techniques used. The first, social network visualisation, is used for presenting useful information about the connections between multiple e-mail accounts and also the strength of the connections, based on the number of e-mail messages exchanged. The second technique, time-series visualisation, is used for presenting information about variations in the volume of e-mail traffic sent or received by particular e-mail accounts. Both of these techniques were noted to be useful for visual exploration of the data, which can help the user/analyst to explore and understand the data. However, there were some limitations noted with the use of social network visualisation and time-series visualisation, which was related to the time and effort required to search for unusual or abnormal behaviour. To overcome these limitations, feature extraction techniques are considered for aiding the user/analyst to find these behaviours.

Feature extraction techniques were the second type of computational technique described in this chapter. The purpose of using feature extraction techniques is to locate certain data records that possess particular features or patterns. The use of feature extraction techniques is considered for e-mail traffic analysis, in order to aid the user/analyst to search for unusual or abnormal changes in e-mail traffic behaviour. For this research, there were two types of feature extraction techniques used to analyse the behaviour of suspect e-mail accounts. The first, decision tree classification, is used to identify occurrences of “unusual” changes in e-mail traffic behaviour between a suspect e-mail account and their associates. These unusual changes in e-mail traffic behaviour are found by comparing the leaf nodes of the decision tree, to determine how a suspect’s e-mail traffic interaction behaviour has changed over time. The second technique, hierarchical fuzzy inference, is used to provide abnormality ratings for a suspect e-mail account’s communication links. A hierarchical fuzzy inference system was developed to fuse information from nine behaviour change measurements to determine the degree of change in each communication link. The final output from the hierarchical fuzzy inference system is used as the abnormality rating, summarising the abnormal change in communication behaviour between a suspect and a particular associate. Both of the feature extraction techniques described in this chapter can extract features or patterns from the data, in order to indicate to the user/analyst where they are located.

Overall, this chapter described two types of computational techniques that can be used together as a set for computational intelligence. This computational intelligence approach of using a set of techniques, suggests that the user/analyst may be able to use computational intelligence to view and understand e-mail traffic behaviour from a variety of perspectives. In the next chapter, it will be described how the computational techniques mentioned in this chapter are used together as a set to analyse e-mail traffic behaviour.

Chapter 4

Development of the E-mail Traffic Analysis System

4.1 Introduction

In the previous two chapters, a coverage was provided on types of computational techniques and methods used for analysing e-mail traffic behaviour. Chapter 2 reviewed a range of current methods used for analysing e-mail traffic behaviour and categorised each of these methods according to the level of analysis or level of detail provided to the user/analyst. Chapter 3 followed on by defining the meaning of computational intelligence for this thesis and described it as an approach for utilising a set of computational techniques to analyse e-mail traffic behaviour. Chapter 3 also described each of the visualisation and feature extraction techniques used in this research for analysing e-mail traffic behaviour. This chapter describes the development of the e-mail traffic analysis system and how it is used to integrate the set of visualisation and feature extraction techniques previously described in Chapter 3, for computational intelligence.

In Section 4.2, an overview is provided on the architecture of the e-mail traffic analysis system. The first part of this section covers the design of the e-mail traffic analysis system and overviews the architecture and the main components of the system. The second part of this section covers the implementation of the e-mail traffic analysis system and describes the software used to develop certain components of the system. The next section of this chapter, Section 4.3, describes the types of e-mail traffic data used for evaluating the e-mail traffic analysis system. Section 4.3.1 describes the purpose of using simulated e-mail traffic data for the research and describes the conceptual simulation model created for

simulating an e-mail system. Section 4.3.2 describes the use of the Enron e-mail dataset and why it has been chosen for the research. The final section, Section 4.4, provides an overview of two different approaches in which the e-mail traffic analysis system can be used for analysing e-mail traffic data.

4.2 System Architecture

The e-mail traffic analysis system is a conceptual system that has been developed to explore the use of different computational techniques for extracting information about the traffic behaviour of e-mail users [94, 95]. The overall purpose for designing this system is to apply computational intelligence by integrating both the visualisation and feature extraction techniques previously described in Chapter 3. The aim of integrating different sets of computational techniques is to enable the user/analyst to consider a range of perspectives for examining and understanding the communication behaviour of suspect e-mail accounts.

To design the e-mail traffic analysis system, there were two important elements considered for the architecture of the system. The first is the use of a modular architecture, to allow for components to be added to the system as individual modules. The purpose of using this design approach is that new functionality can be added at various stages during development of the system, without requiring a total redesign or restructuring of the system. This type of modular approach is useful for if additional computational techniques are to be added to the system, which will provide additional perspectives to the user/analyst for analysing e-mail traffic behaviour.

The second important element in the architecture of the system is allowing the user to control particular parameters used for analysing the traffic behaviour of e-mail users. The reason for providing this is to allow the user or analyst a way of specifying which portions of the data are to be analysed (e.g. specifying particular e-mail accounts or time periods), so that the user/analyst is in control of the analysis process. This is considered important since it enables the system to make use of the user/analyst's knowledge of the investigation task and allows the system to focus on tasks that are difficult for the user/analyst to perform (e.g. identifying particular types of communication behavioural patterns). It also allows the user/analyst to use their judgement and experience to determine whether enough useful information is presented by the system, or whether further investigation is required.

4.2.1 Overview of the System

The architecture of the e-mail traffic analysis system developed for this research is shown in Figure 4.1. The system comprises of a number of components, each of which serve different purposes for processing e-mail traffic data. In terms of the overall data processing for the system, the data is processed through two main stages in order to extract information on the traffic behaviour of e-mail users. The two stages used are: the "Data Acquisition and Filtering Stage" and the "Information Extraction Stage".

At the first stage, "Data Acquisition and Filtering Stage", e-mail data is collected from the e-mail system and cleaned/filtered to fill in missing values, remove noise, and fix up inconsistent data. Once the e-mail data is filtered and cleaned it is stored into the e-mail traffic database. At the second stage, "Information Extraction Stage", e-mail traffic data is extracted and processed through different components to provide information on e-mail traffic behavioural patterns and also on the location of unusual or abnormal behaviour. The information extracted and processed are:

- The social connections between e-mail users ("Social Network Data Processor").
- The level or volume of traffic generated by e-mail users ("Time-Series Data Processor").
- Unusual variations in interactions between an e-mail user and their associates ("Decision Tree Classification Output").
- The degree of change in communication link behaviour between an e-mail user and each of their associates ("Anomaly Detection Unit" and "Hierarchical Fuzzy Inference Module").

After these e-mail traffic behavioural patterns and unusual behaviour information have been extracted and processed, they are visualised and presented to the user for analysis. The next sections describe the role of each of the major components in the e-mail traffic analysis system.

E-mail System Data Collection

This part of the e-mail traffic analysis system obtains traffic information from the e-mail system by analysing the header section of each e-mail message to extract

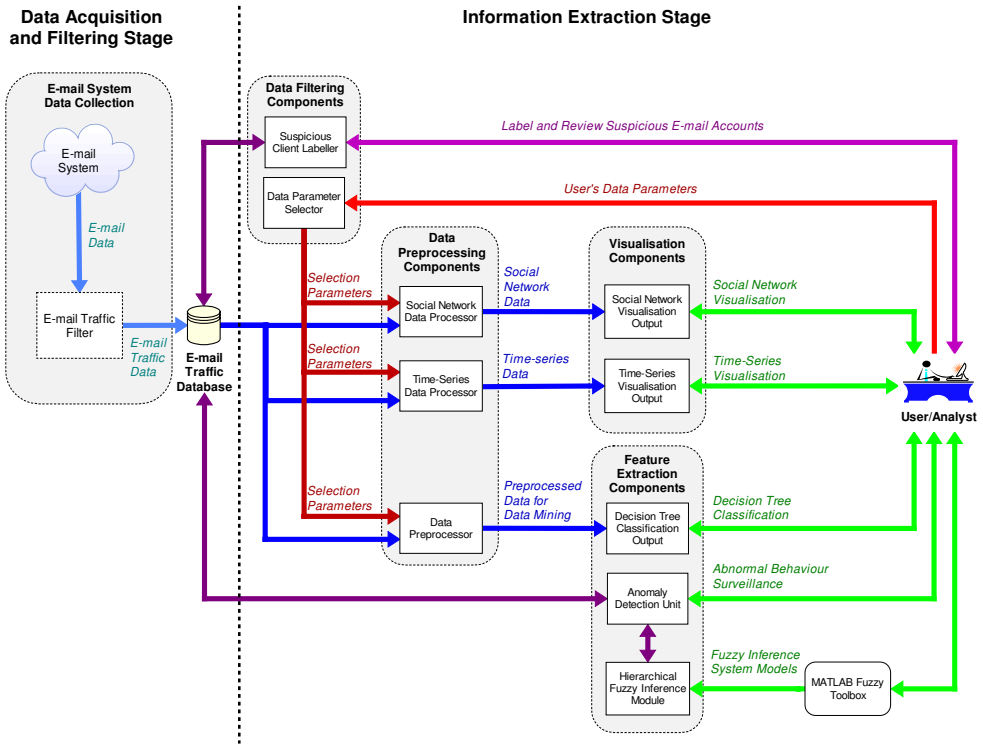


Figure 4.1: Overview of the e-mail traffic analysis system.

the sender, receiver, and date/time information. All other information from the original e-mail data such as the message body or content of e-mail messages is excluded through the e-mail traffic filter. Prior to entering data into the e-mail traffic database, the data is cleaned and filtered to massage the data into the correct format for the analysis. For the type of analyses performed by the e-mail traffic analysis system, all messages with multiple recipients are treated as individual messages sent to multiple recipients at the same time. This type of information is required by the decision tree classification and hierarchical fuzzy inference techniques, as previously mentioned in Sections 3.4.1 and 3.4.2, in order to allow those techniques to analyse the individual interactions between a suspect and each of their associates.

After entering the basic sender, receiver, and date/time information into the e-mail traffic database, the data is further processed by computing the sending delays (i.e. delays between each e-mail message sent) and replying delays (i.e. delays in responding back to messages received from associates) for each e-mail account. These sending delay and replying delay measurements are computed and stored in the database, in order to provide pre-computed measurements for the “Anomaly Detection Unit” and “Hierarchical Fuzzy Inference Module” components. These measurements are stored in order to improve the computation

time used by the “Anomaly Detection Unit” component.

Data Filtering Components

The data filtering components have an important role in the e-mail traffic analysis system, by allowing the user/analyst to specify parameters that control what is analysed by the system. This enables the system to only analyse a small subset of the e-mail traffic data from the original e-mail system. This approach is considered useful, since it enables the user/analyst to steer the direction of the analysis process, making the system only analyse parts of the data that are relevant to the investigation task.

There are two components used for data filtering in the e-mail traffic analysis system. The first is the “Data Selection Parameter” component, which is used to control what is analysed by the social network visualisation, time-series visualisation, and decision tree classification techniques. The purpose of this component is to provide a way of filtering out information that do not need to be visualised by the visualisation techniques and to also filter out data not required to be analysed by the decision tree classification technique. Table 4.1 provides an example of the parameters that are used by the “Data Selection Parameter” component to filter out information processed by the e-mail traffic analysis system.

The second data filtering component used in the e-mail traffic analysis system is the “Suspicious Client Labeller” component. This particular component is used in conjunction with the “Anomaly Detection Unit” component to label e-mail accounts that are deemed ‘suspicious’. This is performed by the user/analyst specifying the suspicion level of particular e-mail accounts, by assigning numbers between the range of 0 to 1 for those e-mail accounts. Numbers close to 0 indicate a low suspicion value, while numbers close to 1 indicate a high suspicion value. The labelled e-mail accounts are then stored in the e-mail traffic database, which are later referred to by the Anomaly Detection Unit. The Anomaly Detection Unit can then be used to select and profile the behaviour of e-mail accounts that have been assigned a suspicion value between a particular range (e.g. between the range of 0.5 to 0.9).

Preprocessing Components

The purpose of the preprocessing components in the e-mail traffic analysis system is to preprocess data into a suitable format for analysis by the visualisation

Table 4.1: Filtering options available through the use of the Data Parameter Selector component.

Filtering Parameter	**Description**
Time Span	Selects a specific period of time between a given start date/time and end date/time for the analysis.
Periodic Snapshots	Specifies that periodic snapshots are used for a time span by dividing the time span into a number of intervals. The effect on the visualisation outputs is a series of snapshots displaying how e-mail traffic behaviour changes over time.
Time-Series Resolution	Specifies the time resolution used to sample and analyse time-series data. Resolution can be specified either as hours, days, weeks, or months.
Selection of Specific E-mail Users	Filters the data so that only the selected list of e-mail users are analysed by the system.
Degrees of Separation	Extends the "Selection of Specific E-mail Users" parameter by adding closely connected e-mail users surrounding the current selection. Additional e-mail users are added based on the number of 'hops' between the currently selected e-mail users and other connected users (e.g. degree of 1 adds surrounding e-mail users within 1 hop, degree of 2 adds surrounding e-mail users within 2 hops). The concept is based upon the idea of "degrees of separation" [53].
Node Selection Constraint	Extends the "Selection of Specific E-mail Users" parameter by specifying whether to analyse only the communications between the selected e-mail users (i.e. exclude communications with users outside the selected group) or whether to analyse all communications with the selected e-mail users (i.e. include communications with users outside the selected group).
Social Link Connection Intensity	Specifies whether to display simple lines as the communication ties for the social network visualisation output, or to display lines with varying intensities/widths as the communication ties for the social network visualisation output.
E-mail Traffic Direction	Specifies the direction of e-mail traffic to analyse for each e-mail user. This can be specified either as incoming traffic, outgoing traffic, or both directions of traffic.

and feature extraction components of the system. This is required in order for information to be extracted on particular aspects of an e-mail user's communication behaviour. The preprocessing components shown in Figure 4.1 follow the preprocessing steps that were described previously in Sections 3.3.1, 3.3.2, and 3.4.1, in which the original e-mail traffic data is formatted for the social network visualisation, time-series visualisation, and decision tree classification techniques.

Visualisation Components

The visualisation components in the e-mail traffic analysis system enable the user/analyst to understand the communication behaviour of e-mail accounts through the visual exploration of the data. Social network visualisation component provides an overview of the connections between e-mail accounts and allows for examination of the strength of communication ties between particular e-mail accounts. Time-series visualisation component alternatively provides information on the level of e-mail traffic exhibited by e-mail accounts over time and also provides information on the temporal variations in interaction between particular e-mail users.

It should be noted that both the social network visualisation and time-series visualisation techniques can be used in collaboration to provide different perspectives on e-mail traffic behaviour. Social network visualisation can be used to provide a quick overview of the e-mail social network, while time-series visualisation can be used to zoom in on certain communication links to examine the variations in traffic volume exchanged between e-mail accounts. The overall effect of this combined use is that user/analyst has a broader understanding of e-mail users' traffic behaviour by viewing their behaviour through two different perspectives. Additional visualisation techniques may be later added to the e-mail traffic analysis system, to provide other perspectives for the user/analyst to understand e-mail traffic behaviour.

Feature Extraction Components

The feature extraction components in the e-mail traffic analysis system provide different methods for finding unusual or abnormal changes in communication behaviour. The decision tree classification component provides a way of analysing particular suspect e-mail accounts for unusual variations in communication behaviour between the suspect and their associates. The "Anomaly Detection Unit"

and the “Hierarchical Fuzzy Inference Module” components are used together to analyse suspect e-mail accounts for abnormal changes in e-mail traffic behaviour between the suspect and each associate.

Both of these feature extraction components can be used in collaboration with the visualisation components to examine e-mail traffic behaviour. This can be performed by using either of the feature extraction components to locate areas in the data where there is unusual or abnormal changes in communication behaviour. These occurrences of unusual/abnormal behaviour can then be further investigated by the user through the use of the visualisation components to understand the nature of the unusual/abnormal behaviour found. The steps used for applying this computational intelligence approach with feature extraction and visualisation techniques is explained further in Section 4.4.

4.2.2 Implementation

The e-mail traffic analysis system has been implemented using the Python programming language [96] to develop various components of the system and MySQL [97, 98] to implement the e-mail traffic database. A diagram of the actual implementation of the e-mail traffic analysis system is shown in Figure 4.2. In comparison to the original diagram of the system in Figure 4.1, Figure 4.2 shows that certain components of the system, such as the Social Network Visualisation Output, Time-Series Visualisation Output, and Decision Tree Classification Output components, were based on existing software applications. The types of existing software used for these components were: GUESS [75] for social network visualisation, TimeSearcher 2 [78] for time-series visualisation, and WEKA (Waikato Environment for Knowledge Analysis) [84] for decision tree classification. The decision tree classification algorithm used from the WEKA data mining program is the J4.8 decision tree algorithm, which is WEKA’s implementation of the C4.5 revision 8 decision tree algorithm [99].

The reason existing software applications are used for these components of the system, is to allow for more focus on evaluating the information presented by each computational technique rather than spending time on the development work required to implement each computational technique. This helped to reduce the amount of time required to implement the visualisation and feature extraction components of the system and allowed for more focus on implementing the back end components that perform the data processing (e.g. the data filtering and preprocessing components). It also allowed for experimentation with dif-

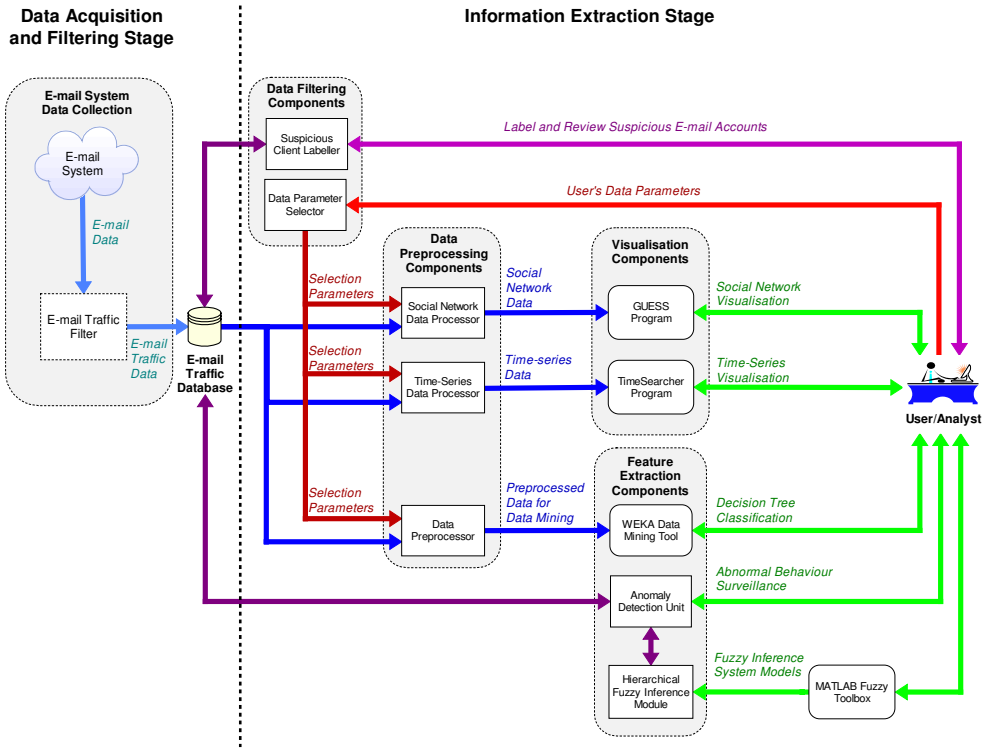


Figure 4.2: Implementation of the e-mail traffic analysis system.

ferent software applications, to determine which programs provided the types of features needed for exploring the data.

For the “Hierarchical Fuzzy Inference Module” component, this has been implemented using a combination of C code obtained from the MATLAB Fuzzy Toolbox [100] and extending the functionality of the original fuzzy inference engine by using Python [96]. The original fuzzy inference engine from MATLAB has been extended through Python, to allow for easy construction of the architecture required for using hierarchical fuzzy inference systems. To develop each of the low-dimensional fuzzy system modules, the MATLAB Fuzzy Toolbox has been used to define the rules and membership functions for each module in the hierarchical fuzzy system. Each of the fuzzy system modules created in MATLAB are then exported and read by the “Hierarchical Fuzzy Inference Module” component to process the data obtained through the “Anomaly Detection Unit”, using the hierarchical architecture defined in Section 3.4.2.

4.3 E-mail Traffic Data

For this research, careful consideration was given toward the type of data used to evaluate the e-mail traffic analysis system. The things considered for the e-mail

traffic data were: using data with known behavioural characteristics and using sufficient amounts of e-mail data for analysis. The knowledge of behavioural characteristics in the data means that it can allow for verification of whether the correct type of communication behaviour is being observed for each e-mail user. It also allows for determination of whether the e-mail traffic analysis system can correctly detect "unusual" e-mail traffic behavioural patterns that are present in the e-mail data. Without some prior knowledge about the nature of the data collected (e.g. if certain e-mail users have been known to conduct criminal or terrorist activities), it will be difficult to test whether the system is able to detect the correct type of e-mail traffic behavioural patterns.

The use of sufficient amounts of e-mail data for analysis is also considered important, since the data ideally should provide enough information on each e-mail user's typical communication behaviour and include all the communication links between the suspect e-mail accounts and their associates. In addition to this, the data ideally should also contain a large number of e-mail users, to help demonstrate that the e-mail traffic analysis system may be applicable to analysing data from large e-mail systems (e.g. systems that contain around 10,000 e-mail users). Without enough data for analysis, there would be incomplete social network information (i.e. missing nodes/e-mail users from the network) and there would be missing traffic information which could be available at the missing network nodes.

There are two types of e-mail traffic data used in this research: simulated e-mail traffic data and the Enron e-mail dataset. Both types of data have known behavioural characteristics and also provide sufficient amounts of data to be analysed by the e-mail traffic analysis system. Each of type of e-mail traffic data are described next in Sections 4.3.1 and 4.3.2.

4.3.1 Simulated E-mail Traffic Data

A simulation tool has been developed for this research to allow for simulation of an e-mail system and to generate simulated e-mail traffic data for the e-mail traffic analysis system [94]. The purpose of the developing the simulation tool is to create and use e-mail traffic data with known behavioural characteristics. This allows for evaluation of whether the e-mail traffic analysis system detects the expected type of behaviour that was known to be entered into the simulation model. The other purpose of the simulation tool is to provide the ability to generate an e-mail system of any desired size and to enter different configurations for

the behaviour of e-mail users. This enables one to experiment with e-mail systems containing different numbers of e-mail users and to also experiment with different behaviour configurations for each e-mail user.

4.3.1.1 Conceptual E-mail System Model

The simulated e-mail traffic data used for the research is created from a conceptual simulation model of the e-mail system [94], which has been modelled as a discrete-event simulation system [101, 102]. The conceptual simulation model comprises of two main types of entities: e-mail clients and behaviour models. The e-mail client entities represent the e-mail accounts of e-mail users, through which e-mail messages are sent from and received by each e-mail account. The diagram in Figure 4.3 shows that the attributes of each e-mail client consist of an e-mail address, an e-mail server name, a list of social contact e-mail addresses, an e-mail mailbox, and an assigned behaviour model. The behaviour model entities, also shown in Figure 4.3, represent the behavioural attributes of different individuals and defines how an individual will interact and respond to others via e-mail. A behaviour model is assigned to different e-mail clients, to define how the client's e-mail account will behave. The attributes of each behaviour model is represented by a set of five personality traits, which are based upon the five personality trait dimensions described in [103, 104, 105].

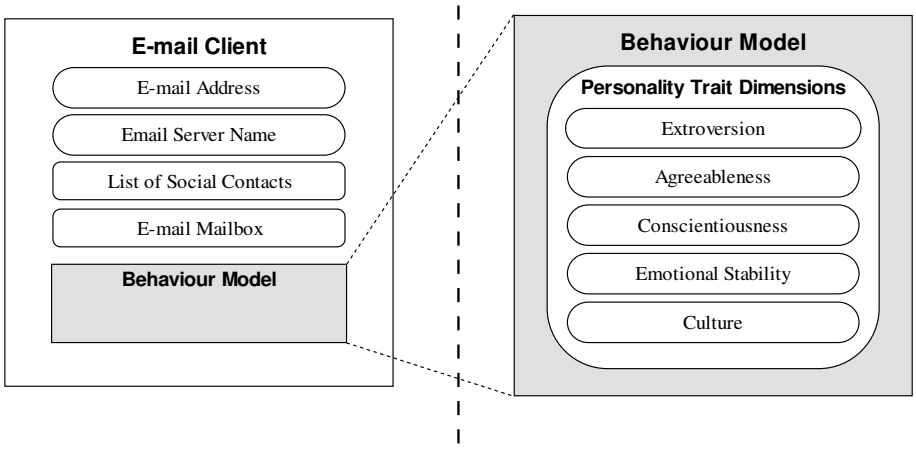


Figure 4.3: The entities making up the e-mail system model.

The basis for modelling the e-mail system as e-mail client and behaviour model entities, is to allow for focus on the modelling of interactions between different e-mail clients, rather than trying to model the e-mail client and e-mail server interactions. This is because the interest of the analysis for this work is the interaction behaviour of individual e-mail users, in terms of how often they send

e-mail messages to other users (e.g. several times per day, or once every few days) and also how they respond to e-mail messages received from other e-mail users (e.g. usually replying back on the same day, or usually replying back within one week of receiving an e-mail message). In this way, the e-mail system model is set up as a series of end-to-end communication points between different e-mail clients, as opposed to a more centralised set-up where the e-mail servers are situated in between the lines of communication between e-mail clients.

The purpose of modelling behaviour as a separate entity from the e-mail client and its e-mail account details, is to treat each behaviour model entity as a type of behavioural profile that specifies the unique behaviour characteristics of particular individuals. In this way, each individual modelled generates a unique set of e-mail traffic behavioural patterns that can be linked back to the attributes that make up the individual's behavioural profile. For example, the behaviour for one individual may be that they send out several e-mails per day and respond to e-mails within one or two days after receiving the original e-mail message. This type of e-mail traffic behaviour pattern can be linked back to attributes of the individual's behavioural profile, which describe the individual as 'outgoing' and 'conscientious'. Alternatively, there could be another individual that has a different set of e-mail traffic behavioural patterns, in which that individual also sends out several e-mails per day, but always takes at least one week to respond to e-mails they receive or does not respond at all. This type of behavioural traffic pattern can also be linked back to the individual's behavioural profile, which describes this individual as 'outgoing', but 'lazy'.

These examples show that focusing on the behaviour of each e-mail user also provides a useful way for describing the behavioural attributes that cause a person to exhibit certain types of e-mail traffic behaviour. As a result of modelling the behaviour model and e-mail client as separate entities, a behaviour model can be assigned to multiple e-mail accounts to represent how some people use several e-mail accounts, but still exhibit similar types of e-mail traffic behavioural patterns through different e-mail accounts. Through this modelling approach, a model of the e-mail system can be created that consists of individuals with distinct types of behaviour profiles, using various e-mail accounts.

4.3.1.2 Personality Traits of the Behaviour Model

The behaviour model in the e-mail system model is made up of five personality trait dimensions, consisting of the dimensions of extroversion, agreeableness,

conscientiousness, emotional stability, and culture. Each of these personality trait dimensions is used to describe a particular aspect of person's personality and explain how a person's underlying traits is linked to their observed behaviour [103, 104, 105]. For example, if a person is described as being quite extroverted, it implies that they are observed to be very outgoing, social, and talk to a lot of people. Likewise, if a person is described as being introverted (the opposite of being extroverted), it implies that they are observed to be very shy, non social, and don't talk a lot to people. A second example is if a person is described as being very conscientious, it would imply that their behaviour is helpful, hard working, and dependable. For a person described as low conscientiousness, it implies their behaviour is careless, lazy, and not dependable. In Table 4.2, adapted from [104, 105], each of the personality trait dimensions is described by trait pair examples, which are used to describe the nature of a person's underlying tendencies and imply about their observed behaviour.

Personality Trait Dimensions	**Trait Pair Examples**
Extroversion-introversion	Talkative-silent; frank-secretive; adventurous-cautious; sociable-reclusive;
Emotional Stability	Calm-anxious; composed-excitable; poised-nervous;
Conscientiousness	Tidy-careless; responsible-undependable; hard working-lazy; persevering-quitting;
Culture	Creative-uncreative; intellectual-nonreflective; well educated-crude; prefers variety- prefers routine;
Agreeableness	Good natured-irritable; gentle-headstrong; cooperative-negativistic; not jealous-jealous;

Table 4.2: Description of the personality trait dimensions, through the use of trait pair examples.

Each of the personality trait dimensions in the behaviour model is assigned a degree value between 0 to 1. These degree values provide a scale of how strongly each personality trait dimension affects a person's behaviour [104], with values

close to 1.0 (high degree values) indicating a strong influence by the personality trait dimension and values close to 0.0 (low degree values) indicating a weak influence by the personality trait dimension. So for example, a degree value of 0.9 for extroversion and 0.7 for conscientiousness trait dimensions, describes a person that is extremely outgoing, hard working, and fairly responsible. Alternatively, a degree value of 0.2 for extroversion and 0.1 for conscientiousness trait dimensions, describes a person that is rather withdrawn, lazy, and unreliable. By assigning different degree values to each set of personality trait dimensions, unique behavioural profiles can be created that represent different individuals in the e-mail system model, each of whom exhibit varying degrees of behaviour for each of the personality trait dimensions.

Since it has been shown that personality trait dimensions and trait degree values provide a way of conceptualising the effects of personality trait dimensions on general behaviour, one can now describe how the personality trait dimensions can be modelled to affect e-mail communication behaviour. However, given that there has been few empirical studies conducted that examine the effects of personality traits on e-mail communication behaviour, some intuitive assumptions were made on how the personality trait dimensions affect e-mail behaviour. Table 4.3 provides a summary of how some of the personality trait dimension attributes from the behaviour model will be interpreted to affect e-mail communication behaviour, based on these intuitive assumptions.

For the extroversion trait dimension, it has been assumed that this will affect how an individual engages in communication with others by e-mail. This means that it will be interpreted as affecting the frequency of e-mails sent out by an e-mail client and affecting the speed that an e-mail client replies to received e-mails. In the conscientiousness trait dimension, it has been assumed that this will affect the how an individual acknowledges and responds to others via e-mail. This will be interpreted as affecting the probability that an e-mail client will send a reply message when they receive an e-mail message from other e-mail clients. For the emotional stability trait dimension, it has been assumed that this will determine how the emotional element of an individual will affect the individual's ability to be consistent with their communication behaviour (i.e. it is assumed that an emotionally stable person can think more clearly and hence act more consistently, compared to a person who is emotionally less stable and acts more erratically due to their inability to think clearly or rationally). In the case of the emotional stability trait dimension, it will be interpreted as affecting the variability in the time delays between e-mails sent and affecting the variability

Personality Trait Degree Values	Effect on Sending Out Of New E-mails	Effect on Replying To Received E-mails
High Degree of Extroversion	Higher frequency of e-mails sent	More likely to receive a fast reply
Low Degree of Extroversion	Lower frequency of e-mails sent	More likely to receive a slow reply
High Degree of Conscientiousness	-	Higher probability of returning a reply
Low Degree of Conscientiousness	-	Lower probability of returning a reply
High Degree of Emotional Stability	Lower variability in time delays between e-mails sent	Lower variability in replying delay time
Low Degree of Emotional Stability	Higher variability in time delays between e-mails sent	Higher variability in replying delay time
High Degree of Agreeableness	-	-
Low Degree of Agreeableness	-	-
High Degree of Culture	-	-
Low Degree of Culture	-	-

Table 4.3: The relationship between the effect of personality trait dimensions on e-mail communication behaviour, based on intuitive assumptions.

in the time delay for a reply message.

It has been considered that the personality trait dimensions of the behaviour model could also be modelled to affect other types of e-mail communication behaviour such as the addition of e-mail attachments, sending e-mails to multiple recipients, and forwarding received e-mails to other e-mail clients. However, it was decided that the conceptual model of the e-mail system should be kept as simple as possible, to allow us to focus on the essential communication behaviours required for e-mail communication interaction. As a result, a behaviour model has been developed that affects the sending and replying behaviour of e-mail clients, with each interaction consisting of one sender and one recipient.

4.3.1.3 Sending and Replying Delay Distributions

To implement the relationships shown in Table 4.3 during the simulation of the conceptual e-mail system model, a normal distribution is used to generate the time delays between each e-mail message sent by e-mail clients and also to generate the time delays when e-mail clients reply to e-mail messages received from other e-mail clients. Each of the normal distributions used is controlled by the personality trait dimension degree values taken from the behaviour model as-

signed to each e-mail client. For the sending delay normal distribution, this is controlled by the extroversion and emotional stability trait dimension degree values. Similarly, the replying delay normal distribution is controlled by the extroversion, emotional stability, and conscientiousness trait dimension degree values.

The sending delay and replying delay normal distributions both use a 7-day window interval (see Figures 4.4 and 4.5), to limit the possible values for the sending and replying delays to within one week. The use of the 7-day window interval was based on a study from [106], where participants were found to reply to e-mails within one day or within one week. Apart from the study conducted by [106], there have been few other empirical studies that cover either the sending delays or replying delays of e-mail users. So using the results shown by [106], the model implements a delay of up to one week.

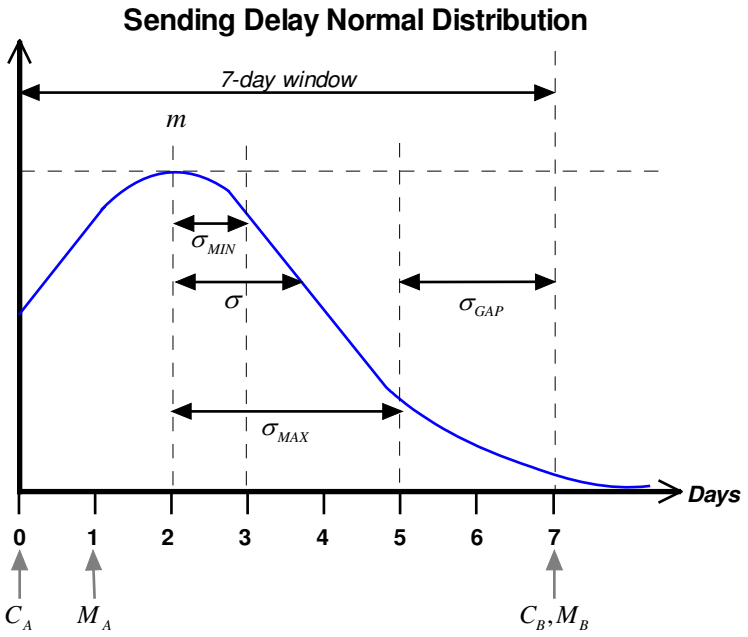


Figure 4.4: Layout for the sending delay normal distribution.

For the sending delay normal distribution, this is defined by the normal distribution function $N(m, \sigma)$, represented in Figure 4.4, where m is the mean of the normal distribution and σ is the standard deviation of the normal distribution. The mean of the distribution is given by:

$$m = M_B - D_{EX}(M_B - M_A) \tag{4.1}$$

where M_A and M_B represent the start and end range for mean values $[M_A = 1, M_B = 7]$, and D_{EX} is the degree of extroversion with a value between the

interval $[0, 1]$. For the standard deviation of the distribution, this is determined by:

$$\sigma_{MAX} = (C_B - C_A)/2 - \sigma_{GAP} \tag{4.2}$$

$$\sigma = \sigma_{MAX} - D_{ES}(\sigma_{MAX} - \sigma_{MIN}) \tag{4.3}$$

where C_A and C_B represent the start and end cut-off points for the 7-day window interval $[C_A = 0, C_B = 7]$, σ_{GAP} is the gap allowance between the standard deviation and the cut-off point for the 7-day window, σ_{MIN} and σ_{MAX} represent the minimum and maximum values for the standard deviation, and D_{ES} is the degree of emotional stability with a value between the interval $[0, 1]$.

To select a sending delay time from the normal distribution, a random number x_S is picked from the distribution $N(m, \sigma)$, such that x_S is within the 7-day window interval $[C_A, C_B]$, where $C_A = 0$ days and $C_B = 7$ days. Once x_S is selected from $N(m, \sigma)$, x_S is used as the sending delay between the previous e-mail message and next e-mail message to be sent by an e-mail client. A new sending delay value x_S is picked from the distribution each time before an e-mail client sends a new e-mail message, to determine when the new e-mail message will be sent.

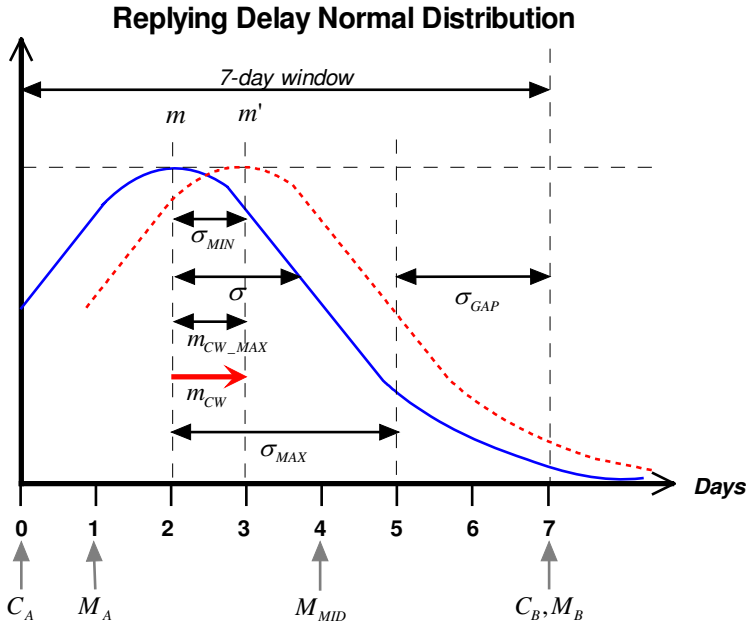


Figure 4.5: Layout for the replying delay normal distribution.

For the replying delay normal distribution, this is defined by a similar normal distribution function $N(m', \sigma)$, as shown in Figure 4.5, where m' is the mean and σ is the standard deviation. The mean for the replying delay normal distribution

is determined in two steps. The first step calculates the mean in the same way as for the sending delay distribution in Eq. (4.1). The second step makes use of the conscientiousness trait degree value and calculates the resulting mean m' as follows:

$$M_{MID} = \frac{M_B - M_A}{2} + M_A \tag{4.4}$$

$$m_{CW} = \begin{cases} \frac{D_C - 0.5}{0.5} \times m_{CW\_MAX} \times \frac{M_{MID} - m}{M_{MID} - M_A}, & m < M_{MID} \\ -\frac{D_C - 0.5}{0.5} \times m_{CW\_MAX} \times \frac{m - M_{MID}}{M_B - M_{MID}}, & m \geq M_{MID} \end{cases} \tag{4.5}$$

$$m' = m + m_{CW} \tag{4.6}$$

where M_{MID} represents the middle of the interval $[M_A, M_B]$, D_C is the degree of conscientiousness value with a value between the interval $[0, 1]$, m_{CW} is the conscientiousness weight factor for the mean, m_{CW_MAX} is the maximum value for the conscientiousness weight factor, and m' is the mean of the normal distribution after the conscientiousness weight factor m_{CW} has been added to the original mean m. For the standard deviation of the distribution, this is determined in the same way as for the sending delay normal distribution as shown in Eq. (4.3).

To select a replying delay time from the normal distribution, a random number x_R is picked from the distribution $N(m', \sigma)$, such that x_R falls anywhere along the $N(m', \sigma)$. When $x_R < C_A$ and $x_R > C_B$, it will be considered that the e-mail client will not be replying to a received e-mail message, given that x_R is outside the 7-day window interval. When $C_A \leq x_R \leq C_B$, then x_R will be used as the replying delay time value, since x_R is within the 7-day window interval. A new replying delay value x_R is picked from $N(m', \sigma)$ each time an e-mail client receives an e-mail message from another client.

4.3.1.4 Generation of Simulated Data

The process for simulating the conceptual e-mail system model consists of several steps, as shown in Figure 4.6. In the first stage "Creation of E-mail System Model Data" (steps 1 to 4), the e-mail system model generator program is used to create the set-up parameters for the conceptual e-mail system model (as seen in Figures 4.7 and 4.8). The first step consists of generating the e-mail clients and behaviour models, where the number of e-mail clients and behaviour models

created is such that:

$$N_e \quad \geq \quad N_b \tag{4.7}$$

where N_e represents the number of e-mail clients and N_b represents the number of behaviour models. In step 2, all behaviour models are assigned to e-mail clients, so that each behaviour model is at least assigned to one e-mail client each. For the case where $N_e > N_b$, some behaviour models may be assigned to more than one e-mail client. Steps 1 and 2 can be repeated as necessary where the e-mail system model generator is used to automatically assign random attribute values for the e-mail clients and behaviour models. Alternatively the user can use the e-mail system model generator program to manually adjust the attribute values of the e-mail clients and behaviour models. Once the attribute settings and behaviour models have been allocated to e-mail clients, the third step consists of generating the social connections between e-mail clients, so that each e-mail client has a list of e-mail addresses of social contacts they will communicate with during the simulation. The final step for the first stage, step 4, is where the e-mail system model data is saved by the e-mail system model generator program to store the parameter settings and configuration of the e-mail system model.

The next stage of the simulation process, "Simulation of E-mail System Model" (steps 5 to 7), uses the e-mail system simulator program to simulate the conceptual e-mail system model. In step 5, the saved e-mail system model data is loaded into the e-mail system simulator program to read the set-up parameters and configuration settings of the e-mail system model. After the data is loaded into the simulator, the simulation run is started in step 6.

During the simulation run, two different types of events will be generated by the e-mail clients: 'sending' events (new e-mail messages sent by e-mail clients to randomly selected social contacts) and 'replying' events (e-mail messages sent in reply to a previously received e-mail message). A representation of the events generated by the e-mail clients is shown in Figure 4.9. This shows how 'sending' events are continuously generated by e-mail clients, whereas 'replying' events can only be triggered by a 'sending' event or another 'replying' event as a result of an e-mail client receiving an e-mail message. Every time a 'sending' or 'replying' event is executed by the simulator, the simulator records a log of the message in the mailbox of each e-mail client involved in the message interaction. The scheduling of delays between messages sent by each e-mail client and

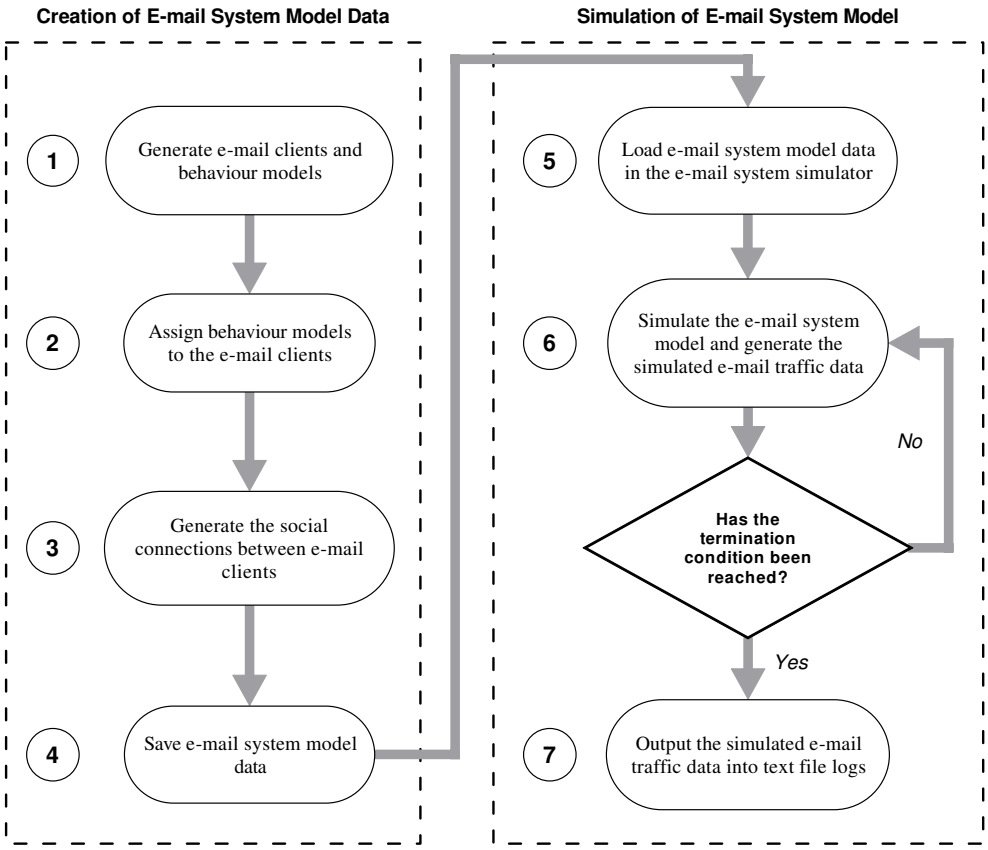


Figure 4.6: Flow diagram of the simulation model set-up and simulation process.

the delay for an e-mail client to reply to a previously received message, is determined by the sending delay and replying delay normal distributions described in section 4.3.1.3.

To terminate the simulation run, the simulator will check whether the termination condition for the simulation has been reached. The termination condition can be specified either as being dependent on M total number of e-mail messages being sent by e-mail clients (e.g. $M = 10,000$ messages), or as being dependent on when the simulation time reaches the specified value T (e.g. $T = 120$ simulation days). After the simulation run is terminated, in step 7 the simulated e-mail traffic data from each e-mail client's mailbox is stored into text files. These text files are then filtered and entered into the e-mail traffic database as shown in Figure 4.10, ready to be analysed by the e-mail traffic analysis system.

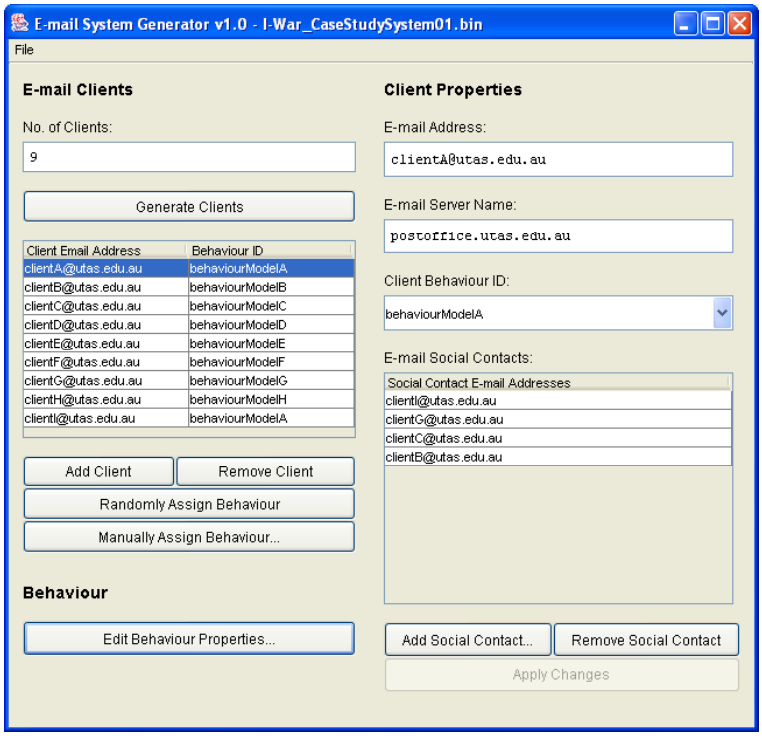


Figure 4.7: The main graphical user interface for the e-mail system model generation program.

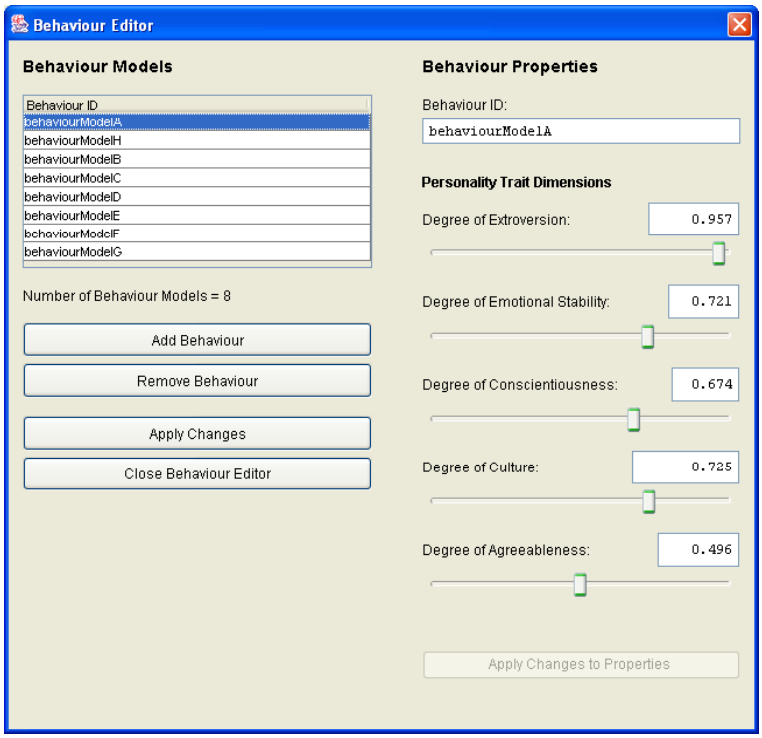


Figure 4.8: The behaviour editor part of the e-mail system model generation program.

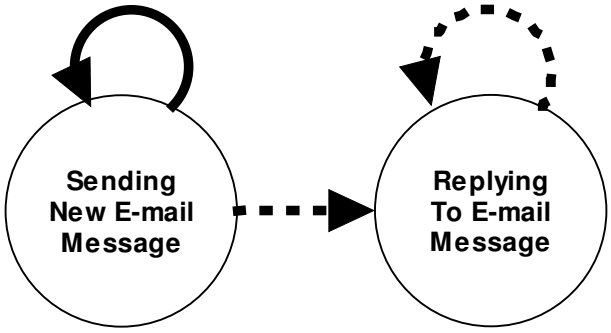


Figure 4.9: Events diagram for events in the conceptual e-mail system model.

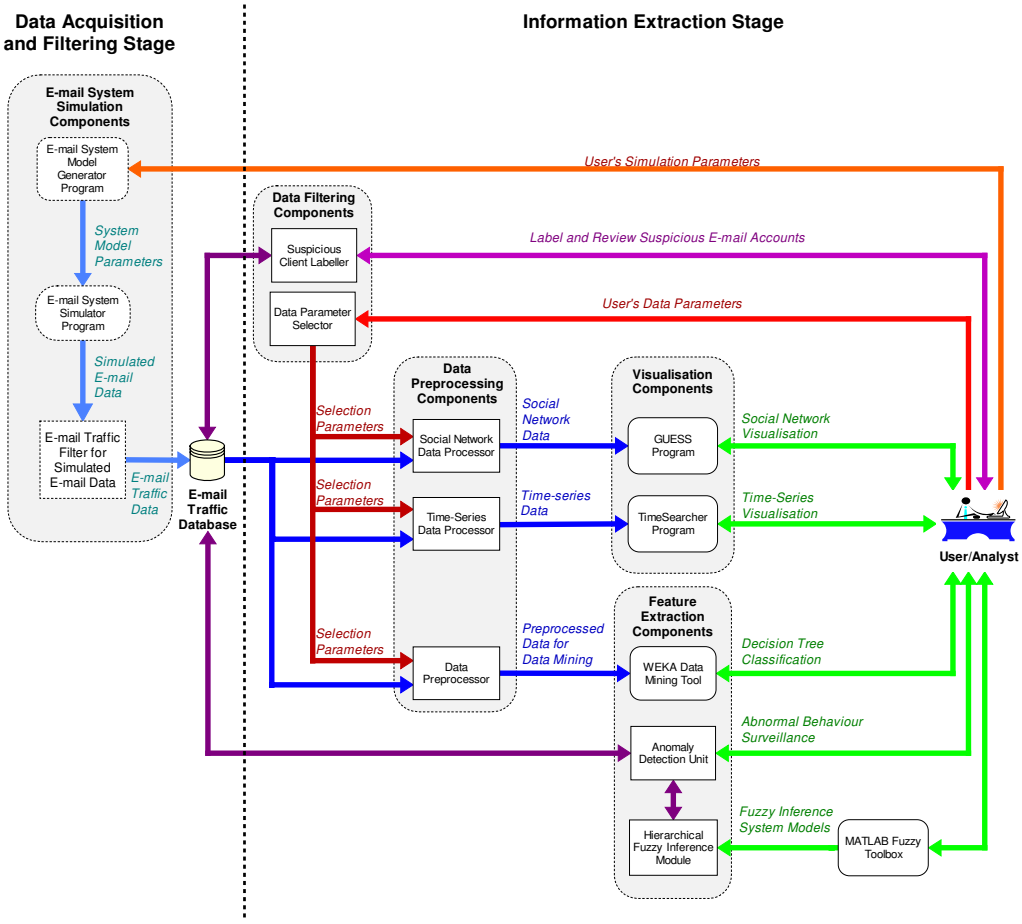


Figure 4.10: Overview of how simulated data is entered into the e-mail traffic analysis system.

4.3.2 The Enron E-mail Dataset

Enron is best known as the former gas pipeline and energy trading company that collapsed in 2001, as a result of fraudulent accounting practices [107]. When Enron was investigated in the United States in 2002, the Federal Energy Regulatory Commission (FERC) publicly released a corpus of e-mails belonging to some of

the Enron employees. After the release of the original corpus, several different versions of the original Enron e-mail corpus data were created and made available for researchers to use [108]. A raw form of the e-mail data is provided by [109] and other versions of the data based on [109] are provided by [110, 111]. The version of the Enron e-mail data used for this research is the "ISI" Enron e-mail dataset provided by [111].

There are a number of reasons for using the Enron e-mail dataset to evaluate the e-mail traffic analysis system. Firstly, the Enron e-mail dataset is a collection of genuine e-mail data and is representative of a large network of e-mail users. The ISI Enron e-mail dataset used consists of e-mail data from the mailboxes of 151 employees and contains 252,759 e-mail messages [108]. Within this dataset, a simple statistical analysis of the Enron e-mail data showed that there were 75,547 unique e-mail addresses in the e-mail data [112]. This means that even though the data was sampled from 151 former Enron employees, it is representative of a very large network of e-mail users. This makes the dataset ideal for determining whether the e-mail traffic analysis system may be applicable for dealing with data from large e-mail systems.

Another reason for using the Enron e-mail dataset is that some of the people sampled in the Enron e-mail data (e.g. Kenneth Lay and Jeffrey Skilling) were criminally charged and sentenced as part of the criminal investigations into the conduct of former Enron employees [113]. The presence of such individuals in the dataset makes it ideal for examining the communication behaviour characteristics of people whom have performed "criminal" activities. This may also allow one to draw inferences about the way these individuals communicated and whether there may be any common communication patterns associated with their illegal activities.

The third reason for using the Enron e-mail dataset is that it contains rich information on the dynamics and changes in communication that occurred when particular events occurred at Enron. Examples of significant events that occurred are the rapid drop in share prices during 2001 and Enron's filing for Chapter 11 bankruptcy protection in December during the same year [107]. The presence of such events is ideal for the purposes of this research, since the e-mail traffic analysis system can be tested to determine whether it is able to detect and locate unusual or abnormal communication behaviour associated with important events that occurred at Enron.

Finally, the Enron e-mail dataset is now becoming an ideal common dataset for e-mail based research [34]. Earlier work on behaviour analysis of e-mail traf-

fic data only used e-mail data from either the researcher's own university (e.g. [2, 18, 41, 42]) or through their organisation (e.g. [44]). The disadvantage of previous approaches is that it is difficult to compare the performance of different analysis methods if the dataset used is different. With the Enron e-mail dataset, it is a large and readily available dataset that allows for comparison of the performance of different analysis methods.

Processing the Enron E-mail Traffic Data

To process the Enron e-mail dataset for the e-mail traffic analysis system, the 'ISI' version of the Enron e-mail dataset [111] was processed through a series of steps in order to make the original e-mail data suitable for analysis by the e-mail traffic analysis system. In the original 'ISI' version of the Enron dataset, the data had already been formatted for MySQL databases and has a database structure that is suitable for extracting traffic information from the e-mail message headers (e.g. sender, receiver, and date/time information). This made it ideal for this research, since the existing information in the ISI Enron e-mail dataset could be easily used to extract traffic information for the e-mail traffic analysis system.

To obtain the traffic information from the ISI Enron e-mail dataset, the sender, receiver, and date/time information was extracted from the MySQL formatted database created by [111]. While extracting traffic information, the data also was massaged to treat e-mail messages with multiple recipients as individual messages sent to multiple recipients at the same time. Once the basic sender, receiver, and date/time information had been entered into the e-mail traffic database, the sending and replying delay measurements were then computed for each e-mail account in the Enron e-mail dataset.

After completing the filtering and massaging of the data for the e-mail traffic database, the resulting characteristics of the Enron e-mail dataset contained 75,547 unique e-mail addresses, with 2,042,442 sent messages after considering messages with multiple recipients as separate e-mail messages. Most of the e-mail messages sent (2,063,748 or 99.966% of messages) were between 1999 to the end of 2002. Only the remaining percentage of the data was found to be outlier messages with dates outside the range of 1999 and the end of 2002.

Before analysing the Enron e-mail dataset, it is important to note that the Enron e-mail dataset has a number of imperfections, which are in part due to modifications made to the dataset since its original release. Tracing back to the origin of the dataset, there have been a number of important changes and modifications

that need to be considered. Firstly, the original Enron e-mail data was only a sample of e-mail data from 151 former Enron employees, which is just a small percentage of the large number of people whom were formerly employed at Enron. This means that the data does not contain full e-mail traffic information about the other e-mail accounts that communicated with the 151 former Enron employees in the sample. Secondly, the original Enron e-mail dataset obtained from the FERC had a number of integrity issues. The original Enron e-mail data was purchased from the FERC by Kaelbing and the integrity issues were later fixed by Gervasio and her group at SRI International (SRI International) for the CALO project [108].

Thirdly, the Enron e-mail dataset has e-mail messages that have been deleted and e-mail addresses that have been modified. In the Enron e-mail dataset obtained by Carnegie Mellon University (CMU), some e-mail messages specific to certain individuals were removed from the dataset due to privacy and legal reasons [108]. The removal of these messages is documented by [109] at CMU. In addition to this, the ISI Enron dataset [111] (which is based on the CMU dataset) had some messages deleted due to [114]:

- Junk data left over from past attachments.
- Some messages with completely blank information.
- Some messages that were returned by the e-mail system due to transaction failure.

Some e-mail addresses were also renamed by [114] because they could not be resolved properly [114]. The types of renaming changes made by [114] are:

- Messages that contained 'invalid' e-mail addresses were changed to "*no.address@enron.com*".
- Recipients in certain messages were not disclosed, so all such messages were changed to a common format "*undisclosed-recepients@enron.com*".

All of these changes and modifications that have occurred since the obtaining of the original dataset shows that the Enron e-mail dataset contains many imperfections. These imperfections need to be considered when performing analysis of the data, since these may affect the information that is inferred from the data.

4.4 Analysing The Data

To investigate the traffic behaviour of e-mail users, there are two approaches in which the user/analyst can use the e-mail traffic analysis system to analyse the data. These approaches are: the sole use of visualisation techniques, or a combination of both feature extraction and visualisation techniques. Both of these approaches utilise the capabilities of the computational techniques integrated into the e-mail traffic analysis system and demonstrate the use of computational intelligence.

Exploring The Data With Visualisation Techniques

One of the approaches for using the e-mail traffic analysis system is to use visualisation techniques to visually explore the traffic behaviour of e-mail users. This is performed by firstly allowing the user to specify an area of interest in the data (e.g. by selecting particular e-mail accounts and a particular period of time), through the use of the "Data Parameter Selector" component. After selecting an area of the data for analysis, the user/analyst can then either use social network visualisation to gain an overview of connections between e-mail accounts or use time-series visualisation to view the volume of e-mail traffic generated by particular e-mail accounts. Since the visual exploration process can be iterative, the user/analyst may want to continue investigating the behaviour of particular e-mail accounts in further detail or to investigate other areas of the e-mail system that is of interest. The steps that can be used for exploring the data with visualisation techniques is illustrated in Figure 4.11.

There are cases in which visual exploration of e-mail traffic behaviour may be useful to the user/analyst. Such a case may be when user/analyst already knows the suspect e-mail account that they want to examine and need to quickly understand the suspect's communication behaviour. Another case may be when the user/analyst wants a general overview of the communication behaviour between suspect e-mail accounts. However, there are other cases in which visual exploration may not be useful for examining e-mail traffic behaviour. Such cases are when the user/analyst wants to locate unusual or abnormal changes in communication behaviour exhibited suspect e-mail accounts. These types of changes in behaviour are difficult to find using visualisation, as previously discussed at the end of Section 3.3. Cases like these can be better handled by using a combination of feature extraction and visualisation techniques.

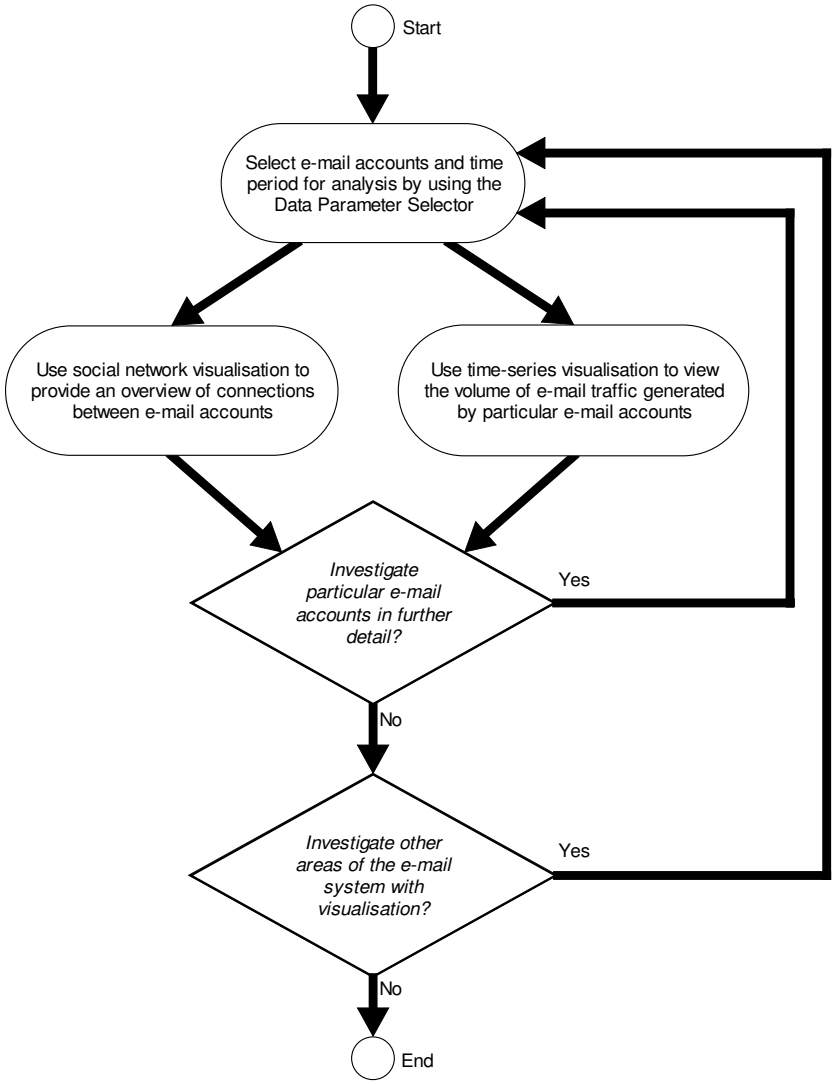


Figure 4.11: Steps for exploring the data using visualisation techniques.

Utilising Both Feature Extraction and Visualisation Techniques

The use of both feature extraction and visualisation techniques is another type of approach for using the e-mail traffic analysis system to examine e-mail traffic data. The purpose of combining these techniques is to utilise the capabilities of each type of technique, in order to better aid the user/analyst in examining and understanding the e-mail traffic behaviour of suspect e-mail accounts. The capabilities of feature extraction techniques can be used to locate unusual changes in communication between a suspect and their associates, while visualisation techniques can be used to provide a better understanding of the details of these unusual changes in communication. This utilisation of different techniques is an example of how the computational intelligence approach can be useful in providing different perspectives for analysing e-mail traffic behaviour.

To analyse the data using both feature extraction and visualisation techniques, there are two ways in which these techniques can be combined. The first method is to use decision tree classification with the visualisation techniques to locate and understand unusual variations in communication behaviour between a suspect and their associates. This is performed by firstly selecting a particular area of the e-mail system with the "Data Parameter Selector" component. After selecting the area of interest, the decision tree classification algorithm is applied to the filtered e-mail traffic data. The resulting decision tree classification output can then be used to locate unusual changes in interaction between a suspect e-mail account and their associates. These unusual changes in interaction behaviour can then be investigated through visual exploration, as indicated by the steps in Figure 4.11. After completing the visual exploration of unusual changes in interaction behaviour, the user/analyst may consider continue investigating other areas of the e-mail system or to conclude the investigation. The steps for using decision tree classification with the visualisation techniques is illustrated in Figure 4.12 .

The second method for combining feature extraction and visualisation techniques, is to use hierarchical fuzzy inference with the visualisation techniques. This combination is used to locate and understand the abnormal changes in behaviour exhibited by the communication links of suspect e-mail accounts. In order to find such changes in behaviour, this is performed by initially selecting a set of suspect e-mail accounts through the "Suspicious Client Labeller" component of the e-mail traffic analysis system. After selecting the set of suspects, the user/analyst then specifies the baseline profiling and surveillance periods through the "Anomaly Detection Unit" to determine what periods of time are used to examine the change in communication behaviour for the suspects. The Anomaly Detection Unit is then used to apply hierarchical fuzzy inference to rate the changes in behaviour observed for each communication link of each suspect. The resulting abnormality rating outputs from the Anomaly Detection Unit can then be used to locate communication links that have exhibited abnormal changes in communication behaviour. These abnormal changes in communication link behaviour can be investigated through visual exploration, as indicated by the steps in Figure 4.11. After completing the visual exploration of communication link behaviour, the user/analyst may consider continue investigating other areas of the e-mail system or to conclude the investigation. The steps for using hierarchical fuzzy inference with the visualisation techniques is illustrated in Figure 4.13.

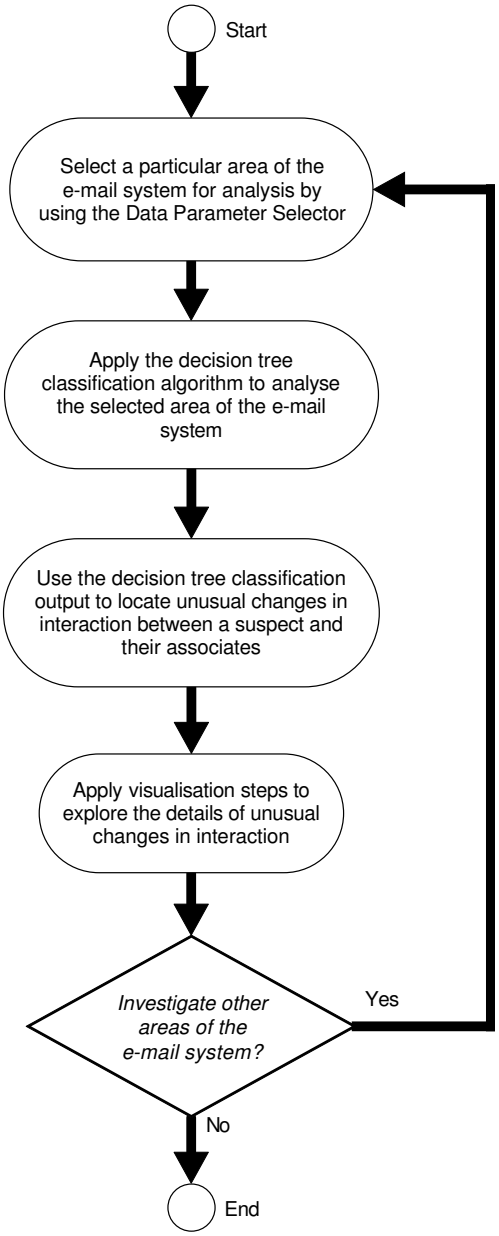


Figure 4.12: Steps for analysing the data with decision tree classification.

Overall, the provision of two different approaches for analysing the data, either through visualisation or combining feature extraction and visualisation, may be useful since it provides flexibility in the way the user/analyst may want to analyse the behaviour of suspect e-mail accounts. The importance of this is that it is not known in advance what type of cases the user/analyst may want to investigate when analysing e-mail traffic behaviour. There may be cases where the user/analyst is only interested in the traffic behaviour of a particular e-mail account, or cases where the user/analyst may want to find out about unusual interactions between a group of suspect e-mail accounts. Providing different

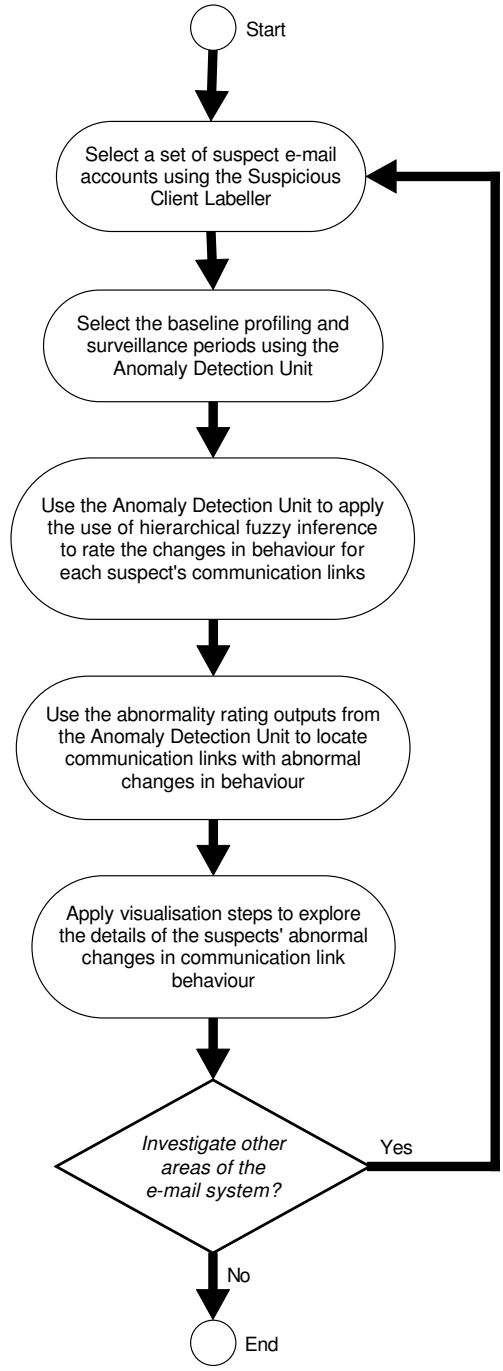


Figure 4.13: Steps for analysing the data with hierarchical fuzzy inference.

approaches for analysing the data in the e-mail traffic analysis system at least allows the user/analyst to have some control in how they want to investigate the behaviour of e-mail users.

4.5 Summary

This chapter described the development of the e-mail traffic analysis system and its use for assisting with the examination of e-mail traffic behaviour. The first section of the chapter described the e-mail traffic analysis system as a conceptual system, which has been developed to explore the use of different computational techniques. The purpose for developing the system is to integrate the visualisation and feature extraction techniques used for the research, in order to demonstrate the use of different perspectives while examining the e-mail traffic behaviour of suspect e-mail accounts.

In the design of the e-mail traffic analysis system, two important design elements were considered. The first design element was the use of a modular architecture to allow for new features to be easily added to the system. This modular architecture is considered useful if additional computational techniques are to be added to the system for more perspectives on the traffic behaviour of e-mail users. The second design element considered was allowing the user to specify parameters that control the portions of the data examined by the e-mail traffic analysis system. This is considered important as well since it focuses the analysis on parts of the data that are relevant to the user/analyst's investigation task and puts the user/analyst in charge of the analysis process.

An overview was then provided on the architecture and implementation of the e-mail traffic analysis system. This described how data was processed through the system and described the role of each of the major components of the system. The keys parts of the system that perform an important role in aiding the user/analyst to examine e-mail traffic behaviour are:

- **Data Filtering Components** - these allow the user/analyst to specify parameters that control the portions of the data analysed by the system.
- **Visualisation Components** - these allow the user/analyst to visually explore and understand the data using visualisation techniques.
- **Feature Extraction Components** - these allow the user/analyst to locate unusual or abnormal changes in communication behaviour.

For the implementation of the e-mail traffic analysis system, the system was developed using Python and MySQL for the database. The social network visualisation, time-series visualisation, and decision tree classification components of the system were implemented using existing software applications, in order

to focus more on evaluating the information presented by these techniques. The hierarchical fuzzy inference component of the system was implemented using a combination of C code from MATLAB and extending the functionality using Python.

The second section of the chapter described the e-mail traffic data used to evaluate the e-mail traffic analysis system. The important things considered for the data used in the research were: knowledge of the behavioural characteristics of the data and the use of sufficient amounts of data which represents each e-mail user's typical communication behaviour. Based on these two important considerations, two types of e-mail traffic data were used in the research: simulated e-mail traffic data and the Enron e-mail dataset. Simulated e-mail traffic data was used in the research to provide data with known behavioural characteristics, which are determined by a set of pre-defined behaviour models. The purpose of this was to allow for the observation of e-mail traffic behaviour and to enable the observed behaviour to be related back to the individual who generated it. The simulated e-mail traffic data is provided through a conceptual simulation model of the e-mail system, which uses personality trait dimensions to determine the traffic behaviour of e-mail users.

The second type of e-mail traffic data used, the Enron e-mail dataset, originated from a collection of e-mails obtained from Enron, a former gas pipeline and energy trading company that collapsed in 2001 as a result of fraudulent accounting practices. This collection of e-mails was chosen for the research since it contained a number of characteristics that made the dataset suitable for analysis. Firstly, the Enron e-mail dataset is a large collection genuine e-mail data, which provides examples of communications sampled from 151 former Enron employees. The next suitable characteristic was that the dataset contained examples of individuals known to have been criminally charged for their involvement in fraudulent accounting practices at Enron. Finally, the dataset also contained rich information on the dynamics and changes in communication that occurred when particular events happened at Enron. These characteristics made the Enron e-mails an ideal dataset for testing the capabilities of the e-mail traffic analysis system, in order to locate unusual or abnormal changes in communication behaviour.

The final section of the chapter described how the e-mail traffic analysis system can be used to analyse e-mail traffic data. Two types of approaches were outlined for analysing e-mail traffic data. The first approach consists of the sole use of visualisation techniques to explore and understand the traffic behaviour

of e-mail users. The second approach applies the use of both feature extraction and visualisation techniques to locate and understand unusual changes in e-mail traffic behaviour exhibited by suspect e-mail accounts. The second approach is an example of how computational intelligence can be used to provide different perspectives on e-mail traffic behaviour to the user/analyst. In the next chapter, it will be demonstrated how the computational intelligence approach is used to analyse the e-mail traffic behaviour of suspect e-mail accounts.

Chapter 5

The E-mail Traffic Analysis System Evaluation and Case Studies

5.1 Introduction

Previous Chapters 3 and 4 described the computational techniques used for analysing e-mail traffic behaviour and the e-mail traffic analysis system developed to integrate each of the techniques for computational intelligence. This chapter focuses on the evaluation the developed e-mail traffic analysis system through two sets of case studies. The purpose of each set of case studies is to demonstrate how the e-mail traffic analysis system is used to investigate the e-mail traffic behaviour of known suspect e-mail accounts. This is to show how particular suspect e-mail accounts can be analysed by using a combination of visualisation and feature extraction techniques, so that the user/analyst obtains a broader understanding about those e-mail accounts' traffic behaviour.

The two sets of case studies presented in this chapter are based on the simulated e-mail traffic data and the existing Enron e-mail dataset. The first set of case studies uses simulated e-mail traffic data and evaluates the use of the e-mail traffic analysis system to examine the traffic behaviour of a small group of e-mail users from a simulated e-mail system. The second set of case studies focuses on the use of the Enron e-mail dataset and evaluates how the e-mail traffic analysis system can be used to investigate the traffic behaviour patterns of a former Enron employee.

5.2 E-mail Traffic Data Simulation and Analysis

The case studies presented in this section focus on the evaluation of the e-mail traffic analysis system using simulated e-mail traffic data. The purpose of this is to demonstrate how the e-mail traffic analysis system can be used to investigate the e-mail traffic behaviour patterns of individuals from a simulated e-mail system. The simulation tool discussed in Section 4.3.1 is used in the case studies to create a simulated e-mail system consisting of a small number of e-mail clients and with different behavioural profiles assigned to each of the simulated e-mail users. After simulating the e-mail traffic interactions of the e-mail users, the computational techniques from the e-mail traffic analysis system are then used to examine the e-mail traffic behaviour of particular individuals. The type of behaviours investigated in the two case studies is the detection of unusual changes in interaction behaviour between simulated e-mail users.

5.2.1 Case Study 1

E-mail System Model Setup

For this case study, an e-mail system simulation model was created comprising of 10 e-mail clients and 9 behaviour models. In this e-mail system model each of the e-mail clients is labelled in the form "client<letter>@utas.edu.au" and each of the behaviour models is labelled in the form "behaviour<letter>", where '<letter>' denotes an alphabetical letter identifier. When referring to an e-mail client in this case study, each e-mail client will be referred to by the client's e-mail address user name (the part before the '@' symbol).

To provide each e-mail client a behavioural profile, behaviour models were allocated to each e-mail client as shown in Figure 5.2. Each of the stacked columns shown in Figure 5.2 represent the relative values of the personality trait degrees assigned to each behaviour model. It should be noted from Figure 5.2 that each of the e-mail clients were allocated a unique behaviour model, except for *clientA* and *clientJ*, which were both assigned the behaviour model labelled 'behaviourA'. In Figure 5.3, the column chart provides an overview of the five personality trait dimensions making up the behavioural profiles of e-mail clients and shows a comparison of the personality trait degree values assigned among the population of e-mail clients. For the social connections between e-mail clients, the associates for each of the e-mail clients were randomly assigned and given the configuration shown in Figure 5.1.

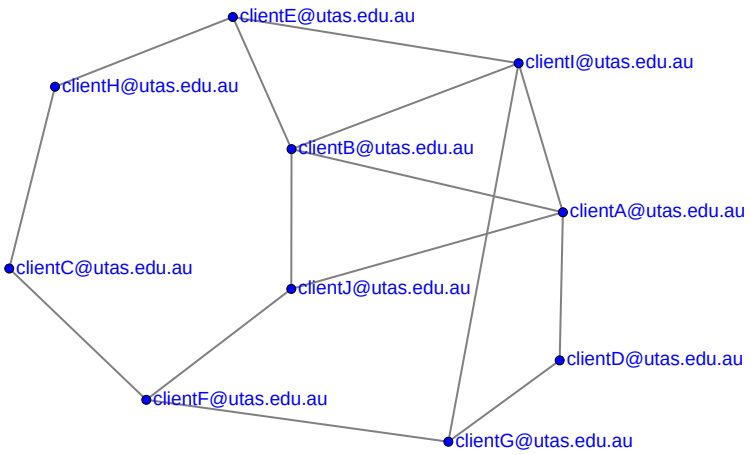


Figure 5.1: The setup configuration given for the social connections between e-mail clients.

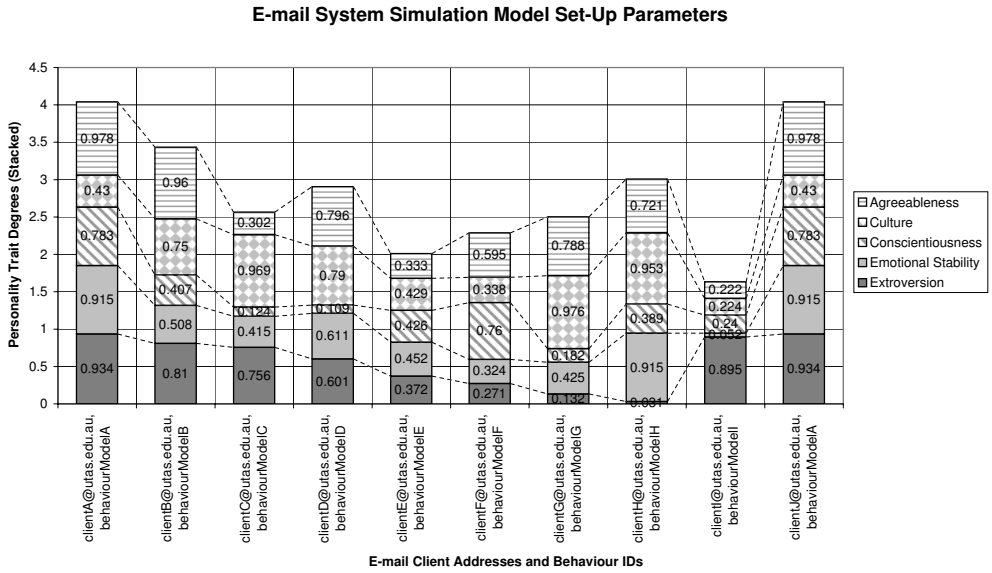


Figure 5.2: Stacked column chart showing the behavioural profiles of each e-mail client and their behaviour model ID.

Simulation Run

The e-mail system model for this case study was simulated over a period of 120 simulation days. This simulated period of time was chosen as the duration of the simulation, to allow enough time for the e-mail clients to start up their interactions from the initial conditions of the simulation (starting with mostly 'sending' events generated by e-mail clients) to a stable level of interactions ('sending' and 'replying' events generated by e-mail clients). The outcome from simulating the e-mail system model over 120 simulation days resulted in 2748 e-mail messages in total being sent by the 10 e-mail clients. The bar chart in Figure 5.4 presents a summary of the total number of messages sent and received

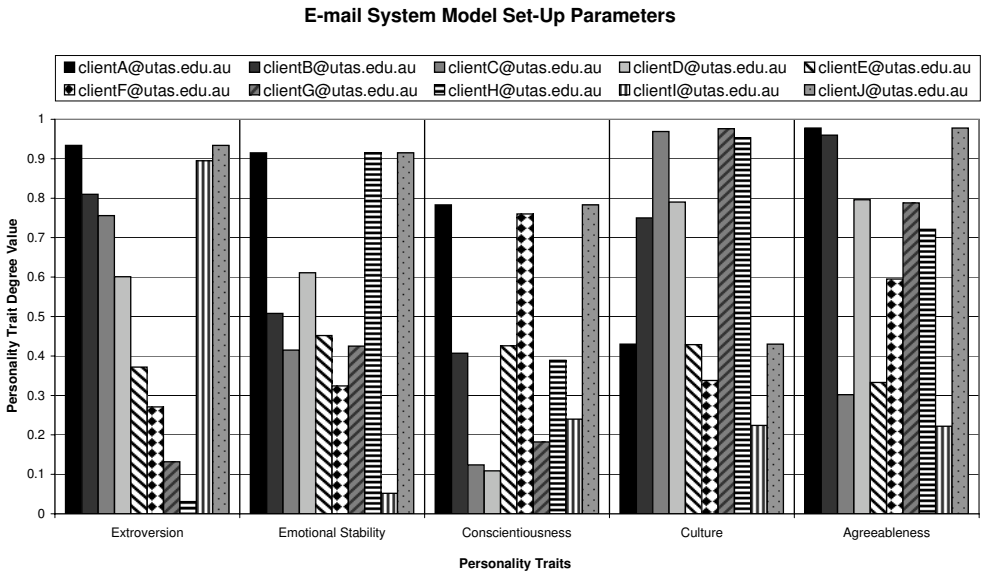


Figure 5.3: Column chart showing the personality trait degree values of all the e-mail clients.

by each of the e-mail clients over the 120 days simulated.

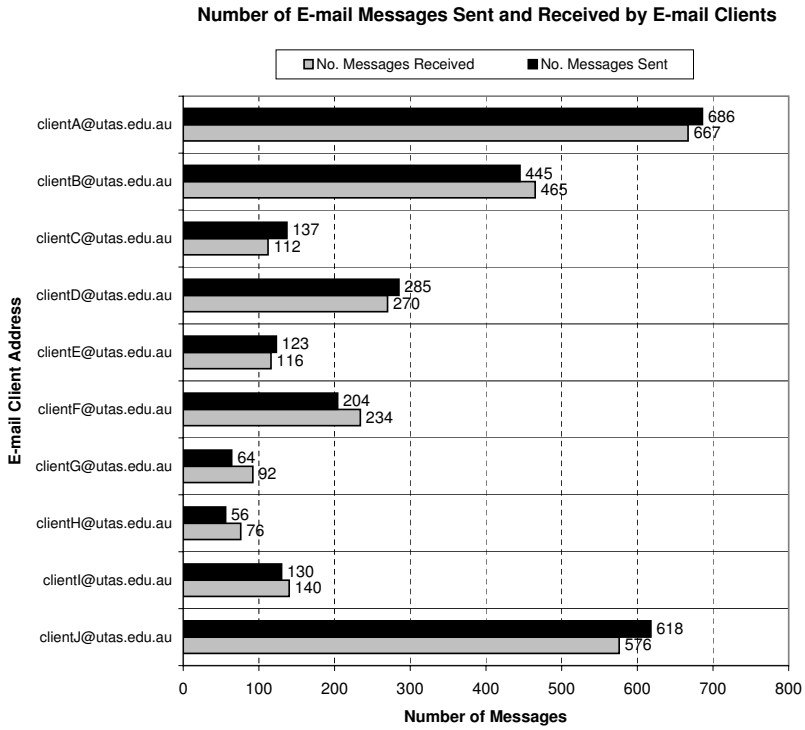


Figure 5.4: Number of e-mail messages sent and received by the e-mail clients over 120 simulation days.

General Overview of the Data

The bar chart in Figure 5.4 provides basic statistical information about the number of e-mail messages sent and received by each simulated e-mail client. This only shows a limited perspective on the e-mail traffic behaviour of each e-mail client. If other perspectives are considered, such as those offered by social network visualisation and time-series visualisation, these present additional information that provide a broader overview of each e-mail client's traffic behaviour. The social network diagram in Figure 5.5 and time-series diagram in Figure 5.6 each show how visualisation techniques can be used obtain a general overview of different aspects of each e-mail client's traffic behaviour. The social network diagram in Figure 5.5, generated by GUESS [75], provides an overview of connections between e-mail clients, as well as the strength of communication ties between clients. The relative strength of the communication ties in Figure 5.5 is indicated by thin lines representing weak communication (i.e. small number of e-mails sent, less than 100 e-mails) and thick lines representing strong communication (i.e. a large number of e-mails sent, more than 200 e-mails). The time-series diagram in Figure 5.6 presents an overview of the volume of e-mail traffic sent or received by the e-mail clients on a weekly basis throughout the simulation. The darker line in the time-series graph represents the time-series data for an e-mail client selected in TimeSearcher 2 [78].

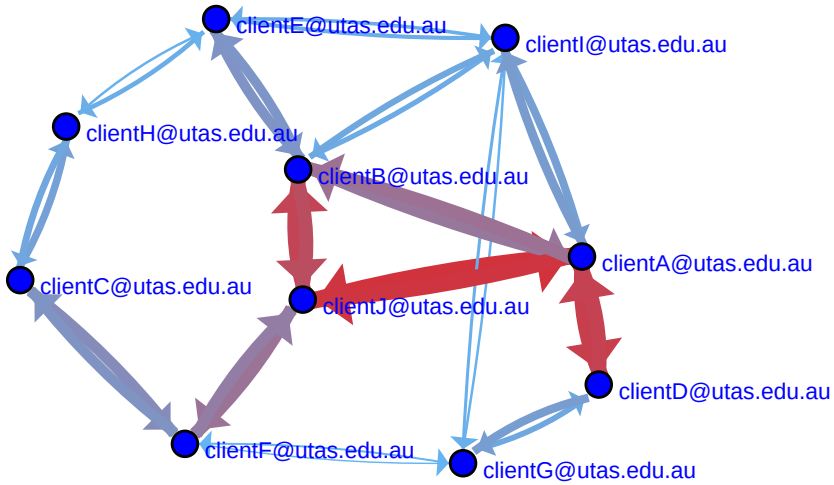


Figure 5.5: Overview of communications between the 10 e-mail clients.

However, using visualisation techniques to explore the data for unusual changes in communication behaviour is difficult. As shown in Figures 5.5 and 5.6, the output of the visualisation techniques used does not provide any immediate visual indication of where an unusual change in behaviour is occurring, even though

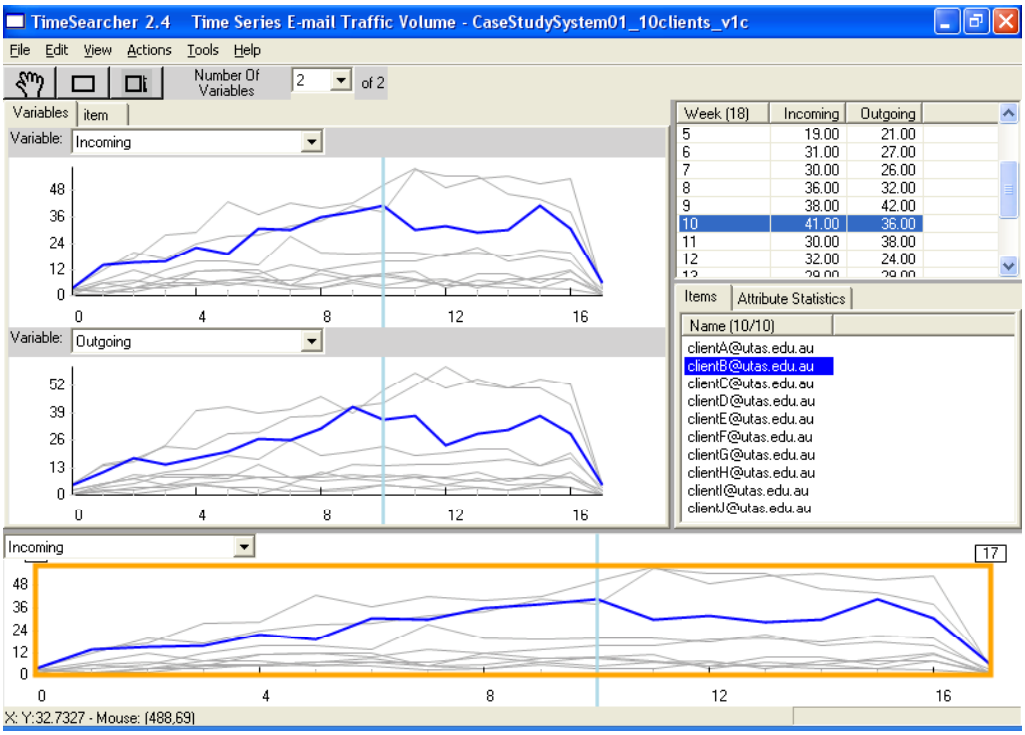


Figure 5.6: Weekly time-series overview of the 10 e-mail clients, with *clientB* selected.

they provide useful information describing general e-mail traffic behaviour. Due to this, it may take a great deal of time and effort on the user's part to find unusual changes in e-mail traffic behaviour through visual exploration. This is where referring to the use of feature extraction techniques to locate unusual changes in communication behaviour can be useful.

Locating Unusual Changes in Interaction Behaviour

The feature extraction technique used for this case study is decision tree classification. Decision tree classification is used here to locate unusual changes in interaction behaviour between the simulated e-mail clients and their associates. The J48 decision tree classification algorithm, supplied by WEKA [84], was applied to analyse the incoming and outgoing traffic logs of each e-mail client in the simulated e-mail system. The resulting two decision tree classification outputs are provided in Appendix F.1, which show the unusual changes in incoming and outgoing interactions found for each e-mail client. The decision tree outputs that had branching information containing date/time information was tabulated and put into Table 5.1. The "Incoming Interactions" column in Table 5.1 presents results from using the 'From' field of e-mail message headers as the decision tree class, while the "Outgoing Interactions" column in Table 5.1 presents re-

Table 5.1: Unusual changes in interaction behaviour found by decision tree classification.

E-mail Account of Interest	**Incoming E-mail Traffic Interactions**	**Outgoing E-mail Traffic Interactions**
clientA@utas.edu.au	clientD to clientA, where date <= day 75, 132 messages	-
	clientJ to clientA, where date > day 75, 149 messages	-
clientF@utas.edu.au	clientC to clientF, where date <= day 41.15, 28 messages	clientF to clientC, where date <= day 66.36, 40 messages
	clientJ to clientF, where date > day 41.15, 119 messages	clientF to clientJ, where date > day 66.36, 84 messages
clientG@utas.edu.au	clientI to clientG, where date <= day 37.23, 13 messages	clientG to clientI, where date <= day 47.46, 9 messages
	clientD to clientG, where date > day 46.84, 43 messages	clientG to clientD, where date > day 47.46, 30 messages
	clientF to clientG, where date > day 37.23 and date <= day 46.84, 5 messages	-
clientI@utas.edu.au	-	clientI to clientG, where date <= day 33.58, 12 messages
	-	clientI to clientA, where date > day 33.58, 44 messages
clientJ@utas.edu.au	clientB to clientJ, where date <= day 82.5, 136 messages	clientJ to clientB, where date <= day 78.6, 139 messages
	clientA to clientJ, where date > day 82.5, 121 messages	clientJ to clientA, where date > day 78.6, 135 messages

sults from using the ‘To’ field of e-mail message headers as the decision tree class.

After compiling the decision tree branch results into Table 5.1, the information from the decision tree classification outputs were then used to identify where unusual changes in interactions are occurring in the e-mail social network. Figures 5.7 and 5.8 show the areas of the e-mail social network where unusual changes in interactions have been identified to be occurring. It is noted that some of the unusual changes in interactions were found from both the inboxes and outboxes of e-mail clients (Figure 5.7), indicating that the interactions were of particular significance for both the incoming and outgoing e-mail traffic of those e-mail clients. Other unusual changes in interactions were found only from either the inbox or outbox of some e-mail clients (Figure 5.8), indicating that the interactions were of particular significance only for either the incoming or outgoing e-mail traffic of those e-mail clients.

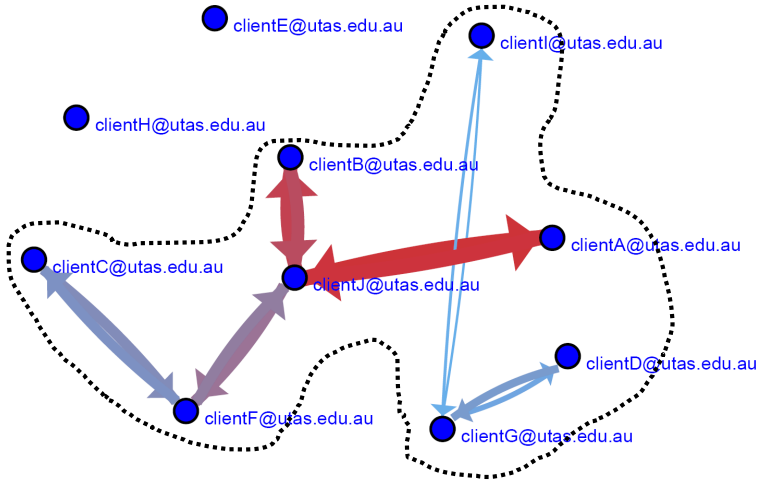


Figure 5.7: Unusual changes in interactions found from both the inboxes and outboxes of e-mail clients (i.e. changes found in both directions).

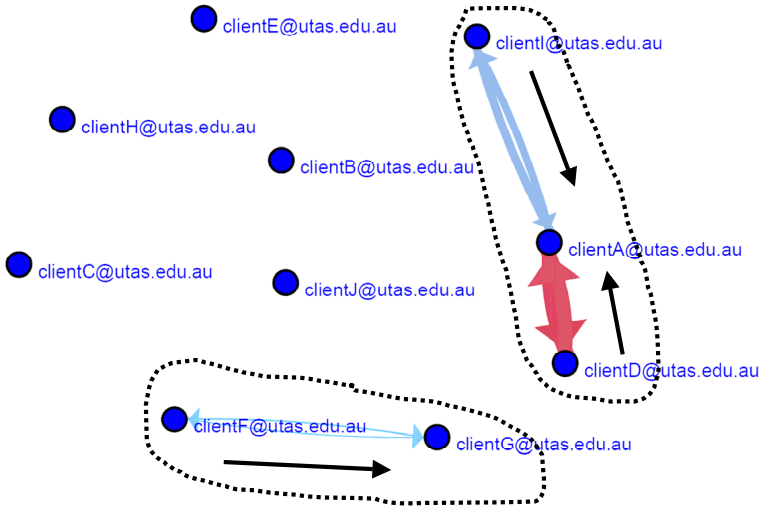


Figure 5.8: Unusual changes in interactions found from either the inbox or outbox of e-mail clients (i.e. changes found only in one direction).

Further Investigation of Unusual Changes in Interaction Behaviour

The unusual changes in interaction behaviour identified in Table 5.1 can be further investigated by using the e-mail traffic analysis system to "drill down" into the details of the interactions belonging to a particular e-mail client. The purpose of this is to reveal more information about the unusual changes in interaction behaviour identified by the decision tree classification output. For this case study, *clientG* has been selected as the e-mail client for further investigation.

The decision tree classification outputs provided by WEKA for *clientG* are shown in Figures 5.9 and 5.10 as visualised decision trees. These decision tree diagrams show exactly the same information as those tabulated for *clientG* in Table 5.1. It

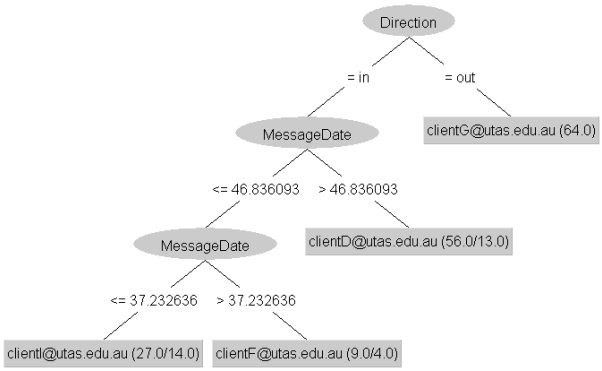


Figure 5.9: Decision tree for *clientG* showing unusual changes in incoming interactions.

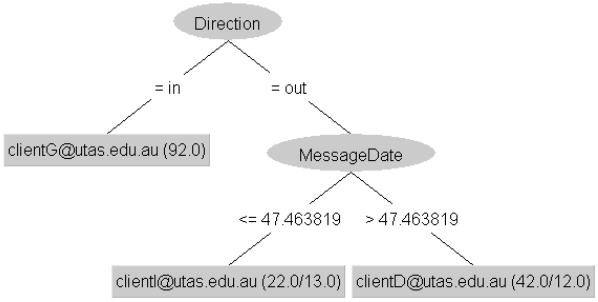


Figure 5.10: Decision tree for *clientG* showing unusual changes in outgoing interactions.

is seen from these visualised decision trees that there is a lot of communication between *clientG* and *clientI* prior to day 47. However, after day 47, it appears that *clientG* communicates more with *clientD*. A close inspection of the weekly time-series data for the interaction between *clientG* and *clientI* in Figure 5.11 shows there is an drop in outgoing e-mail traffic activity after week 6 (i.e. after day 47). At around the same time, it is revealed in Figure 5.12 that there is an increase in outgoing e-mail traffic activity between *clientG* and *clientD* after week 6. The information presented in Figures 5.11 and 5.12 both correspond with the features identified by the visualised decision tree in Figure 5.10, where *clientG* changes from communicating more with *clientI* to communicating more with *clientD*.

In Figure 5.9, it shows that there is some communication between clientG and clientF during the time period of day 37 to day 47, where 5 e-mail messages are sent from clientF to clientG. An examination of the daily time-series data in Figure 5.13, shows that there is a increase in incoming e-mail traffic from *clientF* to *clientG* during the period between days 35 to 56. This confirms from what was observed from Table 5.1 and Figure 5.8, where it was found that the

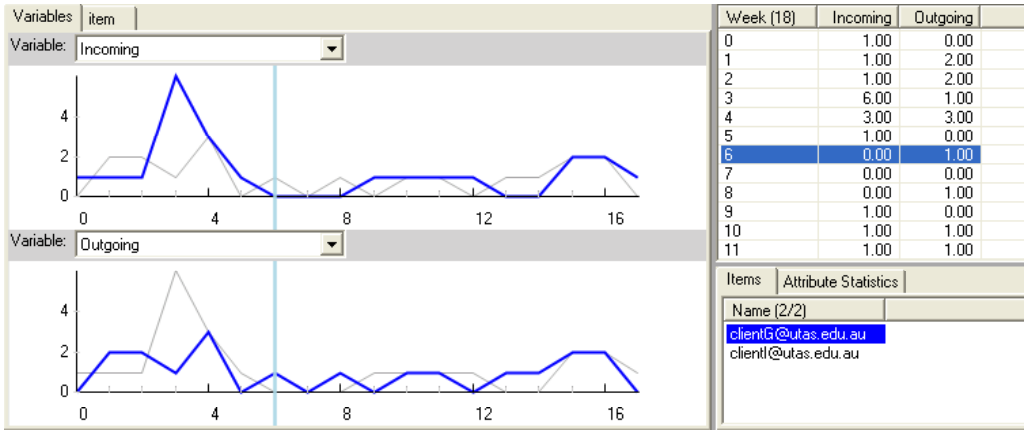


Figure 5.11: Weekly time-series data of e-mail traffic between *clientG* and *clientI*, showing a drop in traffic after week 6 or around day 47.

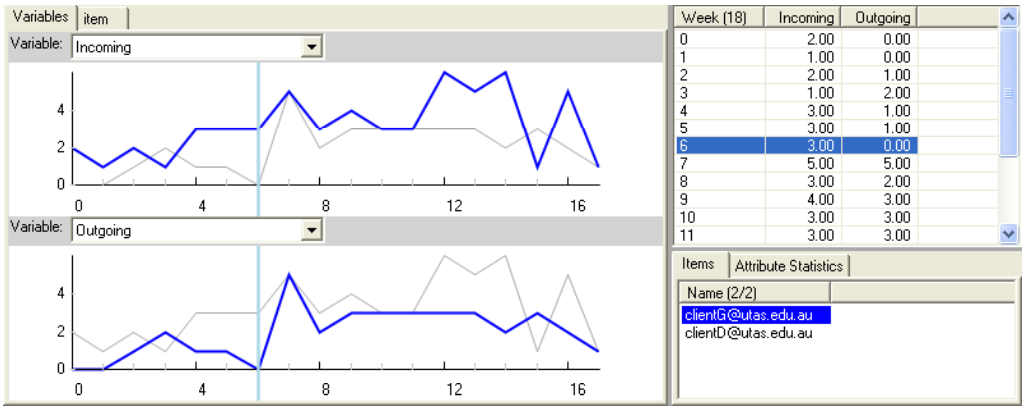


Figure 5.12: Weekly time-series data of e-mail traffic between *clientG* and *clientD*, showing an increase in traffic after week 6.

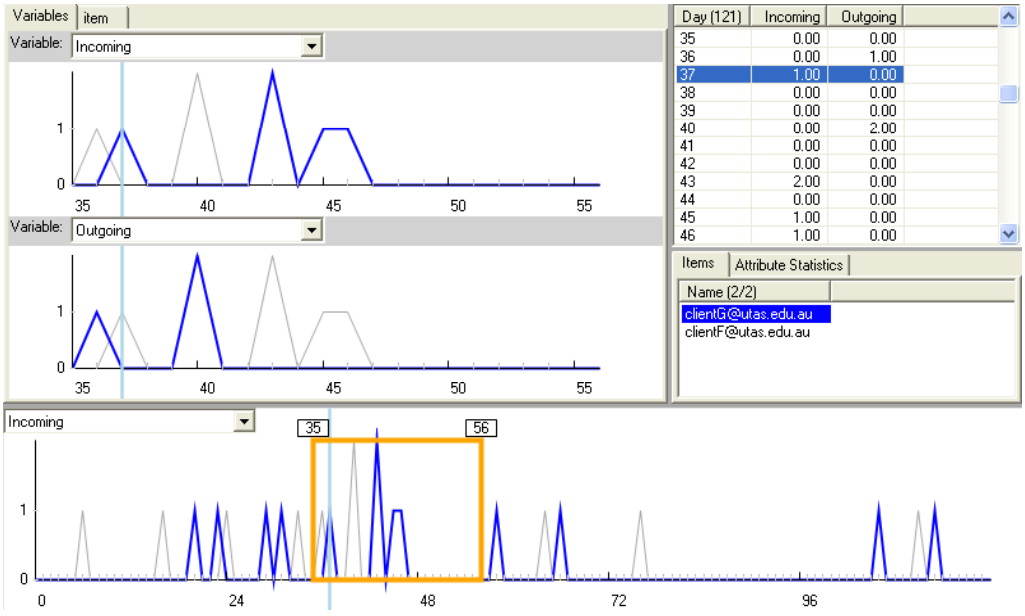


Figure 5.13: Daily time-series data for the interaction between *clientG* and *clientF* from days 35 to 56 (i.e. weeks 5 to 8).

change in communication interaction between *clientG* and *clientF* was mostly in one direction from *clientF*.

5.2.2 Case Study 2

E-mail System Model Setup

In the second case study, the e-mail traffic analysis system is again evaluated through the use of simulated e-mail traffic data created by the simulation tool [94, 95, 115]. For this particular case study, an e-mail system simulation model comprising of 9 e-mail clients and 8 behaviour models is considered. Each of the e-mail clients in the simulation model is again given a label of the form "client<letter>@utas.edu.au" and each behaviour model is given a label of the form "behaviour<letter>", where '<letter>' denotes an alphabetical letter identifier. The e-mail clients will be referred to by the username part of their e-mail address (i.e. the part before the '@' symbol in the e- mail address).

To assign the behavioural profiles, each e-mail client was allocated a behaviour model as shown in Figure 5.14. It should be noted that each e-mail client was assigned to a different behaviour model, except for *clientA* and *clientI*, which were both assigned the behaviour model 'behaviourA'. For the social connections between e-mail clients, the associates for each of the e-mail clients were randomly assigned and given the configuration shown in Figure 5.15.

Simulation Run

The e-mail system model for this second case study was simulated over a period of 182 simulation days or 26 simulation weeks. During this simulated period of time, a total of 3257 e-mail messages were sent by all e-mail clients. The bar chart in Figure 5.16 presents a general summary of the number of messages sent and received by each e-mail client over the 182 simulation days. The social network diagram from GUESS [75] in Figure 5.17 presents a different perspective on the general behaviour of the e-mail clients, by giving an indication of the relative strength of the communication ties that occurred between e-mail clients. Figure 5.17 indicates the strength of each communication link, with thin lines representing weak communication (i.e. small number of e-mails sent, less than 100 e-mails) and thick lines representing strong communication (i.e. a large number of e-mails sent, more than 200 e-mails).

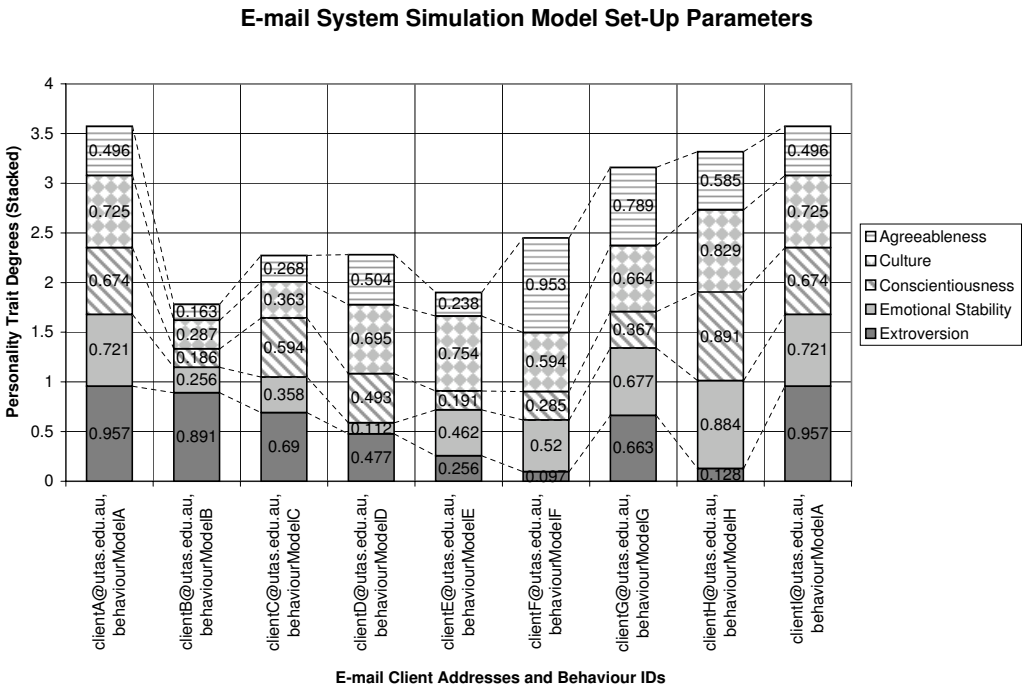


Figure 5.14: The behavioural profiles of each e-mail client in the simulation model.

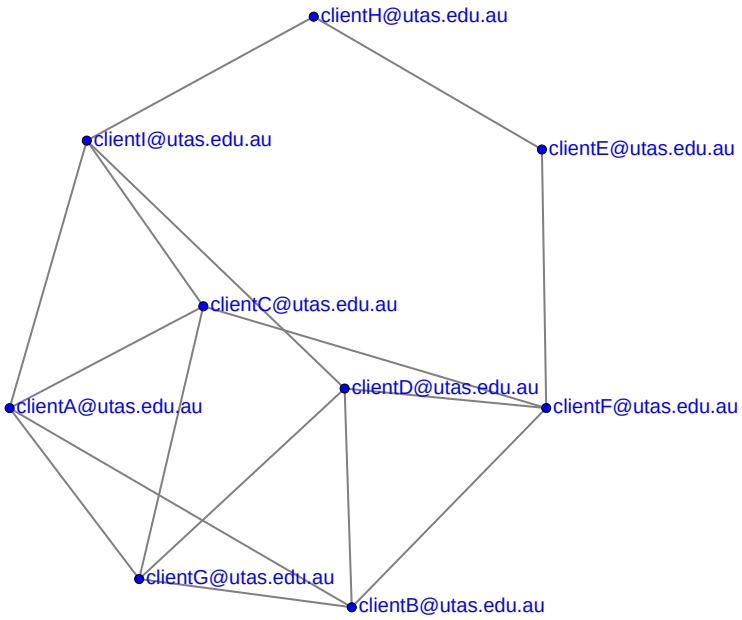


Figure 5.15: The setup configuration given for the e-mail clients' social connections for case study 2.

Locating Unusual Changes in Interaction Behaviour

Decision tree classification is again used for this case study, to detect unusual changes in interaction behaviour between the simulated e-mail users and their associates. The J48 decision tree classification algorithm from WEKA [84] was

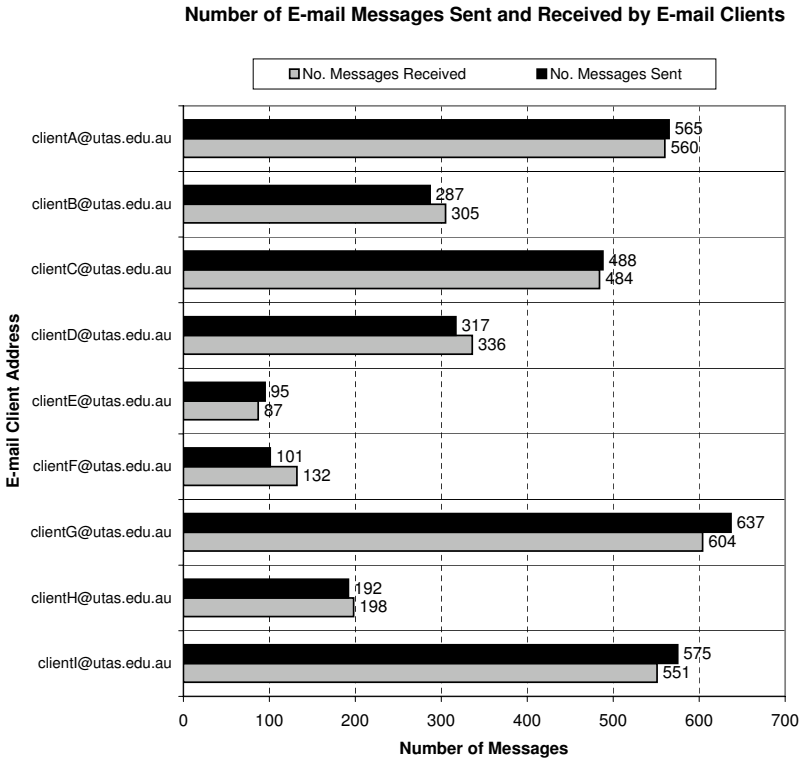


Figure 5.16: Number of e-mail messages sent and received by e-mail clients over 182 simulation days.

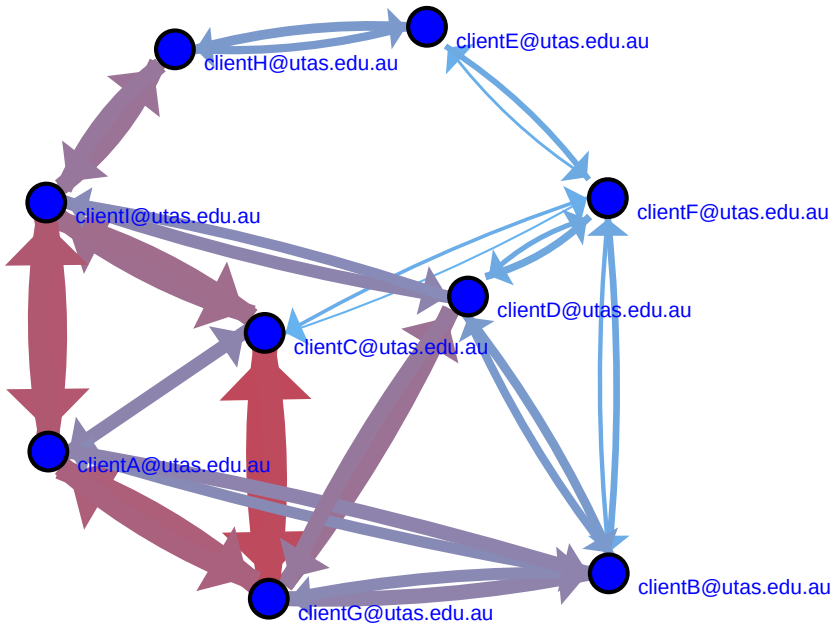
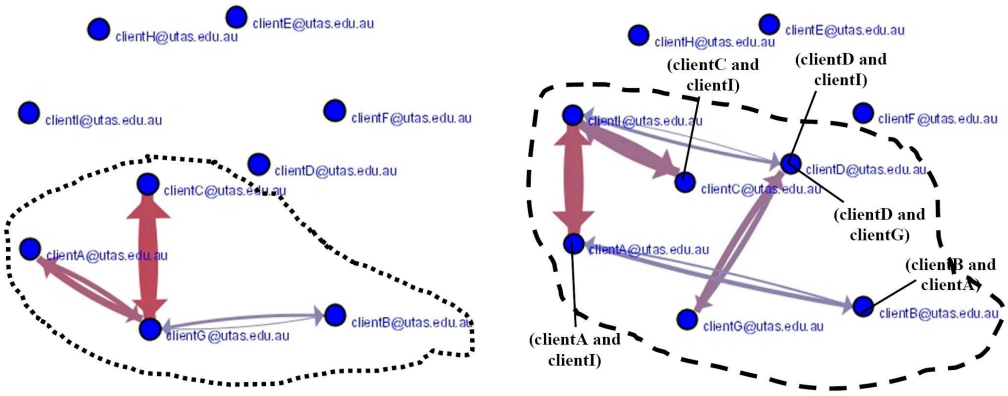


Figure 5.17: Social network overview of the 9 e-mail clients.

Table 5.2: Unusual changes in interactions derived from the two decision tree outputs produced using WEKA.

E-mail Account of Interest	**Incoming Interactions**	**Outgoing Interactions**	**Type of Interaction**
clientA@utas.edu.au	clientI to clientA; where date <= day 24.08; 18 messages	clientA to clientI; where date <= day 22.65; 18 messages	*Two-way -> found in one account
	clientG to clientA; where date > day 24.08 and date <= day 109; 95 messages	clientA to clientG; where date > day 22.65 and date <= day 108.52; 93 messages	**Two-way -> found in two accounts
	clientI to clientA; where date > day 109; 92 messages	clientA to clientI; where date > day 108.52; 92 messages	*Two-way -> found in one account
clientB@utas.edu.au	clientG to clientB; where date <= day 110.6; 68 messages	clientB to clientG; where date <= day 107.34; 62 messages	**Two-way -> found in two accounts
	clientA to clientB; where date > day 110.6; 72 messages	clientB to clientA; where date > day 107.34; 67 messages	*Two-way -> found in one account
clientC@utas.edu.au	clientI to clientC; where date <= day 106.82; 93 messages	clientC to clientI; where date <= day 109.08; 93 messages	*Two-way -> found in one account
	clientG to clientC; where date > day 106.82; 161 messages	clientC to clientG; where date > day 109.08; 156 messages	**Two-way -> found in two accounts
clientD@utas.edu.au	-	clientD to clientB; where date <= day 22.93; 7 messages	One-way interaction
	-	clientD to clientI; where date > day 22.93 and date <= day 26.13; 4 messages	*Two-way -> found in one account
	-	clientD to clientF; where date > day 26.13 and date <= day 28.93; 2 messages	One-way interaction
	-	clientD to clientB; where date > day 28.93 and date <= day 31.47; 2 messages	One-way interaction
	clientI to clientD; where date <= day 48.74; 31 messages	clientD to clientI; where date > day 31.47 and date <= day 49.29; 19 messages	*Two-way -> found in one account
	clientG to clientD; where date > day 48.74; 128 messages	clientD to clientG; where date > day 49.29; 119 messages	*Two-way -> found in one account
clientF@utas.edu.au	-	clientF to clientC; where date <= day 103.67; 15 messages	One-way interaction
	-	clientF to clientD; where date > day 103.67; 19 messages	One-way interaction
clientG@utas.edu.au	-	clientG to clientC; where date <= day 7.7; 3 messages	One-way interaction
	clientB to clientG; where date <= day 32.11; 19 messages	clientG to clientB; where date > day 7.7 and date <= day 35.58; 21 messages	**Two-way -> found in two accounts
	clientA to clientG; where date > day 32.11 and date <= day 105.18; 79 messages	clientG to clientA; where date > day 35.58 and date <= day 108.89; 84 messages	**Two-way -> found in two accounts
	clientC to clientG; where date > day 105.18; 161 messages	clientG to clientC; where date > day 108.89; 159 messages	**Two-way -> found in two accounts
clientH@utas.edu.au	-	clientH to clientE; where date <= day 65.82; 27 messages	One-way interaction
	-	clientH to clientI; where date > day 65.82; 104 messages	One-way interaction

used to analyse all incoming and outgoing e-mail traffic for each e-mail client to produce two sets of decision tree outputs, shown in Appendix F.2. The decision tree outputs that had branching information showing the date/time information were then tabulated and compiled together in Table 5.2. The data presented in Table 5.2 shows the unusual changes in interaction behaviour found from the incoming and outgoing e-mail traffic for some of the e-mail clients. The data in Table 5.2 also identifies the type of interaction change found, by indicating whether the unusual change in interaction was found to be a one-way or two-way change in interaction. It also identifies whether the change in interaction was found in the mailbox belonging to one or two e-mail accounts.



(a) Two-way changes in interactions found from two e-mail accounts.

(b) Two-way changes in interactions found from only one e-mail account.

Figure 5.18: Social network diagrams showing where the decision tree identified unusual changes in interaction from two directions.

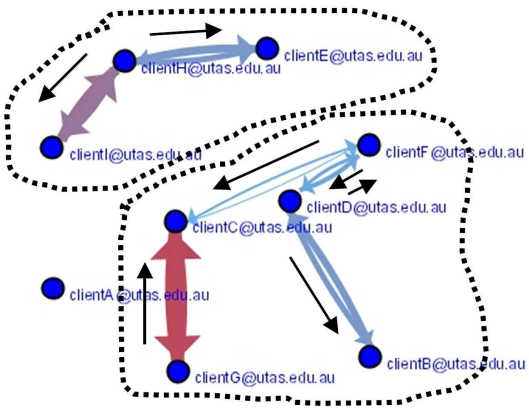


Figure 5.19: Social network diagrams showing where the decision tree identified unusual changes in interaction in one direction.

However, the decision tree information compiled in Table 5.2 does not really provide a sense of the nature of the change in interactions or where these changes in interactions are located in the e-mail system. A better interpretation of the results

can be obtained by using the diagrams produced by social network visualisation to show where the unusual changes in interaction behaviour are occurring, as illustrated in Figures 5.18 and 5.19. From these diagrams, it is observed that many of the unusual changes in interactions involve *clientA*, *clientI*, and *clientG*, whom are highly extroverted individuals (see Figure 5.14). This type of result was what was expected from the assigning those behavioural profiles.

Further Investigation of Unusual Changes in Interaction Behaviour

The interactions found in Table 5.2 can be further investigated using visualisation techniques. For this case study, *clientG* is selected for further investigation. The diagrams and screenshots from Figures 5.20 to 5.26 provide an account of the e-mail traffic interactions that *clientG* has been involved in over the 182 simulation days. An overview of *clientG*'s e-mail traffic activities shown in Figure 5.20 does not reveal anything noticeably 'unusual'. However, if the decision tree in Figure 5.22 is referred to and the interaction between *clientG* and *clientA* between day 35 and day 108 is selected, a close inspection of the outgoing e-mail traffic of *clientG* and *clientA* in Figure 5.23 reveals there is a rather unusual drop in outgoing e-mail traffic for *clientG* on week 14 (prior to week 15 or day 108). The same time-series graph also shows a general downturn in outgoing e-mail traffic between *clientG* and *clientA* after week 15.

The decision tree in Figure 5.22 indicates there is also some unusual change in interaction occurring between *clientG* and *clientC* after day 108. An inspection of the weekly and daily e-mail traffic between *clientG* and *clientC* in Figures 5.24 and 5.25 reveals that there was a very unusual sharp rise in outgoing e-mail traffic between *clientG* and *clientC* after day 108. The social network diagram in Figure 5.26 confirms this change in interaction between *clientG* and *clientC* before and after day 108. These details show there were unusual changes in interactions occurring between *clientG*<->*clientA* and *clientG*<->*clientC* around week 15, suggesting that *clientG* must have had a significant influence on the nature of these interactions.

5.2.3 Discussion

E-mail Traffic Analysis System

The two simulation case studies demonstrated how the computational intelligence approach can be useful for analysing the e-mail traffic behaviour patterns

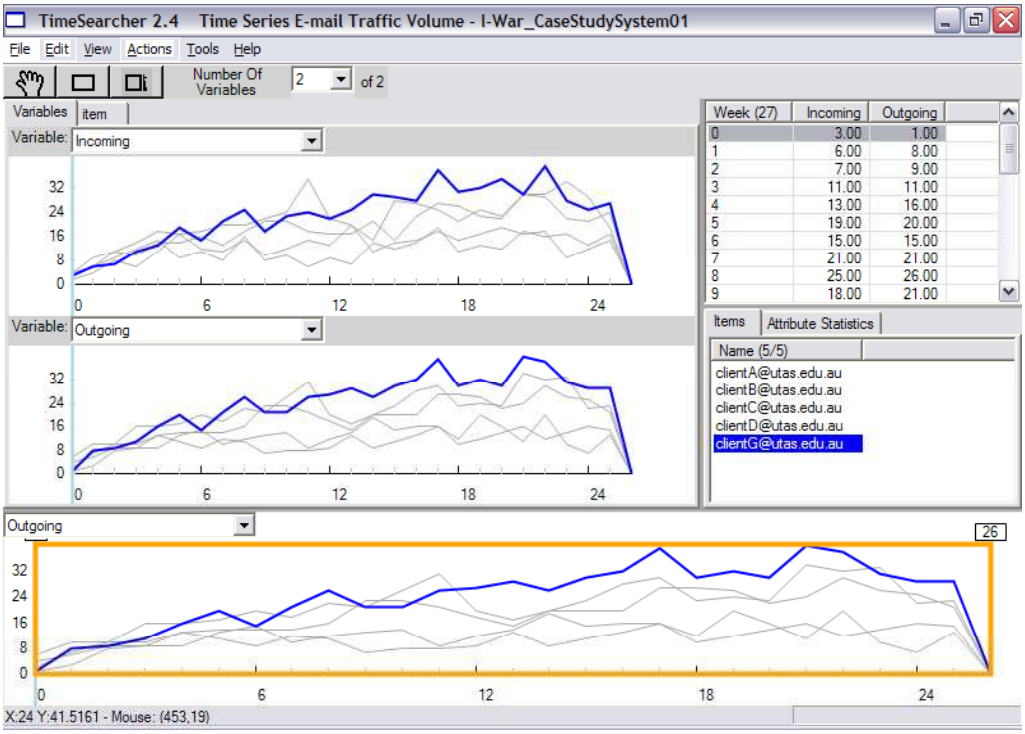


Figure 5.20: Weekly time-series data for *clientG*, showing the general e-mail traffic activity.

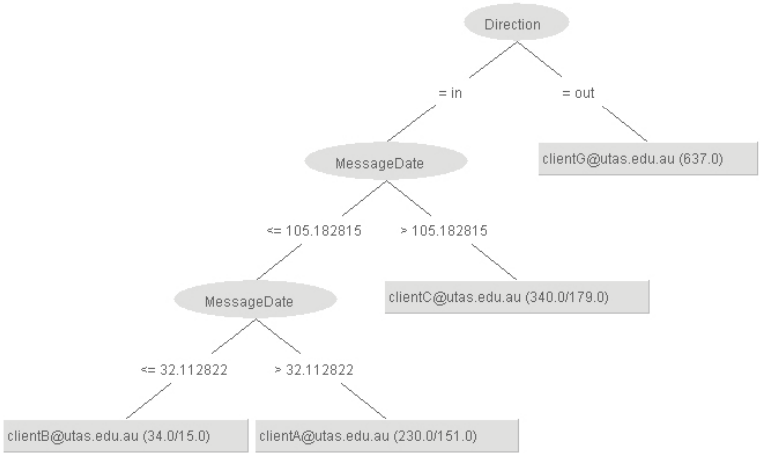


Figure 5.21: Decision tree classification output for *clientG*'s unusual changes in incoming e-mail traffic interactions.

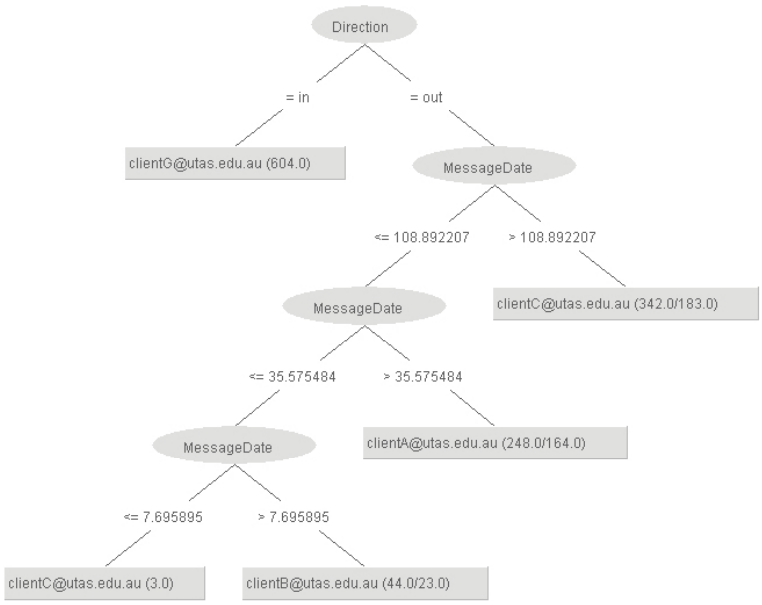


Figure 5.22: Decision tree classification output for *clientG*'s unusual changes in outgoing e-mail traffic interactions.

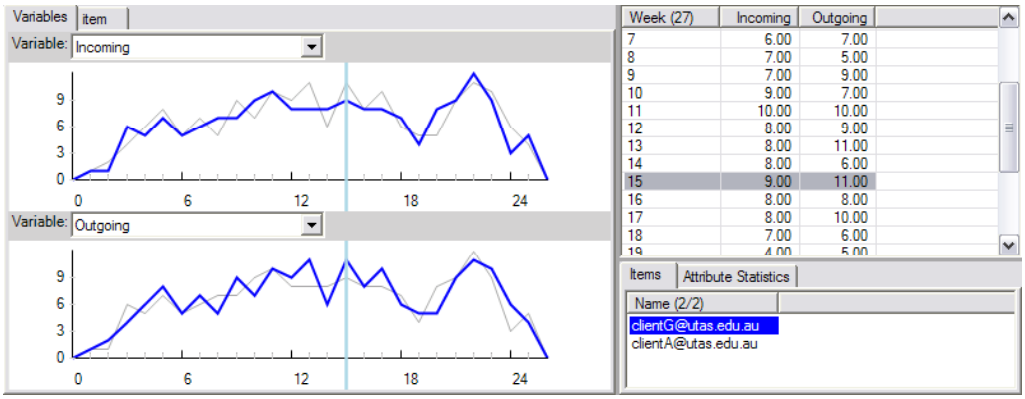


Figure 5.23: Weekly time-series of traffic between *clientG* and *clientA*, highlighting the week 15 (≈ day 108) with a drop in number of outgoing e-mails on week 14.

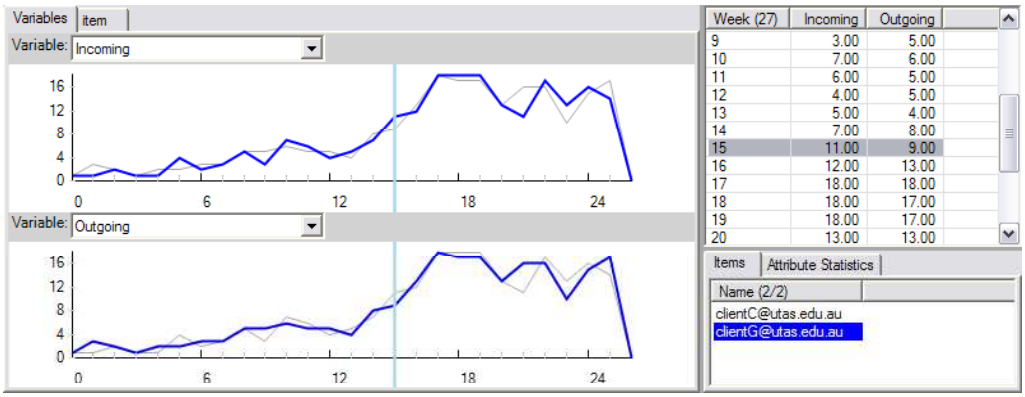


Figure 5.24: Weekly time-series data for *clientG* and *clientC*, highlighting the week 15 (≈ day 108) with a significant rise in number of e-mails from *clientG* to *clientC*.

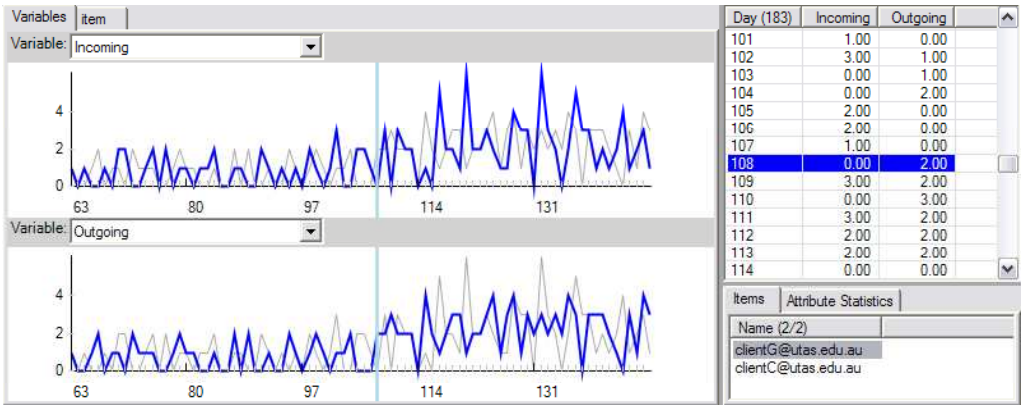
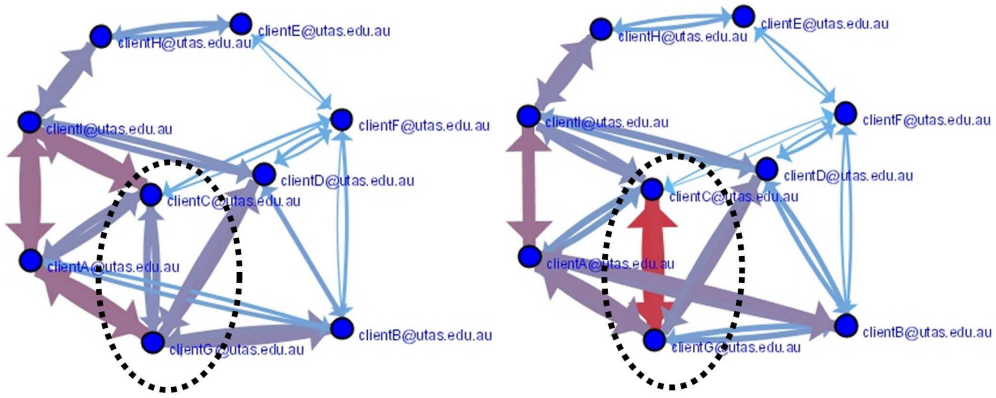


Figure 5.25: Daily time-series data for *clientG* and *clientC*, showing an unusual increase in daily e-mail traffic activity after day 108.



(a) E-mail social network prior to day 108. (b) E-mail social network after day 108.

Figure 5.26: Social network diagrams highlighting the change in communications between *clientG* and *clientC* before and after day 108.

of e-mail users. The utilisation of time-series visualisation, social network visualisation and decision tree classification techniques presented different ways of examining e-mail traffic behaviour. The visualisation techniques provided a useful method for obtaining a general overview of e-mail traffic behaviour of the e-mail users, as shown in Figures 5.5, 5.6, and 5.17. However, the visualisation techniques did not provide any immediate visual indications that inform the user about the location of unusual changes in e-mail traffic behaviour.

Decision tree classification provided a another method for analysing the data, by presenting information that "pinpoint" the exact locations of unusual changes in interaction behaviour between an e-mail client and their associates. While the information presented by the decision tree classification output is intuitive and easy to interpret, as shown in Figures 5.9, 5.10, 5.21 and 5.22, it is not immediately clear what type of changes in interaction behaviour has occurred between an e-mail client and their associates. With the assistance of social network

visualisation and time-series visualisation techniques, these aided in verifying the changes in interaction behaviour found through decision tree classification. The investigation of unusual changes in interaction behaviour for *clientG* in case study 1 (Figures 5.11, 5.12 and 5.13) and *clientG* in case study 2 (Figures 5.23, 5.24, 5.25, and 5.26), showed that visualisation techniques can be useful for understanding the details of the changes in interaction found by the decision tree classification output. Thus, this shows how the computational intelligence approach useful for understanding the e-mail traffic behaviour patterns of the simulated e-mail users.

While the visualised decision tree outputs (see Figures 5.9, 5.10, 5.21, and 5.22) provide an intuitive way of describing and locating unusual changes in e-mail traffic interactions between a suspect e-mail account and their associates, there may be difficulty in interpreting the visualised decision tree output as the size of the tree increases in height and width (i.e. contains a larger number of branch and leaf nodes). Decision tree classification is currently used in this research to identify clusters of e-mail messages sent to or received from a suspect e-mail account's associates. The number of associates allocated for each e-mail client in the two simulation case studies is quite low, within the range of 2 to 4 associates per e-mail client. This results in a relatively small decision tree output that is generally able to separate messages sent to or received from associates into a small number of leaf nodes (e.g. around 4 or 5 leaf nodes as seen in Figures 5.9, 5.10, 5.21, and 5.22). However, as the number of associates for a suspect e-mail account is increased, this may likely result in a decision tree output that contains a much larger number of leaf nodes (e.g. more than 10 leaf nodes). This is because the decision tree classification algorithm still has to attempt to partition and classify clusters of e-mail messages sent to or received from each associate. The larger decision tree output may be more difficult for the user/analyst to interpret, but this something that is still yet to be considered in order to determine the practicality of using decision tree classification with genuine e-mail traffic data.

Conceptual Simulation Model

The conceptual simulation model used for the two simulation case studies showed how the creation of simulated e-mail clients with different behavioural profiles allowed for the observation of the linkage between an e-mail client's assigned behaviour profile and their e-mailing behaviour. By observing the personality trait degree values assigned to the behaviour models of each e-mail client and the total number of e-mail messages sent by clients during the case study simu-

lation, it was easy to examine the effects that certain personality traits, such as extroversion, had on overall e-mail traffic behaviour (e.g. extroverted individuals generally sent more e-mails, introverted individuals generally sent less e-mails). However, the number of e-mails sent by e-mail clients is not just dependent on how extroverted an e-mail client is, but also dependent on to whom the e-mail client is socially connected and how they respond to e-mail messages sent by the more extroverted individuals.

This type of behaviour can be observed by simply examining the number of e-mail messages sent and received by e-mail clients socially connected to *clientA* and *clientJ* in case study 1 (Figure 5.5), and those connected to *clientA*, *clientG*, and *clientI* in case study 2 (Figure 5.17). For case study 1, it can be seen from Figures 5.4 and 5.5 that *clientB*, *clientF*, and *clientG* are influenced in some way by their social connections to *clientA* and *clientJ*. Similarly for case study 2, it can be seen from Figures 5.16 and 5.17 that *clientD* and *clientH* are influenced in some way by their social connections to *clientA*, *clientG*, and *clientI*. An exception to this assumption is *clientI* in case study 1 and *clientB* in case study 2. These e-mail clients are supposed to be highly extroverted but do not send or receive as many e-mail messages as other clients connected to *clientA* and *clientJ* for case study 1 or *clientA*, *clientG*, and *clientI* for case study 2. A suggestion for this exception in behaviour is that clientI from case study 1 and clientB from case study 2 have very low emotional stability trait degree values ($D_{ES} = 0.052$ for *clientI* and $D_{ES} = 0.256$ for *clientB*), indicating that *clientI* and *clientB* would be considered to be highly temperamental and emotional, hence not always communicating with other clients.

What was not easy to predict and understand was the individual interactions between different e-mail clients. With the use of decision tree classification, time-series visualisation, and social network visualisation techniques, the combination of these techniques provided a way for better observation of the e-mail traffic behaviour of e-mail clients and linking the behaviour back to the e-mail clients' behaviour profile. For example, clientG in case study 1 is described as having a fairly introverted, reclusive, and lazy behaviour profile ($D_{EX} = 0.132, D_{ES} = 0.425, D_C = 0.182$), explaining why clientG generated a low volume of e-mail traffic and had a drop in e-mail traffic activity on week 6 (Figures5.11 and 5.12). These simulation case studies therefore demonstrate that the e-mail system simulation model generates particular types of e-mail traffic behavioural patterns, each of which can be linked back to the behavioural profiles of each e-mail client.

Although the conceptual e-mail system model provides a way of simulating the

basic e-mail behaviours of sending e-mails and replying to e-mails, there are some limitations to the conceptual simulation model. Firstly, there are other e-mail behaviours that could be modelled to provide other useful metrics for observing e-mailing behaviour. These e-mail behaviours are sending attachments, forwarding e-mails, and sending to multiple recipients, which were described previously in Section 4.3.1.2. Secondly, the conceptual model does not have a specific way of modelling social groups, which are an important factor in communications in real e-mail social networks. An individual's e-mailing behaviour may be affected by the social groups they belong to, or may be affected by their involvement in the certain social groups. For example, if an individual is involved in organising a conference they would spend a lot of time sending e-mail messages to other conference committee members while the conference is being organised. Once the conference has finished, the individual's e-mailing behaviour will change since they no longer need to send e-mails to committee members. Finally, the conceptual model does not model any specific type of e-mailing habits, such as: sending e-mail messages when arriving to work in the morning, not sending e-mails on weekends, and not sending e-mail messages during normal sleeping hours (e.g. between 10pm to 7am). These factors do have an affect on certain individuals' e-mailing behaviour as described by [41, 42], but are not always consistent among all individuals.

Other Considerations

With the current conceptual simulation model of the e-mail system and the e-mail traffic analyser system, it has only been demonstrated how different e-mail traffic behavioural patterns can be generated and demonstrated how the unusual changes in interaction behaviour can be detected by the use of decision tree classification. However, the simulated e-mail system used for the case studies is only an extremely small e-mail system consisting of 9 or 10 e-mail clients. Obviously the 9-client and 10-client e-mail system used for the case studies provide a useful way for explaining how the e-mail traffic analysis system operates. But for a much larger e-mail client consisting of hundreds or thousands of e-mail users, it must be considered whether using decision tree classification will still be effective for finding unusual changes in interaction behaviour. It still remains to be determined whether decision tree classification still provides useful information to the user when analysing much larger e-mail systems.

5.3 Case Study: Enron E-mail Traffic Data

The case study in this section focuses on the evaluation of the e-mail traffic analysis system using the Enron e-mail corpus. The purpose of the case study is to demonstrate how the e-mail traffic analysis system can be used to investigate the e-mail traffic behaviour patterns of a particular individual from Enron. The individual chosen for the case study will be an individual known to have been involved or closely associated with the conduct of illegal activities within Enron. The computational techniques from the e-mail traffic analysis system are used in this case study to investigate abnormal changes in the selected individual's e-mail traffic behaviour patterns.

5.3.1 Analysis of Enron Employee's Traffic Behaviour

Selecting An Individual To Analyse

There were a number of key people associated with the Enron financial crisis in 2001 who were considered for analysis. A list of people associated with setting up the fraudulent financial records were given by [107], some of whom were also part of senior management in Enron. Out of the list of people considered, only a few of them had their full e-mail traffic information collected as part of the sample taken from the 151 former Enron employees. Based on these considerations, the person selected for this case study was Jeffrey Skilling.

Skilling first joined Enron in 1990 as the chief executive to be in charge of developing Enron's trading services, became CEO of Enron in February 2001, then unexpectedly resigned as CEO on 14th August 2001 for "personal reasons". The reason for selecting Jeffrey Skilling is that he was a key person involved in transforming Enron from a traditional gas-line operator to a "new-economy" trading company in the 1990's [107]. He also had a short 6-month run as CEO of Enron before resigning in August 2001, and most of his mailbox information is available as part of the Enron e-mail dataset. Skilling's involvement in the overall management of Enron as well as being associated with illegal activities within Enron [113] makes him a suitable candidate for investigating his e-mail traffic behaviour.

Before analysing Jeffrey Skilling's e-mail traffic, it had to be determined which out of the 75,547 unique e-mail addresses belonged to Jeffrey Skilling. To find Jeffrey Skilling's e-mail addresses, a database search was performed to find e-mail addresses that contain sequences of characters that may resemble parts

of Jeffrey Skilling's name. This was based on the assumption that e-mail addresses containing such sequence of characters may have been likely to belong to e-mail accounts or e-mail addresses used by Jeffrey Skilling. The ISI Enron dataset created by [111] contains a table, "employeelist", that associates each of the 151 Enron employees with one particular e-mail address. The original ISI Enron dataset by [111] only associated Jeffrey Skilling with the e-mail address "*jeff.skilling@enron.com*". However, there may have been more than one e-mail address used by Jeffrey Skilling, which is something considered in this case study. To obtain Jeffrey Skilling's possible e-mail addresses, a wildcard database search was performed for e-mail addresses matching "j%skilli%" and "skilli%", where "%" is the wildcard character for the search. The results of this search returned 15 possible matching e-mail addresses, shown in Table 5.3.

Table 5.3: A listing of e-mail addresses possibly belonging to Jeffrey Skilling.

Possible Matching E-mail Addresses for Jeffrey Skilling
'jeff.skilling@enron.com', 'jeffrey.k.skilling@enron.com', 'jeffrey.skilling@enron.com', 'jeffreyskilling@yahoo.com', 'jeffrey_skilling@enron.com', 'jeff_skilling@enron.com', 'jskilli.enron@enron.com', 'jskilli@ei.enron.com', 'jskilli@enron.com', 'jskilling@enron.com', 'skilli@ei.enron.com', 'skilli@enron.com', 'skilling@enron.com', 'skilling@tribune.com', 'skillingj@enron.com'

Selection Of Time Periods For Analysis

The Enron e-mail dataset mainly covers a time-span from 1999 to end of 2002 with the exception of outlier messages dated at years such as 0001 and 2044, which were excluded from the analysis. During the 1999 - 2002 time-span, there were a number of key events that occurred, ending with the company's filing for Chapter 11 bankruptcy protection in December 2001 [107, 116]. The time periods chosen for analysis of Jeffrey Skilling's e-mail traffic behaviour is during the intervals: 1st January 1999 to 1st August 2000 and 1st February 2001 to 1st September 2001. These intervals of time were chosen for analysis because they reflect different periods in Jeffrey Skilling's career at Enron. Between 1st January 1999 to 1st August 2000, Jeffrey Skilling was Chief Operating Officer (COO) at Enron [107, 117]. After February 2001, Jeffrey Skilling was elevated to the role of Chief Executive Officer (CEO) and was in charge of the overall operations of Enron. These changes in his role at Enron suggests that his observed communication behaviour is expected to change as a result of becoming CEO in

Enron.

General Overview Of The Individual's Communication Behaviour

A general overview of Jeffrey Skilling's e-mail traffic behaviour patterns can be viewed by using social network visualisation and time-series visualisation to reveal information about how Skilling communicated during the time periods selected. Figure 5.27 shows a social network diagram of Jeffrey Skilling's e-mail addresses and the people whom he communicated with during the period 1st January 1999 to 1st August 2000 when he was COO. The social network diagram in Figure 5.28 shows a change in Skilling's social network communication structure during 1st February 2001 to 1st September 2001 when he was CEO of Enron. The weekly time-series diagram in Figure 5.29 presents another perspective on Skilling's e-mail traffic behaviour, by showing the volume of outgoing and incoming e-mail traffic associated with his e-mail accounts from 1st January 1999 to 1st September 2001. The visualisation diagrams shown in Figures 5.27, 5.28 and 5.29 indicate there were variations in Skilling's communication behaviour over the two time periods selected, but do not provide any immediate visual indications that help to spot or identify abnormal changes in communication behaviour. This is where hierarchical fuzzy inference can be used to aid in identifying abnormal changes in the suspect's communication behaviour.

Detection of Abnormal Changes in Communication Behaviour

In order to use hierarchical fuzzy inference to provide information on abnormal changes in e-mail traffic behaviour, the profiling and surveillance periods need to be specified for the Anomaly Detection Unit component of the e-mail traffic analysis system. The profiling period selected is the initial period of 1st January 1999 to 1st August 2000 when Jeffrey Skilling was COO of Enron and the surveillance period selected is the period of 1st February 2001 to 1st September 2001 when Skilling was CEO of Enron. After specifying the profiling and surveillance periods, the Anomaly Detection Unit was used to obtain the nine communication behaviour measurements from both periods of time and then compute the difference in behaviour measurements. These difference measurements were then input into the hierarchical fuzzy inference system.

For this case study, the hierarchical fuzzy inference system from Section 3.4.2 is used for providing information on the suspect's overall change in e-mail traffic patterns. This hierarchical fuzzy system was used to rate each of the suspect's

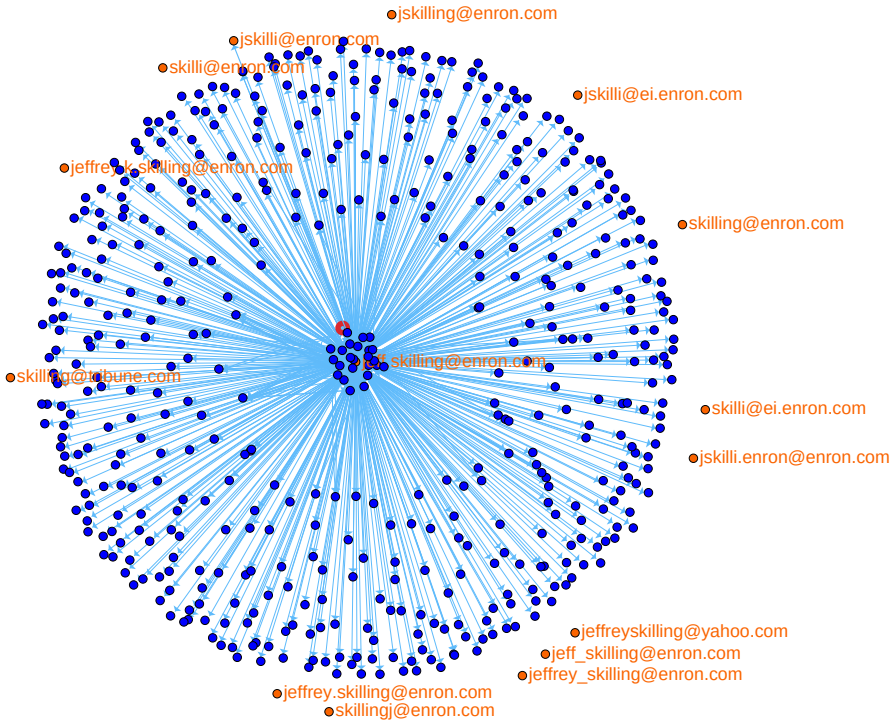


Figure 5.27: Jeffrey Skilling's e-mail addresses (orange) and his circle of associates (blue) from 1st January 1999 to 1st August 2000 (17 months).

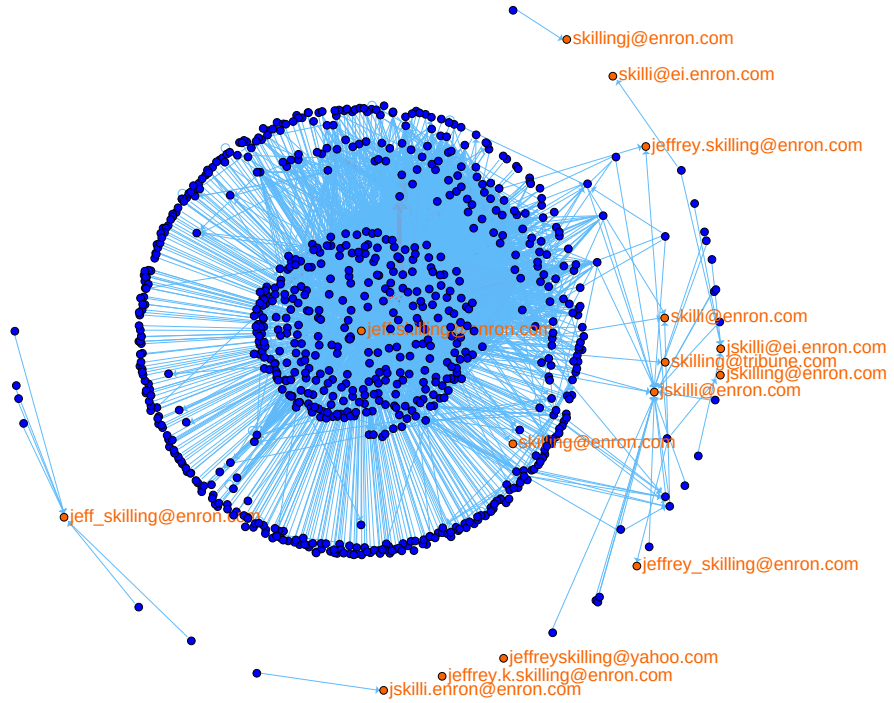


Figure 5.28: Jeffrey Skilling's e-mail addresses (orange) and his circle of associates (blue) from 1st February 2001 to 1st September 2001 (7 months).

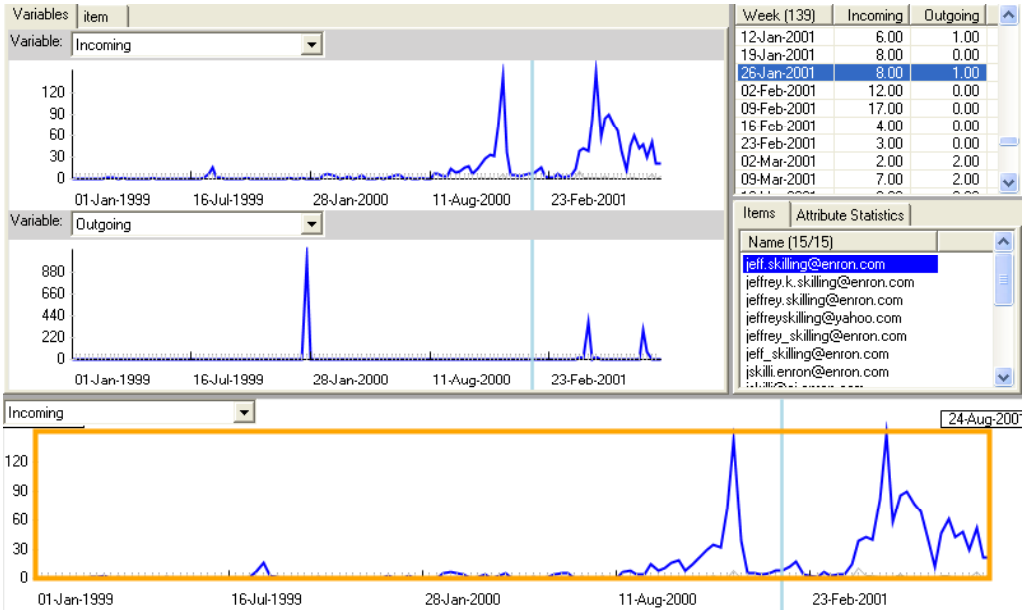


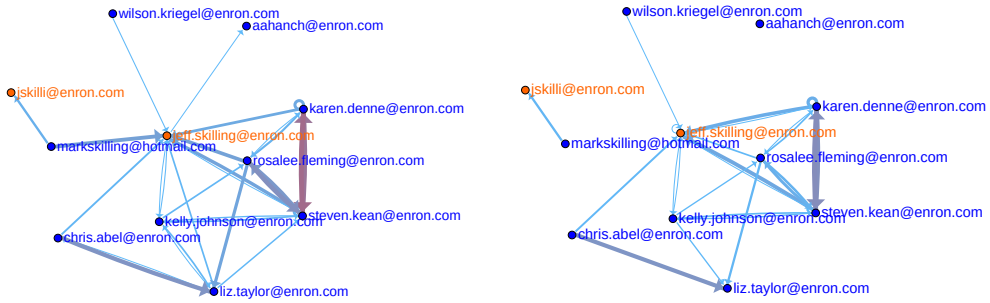
Figure 5.29: Weekly time-series overview of traffic from Jeffrey Skilling's e-mail accounts from 1st January 1999 to 1st September 2001.

Table 5.4: A listing of the top ten abnormality ratings from the communication links analysed.

Communication Link	**Abnormality Rating**
jeff.skilling@enron.com<=>steven.kean@enron.com	0.5
jskilli@enron.com<=>markskilling@hotmail.com	0.5
jeff.skilling@enron.com<=>karen.denne@enron.com	0.3
jeff.skilling@enron.com<=>kelly.johnson@enron.com	0.3
jeff.skilling@enron.com<=>liz.taylor@enron.com	0.3
jeff.skilling@enron.com<=>markskilling@hotmail.com	0.3
jeff.skilling@enron.com<=>wilson.kriegel@enron.com	0.3
jeff.skilling@enron.com<=>chris.abel@enron.com	0.11335529
jeff.skilling@enron.com<=>rosalee.fleming@enron.com	0.092071182
jeff.skilling@enron.com<=>aahanch@enron.com	0.091424688

communication links to provide information about the degree of change in traffic behaviour that has occurred. The top ten abnormality ratings output for each of Jeffrey Skilling's communication links is shown in Table 5.4, sorted in descending order. The full list of abnormality ratings and the difference measurements obtained for each of Jeffrey Skilling's 525 communication links is shown in Appendix G.

What the abnormality ratings in Table 5.4 indicate is a summarised overview



(a) Top ten communication links during 1st January 1999 to 1st August 2000.

(b) Top ten communication links during 1st February 2001 to 1st September 2001.

Figure 5.30: Social network diagrams of Jeffrey Skilling's top ten ranked communication links.

of each of Jeffrey Skilling's communication links and the amount of change in communication behaviour that has occurred over the two selected time periods. From the ratings shown in Table 5.4, it is observed that only a few of Jeffrey Skilling's communication links (the top 8 communication links) showed a reasonable change in communication behaviour. The changes in communication behaviour indicated by the output of the hierarchical fuzzy inference system can be further verified using the visualisation techniques.

The social network diagrams in Figure 5.30 presents an overview of the changes in communication behaviour that have occurred for each of the ten communication links shown in Table 5.4. Another perspective on the abnormality ratings can be obtained by further investigation with time-series visualisation. For the interaction between *jeff.skilling@enron.com* and *rosalee.fleming@enron.com*, this was given an abnormality rating of 0.092071181977. The time-series diagram in Figure 5.31 confirms that there was not much deviation in communication between Jeffrey Skilling and Rosalee Fleming, despite the spike in e-mail traffic that was outside of the profiling and surveillance period.

An investigation of the interaction between *jeff.skilling@enron.com* and *steven.kean@enron.com*, which had an abnormality rating of 0.5, uncovered a reasonable change in communication behaviour during the surveillance period. Figure 5.32 shows increase in e-mail traffic activity from Steven Kean to Jeffrey Skilling, suggesting there might have been a change in relationship during the surveillance period. According to the organisational role spreadsheet provided by [111], Steven Kean was actually the Vice President and Chief of Staff at Enron, which might have explained why he had more communication with Jeffrey Skilling, after Skilling became CEO in February 2001.

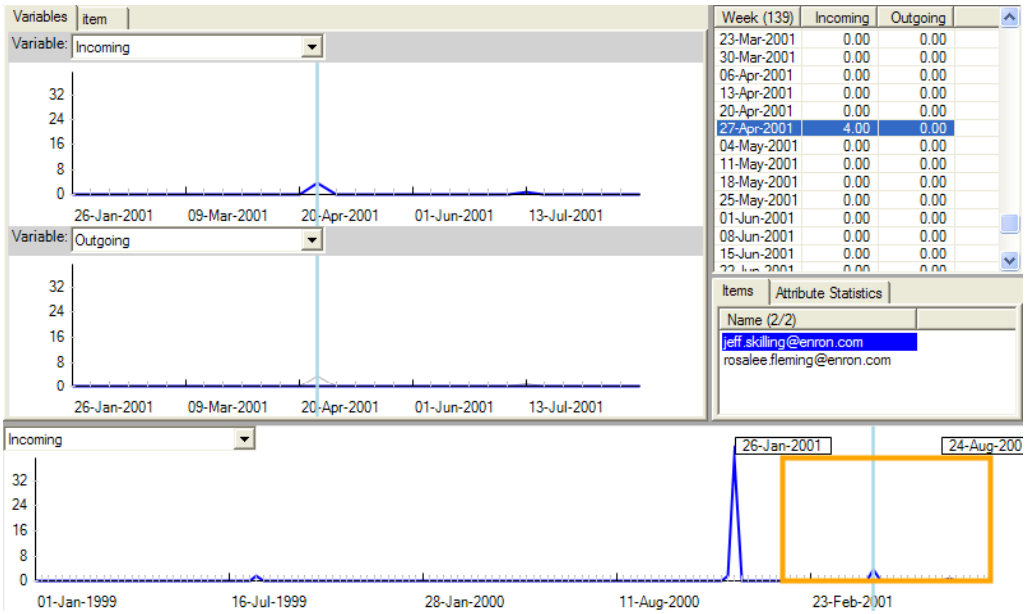


Figure 5.31: Weekly time-series of Jeffrey Skilling and Rosalee Fleming, focusing on the surveillance period.

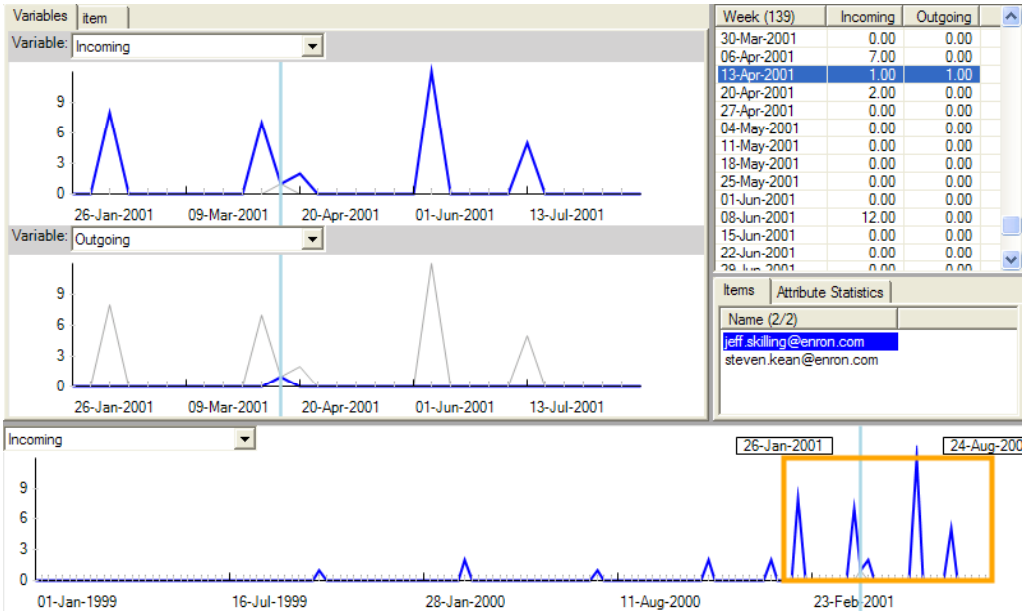


Figure 5.32: Weekly time-series of Jeffrey Skilling and Steven Kean, focusing on the surveillance period.

5.3.2 Discussion

E-mail Traffic Analysis System

The Enron case study demonstrated how the computational intelligence approach can be useful for analysing genuine e-mail traffic data. The combined use of hierarchical fuzzy inference with social network visualisation and time-series visualisaton techniques, showed how useful information can be obtained about the e-mail traffic behaviour patterns of a selected individual. The two visualisation techniques used in the Enron case study provided information about different aspects of a selected individual's traffic behaviour. The diagrams in Figures 5.27, 5.28, and 5.29 provided a general overview of Jeffrey Skilling's e-mail traffic behaviour, by showing how each of his e-mail accounts communicated over different periods of time. While these visualisation diagrams showed how visualisation techniques can be used to provide a general overview of a selected individual's e-mail traffic behaviour, the sole use of visualisation techniques do not reveal details of abnormal changes in communication link behaviour between a suspect e-mail account and a particular associate.

The evaluation of hierarchical fuzzy inference in the Enron case study showed that it is able to provide ratings that indicate the changes in communication behaviour observed for each of the selected individual's communication links, as shown in Table 5.4. These ratings provide a summarised overview of the overall changes in communication behaviour that has occurred for the individual, based on the fusion of nine behaviour change measurements. These ratings can be used to sort the suspect's communication links, as shown in Table 5.4, in order to determine which communication links exhibited the largest or smallest change in communication behaviour. This sorting of communication links can be useful when prioritising which communication links are to be investigated in further detail.

With the support of visualisation techniques, this showed how the ratings provided by hierarchical fuzzy inference can be further investigated. The visualisation diagrams in Figures 5.30, 5.31 and 5.32 showed how visualisation techniques can be used to reveal the details of abnormal changes in communication behaviour found for particular communication links. This shows how it is useful to utilise a number of perspectives on e-mail traffic behaviour, in order to provide a better understanding of the suspect's changes in communication behaviour. Thus, this demonstrates how the computational intelligence approach can be useful for analysing genuine e-mail traffic data.

It should be noted however, that there are some limitations with the use of hierarchical fuzzy inference to detect abnormal changes in communication behaviour. Although fuzzy inference helps to summarise the anomaly detection results, one of its limitations is that a fuzzy inference system is complex to design, and takes a great deal of effort to build and fine-tune its performance. It is often said that: "improving the system becomes rather an art than engineering" [66], meaning that it often takes some trial and error and experience to determine if the system is performing the way it is expected. Another limitation is that the design of the fuzzy rules and fuzzy sets were manually constructed, based on some empirical knowledge of e-mail traffic. However, there are ways of automating some of the design process when developing fuzzy inference systems. An example of this is where [118] used a fuzzy C-means algorithm [119] to automate the part of the process of designing the fuzzy sets for their network intrusion detection system.

One other limitation to note is that the hierarchical fuzzy inference architecture used for this case study, only demonstrates one of the possible groupings for combining the inputs for fuzzy inference. This may not necessarily be the best possible grouping, so the other input combinations will need to be tested to see if they affect the results given by the hierarchical architecture. This is because it is difficult to determine which of the input behaviour change measurement variables are having the most influence on the hierarchical fuzzy inference system's output.

Enron E-mail Dataset

The Enron e-mail dataset used for this case study demonstrated that using a dataset with known characteristics helps with verifying the e-mail traffic behaviour patterns observed in the data by the e-mail traffic analysis system. The known characteristics in the Enron e-mail dataset are: known events that occurred at Enron that are covered by the dataset (e.g. the appointment of Jeffrey Skilling as CEO in February 2001 and Skilling's resignation in August 2001 [107]) and also knowledge about some of the individuals whom were sampled as part of the dataset (e.g. people with executive or managerial positions at Enron, such as Jeffrey Skilling and Steven Kean [111]). These known characteristics in the Enron e-mail dataset provide useful reference points for finding changes in communication behaviour.

With the use of hierarchical fuzzy inference, social network visualisation and time-series visualisation techniques, these revealed useful information about the

changes in communication behaviour that occurred before and after Jeffrey Skilling was appointed CEO of Enron. The visualisation diagrams in Figures 5.27, 5.28, and 5.29 reveal there were major changes in Jeffrey Skilling's e-mail traffic behaviour after he was appointed CEO in February 2001. Closer inspection of Jeffrey Skilling's communication links with hierarchical fuzzy inference in Table 5.4 revealed that there were only up to 8 communication links that showed a reasonable change in communication behaviour. These observations correlate with what was known to have occurred at Enron during the time periods of 1st January 1999 to 1st August 2000 and 1st February 2001 to 1st September 2001. Thus, this demonstrates how the Enron e-mail dataset is useful for verifying the e-mail traffic behaviour patterns observed by the e-mail traffic analysis system.

5.4 Concluding Remarks

The two sets of case studies presented in this chapter demonstrated how the computational intelligence approach is useful for analysing e-mail traffic behaviour. In both case studies, computational intelligence was applied by utilising both feature extraction and visualisation techniques to investigate the changes in communication behaviour exhibited by particular suspect e-mail accounts. It is through the utilisation of different techniques that it is shown how computational intelligence aids in providing a much broader understanding of a suspect's communication behaviour. This was demonstrated in Sections 5.2.1 and 5.2.2 by using decision tree classification with time-series and social network visualisation, and demonstrated in Section 5.3.1 by using hierarchical fuzzy inference with time-series and social network visualisation. Without using a combination of computational techniques, this would only provide a limited understanding about a suspect e-mail account's traffic behaviour.

An important part of the case studies presented is the e-mail traffic data used, since the data contains behavioural characteristics that affect what is observed by the e-mail traffic analysis system. The simulated e-mail traffic data used in Sections 5.2.1 and 5.2.2 had observable e-mail traffic behavioural characteristics that can be linked back to the setup of the simulation model used to generate the data. In the Enron e-mail traffic data used in Section 5.3.1, this had observable e-mail traffic behavioural characteristics that could be verified through knowledge about the role of particular individuals and knowledge about events that occurred within Enron. The two types of data used in the case studies highlight the importance of considering the type of data used for e-mail traffic analysis,

since the computational techniques used for analysing the data can only extract information on behaviour that is present in the data.

Although there were four computational techniques evaluated in this chapter, other computational techniques may be considered for providing additional perspectives on a suspect e-mail account's traffic behaviour. This is because each new technique added may help in providing further understanding of a suspect e-mail account's traffic behaviour by presenting another perspective about the account's behaviour. This would aid in broadening the number of perspectives used by the user/analyst in analysing a suspect e-mail account's traffic behaviour. However, it is still to be determined what would be a suitable number of computational techniques for a user/analyst to utilise while investigating the behaviour of a suspect e-mail account. This may be dependent on what the user/analyst wants to know and the possible types of traffic behaviour that can be measured and analysed from e-mail traffic data.

5.5 Summary

This chapter described the evaluation of the e-mail traffic analysis system through two sets of case studies. Each set of case studies described in this chapter demonstrated how the e-mail traffic analysis system can be used to investigate the e-mail traffic behaviour of known suspect e-mail accounts. The two sets of case studies presented in this chapter were based on the use of simulated e-mail traffic data and the Enron e-mail dataset.

In the first set of case studies, this focused on the use of simulated e-mail traffic data to evaluate the e-mail traffic analysis system. The case studies presented in Sections 5.2.1 and 5.2.2 involved the use of the simulation tool from Section 4.3.1 to create a model of an e-mail system comprising of a small number of e-mail clients (9 or 10 e-mail clients) and assigning different behavioural profiles to each e-mail client. The behavioural profiles assigned enables each e-mail client to exhibit unique e-mail traffic behaviour patterns, which are associated with the personality trait degree values allocated to each e-mail client's behaviour model. Based on behavioural profiles assigned to each e-mail client, the model of the e-mail system was simulated to generate the communication interactions between e-mail clients and to produce the e-mail traffic data resulting from those interactions.

After simulating the interactions between the e-mail clients in the e-mail system

simulation model, the e-mail traffic analysis system was then used to investigate the interaction behaviour of the simulated e-mail clients. Each of the computational techniques from the e-mail traffic analysis system was used to analyse different aspects of the e-mail clients' e-mail traffic behaviour. The visualisation techniques, social network visualisation and time-series visualisation, were used to provide an overview of the traffic behaviour occurring between e-mail clients. It was found that the visualisation techniques provided useful information on the general behaviour of the e-mail clients, but did not provide any immediate visual indications of the presence of unusual changes in interaction behaviour between particular e-mail clients. Decision tree classification was then used in both Sections 5.2.1 and 5.2.2 to analyse the interaction behaviour of each e-mail client from their incoming and outgoing e-mail traffic data. The application of decision tree classification produced two sets of decision tree outputs for each e-mail client, with one decision tree output showing the locations of unusual changes in incoming interactions and the other decision tree output showing the location of unusual changes in outgoing interactions. While the information presented by the decision tree outputs was found to be intuitive and easy to interpret, it was not immediately clear what type of changes in interaction behaviour had occurred between particular e-mail clients and their associates.

To make more sense of the information provided by the decision tree classification outputs, the visualisation techniques were used in Sections 5.2.1 and 5.2.2 to investigate and explore the unusual changes in interactions for a particular e-mail client. It was found that the social network visualisation and time-series visualisation techniques were able to provide useful information about the changes in communication behaviour detected by decision tree classification. Time-series visualisation was found to be useful for describing the changes in e-mail traffic volume occurring between an e-mail client and a particular associate, while social network visualisation was found to be useful for describing the changes in communication intensity between an e-mail client and a particular associate. The combined use of decision tree classification, social network visualisation, and time-series visualisation demonstrated how the computational intelligence approach can be useful for analysing simulated e-mail traffic data.

The second set of case studies focused on the use of the Enron e-mail dataset to evaluate the e-mail traffic analysis system. In contrast to the first set of case studies, the case study in Section 5.3 used genuine e-mail traffic data and focused on analysing the e-mail traffic behaviour patterns of a particular individual from Enron. The individual selected for the case study was Jeffrey Skilling, one of the

former Enron executives who was Chief Operating Officer (COO) of Enron, and was appointed as CEO of Enron in February 2001. Skilling's e-mail traffic data was selected for analysis since he was known to be involved in the overall management of Enron as well as being associated with illegal activities that occurred within Enron. This made him a suitable candidate for investigating his e-mail traffic behaviour patterns.

To analyse Jeffrey Skilling's e-mail traffic data, a wildcard database search was performed to find e-mail addresses that had sequences of characters resembling Jeffrey Skilling's name. Once these e-mail addresses were found, the e-mail traffic analysis system was used to analyse the e-mail traffic behaviour associated with those suspect e-mail addresses for the periods: 1st January 1999 to 1st August 2000 and 1st February 2001 to 1st September 2001. It was found through social network visualisation and time-series visualisation that there were major changes associated with Jeffrey Skilling's e-mail traffic behaviour when comparing his traffic behaviour before and after he became CEO in February 2001. Closer inspection of his traffic behaviour was made by using hierarchical fuzzy inference to rate the abnormal changes in traffic behaviour observed for each of Skilling's communication links. The abnormality ratings provided by hierarchical fuzzy inference provided a useful method for summarising Jeffrey Skilling's overall changes in e-mail traffic behaviour, by describing the degree of behavioural change observed for each communication link. While the abnormality ratings provided a useful method for summarising and sorting Skilling's communication links, it did not reveal much detail about the actual changes the occurred for each communication link. The details of changes in communication link behaviour were investigated using the visualisation techniques and these were found to aid with understanding the rating provided by hierarchical fuzzy inference for particular communication links. This demonstrated how the computational intelligence approach can be useful for analysing genuine e-mail traffic data.

Overall, both sets of case studies demonstrated how the computational intelligence approach can be useful for analysing e-mail traffic behaviour. Each computational technique used presented a different perspective on the traffic behaviour of particular e-mail accounts. While the sole use of a particular technique only provided limited information about certain e-mail accounts, the combined use of different computational techniques, such as the combination of feature extraction with visualisation techniques, enabled unusual/abnormal changes in behaviour to be understood and provided a broader understanding about the e-

mail traffic behaviour of selected suspect e-mail accounts.

Chapter 6

Summary and Further Studies

The research work presented in this thesis has proposed a novel approach for examining and understanding the changes in e-mail traffic behaviour exhibited by a suspected individual's e-mail accounts. This chapter summarises the main discoveries of the thesis and highlights the major contributions of the research presented. It also highlights some of the limitations encountered during the research and discusses suggestions for further studies.

6.1 Investigations Involving E-mail As Evidence

The recent widespread use of computers and information technology into people's daily lives has seen the increasing involvement of computers and information technology in the conduct of illegal or criminal activities [29]. As a result, law enforcement agencies are increasingly having to deal with electronic evidence, which is becoming involved in the investigations of criminal activities. E-mail, has rapidly become a popular form of Internet communications since the early 1990's [9]. Its use has also become involved in various types of illegal or criminal activities. Examples of illegal or criminal activities that may involve the use of e-mail are: financial fraud, identity theft, child pornography distribution, industrial espionage, and stalking.

To examine the electronic evidence associated with criminal investigations, law enforcement agencies rely on the process of computer forensics to extract useful information from electronic data. Computer forensics is needed by law enforcement agencies in order to aid with reconstructing the series of events that have occurred in relation to a crime and to help link suspected individuals to the

crime. Computer forensics is becoming an important part of criminal investigations given growing amount of electronic evidence that is collected through investigations.

In investigations involving e-mail as evidence, law enforcement require ways of analysing e-mail to find useful information that will help provide a more complete understanding of the crime committed. The search for useful evidence from e-mail data is a difficult task, given the massive amounts of e-mail data that may need to be analysed. Therefore, one needs to consider the different types of e-mail analysis methods that may aid computer forensic analysts in extracting useful information from e-mail data.

There are two types of e-mail analysis methods that may be used for computer forensic examination of e-mail data: e-mail content analysis and e-mail traffic analysis. E-mail content analysis is a method that focuses on extracting useful information about the textual content of e-mails. This enables an analyst to obtain information about patterns of word usage or features associated with the linguistic characteristics of the e-mail messages written by particular individuals. E-mail traffic analysis is another method that focuses on extracting useful information related to the transit and delivery of e-mails. Unlike e-mail content analysis, e-mail traffic analysis focuses on obtaining useful information about the exchange of messages and types of e-mails sent, without looking at the content of e-mails. The research in this thesis considers the use of e-mail traffic analysis for investigating the communication activities of a suspected individual.

6.2 Analysis of E-mail Traffic Behaviour

An approach for analysing e-mail traffic is to consider the "behaviour" of the individuals using e-mail. Behaviour analysis of e-mail traffic considers the observable actions performed by individuals using e-mail (e.g. sending e-mails, replying to e-mails, sending attachments) and tries to extract useful information based on those actions performed. There are a variety of ways in which e-mail traffic behaviour can be analysed. One way of categorising e-mail traffic behaviour analysis methods is to consider them through different levels of analysis, whereby each level focuses on a particular level of detail about the e-mail users' behaviour. The levels of analysis that can be considered for analysis of e-mail traffic behaviour are:

- **Individual Behaviour Analysis** - the behaviour of an individual e-mail user.
- **Behaviour Comparison Analysis** - comparing the individual behaviours of multiple e-mail users.
- **Clique Behaviour Analysis** - the behaviour of a small cluster or group of individuals.
- **Network Behaviour Analysis** - the behaviour of the interconnections between a network of e-mail users.

A number of methods [2, 3, 4, 18, 41, 42, 59] have been proposed for detecting unusual or abnormal e-mail traffic behaviour at different levels of analysis. The methods by [2, 3, 18, 41, 42, 59] detect the presence of abnormal e-mail traffic behaviour using one of three approaches. The first approach is detecting abnormal behaviour by comparing the past and present e-mail traffic behaviour profiles of e-mail accounts to determine the presence of abnormal changes in behaviour. The second approach is detecting abnormal e-mail traffic behaviour by comparing the behaviour between different e-mail accounts. This can be done in order to find e-mail accounts that share unusually similar e-mail traffic behaviour to another e-mail account, or to find e-mail accounts that exhibit e-mail traffic behaviour that is significantly abnormal from other e-mail accounts. The third approach is performing a moving window analysis of the e-mail traffic data in order to find unusual bursts or variations in the e-mail traffic behaviour exhibited by particular e-mail accounts. In the method by [4], they propose an algorithm that searches for unusual group communication behaviour by examining the e-mail traffic data for a chain of e-mail messages passed between group members. This algorithmic search is performed in order to find a hidden group structure emerging from the chain of communications occurring amongst group members.

While the present behaviour analysis methods used for e-mail traffic analysis [2, 3, 4, 18, 41, 42, 59] are able to detect different types of abnormal communication behaviour, what has been overlooked is the importance of utilising information provided by different analysis techniques when presenting information to the user or analyst. This is considered important since different analysis techniques can present varying perspectives about the traffic behaviour of the e-mail accounts being analysed. By utilising the perspectives provided by different analysis techniques, this can provide new insight into patterns exhibited by particular e-mail accounts, which were previously hidden in the data.

This thesis proposed "computational intelligence" as an approach for using a set of computational techniques, to extract information from data and present the information to the user/analyst in a useful and intelligent manner. The computational intelligence approach was proposed for analysing e-mail traffic, in order to provide the user or analyst a number of different perspectives while investigating a suspected individual's e-mail traffic behaviour. The purpose of this approach is to allow the user/analyst to correlate patterns observed through different computational techniques, in order to make better sense of an individual's e-mail traffic behaviour. The computational intelligence approach was used in the research to enable the user to examine and understand the overall changes in e-mail traffic behaviour for a suspected individual's e-mail accounts.

6.3 Major Contributions

Through the work on computational intelligence in e-mail traffic analysis, this thesis has presented three major contributions towards the study of e-mail traffic analysis. These contributions are:

- Utilising a combination of techniques to examine e-mail traffic behaviour.
- Development of a personality trait based e-mail traffic simulation tool.
- Development of a novel system architecture for e-mail traffic analysis.

Each of these contributions are described in the following sections.

6.3.1 Utilising A Combination Of Techniques To Examine E-mail Traffic Behaviour

In order to obtain useful information about the changes in traffic behaviour for a suspected individual's e-mail accounts, two types of computational techniques were employed in the research to provide different ways of examining an individual's e-mail traffic behaviour patterns. The first type, visualisation techniques, was considered in order to provide the user the ability to visually explore the data and obtain a quick overview of the traffic behaviour exhibited by a suspected individual's e-mail accounts. The visualisation techniques used were:

- **Social Network Visualisation** - to provide an overview of the connections and the level of communication activity occurring between e-mail users.

- **Time-Series Visualisation** - to provide information on variations in the volume of e-mail messages sent or received by particular e-mail users over time.

The second type of computational technique used, feature extraction techniques, were considered in order to provide the user the ability to specifically locate unusual or abnormal changes in a selected individual's e-mail traffic behaviour. The feature extraction techniques used were:

- **Decision Tree Classification** - to locate unusual changes in interaction behaviour between a suspect e-mail account and its associates, by displaying two decision trees showing when changes in incoming or outgoing interactions occurred.
- **Hierarchical Fuzzy Inference** - to obtain information about the abnormal changes in e-mail traffic behaviour exhibited by a suspect e-mail account, by providing abnormality ratings describing the changes in behaviour observed for each of the suspect's communication links.

Each of the computational techniques selected for the research were chosen specifically to present useful information that can be quickly interpreted and understood by the user/analyst. Visualisation techniques were chosen for their ability to quickly convey information about particular aspects of e-mail traffic behaviour through visual images. With feature extraction techniques, these were chosen for their ability to quickly convey information about the location of data records associated with particular features hidden in e-mail traffic data.

While each of the computational techniques selected can be considered separately, it is only when the techniques are utilised together that the user can obtain the full benefit of computational intelligence. Each of the computational techniques used have their advantages and limitations in the type of information that can be presented to the user. However, limitations in the information presented by one technique can be overcome by the utilising the advantages of another technique. Hence, through the computational intelligence approach the user is able to benefit from utilising a combination of computational techniques, to examine particular aspects of an individual's e-mail traffic behaviour.

6.3.2 Development Of A Personality Trait Based E-mail Traffic Simulation Tool

A simulation tool was developed for the research to enable the generation of simulated e-mail traffic data. The purpose of using the simulation tool was to allow for different behaviour profiles to be assigned to each simulated e-mail user and to be able to observe the effects of those behaviour profiles on the e-mail user's traffic behaviour. This simulation tool provided a method of evaluating computational intelligence by allowing for different setups to be used when investigating the traffic behaviour of particular e-mail users.

To generate the simulated e-mail traffic data, a conceptual e-mail system simulation model was developed as part of the simulation tool to model the traffic interactions occurring between different e-mail users. The conceptual e-mail system simulation model developed comprised of two types of entities: e-mail clients and behaviour models. The e-mail clients were used to represent the e-mail accounts of e-mail users, while the behaviour models were used to represent the behavioural attributes of an e-mail user. The purpose of modelling the e-mail system as e-mail client and behaviour model entities was to focus on the point-to-point interactions that occur between e-mail users, rather than the interaction that occurs between a user's e-mail account and the e-mail server. This modelling approach enabled different behavioural profiles to be assigned to each simulated e-mail user, allowing for the examination of how different behavioural profiles affect the interaction behaviour of particular e-mail users.

The unique aspect of the conceptual e-mail system simulation model developed was the use of personality trait dimensions to represent the behavioural profiles of the simulated e-mail users. The personality trait dimensions presented a simple method of describing each e-mail user's behavioural attributes and allowed for the prediction of how particular e-mail clients are expected to behave through e-mail communications. This aided in evaluating the use of computational intelligence, since the observed e-mail traffic behaviour could be linked back to the personality trait values assigned to each simulated e-mail user.

6.3.3 Development Of A Novel System Architecture For E-mail Traffic Analysis

An e-mail traffic analysis system was developed as a conceptual system to integrate the visualisation and feature extraction techniques used for the research.

The purpose of developing the system was to help demonstrate the use of computational intelligence and to facilitate the investigation of e-mail traffic behaviour. The architecture used for the e-mail traffic analysis system consisted of two important design elements. The first design element was a modular architecture, which allows for new features to be easily added to the system. This was considered important since it can allow for additional computational techniques to be added to the system, providing more perspectives on e-mail traffic behaviour. The second design element was providing the user a set of parameters for controlling what is analysed by the e-mail traffic analysis system. This was considered important since it enabled the system to focus the analysis on particular parts of the e-mail traffic data relevant to the user's investigation task and also places the user in charge of how the data is analysed. These combination of design elements were found to enable the system to focus each of the computational techniques on a particular section of the e-mail traffic data, thereby providing different perspectives on the e-mail traffic behaviour observed.

To evaluate the e-mail traffic analysis system and demonstrate computational intelligence, the system was evaluated using two sets of e-mail traffic data: simulated e-mail traffic data and data from the Enron e-mail corpus. When evaluating the e-mail traffic analysis system with simulated e-mail traffic data, decision tree classification was used with social network visualisation and time-series visualisation to aid in locating unusual changes in a simulated individual's e-mail traffic behaviour. It was found during the simulation data evaluation that decision tree classification provided a useful overview of changes in incoming and outgoing interaction behaviour between a selected e-mail account and their associates. The changes in interaction behaviour detected by decision tree classification could be verified by using both social network and time-series visualisation to provide a better understanding of the changes in behaviour observed.

When evaluating the e-mail traffic analysis system with e-mail traffic data from the Enron e-mail corpus, hierarchical fuzzy inference was used with social network visualisation and time-series visualisation to aid in detecting abnormal changes in a suspect's e-mail traffic behaviour patterns. It was found that hierarchical fuzzy inference was able to provide a useful overview of a suspect's changes in behaviour patterns by providing an abnormality rating for each of the suspect's communication links. This provided a useful summary of the degree of change observed in the communication between a suspect's e-mail account and each of the suspect's associates. To verify the changes in behaviour found by hierarchical fuzzy inference, social network and time-series visualisation were

used to provide additional perspectives on the changes in behaviour found. The effect of using a combination of computational techniques when investigating the Enron e-mail corpus, helped to provide a much better insight into the selected suspect's e-mail traffic behaviour than using the techniques individually by themselves.

6.4 Further Studies

While the research presented in this thesis has shown how computational intelligence can be used to examine the e-mail traffic behaviour of a known suspect e-mail account, there are several areas for further improvements. These areas of improvement relate to the conceptual e-mail system simulation model, evaluation of the decision tree classification technique and evaluation of the hierarchical fuzzy inference technique. The following are suggestions for improvements in these areas and possibilities for extending the research work through further studies.

6.4.1 Extending the Conceptual E-mail System Model

The conceptual e-mail system simulation model described in Section 4.3.1 provided a useful tool for simulating e-mail traffic behaviour, but was limited in only modelling the sending and replying behaviour of e-mail users. The simulation model may be extended to include other types of e-mailing behaviour that are performed by e-mail users. There are a range additional e-mailing behaviours that could be considered for modelling, such as: sending of attachments, forwarding of e-mail messages, sending of messages to multiple recipients, social group behaviour, gender differences in e-mailing behaviour [120, 121] and particular e-mailing habits (e.g. sending messages during particular hours of the day or sending messages during particular days of the week). The addition of these types of behaviour to the simulation model may help in better understanding people's e-mailing behaviour and also allow for experimentation with different types of simulation setups to observe how an individual's e-mail traffic behaviour is affected.

In addition to modelling other types of e-mailing behaviour, consideration may also be given to finding or conducting more empirical studies to support the linkage between personality trait dimensions and e-mail traffic behaviour. When the conceptual e-mail system simulation model was developed [94, 115], there were

few empirical studies conducted that examined the effects of personality traits on e-mail communication behaviour. Due to this, some intuitive assumptions were made on how personality trait dimensions affect e-mail behaviour. It is suggested that further extension of the conceptual e-mail system model could consider finding new empirical studies or consider conducting an empirical study to examine the linkage between personality trait dimensions and e-mail traffic behaviour. This may help in providing better modelling of different individuals' personalities and their simulated e-mail traffic behaviour.

6.4.2 Further Evaluation Of Decision Tree Classification

The decision tree classification technique evaluated in Sections 5.2.1 and 5.2.2 was shown to be able to locate unusual changes in interaction behaviour from the e-mail traffic of a small-sized simulated e-mail system. While this showed that decision tree classification can be used with a small e-mail system of 9 or 10 e-mail clients, it is still yet to be determined whether decision tree classification is still practical for analysing much larger e-mail systems (e.g. with more than 20 e-mail users). It is also still yet to be determined whether decision tree classification will work with genuine e-mail traffic data, such as the Enron e-mail corpus. Further studies involving the use of decision tree classification for analysing e-mail traffic data should consider evaluating decision tree classification with data from much larger e-mail systems and also with genuine e-mail traffic data.

6.4.3 Investigation Of Other Hierarchical Fuzzy Inference Architectures

The hierarchical architecture evaluated in Section 5.3 only demonstrated one of the possible architectures for fusing the e-mail traffic behaviour change measurements. Although other architectures were developed and evaluated in [122], it still remains to be fully understood what type of effect the grouping of particular input behaviour change measurements has on the final output of the hierarchical fuzzy inference system. The difficulty with determining the desired grouping of particular input behaviour change measurements is that the relationship between input variables are currently not well understood. This means that there is further study required to determine the desired architecture for grouping the input variables for the hierarchical fuzzy inference system.

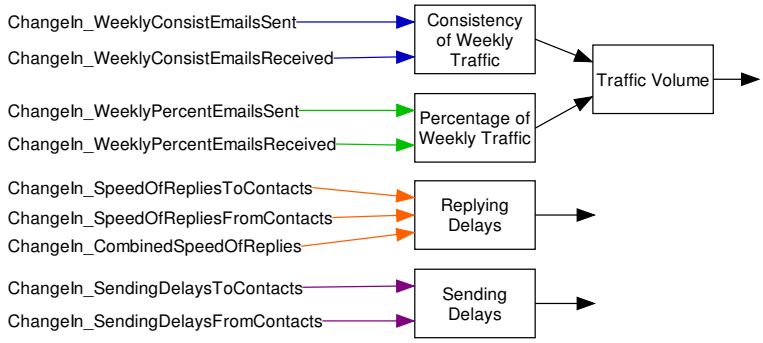


Figure 6.1: A suggestion for a trimmed down version of the hierarchical architecture.

There are three suggestions for extending the work on hierarchical fuzzy inference for finding abnormal changes in e-mail traffic behaviour. The first suggestion is to consider further study of the behaviour change measurements used as the inputs for the hierarchical fuzzy inference system. This could involve studying the relationships between particular variables by using statistical analysis techniques such as principal components analysis [68, 81, 123, 124] or maximum covariance analysis [51] as a starting point for understanding the relationships between variables. A statistical analysis method like principal components analysis can be used to analyse multidimensional data in order to reveal previously unsuspected relationships between variables [81]. Maximum covariance has already been used by [3] to explore the importance of particular e-mail traffic behaviour measurement variables for detecting e-mail based computer viruses. When there is a better understanding of the relationship between particular e-mail traffic behaviour measurement variables, this may then aid in better understanding the type of input groupings to use for the hierarchical fuzzy inference architecture.

A second suggestion is that one may consider creating a set of benchmarked training data [89] in which the hierarchical fuzzy inference system or parts of the hierarchical fuzzy inference system is trained on recognising changes in an individual's e-mail traffic behaviour. Through this method, the architecture for the hierarchical fuzzy inference could be automatically defined by the training process, rather than being defined manually by the system developer. Selection of a suitable e-mail dataset for the training data would be quite important, since it will affect the type of architecture that results from training process.

The third suggestion for further study is to approach the problem differently and consider reducing the number of levels in the hierarchy, so that the meaning of the numerical output from the hierarchical fuzzy inference system is more ob-

vious. The architecture used in the Enron case study (Section 5.3) produces a single numerical output that just describes the degree of change in communication link behaviour, but does not describe exactly what changed. A suggestion is to consider a trimmed down version of the hierarchical architecture as shown in Figure 6.1. This suggested architecture instead produces three numerical outputs, each of which describes changes in particular aspects of an individual's communication link behaviour. While this produces more numerical outputs, it still fuses together a number of behaviour change measurements and summarises some of the e-mail traffic behaviour change information. This may help to produce numerical outputs from the hierarchical fuzzy inference system that may provide a clearer meaning to the user or analyst when describing what particular aspects of an individual's e-mail traffic behaviour has changed.

6.4.4 Correlating E-mail Content Analysis And E-mail Traffic Analysis Results

The research presented in this thesis demonstrated the computational intelligence approach by using a set of computational techniques to provide different perspectives on an individual's e-mail traffic behaviour. An extension of the computational intelligence approach would be to consider adding e-mail content analysis techniques to the set of e-mail traffic analysis techniques used, in order to provide additional insight into the topics or discussions occurring between particular individuals. While in some circumstances the use of both e-mail traffic analysis and content analysis techniques may be considered to be quite intrusive into an individual's privacy, it may still help to provide a better understanding about people's e-mailing behaviour. The correlation of changes in e-mail traffic behaviour with the analysis of topics discussed, may enable the prediction about the types of topics that cause people to change their e-mail communication behaviour. Likewise, it may also enable understanding about why a change in e-mail traffic behaviour was observed when investigating the e-mail communications of a suspected individual. A suggested dataset to use for further studies in this area is the Enron e-mail corpus, since it is a well-known dataset and has characteristics that are well documented by various sources. Use of other e-mail datasets may also be desirable, in order to determine similarities or differences in correlation of traffic analysis and content analysis results.

6.5 Conclusion

Overall, the objective of the research work presented in this thesis has been achieved. The thesis has proposed a computational intelligence approach for analysing e-mail traffic data and has shown how it can be employed to obtain useful information about the overall changes in e-mail traffic behaviour patterns of a suspected individual. The computational intelligence approach described in the thesis has been demonstrated by utilising two types of computational techniques, which provide the user different perspectives about an individual's e-mail traffic behaviour. The computational techniques used were visualisation and feature extraction techniques, which provided different methods of analysing and understanding e-mail traffic behaviour. The feature extraction techniques used, decision tree classification and hierarchical fuzzy inference, both demonstrated ways of detecting unusual or abnormal changes in an individual's e-mail traffic behaviour. The visualisation techniques used, social network visualisation and time-series visualisation, were shown to be useful for investigating the details of the changes in behaviour detected by the feature extraction techniques.

During the investigation of the computational intelligence approach, two novel types of software tools were developed to assist with evaluating the computational intelligence approach. The first tool developed was a personality trait based e-mail traffic simulation tool, which was used to experiment with assigning different behaviour profiles to each simulated e-mail user in order to observe their effect on e-mail traffic behaviour. The second tool developed was an e-mail traffic analysis system, which was used to integrate each of the computational techniques used in the research and to help facilitate the investigation of e-mail traffic behaviour.

Prior to this research, there had been no clear demonstration of the usefulness of utilising a limited number of computational or analysis techniques to investigate an individual's e-mail traffic behaviour. This research has clearly demonstrated the usefulness of this approach and has shown the importance of utilising different perspectives on e-mail traffic behaviour.

References

[1] S. J. Stolfo, S. Hershkop, K. Wang, O. Nimeskern, and C. W. Hu, "A behavior-based approach to securing email systems," in *Computer Network Security*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag Berlin, 2003, vol. 2776, p. 80, figure 5.

[2] S. J. Stolfo, S. Hershkop, H. Chia-Wei, L. Wei-Jen, O. Nimeskern, and W. Ke, "Behavior-based modeling and its application to email analysis," *ACM Transactions on Internet Technology*, vol. 6, no. 2, pp. 187–221, 2006, http://doi.acm.org/10.1145/1149121.1149125.

[3] S. Martin, A. Sewani, B. Nelson, K. Chen, and A. D. Joseph, "Analyzing behaviorial features for email classification," in *Second Conference on Email and Anti-Spam (CEAS 2005)*, Stanford University, Palo Alto, CA, 2005.

[4] J. Baumes, M. Goldberg, M. Hayvanovych, M. Magdon-Ismail, W. Wallace, and M. Zaki, "Finding hidden group structure in a stream of communications," in *Intelligence and Security Informatics 2006 (ISI 2006) Conference*, ser. Lecture Notes in Computer Science, S. Mehrotra, D. D. Zeng, H. Chen, B. Thuraisingham, and F.-Y. Wang, Eds., vol. 3975. San Diego, California, USA: Springer-Verlag Berlin, 2006, pp. 201–212.

[5] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*, 3rd ed. Kitware, 2004.

[6] P. Vervest, *Innovation In Electronic Mail : Towards Open Information Networks - Perspectives On Innovation Policy*. Amsterdam: Elsevier Science Publishers B.V., 1987.

[7] J. B. Postel, *RFC 821: Simple Mail Transfer Protocol*. IETF, 1982, viewed 18 June 2007, ftp://ftp.rfc-editor.org/in-notes/rfc821.txt.

[8] D. H. Crocker, *RFC 822: Standard For The Format Of ARPA Internet Text Messages*. IETF, 1982, viewed 18 June 2007, ftp://ftp.rfc-editor.org/in-notes/rfc822.txt.

[9] A. S. Tanenbaum, *Computer Networks*, 4th ed. Upper Saddle River, NJ: Prentice Hall PTR, 2003.

[10] J. Klensin, *RFC 2821: Simple Mail Transfer Protocol*. IETF, 2001, viewed 18 June 2007, ftp://ftp.rfc-editor.org/in-notes/rfc2821.txt.

[11] P. Resnick, *RFC 2822: Internet Message Format*. IETF, 2001, viewed 18 June 2007, ftp://ftp.rfc-editor.org/in-notes/rfc2822.txt.

[12] D. Stoker, "The benefits and curses of electronic mail," *Journal of Librarianship and Information Science*, vol. 31, no. 4, pp. 185–187, 1999.

[13] E. Blanzieri and A. Bryl, *A Survey Of Anti-Spam Techniques*. Department of Information and Communication Technology, University of Trento, 2006, viewed 19 Dec 2007, http://eprints.biblio.unitn.it/archive/00001070/01/056.pdf.

[14] J. Carpinter and R. Hunt, "Tightening the net: A review of current and next generation spam filtering tools," *Computers & Security*, vol. 25, no. 8, pp. 566–578, 2006.

[15] D. Cook, "Catching spam before it arrives," Honours, University of Tasmania, 2005.

[16] D. Cook, J. Hartnett, K. Manderson, and J. Scanlan, "Catching spam before it arrives: Domain specific dynamic blacklists," in *Fourth Australasian Information Security Workshop (AISW-NetSec 2006)*, ser. Conferences in Research and Practice in Information Technology (CRPIT), vol. 54. Hobart, Tasmania, Australia: Australian Computer Society, 2006, pp. 193 – 202.

[17] T. Jiang, W. Kim, K. Lhee, and M. Hong, "E-mail worm detection using the analysis of behavior," in *Distributed Computing and Internet Technology, Proceedings*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag Berlin, 2005, vol. 3816, pp. 348–356.

[18] S. Hershkop, "Behavior-based email analysis with applications to spam detection," Doctoral Thesis, Columbia University, 2006.

[19] H. Berghel, J. Carpinter, and J.-Y. Jo, "Phish phactors: Offensive and defensive strategies," in *Advances in Computers*, ser. Advances in Computers. San Diego, CA: Elsevier, 2007, vol. 70, pp. 223–268.

[20] D. Schuff, O. Turetken, J. D'Arcy, and D. Croson, "Managing e-mail overload: Solutions and future challenges," *Computer*, vol. 40, no. 2, pp. 31–36, February 2007 2007.

[21] D. Sow, M. Ebling, R. P. Lehmann, J. Davis, and L. Bergman, "Scout contextually organizes user tasks," in *IEEE International Conference on e-Business Engineering*. IEEE, 2005, pp. 94–101.

[22] G. M. Mohay, *Computer and Intrusion Forensics*, ser. Artech House Computer Security Series. Boston: Artech House, 2003.

[23] J. J. McLean, "Homicide and child pornography," in *Handbook of Computer Crime Investigation : Forensic Tools and Technology*, E. Casey, Ed. San Diego, CA: Academic Press, 2002, pp. 361 – 373.

[24] D. L. Shinder and E. Tittel, "Scene of the cybercrime: Computer forensics handbook." Rockland, MA: Syngress Media, 2002, p. 325.

[25] A. Goldsmith, M. Israel, and K. Daly, *Crime and Justice: An Australian Textbook in Criminology*, 2nd ed. Pyrmont, NSW, Australia: Lawbook Co, 2003.

[26] E. Carrabine, P. Iganski, M. Lee, K. Plummer, and N. South, *Criminology: A Sociological Introduction*, 1st ed. London: Routledge, 2004.

[27] E. Casey, *Handbook of Computer Crime Investigation : Forensic Tools and Technology*. San Diego, CA: Academic Press, 2002.

[28] R. McKemmish, "What is forensic computing?" Australian Institute of Criminology, Tech. Rep. 118, June 1999.

[29] B. Etter, *The forensic challenges of e-crime*. Australasian Centre for Policing Research, 2001, viewed 21 Mar 2007, http://www.acpr.gov.au/publications2.asp?Report_ID=115.

[30] B. Nelson, A. Phillips, F. Enfinger, and C. Steuart, *Guide to Computer Forensics and Investigations*. Australia: Thomson/Course Technology, 2004.

[31] A. Rees, *Technology Environment Scan*. Payneham, South Australia: Australasian Centre for Policing Research, 2000, viewed 14 December 2007, http://www.acpr.gov.au/pdf/ACPR133_1.pdf.

[32] W. W. Cohen, "Learning rules that classify e-mail," in *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, California, 1996.

[33] J. D. Brutlag and C. Meek, "Challenges of the email domain for text classification," in *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 103 – 110.

[34] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Machine Learning: ECML 2004: 15th European Conference on Machine Learning*, ser. Lecture Notes in Computer Science, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds., vol. 3201. Pisa, Italy: Springer-Verlag, 2004, pp. 217–226.

[35] J. Mena, *Investigative Data Mining for Security and Criminal Detection*, 1st ed. Butterworth-Heinemann, 2003.

[36] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications : Text Retrieval, Extraction, and Categorization*. Amsterdam ; Philadelphia, PA: John Benjamins Publishers, 2002.

[37] I. B. Crabtree and S. J. Soltysiak, "Identifying and tracking changing interests," *International Journal on Digital Libraries*, vol. 2, no. 1, pp. 38–53, 1998.

[38] P. S. Keila and D. B. Skillicorn, "Structure in the enron email dataset," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 183–199, 2005.

[39] J.-Y. Yeh and A. Harnly, "Email thread reassembly using similarity matching," in *Third Conference on Email and Anti-Spam (CEAS 2006)*, Mountain View, CA, 2006, pp. 64–71.

[40] O. de Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *SIGMOD Record*, vol. 30, no. 4, pp. 55–64, 2001.

[41] S. J. Stolfo, S. Hershkop, K. Wang, O. Nimeskern, and C. W. Hu, "Behavior profiling of email," in *Intelligence and Security Informatics, Proceedings*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag Berlin, 2003, vol. 2665, pp. 74–90.

[42] ——, "A behavior-based approach to securing email systems," in *Computer Network Security*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag Berlin, 2003, vol. 2776, pp. 57–81.

[43] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review E*, vol. 68, no. 6, 2003.

[44] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, "E-mail as spectroscopy: Automated discovery of community structure within organizations," *Information Society*, vol. 21, no. 2, pp. 133–141, 2005.

[45] M. A. Caloyannides, *Computer Forensics and Privacy*, ser. Artech House computer security series. Boston, MA: Artech House, 2001.

[46] B. Etter, T. Rohl, A. Ross, P. Wilkins, M. Sim, J. Geurts, K. Webster, M. Wieszyk, I. McCartney, S. Jiggins, and A. Rees, *The Virtual Horizon : Meeting The Law Enforcement Challenges : Developing an Australasian Law Enforcement Strategy for Dealing with Electronic Crime*. Payneham, South Australia: Australasian Centre for Policing Research, 2000, viewed 14 December 2007, http://www.acpr.gov.au/pdf/ACPR134_1.pdf.

[47] S. M. Cherry, "Remailers elude e-mail surveillance," *IEEE Spectrum*, vol. 38, no. 11, pp. 69–69, November 2001.

[48] B. Schneier, *E-mail security : how to keep your electronic messages private*. New York: Wiley, 1995.

[49] B. B. Lahey, *Psychology: An Introduction*, 9th ed. New York: McGraw-Hill, 2007.

[50] I. M. Chakravarti, R. G. Laha, and J. Roy, *Handbook of Methods of Applied Statistics*. John Wiley and Sons., 1967.

[51] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press, 2004.

[52] J. Scott, *Social Network Analysis: A Handbook*, 2nd ed. London: SAGE Publications, 2000.

[53] D. J. Watts, *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton, 2003.

[54] M. E. J. Newman, "The structure and function of complex networks," *Siam Review*, vol. 45, no. 2, pp. 167–256, 2003.

[55] M. E. J. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 66, no. 3, pp. 35 101–1–4, 2002.

[56] J. P. Eckmann, E. Moses, and D. Sergi, "Entropy of dialogues creates coherent structures in e-mail traffic," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 40, pp. 14 333–14 337, 2004.

[57] G. Caldarelli, F. Coccetti, and P. De Los Rios, "Preferential exchange: strengthening connections in complex networks," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 70, no. 2, pp. 27 102–1–4, 2004.

[58] A. Chapanond, M. S. Krishnamoorthy, and B. Yener, "Graph theoretic and spectral analysis of enron email data," in *Workshop on Link Analysis, Counterterrorism and Security at SIAM International Conference on Data Mining*, Newport Beach, CA, USA, 2005, pp. 15–22.

[59] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, "Scan statistics on enron graphs," in *Workshop on Link Analysis, Counterterrorism and Security at SIAM International Conference on Data Mining*, Newport Beach, CA, USA, 2005, pp. 23–32.

[60] D. A. Keim, F. Mansmann, C. Panse, J. Schneidewind, and M. Sips, "Mail explorer - spatial and temporal exploration of electronic mail," in *EuroVis05: Joint Eurographics - IEEE VGTC Symposium on Visualization*, K. W. Brodlie, D. J. Duke, and K. I. Joy, Eds. Leeds, United Kingdom: Eurographics Association, 2005, pp. 247–254.

[61] A. P. Engelbrecht, *Computational Intelligence : An Introduction*. Chichester, England: John Wiley & Sons, 2002.

[62] T. M. Khoshgoftaar, *Software Engineering With Computational Intelligence*. Norwell, Massachusetts: Kluwer Academic Publishers, 2003.

[63] F. J. Gravetter and L. B. Wallnau, *Statistics for the Behavioral Sciences*, 6th ed. Australia: Thomson/Wadsworth, 2004.

[64] E. R. Tufte, *The visual display of quantitative information*, 1st ed. Cheshire, Connecticut: Graphics Press, 1983.

[65] M. A. Boden, *Artificial Intelligence and Natural Man*. New York: Basic Books, 1977.

[66] M. Negnevitsky, *Artificial Intelligence: A Guide to Intelligent Systems*, 2nd ed. Essex: Addison Wesley, 2004.

[67] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*. New York: Blaisdell, 1969.

[68] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson Education, 2006.

[69] L. C. Freeman, "Visualizing social networks," *Journal of Social Structure*, vol. 1, no. 1, 2000, http://www.cmu.edu/joss/content/articles/volume1/Freeman.html.

[70] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, 2004.

[71] J. Xu and H. C. Chen, "Untangling criminal networks: A case study," in *Intelligence and Security Informatics, Proceedings*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag Berlin, 2003, vol. 2665, pp. 232–248.

[72] W. S. Torgerson, "Multidimensional scaling: Theory and method," *Psychometrika*, vol. 17, pp. 401–419, 1952.

[73] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Information Processing Letters*, vol. 31, no. 1, pp. 7–15, 1989.

[74] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *SOFTWARE–PRACTICE AND EXPERIENCE*, vol. 21, no. 11, pp. 1129–1164, 1991.

[75] E. Adar, "Guess: A language and interface for graph exploration," in *CHI 2006: The SIGCHI Conference on Human Factors in Computing Systems*, R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson, Eds. Montréal, Québec, Canada: ACM Press, 2006, pp. 791–800.

[76] C. Chatfield, *The Analysis of Time Series: An Introduction*, 5th ed., ser. Texts in statistical science. London: Chapman and Hall, 1996.

[77] J. Han and M. Kamber, *Data mining: concepts and techniques*, ser. Morgan Kaufmann series in data management systems. San Francisco, CA: Morgan Kaufmann, 2001.

[78] A. Aris, A. Khella, P. Buono, B. Shneiderman, and C. Plaisant, *TimeSearcher 2*. Human-Computer Interaction Laboratory, Computer Science Department, University of Maryland, 2005, viewed 8th February 2007, http://www.cs.umd.edu/hcil/timesearcher/.

[79] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[80] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.

[81] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2006.

[82] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[83] ——, *RuleQuest Research*, 2006, viewed 15 July 2007, http://www.rulequest.com.

[84] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, Calif.: Morgan Kaufmann, 2005.

[85] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[86] E. Turban and J. E. Aronson, *Decision Support Systems and Intelligent Systems*, 6th ed. Upper Saddle River, NJ: Prentice Hall, 2001.

[87] C. Bagnoli and H. C. Smith, "The theory of fuzzy logic and its application to real estate valuation," *Journal of Real Estate Research*, vol. 16, no. 2, 1998.

[88] G. V. S. Raju, J. Zhou, and R. A. Kisner, "Hierarchical fuzzy control," *International Journal of Control*, vol. 54, no. 5, pp. 1201–1216, 1991.

[89] V. Torra, "A review of the construction of hierarchical fuzzy systems," *International Journal of Intelligent Systems*, vol. 17, no. 5, pp. 531–543, 2002.

[90] R. Bace and P. Mell, *NIST Special Publication 800-31: Intrusion Detection Systems*. National Institute of Standards and Technology (NIST), 2001, viewed 26 February 2004, http://csrc.nist.gov/publications/nistpubs/800-31/sp800-31.pdf.

[91] J. Gomez, F. Gonzalez, and D. Dasgupta, "An immuno-fuzzy approach to anomaly detection," in *12th IEEE International Conference on Fuzzy Systems*, vol. 2, 2003, pp. 1219–1224.

[92] S. B. Cho, "Incorporating soft computing techniques into a probabilistic intrusion detection system," *IEEE Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, vol. 32, no. 2, pp. 154–160, 2002.

[93] N. J. Salkind, *Statistics for people who (think they) hate statistics*, 2nd ed. Thousand Oaks, CA: Sage Publications, 2004.

[94] M. J.-H. Lim, M. Negnevitsky, and J. Hartnett, "Personality trait based simulation model of the e-mail system," *International Journal of Network Security*, vol. 3, no. 2, pp. 164–182, 2006.

[95] M. J. Lim, M. Negnevitsky, and J. Hartnett, "Tracking and monitoring e-mail traffic activities of criminal and terrorist organisations using visualisation tools," *Journal of Information Warfare*, vol. 5, no. 2, pp. 46 – 60, 2006.

[96] *Python Programming Language*, 2005, viewed 24 February 2005, http://www.python.org/.

[97] V. Vaswani, *MySQL: The Complete Reference*. McGraw-Hill/Osborne, 2004.

[98] *MySQL*, 2007, viewed 8 October 2007, http://www.mysql.org.

[99] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, Calif.: Morgan Kaufmann, 2000.

[100] Mathworks, *MATLAB Fuzzy Toolbox*, 2006, viewed 12 October 2006, http://www.mathworks.com.

[101] J. Banks, J. Carson, B. L. Nelson, and D. Nicol, *Discrete-Event System Simulation*, 4th ed. Upper Saddle River, NJ: Prentice Hall, 2005.

[102] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*, 3rd ed. McGraw-Hill, 2000.

[103] W. T. Norman, "Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings," *Journal of Abnormal and Social Psychology*, vol. 66, no. 6, pp. 574 – 583, 1963.

[104] I. Ajzen, *Attitudes, personality, and behavior*, ser. Mapping social psychology. Stony Stratford, Milton Keynes: Open University Press, 1988.

[105] S. C. Cloninger, *Theories of personality: understanding persons*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall, 1996.

[106] A. Lantz, "Does the use of e-mail change over time?" *International Journal of Human-Computer Interaction*, vol. 15, no. 3, pp. 419–431, 2003.

[107] P. C. Fusaro and R. M. Miller, *What Went Wrong at Enron: Everyone's Guide to the Largest Bankruptcy in U.S. History*. Hoboken, NJ: John Wiley & Sons, 2002.

[108] J. Diesner, T. L. Frantz, and K. M. Carley, "Communication networks from the enron email corpus "it's always about the people. enron is no different"," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 201 – 228, 2005.

[109] W. W. Cohen, *The CMU Enron Email Dataset*, 2004, viewed 12 October 2006, http://www.cs.cmu.edu/ enron/.

[110] A. Fiore and J. Heer, *UC Berkeley Enron Email Analysis*, 2005, viewed 12 October 2006, http://bailando.sims.berkeley.edu/enron_email.html.

[111] J. Shetty and J. Adibi, *The ISI Enron Database Schema and Brief Statistical Report*, 2005, viewed 12 October 2006, http://www.isi.edu/ adibi/Enron/Enron.htm.

[112] M. J.-H. Lim, M. Negnevitsky, and J. Hartnett, "A fuzzy approach for detecting anomalous behaviour in e-mail traffic," in *4th Australian Digital Forensics Conference*, C. Valli and A. Woodward, Eds. Perth, Western Australia: School of Computer and Information Science, Edith Cowan University, 2006, pp. 36 – 49.

[113] *The Houston Chronicle*, 2006, viewed 5th February 2007, http://www.chron.com/news/specials/enron/.

[114] J. Shetty and J. Aldibi, *The Enron Dataset Database Schema and Brief Statistical Report*, 2006, viewed 7 Aug 2006, http://www.isi.edu/ adibi/Enron/Enron_Dataset_Report.pdf.

[115] M. J. Lim, M. Negnevitsky, and J. Hartnett, "Tracking and monitoring e-mail traffic activities of criminal and terrorist organisations using visualisation tools," in *6th Australian Information Warfare & Security Conference*, G. Pye and M. Warren, Eds. Geelong, Victoria, Australia: School of Information Systems, Deakin University, 2005, pp. 112 – 124.

[116] L. Fox, *Enron: The Rise and Fall*. Hoboken, NJ: John Wiley & Sons, 2003.

[117] *Enron Corp. - Timeline of the rise and fall of Enron Corp*. The Houston Chronicle, 2006, viewed 5th February 2007, http://www.chron.com/news/specials/enron/timeline.html.

[118] J. E. Dickerson, J. Juslin, O. Koukousoula, and J. A. Dickerson, "Fuzzy intrusion detection," *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, vol. 3, pp. 1506–1510, 2001.

[119] J. C. Bezdek, *Pattern Recognition with Fuzzy objective Function Algorithms*. New York: Plenum Press, 1981.

[120] B. Boneva, R. Kraut, and D. Frohlich, "Using e-mail for personal relationships: The difference gender makes," *American Behavioral Scientist*, vol. 45, no. 3, pp. 530–549, 2001.

[121] B. Boneva and R. Kraut, "Email, gender, and personal relationships," in *The Internet in everyday life*, B. Wellman and C. A. Haythornthwaite, Eds. Malden, MA: Blackwell Publishing, 2002, pp. 372–403.

[122] M. J.-H. Lim, M. Negnevitsky, and J. Hartnett, "Detecting abnormal changes in e-mail traffic using hierarchical fuzzy systems," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2007)*. Imperial College, London, UK: IEEE, 2007, pp. 1309–1314.

[123] K. Cios, W. Pedrycz, and R. Swiniarski, *Data Mining Methods for Knowledge Discovery*. Norwell, MA: Kluwer Academic, 1998.

[124] D. J. Hand, P. Smyth, and H. Mannila, *Principles of Data Mining*, ser. Adaptive computation and machine learning. Cambridge, Mass.: MIT Press, 2001.

[125] M. Tolson and K. Feser, *Jeff Skilling's spectacular career*. Chron.com, 2004, viewed 11 January 2008, http://www.chron.com/disp/story.mpl/jenny/2404018.html.

[126] P. S. Keila and D. B. Skillicorn, "Detecting unusual email communication," in *2005 Conference of the Centre For Advanced Studies on Collaborative Research*, J. R. Cordy, A. W. Kark, and D. A. Stewart, Eds. Toronto, Ontario, Canada: ACM Press, 2005, pp. 117 – 125.

[127] *Findlaw Legal News: Special Coverage: Enron*. Findlaw.com, 2007, viewed 16 April 2007, http://news.findlaw.com/legalnews/lit/enron/.

[128] *Indictment (US vs. Richard Causey)*. Findlaw.com, 2004, viewed 16 April 2007, http://news.findlaw.com/hdocs/docs/enron/uscausey12104ind.pdf.

[129] *Indictment (US vs. Fastow)*. Findlaw.com, 2002, viewed 22 Jan 2008, http://news.findlaw.com/hdocs/docs/enron/usfastow103102ind.pdf.

Appendix A

E-mail Traffic Database Schema

The e-mail traffic database of the e-mail traffic analysis system contains a number of tables that are used to store e-mail traffic information and also behaviour profiles used for anomaly detection. The database has been designed to enable the e-mail traffic data from a number of different e-mail systems to be entered into the same database. The schema for the e-mail traffic database follows the layout shown in Figure A.1. A description of each table and their associated fields is given in the following sections.

NOTE: The date/time information used in the database is stored as double floating point numbers, to enable the real-world date/time to be computed relative to a reference date (e.g. 1/1/2001 4:55:00). Alternatively, the double floating point numbers can also used to represent the time since the beginning of the simulation for a simulated e-mail system (e.g. 201 days, 51 minutes, and 44 seconds). Each unit used for the double floating point date/time representation is equal to 1 minute (e.g. 101.067 time units is equal to 101 minutes and 4 seconds).

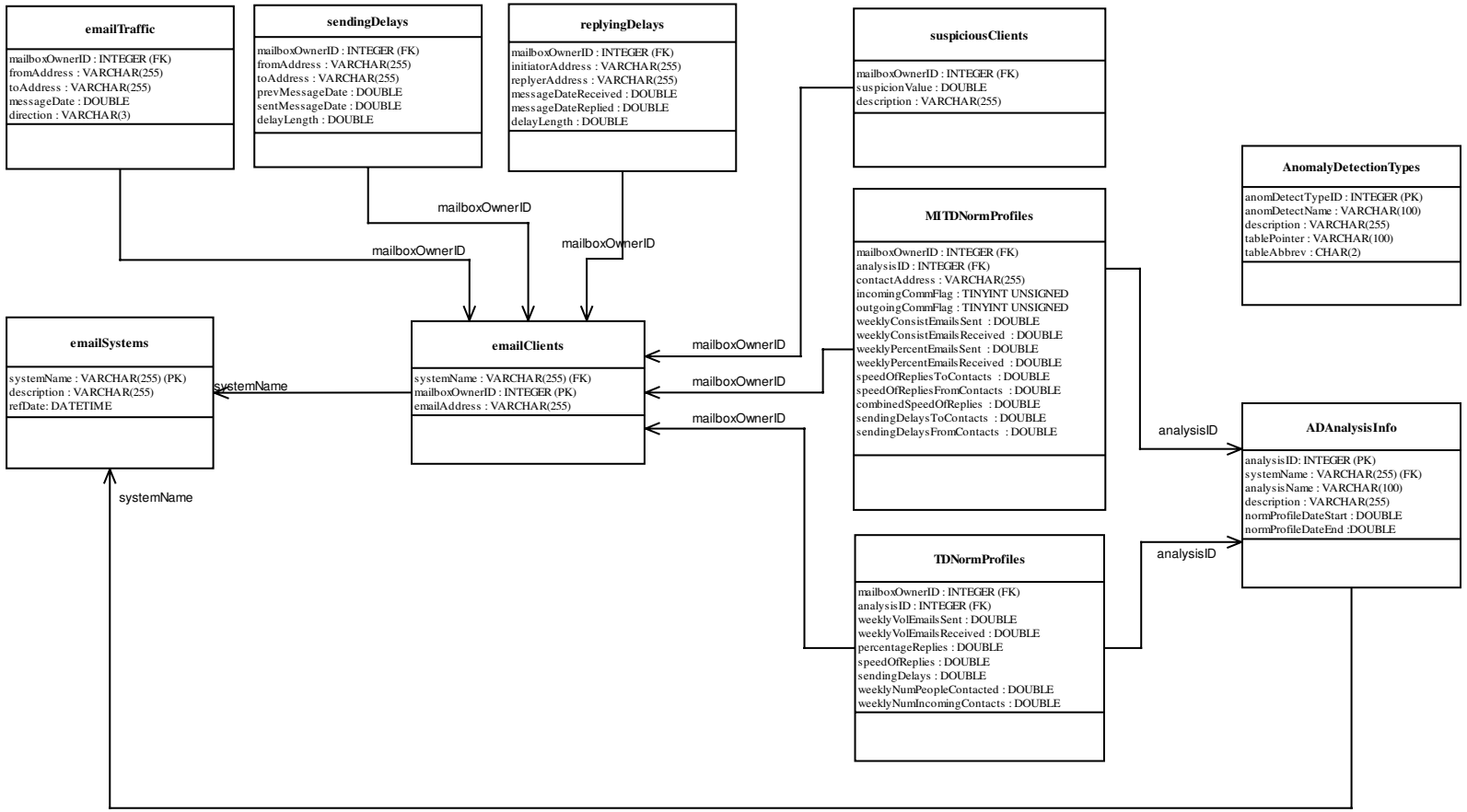


Figure A.1: Diagram of the schema layout for the e-mail traffic database.

A.1 'emailSystems' Table

This table registers information about each of the e-mail systems that have their traffic data stored in the e-mail traffic database. The definition of each field for this table are as follows:

- **systemName** - The name of the e-mail system.
- **description** - A description about the e-mail system.
- **refDate** - A reference date and time that is used to compute the real-world date and time (e.g. 1/1/2001 4:55:00). If this field contains NULL, then it represents that there will be no conversion to a real-world date and time. For example: in the case of data from a simulated e-mail system, there will be no real-world date/time reference since the data is artificially created from a simulation model.

A.2 'emailClients' Table

This table stores records on each of the e-mail accounts that are used in a particular e-mail system. The definition of the fields for this table are as follows:

- **systemName** - The name of the e-mail system.
- **mailboxOwnerID** - An identifier number that uniquely associates an e-mail account as belonging to particular e-mail system.
- **emailAddress** - The e-mail address used by a particular e-mail account (e.g. *name@yahoo.com.au*).

A.3 'emailTraffic' Table

This table stores traffic information obtained from each e-mail message sent or received by an e-mail account from a particular e-mail system. Each record in this table is considered as a single e-mail message to an individual recipient. The fields for this table are defined as follows:

- **mailboxOwnerID** - An identifier number that uniquely associates an e-mail account as belonging to particular e-mail system.

- **fromAddress** - The sender’s e-mail address.
- **toAddress** - The recipient’s e-mail address.
- **messageDate** - The date/time the e-mail message was sent.
- **direction** - An indication of whether the e-mail message was sent or received by the e-mail account’s owner. The value of ‘in’ indicates the message was received by the account owner and the value of ‘out’ indicates the message was sent by the account owner.

A.4 ‘sendingDelays’ Table

Stores information about the sending delays between e-mail messages sent by an e-mail account to one of its associates. The fields in this table are defined as follows:

- **mailboxOwnerID** - An identifier number that uniquely associates an e-mail account as belonging to particular e-mail system.
- **fromAddress** - The sender’s e-mail address (i.e. the e-mail account owner’s e-mail address).
- **toAddress** - The recipient’s e-mail address.
- **prevMessageDate** - The date/time for the previous e-mail message.
- **sentMessageDate** - The date/time for the current e-mail message.
- **delayLength** - The computed time delay between the two e-mail messages.

A.5 ‘replyingDelays’ Table

Stores information about the replying delays between e-mail messages received by an e-mail account and the reply messages sent back to the original sender. The fields in this table are defined as follows:

- **mailboxOwnerID** - An identifier number that uniquely associates an e-mail account as belonging to particular e-mail system.

- **initiatorAddress** - The original sender's e-mail address.
- **replyerAddress** - The e-mail address of the e-mail account that is replying to the original message.
- **messageDateReceived** - The date/time the original message was received.
- **messageDateReplied** - The date/time the reply message was sent.
- **delayLength** - The computed time delay between the original and reply message.

A.6 'suspiciousClients' Table

The purpose of this table is to store information about suspicion values assigned by the user or analyst to particular e-mail accounts. This table is used by the anomaly detection unit of the e-mail traffic analysis system to determine which e-mail accounts are to be analysed for abnormal changes in communication behaviour. The selection e-mail accounts for analysis can be made based on the suspicion value assigned to particular e-mail accounts. The fields for this table are assigned as follows:

- **mailboxOwnerID** - An identifier number that uniquely associates an e-mail account as belonging to particular e-mail system.
- **suspicionValue** - A suspicion value that is arbitrarily set by the user to determine the degree of 'suspicion' for a particular e-mail account. The suspicion value can be set between the values of 0 to 1.
- **description** - A description about why the suspicion value was assigned to a particular e-mail account (e.g. 'associated with Al Qaeda', 'part of a criminal drug ring').

A.7 'MITDNormProfiles' Table

This table stores behaviour profile information for the anomaly detection unit of the e-mail traffic analysis system. The behaviour profile information stored relates to behaviour measurements computed from an e-mail account's communication links (i.e. each communication tie between an e-mail account and a

particular associate). The information from this table is used in the e-mail traffic analysis system for threshold anomaly detection and anomaly detection using the hierarchical fuzzy inference system. Each of the fields for this table are described as follows:

- **mailboxOwnerID** - An identifier number that uniquely associates an e-mail account as belonging to particular e-mail system.
- **analysisID** - An identifier number that uniquely associates the defined normal behaviour profiling period with a particular e-mail system. Refer to the description of the ‘ADAnalysisInfo’ table.
- **contactAddress** - The e-mail address of an associate of the e-mail account owner.
- **incomingCommFlag** - A boolean flag indicating whether communications from the associate are normally incoming towards the e-mail client.
- **outgoingCommFlag** - A boolean flag indicating whether communications from the e-mail client are normally outgoing towards the contact.
- **weeklyConsistEmailsSent** - Behaviour profile measurement of the normal consistency of e-mails sent each week to a particular associate.
- **weeklyConsistEmailsReceived** - Behaviour profile measurement of the normal consistency of e-mails received each week from a particular associate.
- **weeklyPercentEmailsSent** - Behaviour profile measurement of the normal percentage of e-mails sent to a particular associate each week.
- **weeklyPercentEmailsReceived** - Behaviour profile measurement of the normal percentage of e-mails received from a particular associate each week.
- **speedOfRepliesToContacts** - Behaviour profile measurement of the normal speed of replies given to a particular associate.
- **speedOfRepliesFromContacts** - Behaviour profile measurement of the normal speed of replies received from a particular associate.
- **combinedSpeedOfReplies** - Behaviour profile measurement of the normal speed of replies exchanged between both the e-mail client and a particular associate.

- **sendingDelaysToContacts** - Behaviour profile measurement of the normal speed of time delays between e-mails sent to a particular associate.
- **sendingDelaysFromContacts** - Behaviour profile measurement of the normal speed of time delays between e-mails received from a particular associate.

A.8 'TDNormProfiles' Table (Obsolete)

This is an obsolete table that is no longer used. The table was originally used for initial work on the anomaly detection unit of the e-mail traffic analysis system, whereby the threshold anomaly detection method was used. The original intention of this table was to store simple behaviour profile information based on measurements computed from the general overall traffic behaviour of particular e-mail accounts. This table has been superceded by the 'MITDNormProfiles' table.

A.9 'ADAnalysisInfo' Table

This table enables multiple normal behaviour profiling periods to be used for the same set of e-mail accounts (e.g. one e-mail account can be profiled between 1st Jan 2001 to 1st Mar 2001 and also from 1st Aug 2001 to 1st Dec 2001). This means that e-mail accounts can have more than one normal behaviour profile recorded for different periods of time. Each of the fields used for this table are as follows:

- **analysisID** - An identifier number that uniquely associates the defined normal behaviour profiling period with a particular e-mail system.
- **systemName** - The name of the e-mail system being analysed.
- **analysisName** - The name given to the defined normal behaviour profiling period (e.g. 'EnronJan1999Aug2000').
- **description** - Description of the defined normal behaviour profiling period (e.g. "Period prior to Jeffrey Skilling's appointment as CEO of Enron").
- **normProfileDateStart** - The date/time specifying the start of the normal behaviour profiling period.

- **normProfileDateEnd** - The date/time specifying the end of the normal behaviour profiling period.

A.10 ‘AnomalyDetectionTypes’ Table

This table enables more than one type of anomaly detection technique to be used by the e-mail traffic analysis system. This means that the analysis data for more than one anomaly detection technique can be stored in the database. It is assumed for this table that the anomaly detection technique used will store a normal behaviour profile for selected e-mail accounts. Each of the fields in this table are defined as follows:

- **anomDetectTypeID** - Unique identifier for the anomaly detection technique listed.
- **anomDetectName** - The name for the anomaly detection technique. No spaces allowed in the name, since the .
- **description** - A description about the anomaly detection technique.
- **tablePointer** - The name of the table that will be used to store the normal behaviour profiles belonging to the associated anomaly detection technique.
- **tableAbbrev** - A 2 character abbreviation for the table that will be used to store the normal behaviour profiles belonging to the associated anomaly detection technique.

Appendix B

Class Diagram For The E-mail System Simulation Model

The UML based class diagram displayed in Figure B.2 shows an overview of the of all the Python classes used for implementing the conceptual e-mail system simulation model. The diagram in Figure B.1 provides a legend for the symbols used in the class diagram.

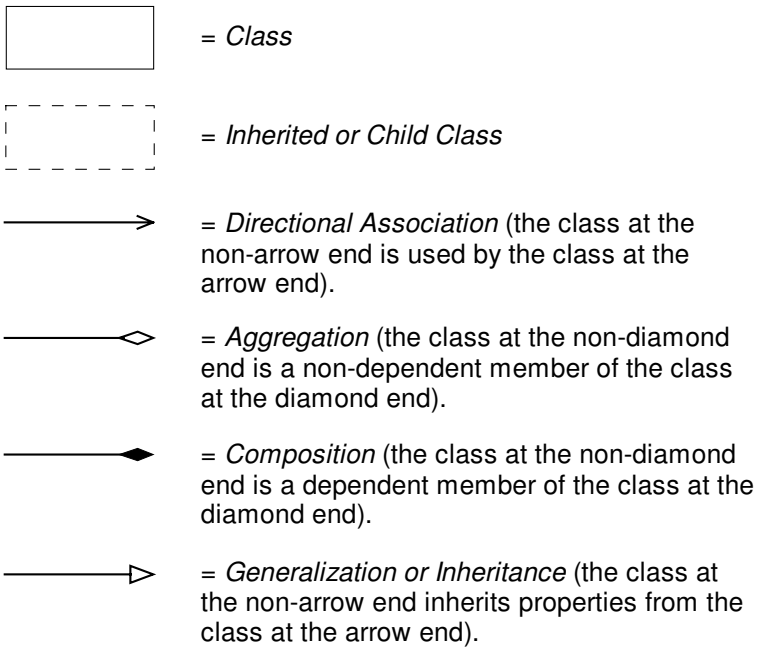


Figure B.1: Legend of the symbols used in the class diagram.

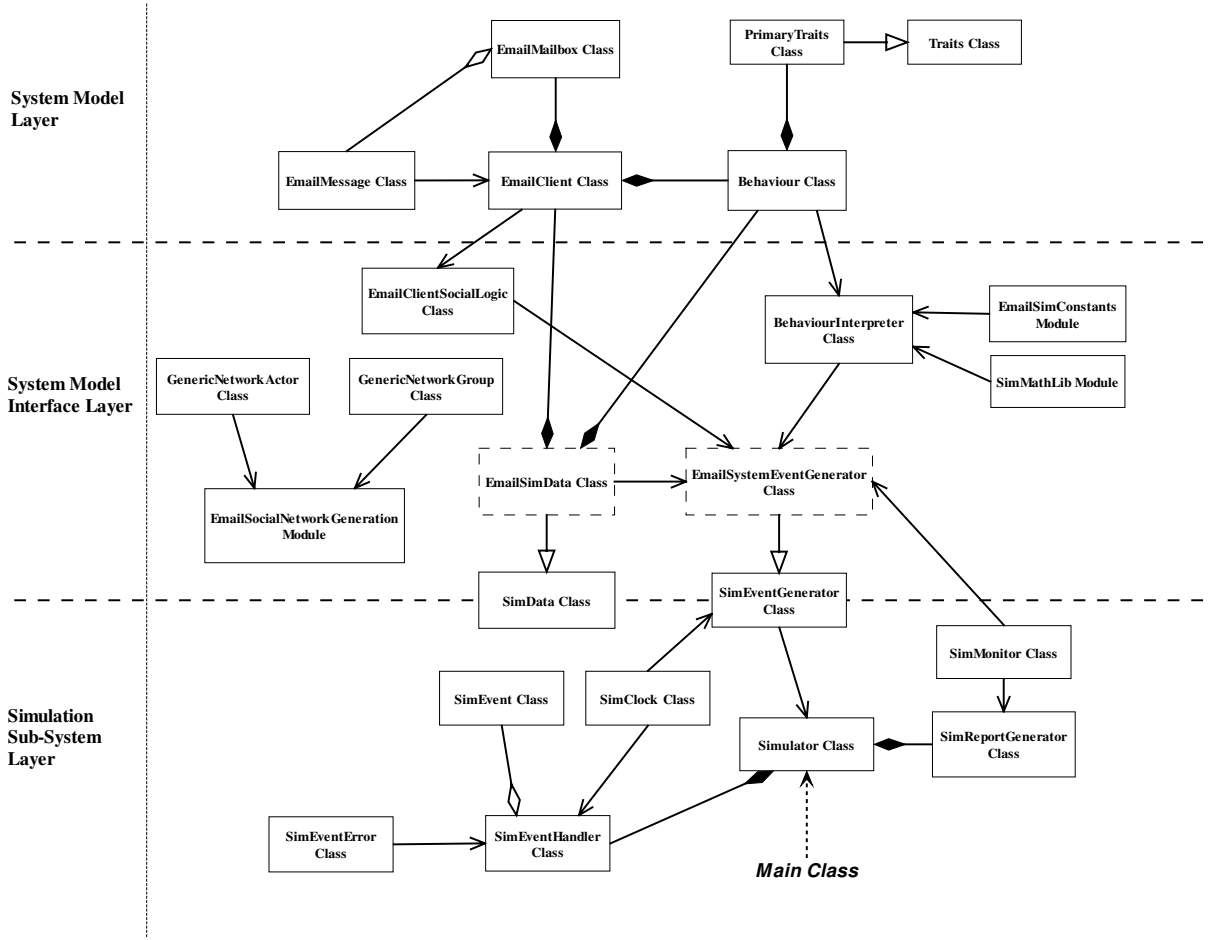


Figure B.2: Class diagram of the implementation of the conceptual e-mail system simulation model.

Appendix C

Enron Events Timeline

The timeline displayed in Figure C.1 shows the events at Enron associated with Jeffrey Skilling's employment and events leading to the eventual collapse of Enron in December 2001. Information provided for this timeline was obtained from [107, 113, 117, 125].

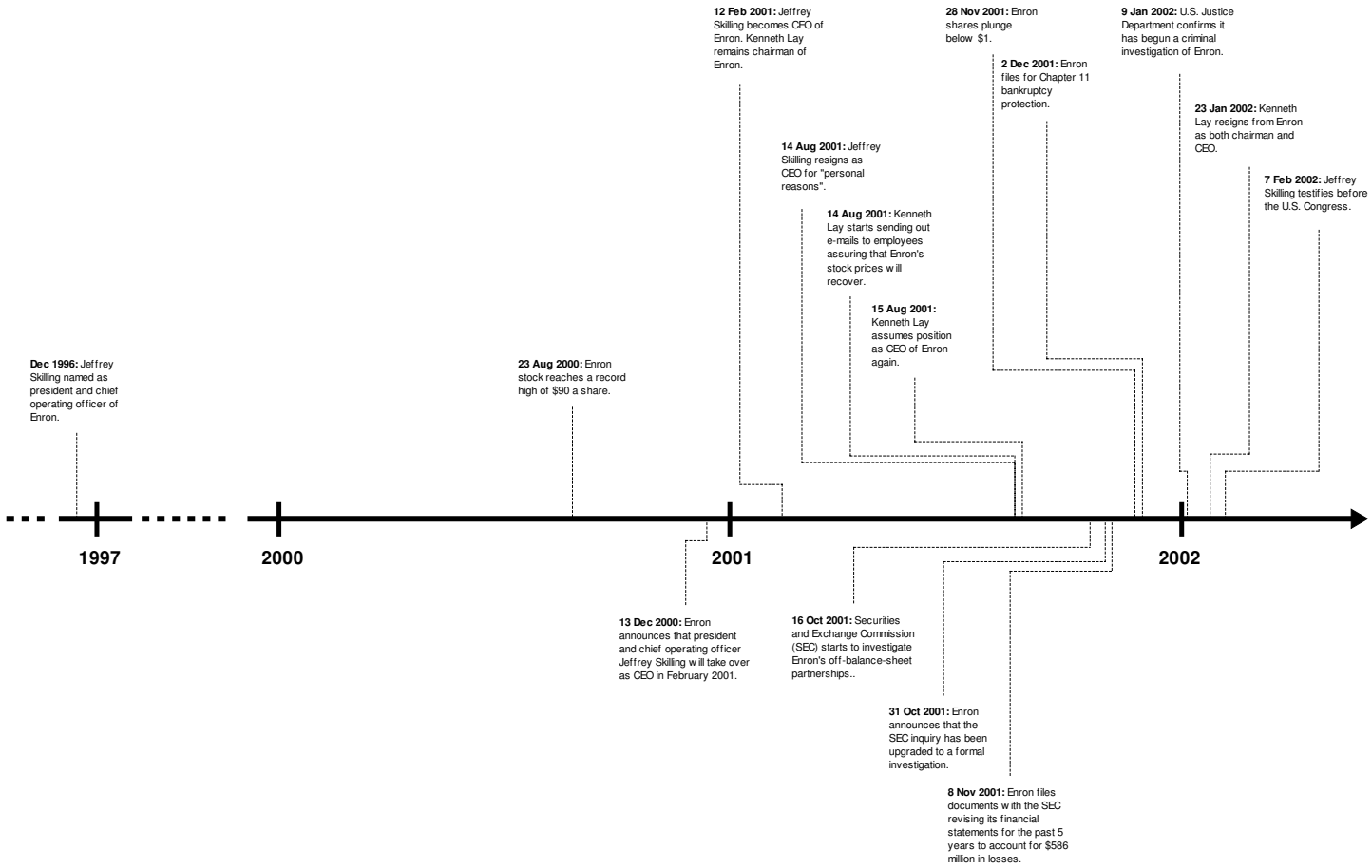


Figure C.1: Events associated with Jeffrey Skilling's employment and the collapse of Enron.

Appendix D

Former Employees of Enron

Tables D.1 and D.2 provides information on some of the former employees of Enron whom had key involvement in the management of the company. The information in these tables identifies the name of the employee, their role in Enron, a brief description about their involvement in Enron, and a listing of possible e-mail addresses that may have been used by these employees. The e-mail addresses listed were obtained from the ISI Enron e-mail dataset [111] by performing a wildcard search for e-mail addresses with sequences of characters resembling parts of the employee's name. The information for the description of each employee were obtained from the following sources: [107, 113, 111, 126, 127, 128, 129].

Table D.1: Information on Kenneth Lay, Jeffrey Skilling, and Andrew Fastow.

Name	Role	Description	Possible E-mail Addresses Used By Former Employee
Kenneth Lay	Founder/Chairman/ CEO of Enron	▪ Kenneth Lay has been with Enron since its beginnings in the mid-1980's. ▪ Lay contributed heavily to political campaigns and political action on both Democrat and Republican sides of politics. ▪ Stepped down as CEO in February 2001, but reassumes CEO position after Jeffrey Skilling resigns on 14th August 2001. ▪ Lay started an e-mail campaign just after 14th August 2001, trying to tout the Enron stock. ▪ Lay resigns as chairman and CEO of Enron on 23rd January 2002.	'k.l.lay@enron.com'; 'k.lay@enron.com'; 'ken.lay-.chairman.of.the.board@enron.com'; 'ken.lay-@enron.com'; 'ken.lay@enron.com'; 'kenlay@enron.com'; 'kenneth.l.lay@enron.com'; 'kenneth.lay@enron.com'; 'kennethlay@enron.com'; 'kenneth_lay@enron.com'; 'kenneth_lay@enron.net'; 'ken_lay@enron.com'; 'ken_lay@enron.net'; 'klay.enron@enron.com'; 'klay@enron.com'; 'kllay@enron.com'; 'k_lay@enron.com'; 'lay.kenneth@enron.com'; 'layk@enron.com';
Jeffrey Skilling	CEO of Enron	▪ Jeffrey Skilling joined Enron in 1990 and became chief executive officer of Enron gas services. ▪ Skilling named as president and chief operating officer of Enron in December 1996. ▪ Became CEO of Enron in February 2001. ▪ Resigned as CEO of Enron on 14th August 2001, for "personal reasons".	`jeff.skilling@enron.com'; `jeffrey.k.skilling@enron.com'; `jeffrey.skilling@enron.com'; `jeffreyskilling@yahoo.com'; `jeffrey_skilling@enron.com'; `jeff_skilling@enron.com'; `jskilling@enron.com'; `skilli@ei.enron.com'; `skilli@enron.com'; `skilling@enron.com'; `skilling@tribune.com'; `skillingj@enron.com';
Andrew Fastow	Chief Financial Officer of Enron	▪ Andrew Fastow was hired by Enron Capital (an Enron subsidiary) in 1990. ▪ Fastow became a friend of Jeffrey Skilling while working at Enron. ▪ As chief financial officer, Fastow had oversight of Enron's financial activities. ▪ Was involved with developing Special Purpose Entities (SPE), which were used to hide Enron's true financial status. ▪ Fastow was ousted from Enron in October 2001.	`andrew.fastow@enron.com'; `andrew.fastow@ljminvestments.com'; `andrew.s.fastow@enron.com'; `andy.fastow@enron.com'

Table D.2: Information on Richard A. Causey and John M. Forney

Name	Role	Description	Possible E-mail Addresses Used By Former Employee
Richard A. Causey	Former Chief Accounting Officer for Enron	▪ From 1986 to 1991, Causey was an employee of Arthur Andersen. ▪ In 1991, Enron hired Causey as Assistant Controller of Enron Gas Services Group. ▪ From 1992 to 1997, Causey served in various positions in a business unit known as Enron Capitol and Trade. ▪ In 1998, Causey was made Chief Accounting Officer of Enron and an Executive Vice President. ▪ He was involved with other Enron executives and senior managers in a wide-ranging scheme to manipulate Enron's financial results and making false and misleading statements about Enron's businesses.	` Rick Causey@ENRON'; `rcausey@enron.com'; `richard.causey@enron.com'; `rick.causey@enron.com';
John M. Forney	Manager of the Real Time Trading Desk at Enron	▪ John Forney was a senior trader at Enron. ▪ Has been accused of creating illegal trading schemes to manipulate energy prices in California during the U.S. winter season of 2000 – 2001. ▪ Forney was the 3rd Enron executive to be arrested.	'forney.john@enron.com'; 'jforney@ect.enron.com'; 'jforney@enron.com'; 'john.forney@enron.com'; 'm..forney@enron.com';

Appendix E

Hierarchical Fuzzy Inference System Development

This section provides information about the architecture, input variable fuzzy sets, rule bases, and the input-output mappings of the hierarchical fuzzy inference system described in Section 3.4.2 and evaluated in Section 5.3. The hierarchical fuzzy inference system described here has been developed using the MATLAB Fuzzy Toolbox [100].

E.1 Architecture

The diagram in Figure E.1 shows the architecture used for the hierarchical fuzzy inference system. This diagram has been labelled to indicate each layer of the hierarchy and has an identifier assigned to each fuzzy system module to represent the hierarchy layer and rule base number of each module (e.g. L1RB1 is the first rule base in layer 1 of the hierarchy).

E.2 Summary Of The Rule Bases

The information shown in Table E.1 provides a summary of information about each of the rule bases used in the hierarchical fuzzy inference system. This table provides information about: the number of input variables, the number of output variables, the total number of input membership functions, the total number of output membership functions, and the total number of rules used in each rule base.

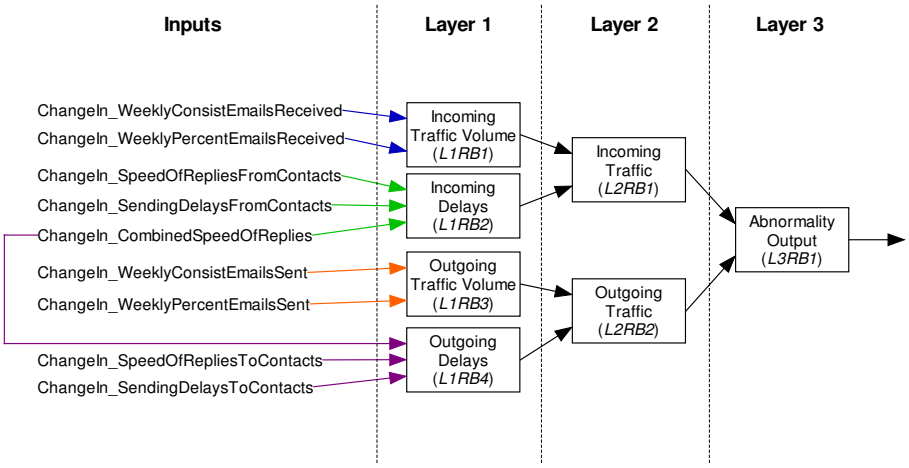


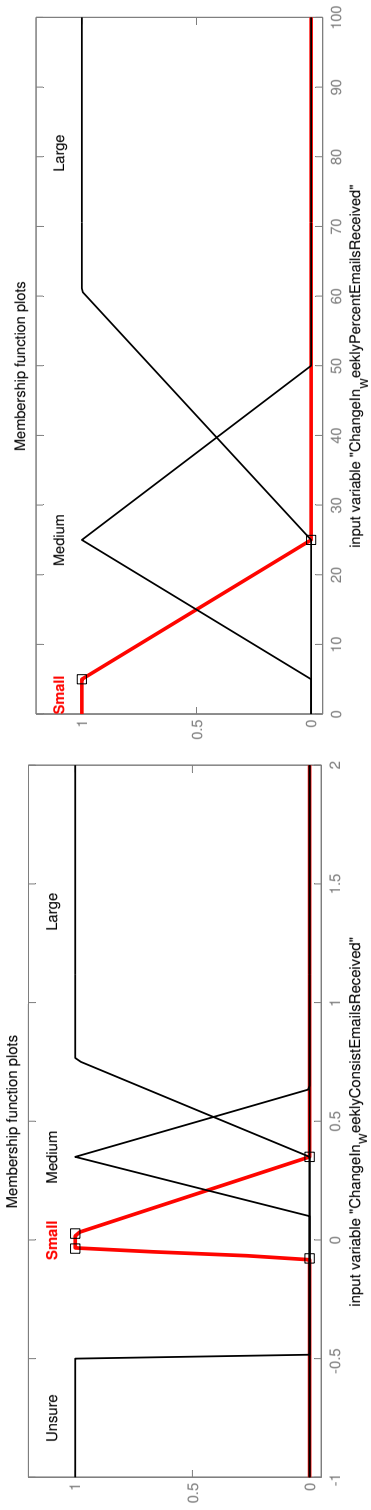
Figure E.1: Architecture of the hierarchical fuzzy inference system.

Table E.1: Summary of the rule bases used for the hierarchical fuzzy inference system.

Rule Base ID	No. of Input Variables	No. of Output Variables	Total Number of Input Membership Functions	Total Number of Output Membership Functions	No. of Rules
L1RB1	2	1	7	5	12
L1RB2	3	1	12	5	45
L1RB3	2	1	7	5	12
L1RB4	3	1	12	5	45
L2RB1	2	1	10	5	25
L2RB2	2	1	10	5	25
L3RB1	2	1	10	5	24

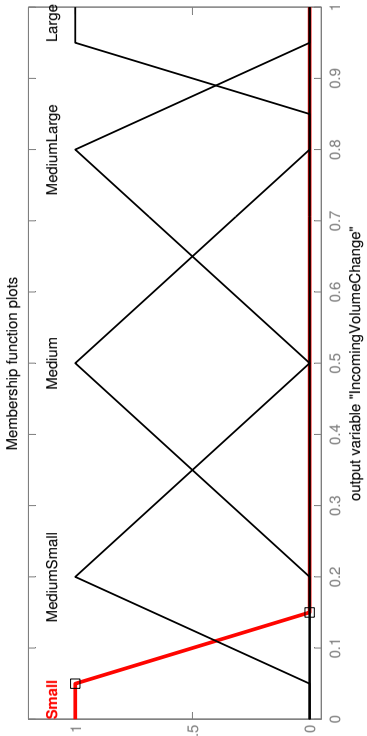
E.3 Input And Output Variable Fuzzy Sets

The diagrams in Figures E.3 to E.3, show the fuzzy sets used for the input and output variables of each fuzzy system module in the hierarchical fuzzy inference system. Each of the fuzzy sets shown are screen printouts from the MATLAB Fuzzy Toolbox [100].



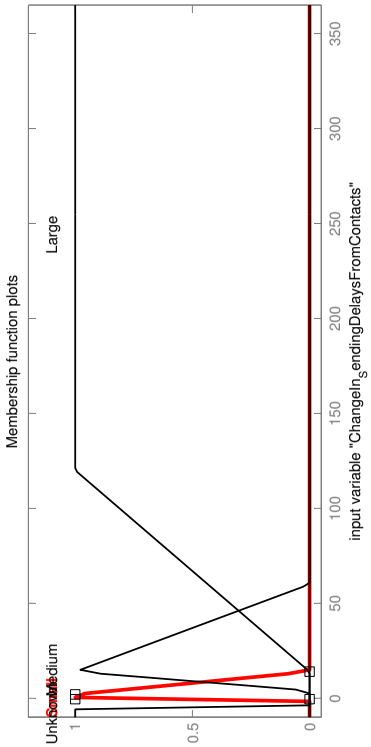
(a) L1RB1 input 1.

(b) L1RB1 input 2.

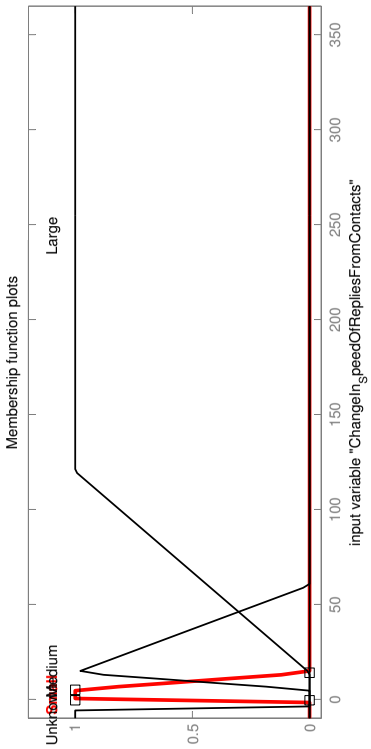


(c) L1RB1 output.

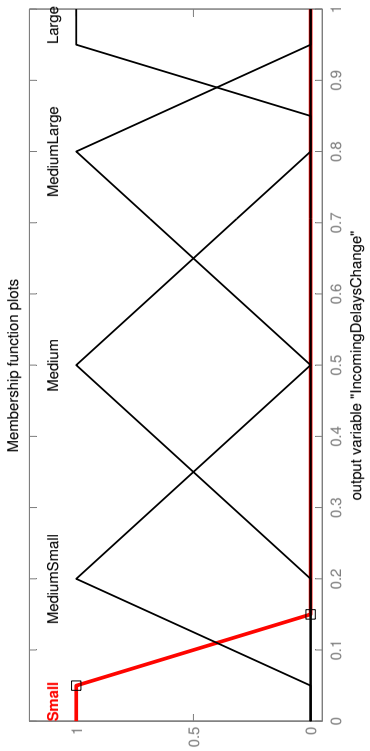
Figure E.2: The fuzzy sets used for the input and output variables of L1RB1.



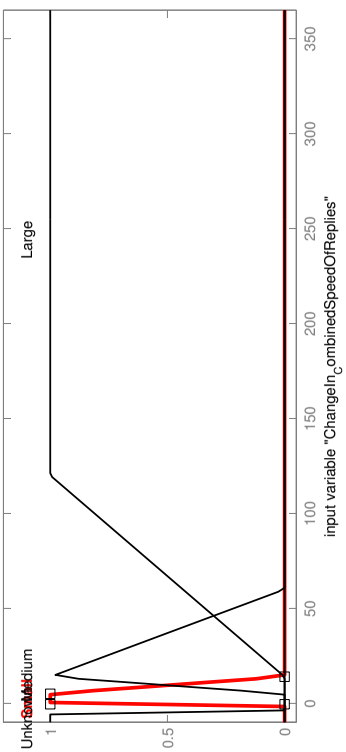
(b) L1RB2 input 2.



(a) L1RB2 input 1.

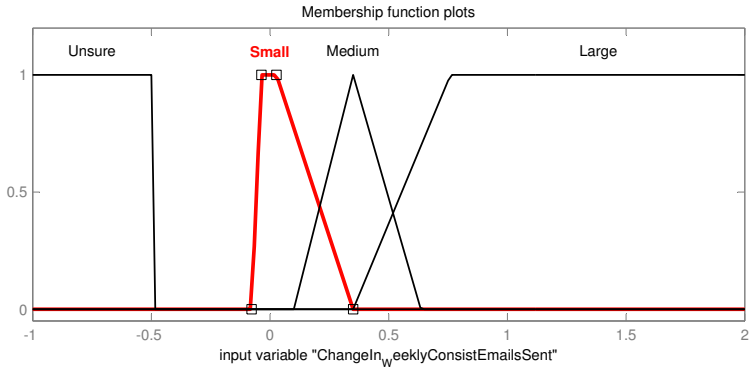


(d) L1RB2 output.

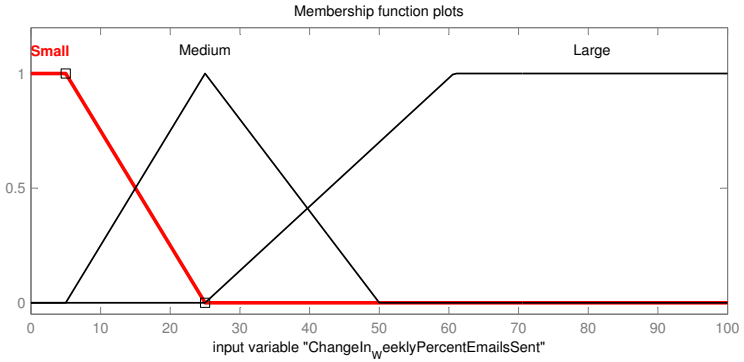


(c) L1RB2 input 3.

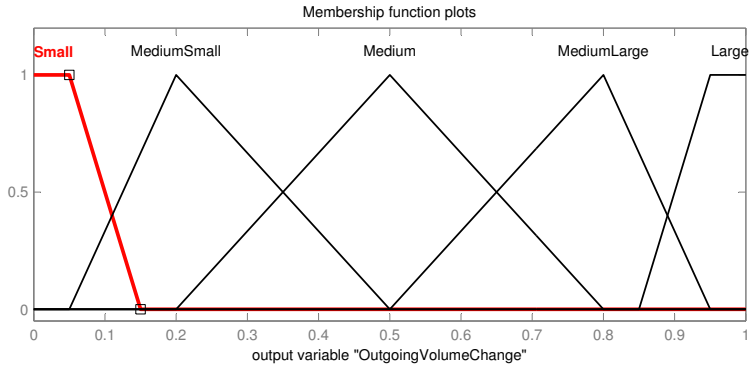
Figure E.3: The fuzzy sets used for the input and output variables of L1RB2.



(a) L1RB3 input 1.

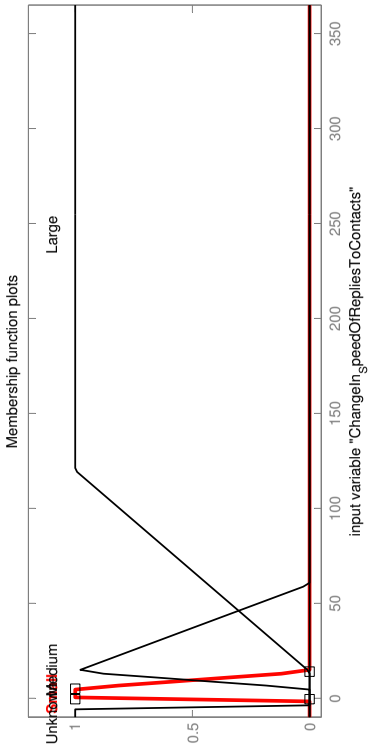


(b) L1RB3 input 2.

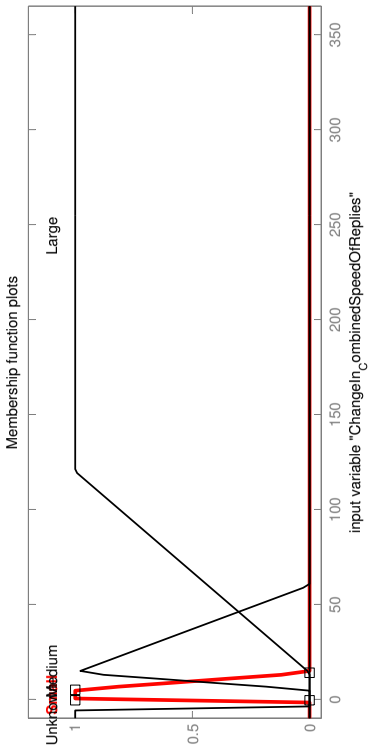


(c) L1RB3 output.

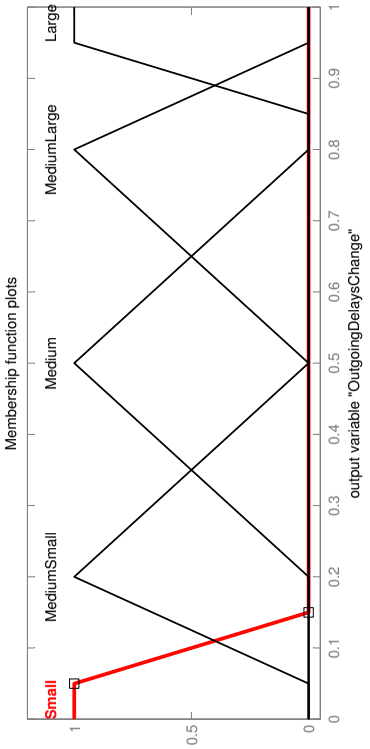
Figure E.4: The fuzzy sets used for the input and output variables of L1RB3.



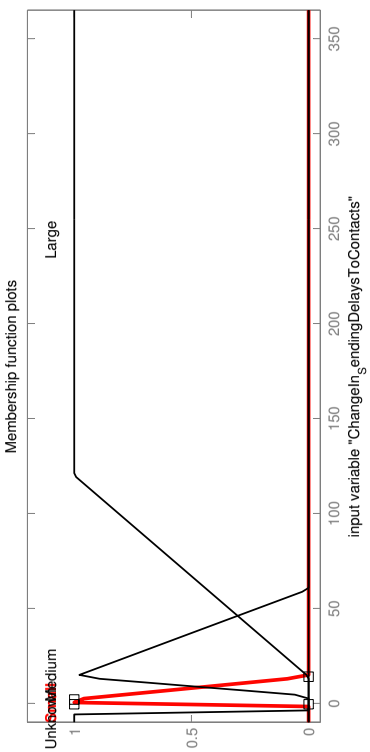
(b) L1RB4 input 2.



(a) L1RB4 input 1.

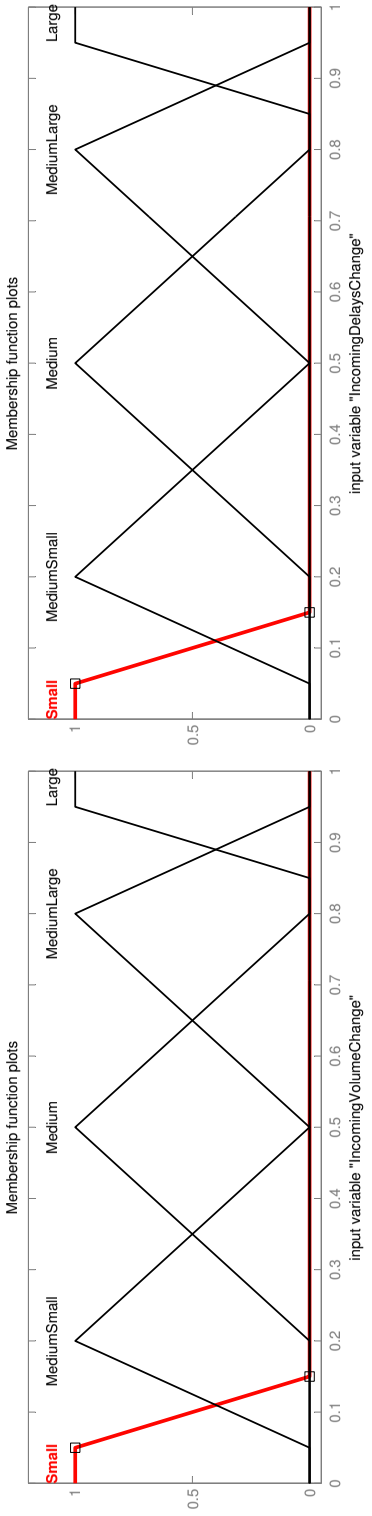


(d) L1RB4 output.



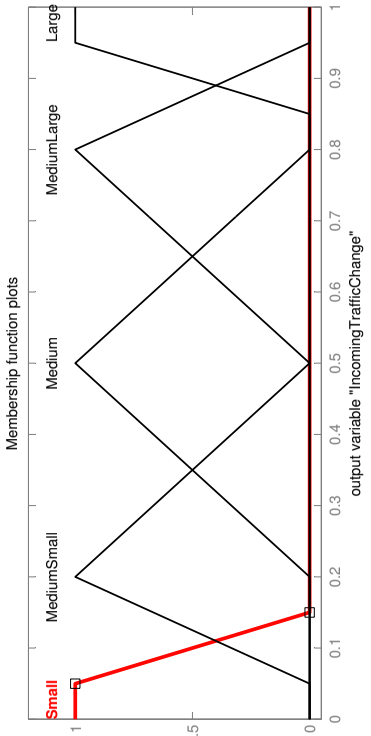
(c) L1RB4 input 3.

Figure E.5: The fuzzy sets used for the input and output variables of L1RB4.



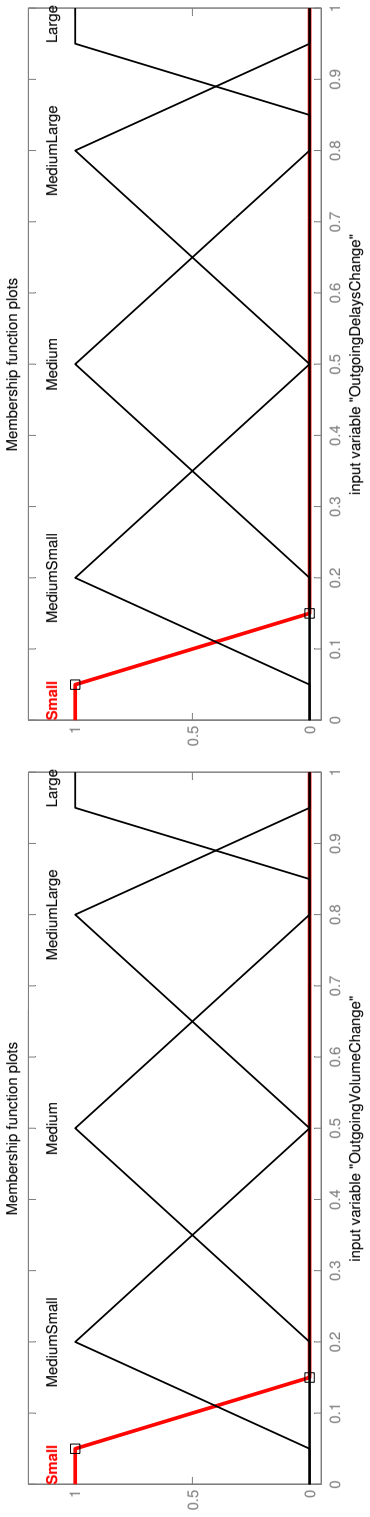
(a) L2RB1 input 1.

(b) L2RB1 input 2.



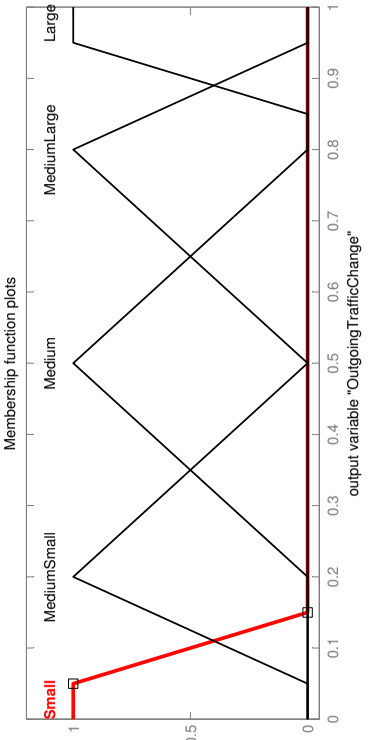
(c) L2RB1 output.

Figure E.6: The fuzzy sets used for the input and output variables of L2RB1.



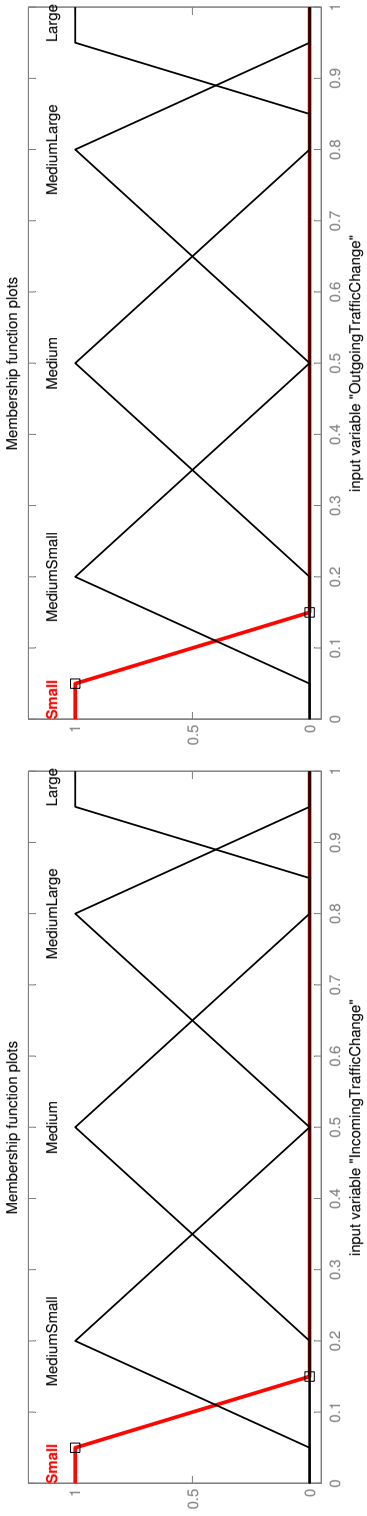
(a) L2RB2 input 1.

(b) L2RB2 input 2.



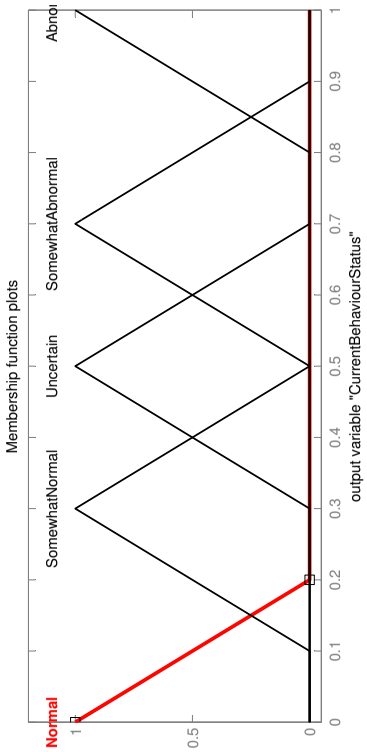
(c) L2RB2 output.

Figure E.7: The fuzzy sets used for the input and output variables of L2RB2.



(a) L3RB1 input 1.

(b) L3RB1 input 2.

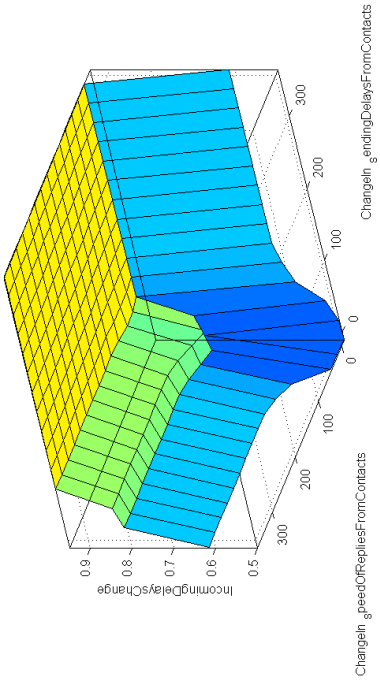


(c) L3RB1 output.

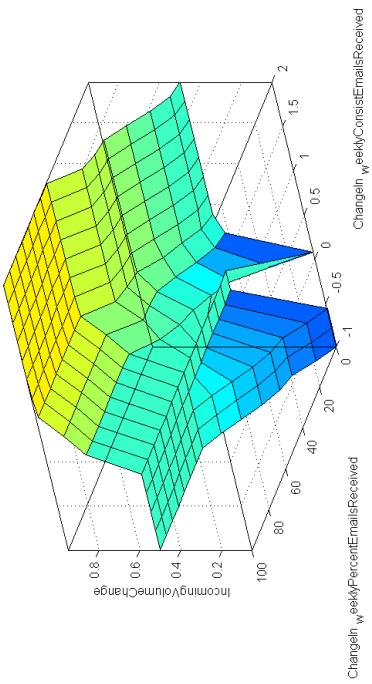
Figure E.8: The fuzzy sets used for the input and output variables of L3RB1.

E.4 Input-Output Mappings Of The Fuzzy System Modules

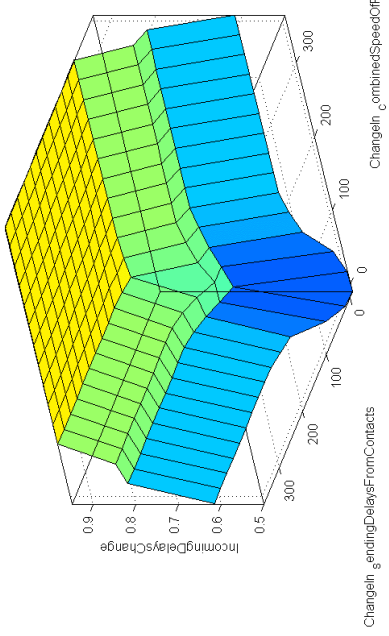
The surface plots in Figures E.4 to E.12 show the input-output mapping of each fuzzy system module in the hierarchical fuzzy inference system. Each of the input-output mappings are determined by the membership functions defined for each input/output variable, and also by the IF-THEN rules used in each fuzzy system module. These surface plots were generated using the MATLAB Fuzzy Toolbox [100].



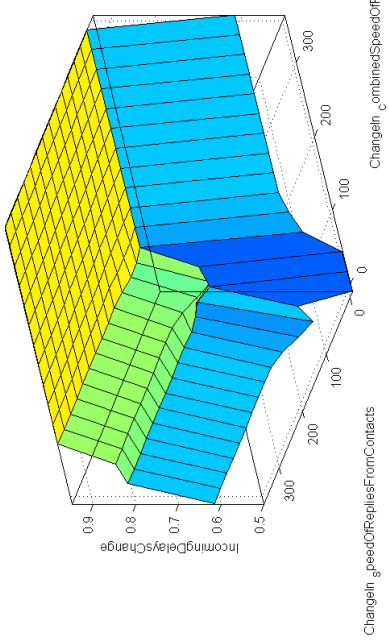
(b) L2RB2, inputs 1 and 2.



(a) L1RB1.

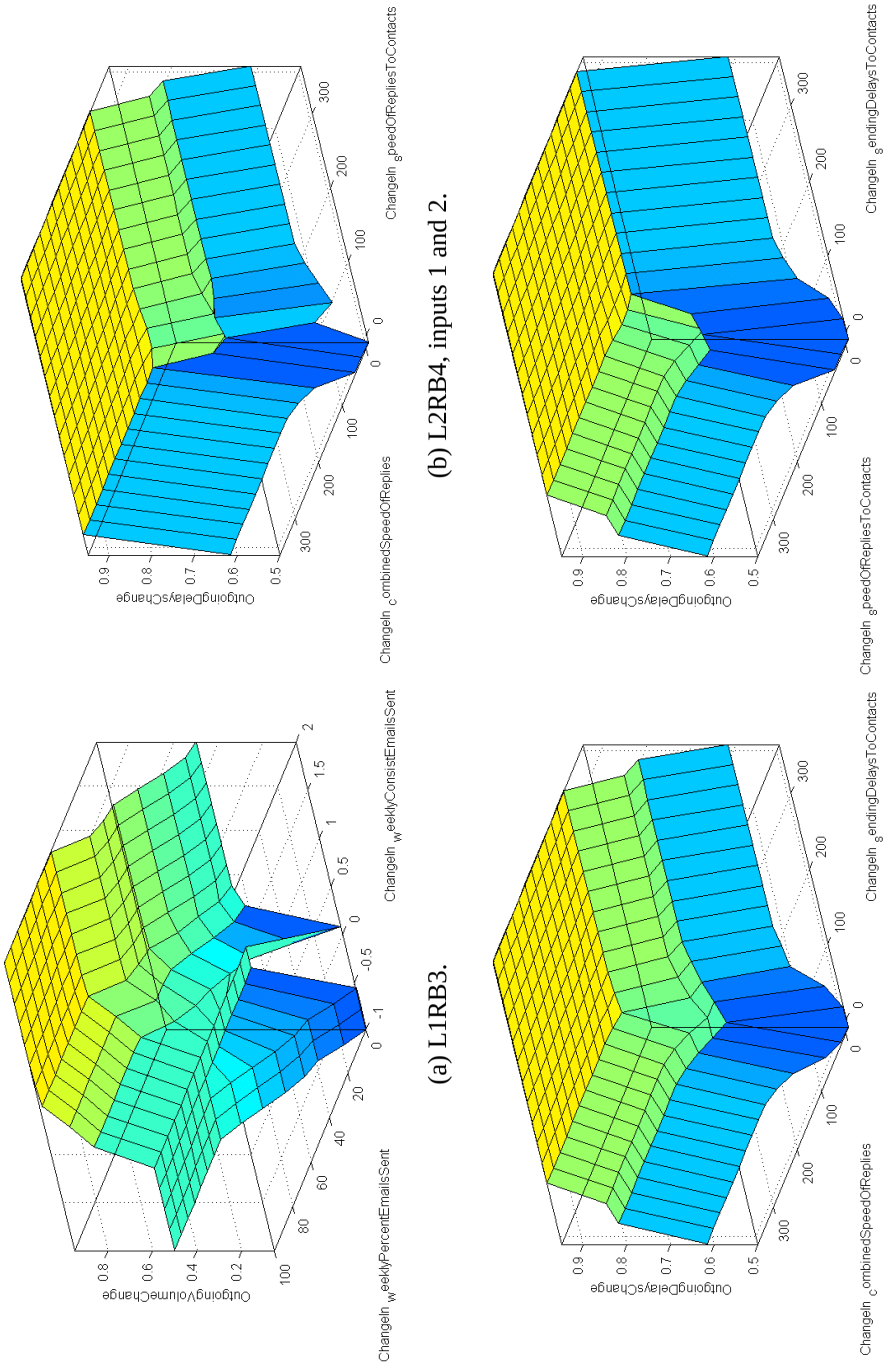


(d) L2RB2, inputs 2 and 3.



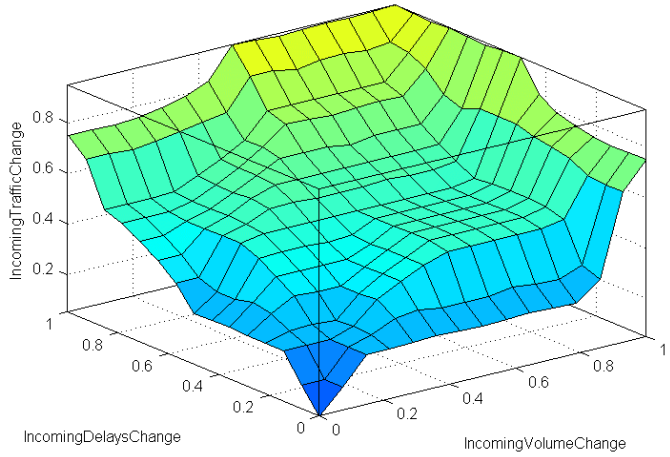
(c) L2RB2, inputs 1 and 3.

Figure E.9: The input-output mappings for the first and second rule bases of layer 1.

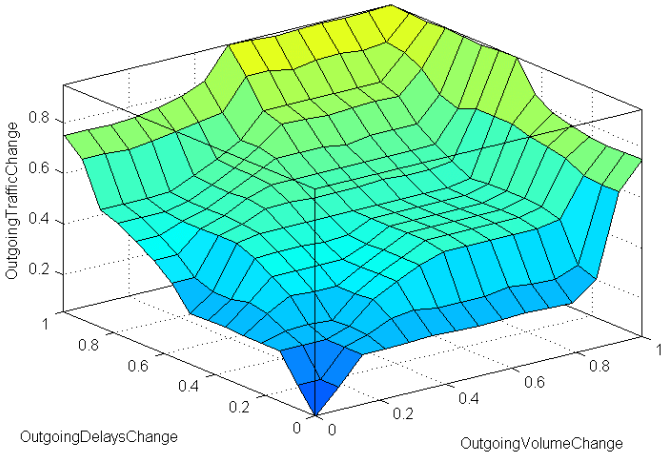


(a) L1RB3. (b) L2RB4, inputs 1 and 2. (c) L2RB4, inputs 1 and 3. (d) L2RB4, inputs 2 and 3.

Figure E.10: The input-output mappings for the third and fourth rule bases of layer 1.



(a) L2RB1.



(b) L2RB2.

Figure E.11: The input-output mappings for layer 2 of the hierarchy.

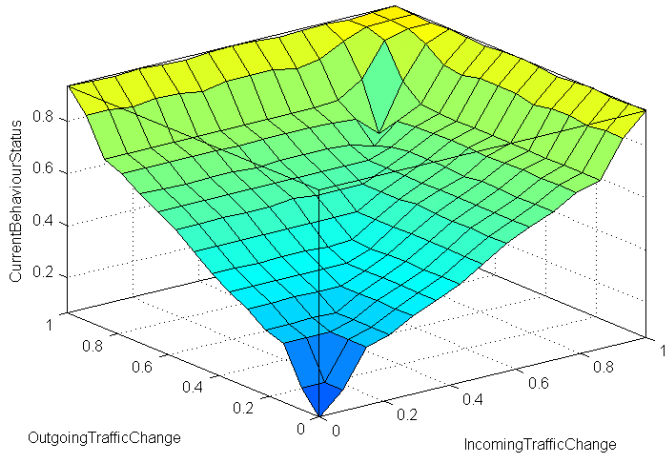


Figure E.12: The input-output mapping for L3RB1 in layer 3 of the hierarchy.

Appendix F

Simulated E-mail Traffic Data Analysis Results

F.1 Case Study 1 Decision Tree Results

This section provides the complete set of decision tree classification results for the e-mail users of simulation case study 1 (Section 5.2.1). The data for these results were obtained from the decision tree models generated by the J4.8 decision tree algorithm (provided by the WEKA data mining program [84]). The decision tree results for simulation case study 1 are shown in Tables F.1 to F.3.

Table F.1: Decision tree results for *clientA* to *clientD* from simulation case study 1.

E-mail Client's Address	Unusual Changes in Incoming Interactions Decision Tree Output	Unusual Changes in Outgoing Interactions Decision Tree Output
clientA@utas.edu.au	Direction = in | MessageDate <= 75.066366: clientD@utas.edu.au (330.0/198.0) | MessageDate > 75.066366: clientJ@utas.edu.au (337.0/188.0) Direction = out: clientA@utas.edu.au (686.0) Number of Leaves : 3 Size of the tree : 5	Direction = in: clientA@utas.edu.au (667.0) Direction = out: clientJ@utas.edu.au (686.0/434.0) Number of Leaves : 2 Size of the tree : 3
clientB@utas.edu.au	Direction = in: clientJ@utas.edu.au (465.0/240.0) Direction = out: clientB@utas.edu.au (445.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientB@utas.edu.au (465.0) Direction = out: clientJ@utas.edu.au (445.0/238.0) Number of Leaves : 2 Size of the tree : 3
clientC@utas.edu.au	Direction = in: clientF@utas.edu.au (112.0/38.0) Direction = out: clientC@utas.edu.au (137.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientC@utas.edu.au (112.0) Direction = out: clientF@utas.edu.au (137.0/49.0) Number of Leaves : 2 Size of the tree : 3
clientD@utas.edu.au	Direction = in: clientA@utas.edu.au (270.0/35.0) Direction = out: clientD@utas.edu.au (285.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientD@utas.edu.au (270.0) Direction = out: clientA@utas.edu.au (285.0/57.0) Number of Leaves : 2 Size of the tree : 3

Table F.2: Decision tree results for *clientE* to *clientH* from simulation case study 1.

E-mail Client's Address	Unusual Changes in Incoming Interactions Decision Tree Output	Unusual Changes in Outgoing Interactions Decision Tree Output
clientE@utas.edu.au	Direction = in: clientB@utas.edu.au (116.0/43.0) Direction = out: clientE@utas.edu.au (123.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientE@utas.edu.au (116.0) Direction = out: clientB@utas.edu.au (123.0/53.0) Number of Leaves : 2 Size of the tree : 3
clientF@utas.edu.au	Direction = in | MessageDate <= 41.150716: clientC@utas.edu.au (52.0/24.0) | MessageDate > 41.150716: clientJ@utas.edu.au (182.0/63.0) Direction = out: clientF@utas.edu.au (204.0) Number of Leaves : 3 Size of the tree : 5	Direction = in: clientF@utas.edu.au (234.0) Direction = out | MessageDate <= 66.360544: clientC@utas.edu.au (84.0/44.0) | MessageDate > 66.360544: clientJ@utas.edu.au (120.0/36.0) Number of Leaves : 3 Size of the tree : 5
clientG@utas.edu.au	Direction = in | MessageDate <= 46.836093 | | MessageDate <= 37.232636: clientI@utas.edu.au (27.0/14.0) | | MessageDate > 37.232636: clientF@utas.edu.au (9.0/4.0) | MessageDate > 46.836093: clientD@utas.edu.au (56.0/13.0) Direction = out: clientG@utas.edu.au (64.0) Number of Leaves : 4 Size of the tree : 7	Direction = in: clientG@utas.edu.au (92.0) Direction = out | MessageDate <= 47.463819: clientI@utas.edu.au (22.0/13.0) | MessageDate > 47.463819: clientD@utas.edu.au (42.0/12.0) Number of Leaves : 3 Size of the tree : 5
clientH@utas.edu.au	Direction = in: clientC@utas.edu.au (76.0/27.0) Direction = out: clientH@utas.edu.au (56.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientH@utas.edu.au (76.0) Direction = out: clientC@utas.edu.au (56.0/18.0) Number of Leaves : 2 Size of the tree : 3

Table F.3: Decision tree results for *clientI* and *clientJ* from simulation case study 1.

E-mail Client's Address	Unusual Changes in Incoming Interactions Decision Tree Output	Unusual Changes in Outgoing Interactions Decision Tree Output
clientI@utas.edu.au	Direction = in: clientA@utas.edu.au (140.0/81.0) Direction = out: clientI@utas.edu.au (130.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientI@utas.edu.au (140.0) Direction = out | MessageDate <= 33.57773: clientG@utas.edu.au (28.0/16.0) | MessageDate > 33.57773: clientA@utas.edu.au (102.0/58.0) Number of Leaves : 3 Size of the tree : 5
clientJ@utas.edu.au	Direction = in | MessageDate <= 82.495611: clientB@utas.edu.au (317.0/181.0) | MessageDate > 82.495611: clientA@utas.edu.au (259.0/138.0) Direction = out: clientJ@utas.edu.au (618.0) Number of Leaves : 3 Size of the tree : 5	Direction = in: clientJ@utas.edu.au (576.0) Direction = out | MessageDate <= 78.600952: clientB@utas.edu.au (313.0/174.0) | MessageDate > 78.600952: clientA@utas.edu.au (305.0/170.0) Number of Leaves : 3 Size of the tree : 5

F.2 Case Study 2 Decision Tree Results

This section provides the complete set of decision tree classification results for the e-mail users of simulation case study 2 (Section 5.2.2). The data for these results were also obtained from the decision tree models generated by the J4.8 decision tree algorithm (provided by the WEKA data mining program [84]). The decision tree results for simulation case study 2 are shown in Tables F.4 to F.6.

Table F.4: Decision tree results for *clientA* to *clientC* from simulation case study 2.

E-mail Client's Address	Unusual Changes in Incoming Interactions Decision Tree Output	Unusual Changes in Outgoing Interactions Decision Tree Output
clientA@utas.edu.au	Direction = in | MessageDate <= 108.997388 | | MessageDate <= 24.07781: clientI@utas.edu.au (30.0/12.0) | | MessageDate > 24.07781: clientG@utas.edu.au (255.0/160.0) | MessageDate > 108.997388: clientI@utas.edu.au (275.0/183.0) Direction = out: clientA@utas.edu.au (565.0) Number of Leaves : 4 Size of the tree : 7	Direction = in: clientA@utas.edu.au (560.0) Direction = out | MessageDate <= 108.524022 | | MessageDate <= 22.647081: clientI@utas.edu.au (30.0/12.0) | | MessageDate > 22.647081: clientG@utas.edu.au (252.0/159.0) | MessageDate > 108.524022: clientI@utas.edu.au (283.0/191.0) Number of Leaves : 4 Size of the tree : 7
clientB@utas.edu.au	Direction = in | MessageDate <= 110.600284: clientG@utas.edu.au (152.0/84.0) | MessageDate > 110.600284: clientA@utas.edu.au (153.0/81.0) Direction = out: clientB@utas.edu.au (287.0) Number of Leaves : 3 Size of the tree : 5	Direction = in: clientB@utas.edu.au (305.0) Direction = out | MessageDate <= 107.341808: clientG@utas.edu.au (141.0/79.0) | MessageDate > 107.341808: clientA@utas.edu.au (146.0/79.0) Number of Leaves : 3 Size of the tree : 5
clientC@utas.edu.au	Direction = in | MessageDate <= 106.819134: clientI@utas.edu.au (224.0/131.0) | MessageDate > 106.819134: clientG@utas.edu.au (260.0/99.0) Direction = out: clientC@utas.edu.au (488.0) Number of Leaves : 3 Size of the tree : 5	Direction = in: clientC@utas.edu.au (484.0) Direction = out | MessageDate <= 109.08375: clientI@utas.edu.au (232.0/139.0) | MessageDate > 109.08375: clientG@utas.edu.au (256.0/100.0) Number of Leaves : 3 Size of the tree : 5

Table F.5: Decision tree results for *clientD* to *clientF* from simulation case study 2.

E-mail Client's Address	Unusual Changes in Incoming Interactions Decision Tree Output	Unusual Changes in Outgoing Interactions Decision Tree Output
clientD@utas.edu.au	Direction = in | MessageDate <= 48.744651: clientI@utas.edu.au (60.0/29.0) | MessageDate > 48.744651: clientG@utas.edu.au (276.0/148.0) Direction = out: clientD@utas.edu.au (317.0) Number of Leaves : 3 Size of the tree : 5	Direction = in: clientD@utas.edu.au (336.0) Direction = out | MessageDate <= 49.286308 | | MessageDate <= 31.475441 | | | MessageDate <= 26.134273 | | | | MessageDate <= 22.932191: clientB@utas.edu.au (14.0/7.0) | | | | MessageDate > 22.932191: clientI@utas.edu.au (4.0) | | | MessageDate > 26.134273 | | | | MessageDate <= 28.934465: clientF@utas.edu.au (3.0/1.0) | | | | MessageDate > 28.934465: clientB@utas.edu.au (2.0) | | MessageDate > 31.475441: clientI@utas.edu.au (36.0/17.0) | MessageDate > 49.286308: clientG@utas.edu.au (258.0/139.0) Number of Leaves : 7 Size of the tree : 13
clientE@utas.edu.au	Direction = in: clientH@utas.edu.au (87.0/23.0) Direction = out: clientE@utas.edu.au (95.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientE@utas.edu.au (87.0) Direction = out: clientH@utas.edu.au (95.0/33.0) Number of Leaves : 2 Size of the tree : 3
clientF@utas.edu.au	Direction = in: clientD@utas.edu.au (132.0/95.0) Direction = out: clientF@utas.edu.au (101.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientF@utas.edu.au (132.0) Direction = out | MessageDate <= 103.66992: clientC@utas.edu.au (52.0/37.0) | MessageDate > 103.66992: clientD@utas.edu.au (49.0/30.0) Number of Leaves : 3 Size of the tree : 5

Table F.6: Decision tree results for *clientG* to *clientI* from simulation case study 2.

E-mail Client's Address	Unusual Changes in Incoming Interactions Decision Tree Output	Unusual Changes in Outgoing Interactions Decision Tree Output
clientG@utas.edu.au	Direction = in | MessageDate <= 105.182815 | | MessageDate <= 32.112822: clientB@utas.edu.au (34.0/15.0) | | MessageDate > 32.112822: clientA@utas.edu.au (230.0/151.0) | MessageDate > 105.182815: clientC@utas.edu.au (340.0/179.0) Direction = out: clientG@utas.edu.au (637.0) Number of Leaves : 4 Size of the tree : 7	Direction = in: clientG@utas.edu.au (604.0) Direction = out | MessageDate <= 108.892207 | | MessageDate <= 35.575484 | | | MessageDate <= 7.695895: clientC@utas.edu.au (3.0) | | | MessageDate > 7.695895: clientB@utas.edu.au (44.0/23.0) | | MessageDate > 35.575484: clientA@utas.edu.au (248.0/164.0) | MessageDate > 108.892207: clientC@utas.edu.au (342.0/183.0) Number of Leaves : 5 Size of the tree : 9
clientH@utas.edu.au	Direction = in: clientI@utas.edu.au (198.0/62.0) Direction = out: clientH@utas.edu.au (192.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientH@utas.edu.au (198.0) Direction = out | MessageDate <= 65.82457: clientE@utas.edu.au (51.0/24.0) | MessageDate > 65.82457: clientI@utas.edu.au (141.0/37.0) Number of Leaves : 3 Size of the tree : 5
clientI@utas.edu.au	Direction = in: clientA@utas.edu.au (551.0/365.0) Direction = out: clientI@utas.edu.au (575.0) Number of Leaves : 2 Size of the tree : 3	Direction = in: clientI@utas.edu.au (551.0) Direction = out: clientA@utas.edu.au (575.0/387.0) Number of Leaves : 2 Size of the tree : 3

Appendix G

Enron E-mail Traffic Data Analysis Results

This section provides the complete set of behaviour measurement results obtained for the Enron case study in Section 5.3. The results comprise of measurements for each of Jeffrey Skilling's 525 communication links. The results are shown in Tables G.1 to G.14 and it uses the following symbols:

x_1 = CombinedSpeedOfReplies (Normal Behaviour Measurement)
y_1 = CombinedSpeedOfReplies (Current Behaviour Measurement)

x_2 = SendingDelaysFromContacts (Normal Behaviour Measurement)
y_2 = SendingDelaysFromContacts (Current Behaviour Measurement)

x_3 = SendingDelaysToContacts (Normal Behaviour Measurement)
y_3 = SendingDelaysToContacts (Current Behaviour Measurement)

x_4 = SpeedOfRepliesFromContacts (Normal Behaviour Measurement)
y_4 = SpeedOfRepliesFromContacts (Current Behaviour Measurement)

x_5 = SpeedOfRepliesToContacts (Normal Behaviour Measurement)
y_5 = SpeedOfRepliesToContacts (Current Behaviour Measurement)

x_6 = WeeklyConsistEmailsReceived (Normal Behaviour Measurement)
y_6 = WeeklyConsistEmailsReceived (Current Behaviour Measurement)

x_7 = WeeklyConsistEmailsSent (Normal Behaviour Measurement)
y_7 = WeeklyConsistEmailsSent (Current Behaviour Measurement)

x_8 = WeeklyPercentEmailsReceived (Normal Behaviour Measurement)
y_8 = WeeklyPercentEmailsReceived (Current Behaviour Measurement)

x_9 = WeeklyPercentEmailsSent (Normal Behaviour Measurement)
y_9 = WeeklyPercentEmailsSent (Current Behaviour Measurement)

UNK = Unknown Delay Flag
NULL = Unknown Consistency Value

Table G.1: Enron case study behaviour measurements, rows 1 to 40.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>steven.kean@enron.com	0.5	UNK	1964	103883.5	27.441667	UNK	UNK	UNK	709	UNK	3219	-0.02393617	-0.01529391	NULL	-0.00729927	1.282051282	16.00496278	0	1.612903226
jskilli@enron.com<=>markskilling@hotmail.com	0.5	UNK	UNK	9	27695	UNK	UNK	UNK	UNK	UNK	UNK	-0.02666667	0.228300511	NULL	NULL	2.564102564	25.80645161	0	0
jeff.skilling@enron.com<=>karen.denne@enron.com	0.3	UNK	31369.783	UNK	300.61667	UNK	UNK	UNK	UNK	UNK	31369.7833	-0.01315789	0.178822809	NULL	-0.00729927	1.282051282	16.97270471	0	1.612903226
jeff.skilling@enron.com<=>kelly.johnson@enron.com	0.3	UNK	87582.717	0	2468.2333	UNK	UNK	UNK	87582.71667	UNK	UNK	-0.01315789	-0.01470588	NULL	-0.00729927	0.854700855	0.64516129	0	3.225806452
jeff.skilling@enron.com<=>liz.taylor@enron.com	0.3	UNK	UNK	23921	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.02666667	0.349932705	NULL	NULL	0.747863248	0	0	0
jeff.skilling@enron.com<=>markskilling@hotmail.com	0.3	UNK	UNK	1007	UNK	UNK	UNK	UNK	UNK	UNK	UNK	0.427702703	0.453219271	NULL	NULL	23.13797314	0	0	0
jeff.skilling@enron.com<=>wilson.kriegel@enron.com	0.3	UNK	UNK	573	600	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.01470588	NULL	NULL	0.641025641	1.612903226	0	0
jeff.skilling@enron.com<=>chris.abel@enron.com	0.11335529	UNK	UNK	UNK	16944	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	0.083491461	NULL	NULL	0.213675214	3.796526055	0	0
jeff.skilling@enron.com<=>rosalee.fleming@enron.com	0.092071182	UNK	UNK	0	0	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	0.040086427	NULL	NULL	0.366300366	2.903225806	0	0
jeff.skilling@enron.com<=>aahanch@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>aalkhay@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>abassis@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>aborgatt@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>abrown4@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>acarroll@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>acolpea@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>adiaz@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>adietz@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>agarg@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>agonzal2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>agupta2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ahamby@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ahernan6@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ahoxha1@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ajackso5@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ajain4@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ajanna@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ajon@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>akhoja@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>akollar@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>akorioth@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>alarsen@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>alberto.gude@enron.com	0.091424688	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.00729927	NULL	NULL	0.854700855	0	0	0
jeff.skilling@enron.com<=>alehner@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>alyon@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>amanceb@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>amiles@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>amiller@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>anahman@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>aperkin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.2: Enron case study behaviour measurements, rows 41 to 80.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>aredric@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>aroman@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>asait@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>asaler@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>athomas@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>awais2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>awu3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bblakel@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bbraddo@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bbradfo2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bcohen@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bcreigh@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bdornie@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>beixman@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bfeschi@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bfreema@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bgarret@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bglisan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bhaufre@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bhayden@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bhendon@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bhull@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bluna@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bmatcha@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bmauritz@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bneff2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bobbie.power@enron.com	0.091424688	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.01470588	NULL	NULL	0.854700855	3.225806452	0	0
jeff.skilling@enron.com<=>bozcan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bpham2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>brad.blesie@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>brenda_f_herod@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>brian.hoskins@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>brogers2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bromine@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>brudy@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bscholt@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bterp@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bvascon@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>bvotaw@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	0	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.01413228	0	0	0.002564103	1.612903226
jeff.skilling@enron.com<=>bwax@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.3: Enron case study behaviour measurements, rows 81 to 120.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>bwillia4@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>calator@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cbarnes2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cbarthe@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cbogle@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cbresla@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cbroido@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cburke@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cchenowe@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cchild2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cclark4@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ccoffma@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ccontre@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ccrawfo@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ccunnin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cday@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cdelace@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cecilia.manzolillo@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cgorric@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chacket@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chad.gronvold@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>charlene.jackson@enron.com	0.091424688	UNK	UNK	0	UNK	UNK	0	UNK	UNK	UNK	UNK	-0.01315789	-0.01321586	NULL	-0.00729927	1.282051282	0	0	1.075268817
jeff.skilling@enron.com<=>charri2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chau-ye.wu@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chelfri@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cheng2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chenriqu@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cherron2@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chor.lim.goh@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chris.stokley@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chuchu.wang@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chuizer@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>chyde2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cingstad@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cjackso@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ckvorni@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>clandry@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>clau@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cloehr@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cmallor@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.4: Enron case study behaviour measurements, rows 121 to 160.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>cmetz@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cpaipan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cpennix@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cpernot@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cray@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cribeiro@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cruiz@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cseigle@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>csheh@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>csimoes@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>csmith2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>csoutha@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cstuart@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ctorn@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cturria@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cuus@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cvarnel@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cvaugha@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cvicens@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cvotaw@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cwalker@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cweber@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>cwinfre@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>darzola@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>david.forster@enron.com	0.091424688	UNK	UNK	1084.5	UNK	UNK	UNK	UNK	UNK	UNK	UNK	0.575874415	0.58499361	NULL	-0.00729927	4.594017094	0	0	1.612903226
jeff.skilling@enron.com<=>david.j.vitrella@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>david.tagliarino@enron.com	0.091424688	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.00729927	NULL	NULL	0.641025641	0	0	0
jeff.skilling@enron.com<=>david_hunker@pgn.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dbrown4@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dbuchana@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dcarnei@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dcase@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dchang@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dconnal@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ddrakes@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ddye2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>deiching@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>derek_mo@enron.net	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dfeather@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dfontan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.5: Enron case study behaviour measurements, rows 161 to 200.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>dganc@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dgorte@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dharvey2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dherman2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dkhamap@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dkhandke@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dlynch@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dmalone@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dmartin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dmarye@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dmathe2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dmu@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dorothy.barnes@enron.com	0.091424688	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.00729927	NULL	NULL	1.282051282	0	0	0
jeff.skilling@enron.com<=>dprofir@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dreck@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>driddle@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dries@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dsander3@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dsewell@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dterlip@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dthames@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dtran@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dvoorhe@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dwindle@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>dyuan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>eallward@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ebass@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ebess@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ebetzer@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ecoats@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ecross2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>efraser@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>egant@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>egore@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ehamb@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ehokmar@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ehowley@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>eleydic@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>elliot.mainzer@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>emason@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.6: Enron case study behaviour measurements, rows 201 to 240.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>emccart@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>epao@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>epeders@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>erainer@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>erice2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>erwin.landivar@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>eschult@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>escott@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>eshim@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>esimpso@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>etellec@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ewood@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>expense.report@enron.com	0.091424688	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.01470588	0	0	0.427350427	0
jeff.skilling@enron.com<=>fahad@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>fbunjam@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>fcohaga@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ffarhan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>fkarbar@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>flalji@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>fzoes@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>garrett_ashmore@enron.net	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gbabbar@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gbenite@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gcalver@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gcarant@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ggupta@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gjohnson@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gjunque@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gkoepke@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gmagee@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gmaisas@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gmargai@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gmarti2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gmonroy@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gpenman@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>grodrigu@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gsohoni@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gsolber@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gsurowi@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gtripp@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.7: Enron case study behaviour measurements, rows 241 to 280.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>gvanmar@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gzimmer@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>gzivic@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>halon2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hang.bui@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>harry.arora@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hbucalo@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hbuchan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hcampos@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hcubill@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hgutier@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	0	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.01413228	0	0	0.002564103	1.612903226
jeff.skilling@enron.com<=>hhellma@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hlin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hmitche@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>hmurill@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>icaplan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ipetroz@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>iqureish@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>isingh@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>itareen@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jallari@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jalthaus@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jay_hawthorn@enron.net	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jbest@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jblachma@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jbratto@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jbuerkl@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jchampi@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jchoi@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jchola@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jcole@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jcornet@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jcoulte@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jdavidso@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jdeffner@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jdematos@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jdesai@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jean-sebastien.fontaine@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jeanie_miller@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jepstei@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.8: Enron case study behaviour measurements, rows 281 to 320.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>jescoba@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jespera@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jewel.meeks@enron.com	0.091424688	UNK	UNK	1909	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.00729927	NULL	NULL	0.641025641	0	0	0
jeff.skilling@enron.com<=>jferrar@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jgodbol@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jgordon@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jgraham@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jguo@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jhoff2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jhoover@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jhopley@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jhowton@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jhuffak@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jking4@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jking5@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jkoop@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jlang@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jlangfo@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jlee7@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jli@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jlin3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmackey@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmarti2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmartin5@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmassey@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmcpher@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmerrit@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmiao@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmillig@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jmyung@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jneil@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jnewbro@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>john.arnold@enron.com	0.091424688	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.00729927	NULL	NULL	1.282051282	0	0	0
jeff.skilling@enron.com<=>joseph.sutton@enron.com	0.091424688	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.01470588	NULL	NULL	0.427350427	0	0	0
jeff.skilling@enron.com<=>jpage@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jpielop@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jpyke@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jrandol@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jreside@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jriley@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.9: Enron case study behaviour measurements, rows 321 to 360.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>jrostan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jroumel@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jryan3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jsamudi@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jscarbo@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jschube@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jseigal@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jshankm@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jslone@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jsnyder@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jsparli@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jstark@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jsurfac@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jthompk@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jthompso@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jthorne@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jthrashe@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jvangeld@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jwang3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jwhiteh@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jwiesep@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jwilli5@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jvazigi@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>jzhang@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kashby@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kbennou@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kchick@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kchow@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kcompea@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ketter@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kfrancis@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kgordon2@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kgreiner@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>khall2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>khand@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kherrma@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kkasravi@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kknop@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kkuyken@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>klucas@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.10: Enron case study behaviour measurements, rows 361 to 400.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>kmccoy@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kmcnell@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kmonk@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kneuner@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kristina_lund@enron.net	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kruffco@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>krusciit@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kshulkl@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ktate@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kwomac2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>kzheng@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lbenavi@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lcampbel@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lconnol@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>leonardo.pacheco@enron.com	0.091424688	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	UNK	UNK	0.702439359	0.715827824	NULL	NULL	2.564102564	0	0	0
jeff.skilling@enron.com<=>lfields@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lgillet@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lho@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lhope@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lhowens@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>libasco@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ling_li@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ljackso@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lmalone@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lmastran@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lmena@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lmendez@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lmiller@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>louise.kitchen@enron.com	0.091424688	UNK	UNK	UNK	UNK	UNK	47.6	UNK	UNK	UNK	UNK	-0.01315789	-0.00729927	NULL	-0.01077122	1.282051282	0	0	2.688172043
jeff.skilling@enron.com<=>lpacheco@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lpan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lrosenb@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lsnowde@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ltham@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lwente@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lworthy@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lxiao@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>lyin2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mabraha@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mallen3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.11: Enron case study behaviour measurements, rows 401 to 440.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>malonso@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>marissa.c.womble@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>martin.lin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mavs@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mbaker2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mbawa@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mcapest@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mcastigl@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mcourtn@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mday@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mdrisc3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mduffy@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mdypiang@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>medmonds@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>medward@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>meichma@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>melanie_king@enron.net	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>meubank@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mflaniga@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mfrancis@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mgandy@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mgarber2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mgimble@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mglassc@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mgonzal3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mhamlin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mharris4@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mhatten@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mheller@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mhender@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mike_shannon@enron.net	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mjachim@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mjohn@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mkolman@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mlafuze@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mlay@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mleblan@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mlenhart@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mlian@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mmakows@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.12: Enron case study behaviour measurements, rows 441 to 480.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>mmarvin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mmata@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mmcgowa@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mmixon@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mnataraj@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mparikh@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mparraca@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mpasad2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mpoulso@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mratner@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mrhee@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mrodrig@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>msabine@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>msacchi@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>msergee@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>msimmon@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>msoness@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>msteven3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mstockt@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mstower@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mthomps@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mvasque@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mvegala@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mvicens@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mwarner@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mwenz@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>mzhang@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nade@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nalvino@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nfoley@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nhiemst@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nhiggin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nla@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>npalcze@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nshah@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nstepha@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nvolcy@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nwill@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nwu@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>nzhu@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0

Table G.13: Enron case study behaviour measurements, rows 481 to 520.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>owathing@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>padair@enron.co.uk	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>paul.trieschman@enron.com	0.091424688	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.01315789	-0.00729927	NULL	NULL	1.282051282	0	0	0
jeff.skilling@enron.com<=>pbennet@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pblanco@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pbroder@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pburkha@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pfissle@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pghosh@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pgregor@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>phayes@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pheintz@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pjaising@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pjeanmar@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pleahy@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pmarkey@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pmcglory@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pmeyers@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ppatel3@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>pramgola@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ptlapek@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ptravis@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ptriesc@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ptucker@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rbisono@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rcarson@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rchetti@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rgrube@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rhart@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rhsu@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rlassan@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rleiber@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rmathew@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rmwongo@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rphilli1@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rrezaei@azurix.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rrizopa@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>rsoares@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>ryzquie@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>sally.beck@enron.com	0.091424688	UNK	172737.97	UNK	600	UNK	0	UNK	UNK	UNK	172737.967	-0.01315789	-0.02678571	NULL	-0.00729927	0.641025641	3.225806452	0	1.075268817

Table G.14: Enron case study behaviour measurements, rows 521 to 525.

Communication Link	Abnormality Rating	x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4	x_5	y_5	x_6	y_6	x_7	y_7	x_8	y_8	x_9	y_9
jeff.skilling@enron.com<=>sherri.reinartz@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.854700855	0
jeff.skilling@enron.com<=>tdutta@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>tnoble2@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0
jeff.skilling@enron.com<=>vince.kaminski@enron.com	0.091424688	UNK	UNK	97912	UNK	UNK	UNK	UNK	UNK	UNK	UNK	-0.02666667	-0.03356449	NULL	NULL	1.282051282	0	0	0
jeff.skilling@enron.com<=>wjennin@enron.com	0.091424688	UNK	UNK	UNK	UNK	0	UNK	UNK	UNK	UNK	UNK	NULL	NULL	-0.01315789	-0.00729927	0	0	0.002564103	0