A Modified Anonymisation Algorithm Towards Reducing Information Loss.

Rose Tinabo
Technological University Dublin

A Modified Anonymisation Algorithm Towards Reducing Information Loss

By

Rose Tinabo *(MSc.IT)*

Supervisors: Dr. Fredrick Mtenzi and Prof. Brendan O'Shea

School of Computing
Dublin Institute of Technology,
Kevin Street, Dublin 8, Ireland.

PhD

A thesis submitted to the Dublin Institute of Technology in fulfillment of the requirements for the degree of Doctor of Philosophy

June, 2013

Abstract

The growth of various technologies in the modern digital world results in the collection and storage of huge amounts of individual's data. In addition of providing direct services delivery, this data can be used for other non-direct activities known as *secondary use*. This includes activities such as doing research, analysis, quality and safety measurement, public health, and marketing. These activities enhance services experiences for individuals, expand knowledge and making appropriate decisions, strengthen understanding about the effectiveness and efficiency of the systems, support public education and aid organisations in meeting customers' needs.

The collected data may contain personal-specific and sensitive information, such as medical records and financial records, that may cause privacy breaches if compromised. The process of ensuring an individual's privacy results in information loss which renders data less useful. This problem is everywhere were data is collected, but the problem is critical in the healthcare domain due to the sensitive nature of the healthcare data and their importance for several secondary uses. Therefore, in order to increase sharing of the collected data, approaches that ensure an individual's privacy with reduced information loss that renders the data useful are needed.

There are number of approaches used to ensure an individual's privacy such as removing Personal Identifiable Information (PII), encryption, and statistical databases. But most of the existing approaches results in substantial information loss or the anonymisation level achieved may still results in the identification of the individual's sensitive information. This research investigates the problem of ensuring an individual's privacy while reducing the amount of information loss. Thus, the research attempts to answer the problem of how the data holders, such as hospitals, private, and government agencies, can ensure an individual's privacy while sharing data which is still useful.

This research proposes an anonymisation algorithm, named *kl-redInfo*, that ensures individual's privacy with a reduced amount of information loss that renders data useful. The *kl-redInfo* algorithm ensures individual's privacy by achieving the main two privacy requirements, *k-anonymity* and *l-diversity*, that aim at ensuring an individual's privacy against both identity and sensitive attribute disclosures. The information loss is reduced by using the three proposed modified approaches that reduce the values of the information loss metrics, which indicate a reduction of the information loss. These approaches are; systematic incorporation of the remaining records in the group that results in lower information loss, using both the group-creation part of the anatomisation approach and cell-based generalisation, and sorting the records according to the attributes that can be linked to identify an individual, also known as quasi-identifier attributes.

The research shows that, each of the proposed modified approaches contribute in reducing the amount of information loss with the approach of systematic incorporation of the remaining records in the group that results in a lower value of the information loss metric being the most important. The research find that, even though each of the proposed modifications contributes in reducing the amount of information loss, the amount of information loss resulting from the application of the combined three proposed modifications is significantly reduced. Therefore, the research uses the three proposed modifications to design the proposed *kl-redInfo* algorithm.

The research shows that, the proposed *kl-redInfo* algorithm results in significant reduction of the information loss compared to the widely used privacy-preserving data publishing algorithms that proved to result in lower information loss. This was indicated by the lower values of the three information loss metrics; Normalised Certainty Penalty (NCP), Discernibility Penalty (DP), and Kullback_Leibler divergence (KL_divergence), that implies reduction in the amount of information loss. The reduction of the information loss resulting from the application of the *kl-redInfo* algorithm was due to the use of the three proposed modified approaches, systematic incorporation of the remaining records in the group that results in a lower amount of information loss; using both group-creation part of the anatomisation approach and cell-based generalisation; and sorting the records according to quasi-identifiers.

Declaration Page

I certify that this thesis which I now submit for examination for the award of Doctor of Philosophy, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for postgraduate study by research of the Dublin Institute of Technology and has not been submitted in whole or in part for another award in any other third level institution.

The work reported on in this thesis conforms to the principles and requirements of the DIT's guidelines for ethics in research.

DIT has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature .... Date ....

Acknowledgements

I would like to express my sincere thanks to my supervisors Dr. Fredrick Mtenzi and Professor Brendan O'Shea. This research could not be of value without their advice and valuable feedback. Their continuous encouragement and support kept me strong. Even though it has been a very challenging life but also was interesting, educational and very rewarding.

I would like to thanks The Institute of Finance Management (IFM) for providing financial support throughout my studies. Also, I would like to thanks all DIT staff, I would like to particularly thanks Dr. Ronan Fitzpatrick for his support and constructive comments towards completion of this thesis. Furthermore, I would also like to thanks IT staff at Muhimbili National Hospital in Tanzania, and my colleagues from IFM who are studying at DIT for their cooperation during my research period.

Lastly, to my family and friends for their love and encouragements throughout the time of this research. That kept me focused, committed, and determined.

List of Abbreviations

DP	Discernibility Penalty
EC	Equivalence Class
KL-divergence	Kullback-Leibler divergence
MNH	Muhimbili National Hospital
NCP	Normalised Certainty Penalty
NIMR	National Institute for Medical Research
PII	Personal Identifiable Information
PPDP	Privacy-Preserving Data Publishing
PPS number	Personal Public Service number
QID	Quasi-identifier Attribute
QIT	Table of Quasi-identifiers
ST	Table of sensitive attributes

Table of Contents

Abstract i

Declaration Page iv

Acknowledgements v

List of Abbreviations vi

List of Figures xii

List of Tables xviii

List of Algorithms xxiii

Chapter 1 INTRODUCTION **1**

1.1 Research background 1

1.1.1 Privacy-Preserving Data Publishing (PPDP) 3

1.1.2 The Anonymisation Approach 8

1.1.3 Assumptions when Anonymising Data 10

1.2 Research Scope 11

1.3 Research Motivation 13

1.4 Statement of the problem 14

1.5 Research Aim and Objectives 16

1.6 Research Methodology 18

1.7 Research Contributions 21

1.8 Research Dissemination 21

1.9 Thesis Organisation 23

Chapter 2 HEALTHCARE DATA IN INFORMATION SOCIETY **24**

2.1 Introduction 24

2.2 Characteristics of Healthcare Data 25

2.2.1 Data Sensitivity 26

2.2.2 Data Diversity 28

2.2.3 Data Volume 28

2.2.4 Data Usefulness 28

2.3 Benefits of Using Healthcare Data 30

2.3.1 Reducing Medical Error 31

2.3.2 Controlling Healthcare Costs 31

2.3.3 Enhancing Quality and Accessibility of the Care Services . . 32

2.3.4 Provision of Information to Physicians 32

2.4 Challenges in Sharing Healthcare Data 33

2.5 Data Protection Laws 34

2.5.1 Data Protection Law in Ireland 34

2.5.2 Data Protection Law in Tanzania 35

2.6 Chapter Summary 37

Chapter 3 PRIVACY-ENHANCING APPROCHES **39**

3.1 Ethical and Legal Frameworks 40

3.2 Policy and Regulatory Frameworks 41

3.3 Privacy-enhancing Technologies (PETs) 41

3.4 Statistical Techniques 42

3.4.1 Query Restriction 43

3.4.2 Input Perturbation 43

3.4.3 Output Perturbation 44

3.5 Cryptographic Techniques 45

3.6 Anonymisation Techniques 46

3.6.1 Privacy Models 46

3.6.2 Anonymisation Approaches 57

3.6.3 Anonymisation Algorithms 65

3.7 Chapter Summary 69

Chapter 4 THE DESIGN OF THE *kl-redInfo* ALGORITHM **70**

4.1 Anatomy Algorithm 71

4.2 The Proposed *kl-redInfo* Algorithm 73

4.3 Algorithm Walkthrough 76

4.4 Key Features of the *kl-redInfo* algorithm 81

4.4.1 Systematic Incorporation of the Remaining Records 81

4.4.2 Using both bucketisation and cell-based generalisation approaches 82

4.4.3 Considers Distribution of the Quasi-identifiers 83

4.5 A solution Architecture 85

4.6 Chapter Summary 86

Chapter 5 EXPERIMENTAL ENVIRONMENT AND SETUP **87**

5.1 Experiment Setup 87

5.2 Datasets and Parameters 88

5.3 Implemented Algorithms 93

5.3.1 The *l-mondrian* Algorithm 94

5.3.2 The *(g-anatomy)* Algorithm 95

5.4 Evaluation Metrics 97

5.4.1 Discernibility Penalty (DP) 98

5.4.2 The Normalised Certainty Penalty (NCP) 99

5.4.3 Kullback-Leibler divergence (KL-divergence) 101

5.5 Software Verification 102

5.5.1 Verification of *kl-redInfo* 102

5.5.2 Verification of *l-mondrian* 105

5.5.3 Verification of *g-anatomy* 108

5.6 Wilcoxon signed-rank test 111

5.7 Chapter Summary 114

Chapter 6 EVALUATION of the *kl-redInfo* **116**

6.1 The Information loss Metrics of the Algorithms without the Proposed Modifications 118

6.2 Proposed Modifications 127

6.2.1 Systematic Incorporation of the Remaining records in the equivalence classes 127

6.2.2 Using both bucketisation and cell-based generalisation approaches 133

6.2.3 Sorting records according to Quasi-identifiers 138

6.3 The Significance of the Proposed Modifications 144

6.4 Comparison of *kl-redInfo* with the Existing Algorithms . . 151

6.5 Other Experiments and Findings 160

6.5.1 Impact of the different datasets size 160

6.5.2 Impact of the different Parameter values 164

6.5.3 Algorithms Perfomance 165

6.6 Chapter Summary 166

Chapter 7 CONCLUSIONS AND FUTURE WORK **167**

7.1 Conclusions 167

7.2 Research Contributions 169

7.3 Benefits from this Research 172

7.4 Future Work 174

REFERENCES 177

Chapter Appendix A Information Classification **190**

Chapter Appendix B Glossary of Terms **192**

Chapter Appendix C The Information loss metrics of the algorithms without the proposed modifications **196**

Chapter Appendix D The information loss metrics of the *kl-redInfo* with sequential and systematic incorporation approaches **201**

Chapter Appendix E The information loss metrics of the *kl-redInfo* with generalisation and with both bucketisation and generalisation **205**

Chapter Appendix F The information loss metrics of the *kl-redInfo* without and with sorting approach on Adult dataset **209**

Chapter Appendix G The information loss metrics of the *kl-redInfo* without and with the proposed modifications on Adult dataset **213**

Chapter Appendix H The information loss metrics of the *kl-redInfo* with proposed modifications, *l-mondrian*, and *g-anatomy* **217**

Chapter Appendix I The information loss metrics of the *kl-redInfo* algorithm, on the different size of the Adult dataset **222**

List of Figures

Figure 1.1 **Aspects of privacy (Stallings and Brown, 2012)** 5

Figure 1.2 **Privacy Preserving Data Publishing Phases (Fung et al., 2010)** 7

Figure 2.1 **Stakeholders of the healthcare data (Appari and Johnson, 2010)** 29

Figure 3.1 **Generalisation hierarchy of the Marital status attribute (Samarati, 2001)** 58

Figure 4.1 **A Solution Architecture (Author, 2013)** 85

Figure 5.1 **Generalisation hierarchies of the QIDs used in this research (LeFevre et al., 2005)** 91

Figure 5.2 **Screenshot of the 10 records PatInfo dataset to be anonymised** 102

Figure 5.3 **Screenshot of the values of the information loss metric resulting from the application of the *kl-redInfo* algorithm on 10 records PatInfo dataset** 103

Figure 5.4 **Screenshot of the values of the information loss metric resulting from the application of the *l-mondrian* algorithm on 10 records PatInfo dataset** 106

Figure 5.5 **Screenshot of the values of the information loss metric resulting from the application of the *g-anatomy* algorithm on 10 records PatInfo dataset** 109

Figure 6.1 **DP resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* algorithm without the proposed modifications on the PatInfo dataset** 119

Figure 6.2 **NCP resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* algorithm without the proposed modifications on the PatInfo dataset** 120

Figure 6.3 **KL-divergence resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* algorithm without the proposed modifications on the PatInfo dataset** 121

Figure 6.4 **DP resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the PatInfo dataset** 129

Figure 6.5 **NCP resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the PatInfo dataset** 130

Figure 6.6 **KL-divergence resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the PatInfo dataset** 131

Figure 6.7 **DP resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the PatInfo dataset** 134

Figure 6.8 **NCP resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the PatInfo dataset** 135

Figure 6.9 **KL-divergence resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the PatInfo dataset** 136

Figure 6.10 **DP resulting from the application of the *kl-redInfo* without and with sorting approach on the PatInfo dataset** . 139

Figure 6.11 **NCP resulting from the application of the *kl-redInfo* without and with sorting approach on the PatInfo dataset** . 140

Figure 6.12 **KL-divergence resulting from the application of the *kl-redInfo* without and with sorting approach on the PatInfo dataset** 141

Figure 6.13 **DP of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modified approaches on the PatInfo dataset** 152

Figure 6.14 **NCP of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modified approaches on the PatInfo dataset** 153

Figure 6.15 **KL-divergence of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modified approaches on the PatInfo dataset** 154

Figure 6.16 **DP resulting from the application of the *kl-redInfo* algorithm on different size of the PatInfo dataset when $k = l = 6$** 161

Figure 6.17 **NCP resulting from the application of the *kl-redInfo* algorithm on different size of the PatInfo dataset when $\boldsymbol{k = l = 6}$** 162

Figure 6.18 **KL-divergence resulting from the application of the *kl-redInfo* algorithm on different size of the PatInfo dataset when $\boldsymbol{k = l = 6}$** 163

Figure A.1 **Information Classification (Al-Fedaghi and Al-Azmi, 2012)** 190

Figure C.1 **DP resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* algorithm without the proposed modifications on the Adult dataset** 196

Figure C.2 **NCP resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* without the proposed modifications algorithms on the Adult dataset** 197

Figure C.3 **KL-divergence resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* without the proposed modifications algorithms on the Adult dataset** 198

Figure D.1 **DP resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset** 201

Figure D.2 **NCP resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset** 202

Figure D.3 **KL-divergence resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset** 203

Figure E.1 **DP resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the Adult dataset** 205

Figure E.2 **NCP resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the Adult dataset** 206

Figure E.3 **KL-divergence resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the Adult dataset** 207

Figure F.1 **DP resulting from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset** . . 209

Figure F.2 **NCP resulting from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset** . . 210

Figure F.3 **KL-divergence resulting from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset** 211

Figure G.1 **DP resulting from the application of the *kl-redInfo* without and with the proposed modifications on the Adult dataset** 213

Figure G.2 **NCP resulting from the application of the *kl-redInfo* without and with the proposed modifications on the Adult dataset** 214

Figure G.3 **KL-divergence resulting from the application of the *kl-redInfo* without and with the proposed modifications on the Adult dataset** 215

Figure H.1 **DP resulting from the application of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modifications on the Adult dataset** 217

Figure H.2 **NCP resulting from the application of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modifications on the Adult dataset** 218

Figure H.3 **KL-divergence resulting from the application of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modifications on the Adult dataset** 219

Figure I.1 **DP resulting from the application of the *kl-redInfo* algorithm on different size of the Adult dataset when $\boldsymbol{k = l = 6}$** 222

Figure I.2 **NCP resulting from the application of the *kl-redInfo* algorithm on different size of the Adult dataset when $\boldsymbol{k = l = 6}$** 223

Figure I.3 **KL-divergence resulting from the application of the *kl-redInfo* algorithm on different size of the Adult dataset when $\boldsymbol{k = l = 6}$** 223

List of Tables

Table 3.1 **Privacy Models (Fung et al., 2010)** 48

Table 3.2 **Patients' Information Table** 49

Table 3.3 **2-Anonymous Table** 50

Table 3.4 **2-diversity Table** 53

Table 3.5 **QIT Table** 63

Table 3.6 **QIT and ST** 63

Table 3.7 **Buckets of the distinct sensitive attributes** 64

Table 3.8 **Characteristics of the Existing Algorithms (Author, 2013)** 65

Table 4.1 **Patients' Information Table** 77

Table 4.2 **Records sorted according to QID values** 78

Table 4.3 **Records grouped according to Sensitive attribute** . . . 79

Table 4.4 **The First Bucket** 80

Table 4.5 **Anonymised records as a result of *kl-redInfo* algorithm** 81

Table 4.6 **Characteristics of the Existing Algorithms compared to the Proposed (Author, 2013)** 84

Table 5.1 **The PatInfo dataset schema** 89

Table 5.2 **The Adult dataset schema** 90

Table 5.3 **Characteristics of the Implemented Algorithms** 97

Table 5.4 **The example of the Wilcoxon signed-rank test results tables** . . . 113

Table 6.1 **The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and *l-mondrian* on the PatInfo dataset** . . . 124

Table 6.2 **The comparison results for the values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and *l-mondrian* on the PatInfo dataset** . . . 124

Table 6.3 **The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and the *g-anatomy* on the PatInfo dataset** . 125

Table 6.4 **The comparison results for the values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modified approaches and *g-anatomy* on the PatInfo dataset** . . . 126

Table 6.5 **The comparison results for the values of the information loss metrics resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the PatInfo dataset** . . . 132

Table 6.6 **The comparison results from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation approaches on the PatInfo dataset** 137

Table 6.7 **The comparison results from the application of the *kl-redInfo* without and with sorting approach on PatInfo dataset**142

Table 6.8 **The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm when records sequentially and systematically incorporated on the PatInfo dataset** 145

Table 6.9 **The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm with generalisation and with both bucketisation and cell-generalisation approaches on the PatInfo dataset** 146

Table 6.10 **The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm without and with sorting approach on the PatInfo dataset** 147

Table 6.11 **The values of the information loss metrics resulting from the application of the *kl-redInfo* without and with the proposed modified approaches on the PatInfo dataset** . . . 149

Table 6.12 **The comparison results from the application of the *kl-redInfo* without and with the proposed modified approaches on the PatInfo dataset** 149

Table 6.13 **The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modified approaches and *l-mondrian* on the PatInfo dataset** 156

Table 6.14 **The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modified approaches and *g-anatomy* on the PatInfo dataset** 157

Table 6.15 **The comparison results from the application of the *l-mondrian* and *kl-redInfo* with the proposed modified approaches on the PatInfo dataset** 158

Table 6.16 **The comparison results from the application of the *g-anatomy* and *kl-redInfo* with the proposed modified approaches on the PatInfo dataset** 159

Table A.1 **Information Classification** 191

Table B.1 **Glossary of Terms** 195

Table C.1 **The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and *l-mondrian* algorithm on the Adult dataset** 199

Table C.2 **The comparison results from the application of the *kl-redInfo* algorithm without the proposed modifications and *l-mondrian* on the Adult dataset** 199

Table C.3 **The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and *g-anatomy* algorithm on the Adult dataset** 200

Table C.4 **The comparison results from the application of the *kl-redInfo* algorithm without the proposed modifications and *g-anatomy* on the Adult dataset** 200

Table D.1 **The values of the information loss metrics resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset** . 204

Table D.2 **The comparison results from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset** 204

Table E.1 **The values of the information loss metrics resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-generalisation approaches on Adult dataset** 208

Table E.2 **The comparison results from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the Adult dataset** 208

Table F.1 **The values of the information loss metrics resulting from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset** 212

Table F.2 **The comparison results from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset** 212

Table G.1 **The values of the information loss metrics resulting from the application of the *kl-redInfo* without and with the proposed modifications on the Adult dataset** 216

Table G.2 **The comparison results from the application of the *kl-redInfo* algorithm without and with the proposed modifications on the Adult dataset** 216

Table H.1 **The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modifications and *l-mondrian* on the Adult dataset** 220

Table H.2 **The comparison results from the application of the *kl-redInfo* algorithm with the proposed modifications and *l-mondrian* on the Adult dataset** 220

Table H.3 **The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modifications and *g-anatomy* on the Adult dataset** 221

Table H.4 **The comparison results from the application of the *kl-redInfo* algorithm with the proposed modifications and *g-anatomy* on the Adult dataset** 221

List of Algorithms

1 **Anatomy algorithm (Xiao and Tao, 2006b)** 71

2 **The proposed *kl-redInfo* algorithm (Author, 2013)** 75

3 ***l-mondrian* algorithm (LeFevre et al., 2008)** 95

4 ***g-anatomy* algorithm (Xiao and Tao, 2006b)** 96

Chapter 1

INTRODUCTION

1.1 Research background

The development of various technologies in the modern digital world result in the collection and storage of huge amounts of individual's data. Different individual's data, such as medical records, financial records, and academic records are collected and stored by both public institutions and private companies. There are several purposes of collecting these data.

A first example of such purpose is related to research. A number of organisations such as medical institutions and statistical agencies collect and disseminate data so that not only themselves but also external analysts can use these data for research purposes, for decision making, and for many other uses. For example, in the healthcare domain the availability of such data helps to prevent medical errors and enhances patient care, healthcare economy and healthcare research (Pommerening and Michael, 2004).

A Business-oriented focus is another example of the purpose of collecting data. Different private companies collect information about clients, about other companies and product, so that they can classify or predict clients' behaviors (data mining), or they can compare themselves with the rival companies. This information helps companies to determine future market strategies (Herranz and Nin, 2010).

Also, the data can be collected for security purposes. Information about people, purchases, trips, and personal communications is stored to decrease or detect possible security risks for the society or implement control policies. Therefore, this data is useful in providing better quality services to individuals and its availability is crucial in several activities such as education, planning and decision-making (Safran et al., 2007).

All these examples illustrate the benefits that a society can obtain by the development of different technologies. However, the collected data may contain information about an individual that can be linked or be linkable to identify an individual (Sweeney, 1997; Samarati and Sweeney, 1998; Klimavicz, 2007). This information, known as Personal Identifiable Information (PII), includes any information that can be used to distinguish or trace an individual's identity, such as name, dates and place of birth, mother's maiden name, biometric records, and any other personal information such as age, gender, race, ethnicity, dates and place of residence (HIPAA, 1996). This data needs to be protected as its disclosure may result in identity theft, embarrassment, or blackmail of an individual, while an organisation may lose its public trust, legal liability, or may result in high costs to handle the breaches.

Protecting PII results in information loss which renders data less useful. Therefore, a balance between the two is important in order to ensure both data protection and data usefulness. Currently more emphasis is on ensuring an individual's privacy while the usefulness of such data has not been well-considered i.e., protection of

PII has received much more attention than the usefulness of the data. As a result, the concept of data usefulness is less considered. When applying data protection techniques, preserving the data is a continuous trade-off between strengthening the protection of PII and maintaining an adequate level of data usefulness.

Protecting PII while reducing the amount of information loss is still a challenging problem in the modern digital world. Several techniques have been proposed to address the problem of ensuring data protection while at the same time supporting legitimate use of the data, ranging from cryptographic approaches (Quantin et al., 2000; Hou and Tu, 2005) to perturbation approaches (Muralidhar et al., 1999, Nunez et al., 2007). But most of the existing techniques provide data protection while causing a high level of information loss, which reduces the usefulness of the data. Therefore, it is important to develop techniques that ensure protection of PII when sharing useful data. This undertaking is called Privacy-Preserving Data Publishing (PPDP).

1.1.1 Privacy-Preserving Data Publishing (PPDP)

The availability of huge numbers of databases which record a large variety of an individual's information, increase concern for privacy in the modern digital world. This makes it possible to discover information about a specific individual by simply connecting a certain number of the available databases.

"Privacy" is a term used with many meanings. Therefore, it is very hard to define; it is commonly used for anything from the state of being alone or undisturbed, our freedom from interference or public attention, up to the right of anonymity (Wright, 2004).

Gavison (1979) defines privacy in three inter-related kinds of privacy: secrecy, anonymity, and solitude. Secrecy concerns information that others may gather about us. Anonymity addresses how much intently we are in the public, and solitude measures the degree to which others have physical access to us.

The term confidentiality and privacy are often used as synonyms, but they are different concepts. Data confidentiality is about difficulty or impossibility of unauthorised users to learn any information about the data. Usually confidentiality is achieved by enforcing an access policy and possibly using crytographic algorithms. Privacy relates to the data that can be safely disclosed without leaking sensitive information about individuals who are subjects of the data.

The most common definition of privacy is the one defined by Westin (1968), "Privacy is the claim of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others". Therefore, in other words, privacy is a 'claim', 'entitlement' or 'right' of an individual to determine what personal information may be communicated to others. This definition of privacy is difficult to be achieved in the modern digital world where data is collected in every action we take.

Therefore, for the purposes of this research, a privacy definition as defined in relation to 'control' over access to an individual's information is adopted. The research defines *privacy* as the provision of control regarding the use and disclosure of individual's information. This is based on the state or condition of limited access to individual data (Walters, 2002).

An individual's privacy can be preserved in many different aspects which differ in scope, properties, and limitation. The main privacy aspects are anonymity, pseudonymity, unlinkability, and unobservability as summaried in Figure 1.1 (Stallings and Brown, 2012). Anonymity means an individual is not identifiable within a set of data subjects; Pseudonymity means it is not possible to identify true identity of an individual; Unlinkability means it is not possible to relate the data with an individual; and Unobservability means it is not possible to identify if an individual's information is on the shared dataset (Pfitzmann and Hansen, 2010).

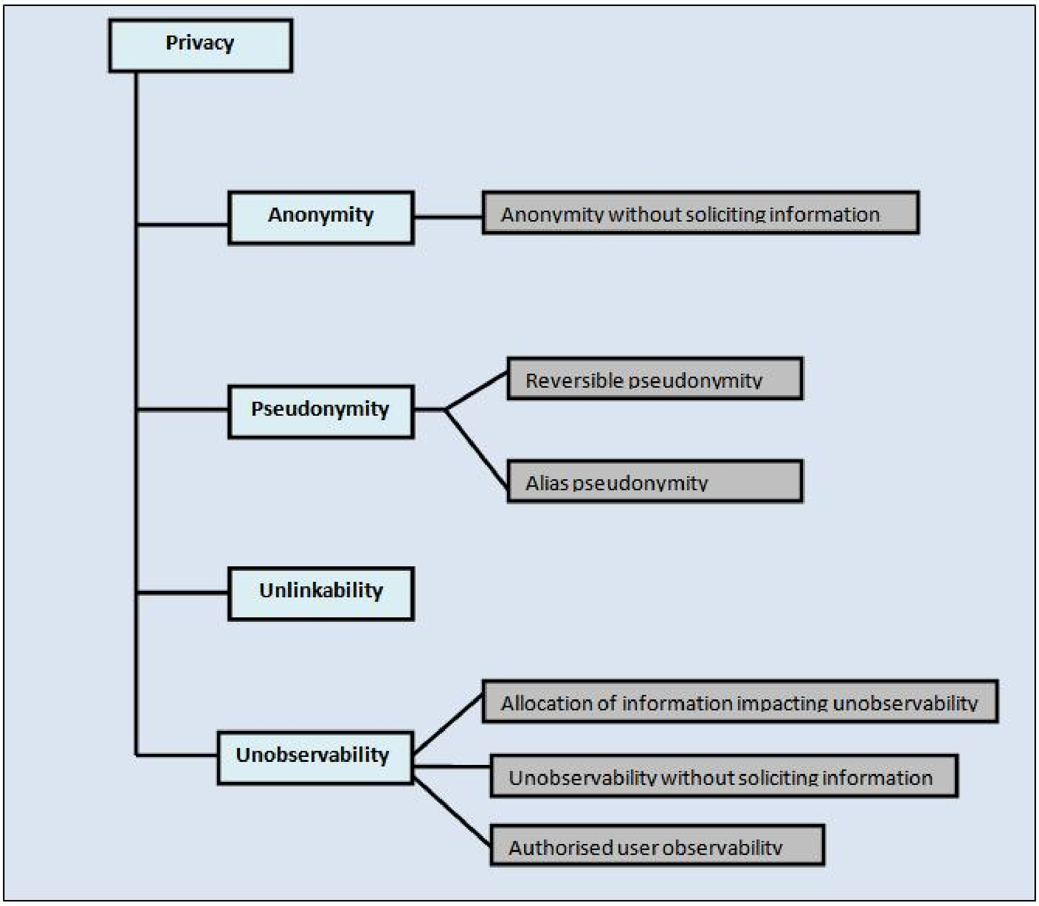

Privacy
Anonymity
Anonymity without soliciting information
Pseudonymity
Reversible pseudonymity
Alias pseudonymity
Unlinkability
Unobservability
Allocation of information impacting unobservability
Unobservability without soliciting information
Authorised user observability

Figure 1.1: Aspects of privacy (Stallings and Brown, 2012)

Privacy-Preserving Data Publishing (PPDP) is one of the broad areas of privacy-preserving that deals with anonymising the data so that its privacy remains preserved when shared for different purposes. When sharing data it is necessary to prevent the sensitive information of the individuals from being disclosed. There are two main types of information disclosure identified in the literature: identity disclosure and sensitive attribute disclosure (Dalenius and Reiss, 1982; Kim, 1986; Lambert, 1993). Identity disclosure occurs when an individual is linked to a particular record in the shared dataset. Sensitive attribute disclosure occurs when new information about some individual is revealed, i.e., the shared data makes it possible to infer the characteristics of an individual more accurately than would be possible before the data is shared.

Privacy Preserving Data Publishing has two main phases; data collection phase and data publishing phase (Fung et al., 2010). Figure 1.2 describes a typical scenario for data collection and publishing. The scenario starts by the data holder to collect data from data subjects (e.g., Esther, Ted, Joseph or Angela) in the data collection phase. This is followed by the data publishing phase where, the data holder releases the collected data to the data recipient, who will then use the shared data for different purposes. For example, a hospital collects data from patients and releases the patient records to an external medical center. In this example, the hospital is the data holder, patients are the data subjects, and the medical center is the data recipient. The medical center could use the shared data for different purposes from a simple count of the number of men/women with a certain disease, to a sophisticated data analysis.

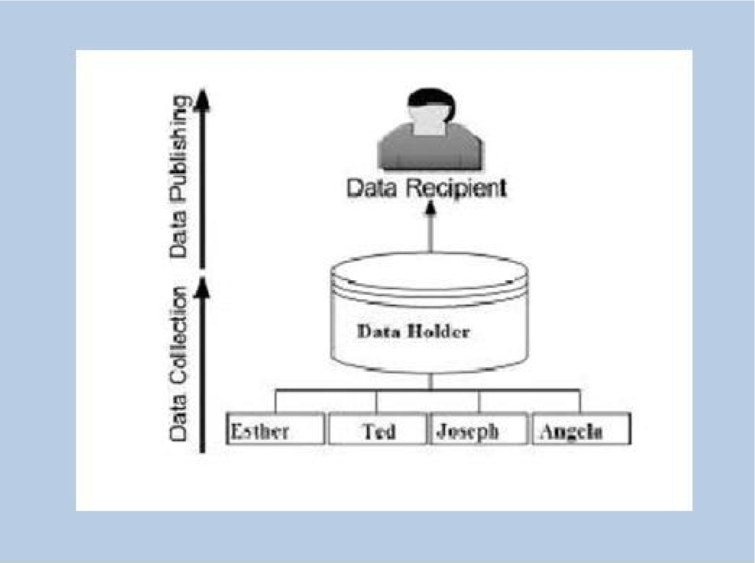


Figure 1.2: Privacy Preserving Data Publishing Phases (Fung et al., 2010)

There are two types of data holders; untrusted and trusted data holders (Gehrke, 2006). An untrusted data holder is not trusted and may attempt to identify sensitive information from the data subject. In this type of data holder, the privacy issues are mainly considered in the data collection phase. Various cryptographic solutions (Quantin et al., 2000; Hou and Tu, 2005); anonymous communications (Chaum, 1981); and statistical approaches (Muralidhar et al., 1999; Nunez et al., 2007) were proposed to collect records anonymously from their data subjects without revealing the data subjects' identity.

In the trusted type of data holder, the data holder is trustworthy and data subjects are willing to provide their personal information to the data holder, for example a doctor. However, the trust is not automatically passed to the data recipient. This research assumes the data holder is trustworthy and thus the privacy issues are considered in the data publishing phase.

In most cases the collected data is stored in a table form consisting of Unique Identifier, Quasi Identifier, Sensitive Attributes, and Non-Sensitive Attributes. The Unique Identifier is a set of attributes containing information that explicitly identifies the data subject, such as name and Personal Public Service number (PPS number) . The Quasi-Identifier (QID) is a set of attributes that could potentially identify the data subject such as gender, race and marital status. The Sensitive Attributes consists of sensitive person-specific information such as disease, salary, and disability status. The Non-Sensitive Attributes contains all attributes that are not considered sensitive by the data subject and whose release is not harmful (e.g., Favorite color) (Samarati, 2001). The classification of these information is presented in Appendix A.

1.1.2 The Anonymisation Approach

This research is based on the specific Privacy-Preserving Data Publishing (PPDP) approach known as anonymisation that deals with removing the association between the Personal Identifiable Information (PII) and the individual person. The anonymisation approach alters data in order to make it impossible to link individuals with their data. The approach seeks to protect the identity and/or the sensitive data of the data subjects when data is shared for different purposes (Gavish and Gerdes Jr, 1998; Aggarwal et al., 2005; Pfitzmann and Hansen, 2008).

Usually the unique identifiers of data subjects are removed before sharing the data. But removing individual's unique identifiers information does not guarantee the protection of the shared data (Samarati, 2001; Sweeney, 2002; Zielinski, 2007). Since the shared data often contains other information, which is known as quasi-identifier (QID) such as gender, marital status and race that can be linked or matched to publicly available information or by looking at unique characteristics found in the fields and records of the database itself to identify an individual.

Sweeney (2002a) showed a real-life privacy threat to the former governor of the state of Massachusetts, William Weld. In Sweeney's example, an individual's name in a public voter list was linked with his record in a published medical database through the combination of zip code, date of birth, and gender. Each of these attributes does not uniquely identify a data subject, but their combination often singles out a unique or a small number of data subjects. Sweeney showed that 87% of the United States of America (USA) population had reported characteristics that likely made them unique based only on zip code, date of birth, and gender. This is known as a *linking attack*, which is currently a serious problem due to the increase in the computational power available and easy accessibility of large amount of information.

To prevent *linking attacks*, the data holder provides an anonymous dataset by applying different anonymisation techniques to the QID attributes in the original dataset. Anonymisation techniques hide some detailed information so that several records become indistinguishable from each other with respect to QIDs. Consequently, if a person is linked to a record through QIDs, that person is also linked to all other records that have the same value for QID, making the linking ambiguous.

Alternatively, anonymisation techniques could generate a synthetic dataset based on the statistical properties, or add noise to the original dataset. The aim of an anonymisation approach is to produce an anonymous dataset that satisfies given privacy requirements determined by the chosen privacy model and to retain as much data as possible. Different information metrics such as Normalised Certainty

Penalty (NCP), Discernibility Penalty (DP), and Kullback-Leibler divergence (KL-divergence), are used to measure the usefulness of the anonymous dataset. Note that the Non-Sensitive Attributes are published if they are important to the purpose.

1.1.3 Assumptions when Anonymising Data

Reducing the information loss while sharing data for different purposes is still a challenging problem in the Privacy Preserving and Data Publishing domain. Anonymising data has become more difficult due to the following four desirable assumptions that have to be achieved (Fung et al., 2010):

- *It is difficult to know how the data recipient will use the data.*

  Sometimes, the data holder does not even know who are the recipients at the time of sharing the data. Therefore, it is difficult to know how the recipients will use the data.

- *The data recipient could be an attacker.*

  The contracts and agreements cannot guarantee later misplacement of the sensitive data which in turn could cause the data to end up in the wrong hands. For example, the data recipient, may be a trustworthy drug research company; however, it is difficult to guarantee that all staff in the company are also trustworthy. This property makes the anonymising process different from the encryption and cryptographic approaches, in which only authorised and trustworthy recipients are given the private key for accessing the cleartext. The major challenge in anonymising data is to simultaneously preserve both the privacy and the information usefulness in the shared data.

- *Data should be shared at an individual level and not the group of individuals.* Releasing microdata (personal data in its raw or non-aggregate form) will have lower information loss than sharing aggregate results. Thus the data will be more useful, but this may result in the breach of individual's privacy.

- *Released data should remain truthful.* In most of the data sharing scenarios, it is important that each shared record corresponds to an existing individual in real life. For example, a pharmaceutical researcher (the data recipient) may need to examine the actual patient records to discover some previously unknown side effects of the tested drug. If a shared record does not correspond to an existing patient in real life, it is difficult to deploy results in the real world. Randomised and synthetic data do not meet this requirement (El Emam, 2008). Although an encrypted record corresponds to a real life patient, the encryption hides the semantics required for acting on the represented patient.

1.2 Research Scope

This research is closely related with Privacy-Preserving Data Mining (PPDM). The main idea of PPDM is to extend data mining techniques to work with the modified data to mask sensitive information (Agrawal and Srikant, 2000). The key issues are how to modify the data and how to recover the data mining results from the modified data. Unlike PPDM solutions that are mainly based on the data mining task under consideration, Privacy-Preserving Data Publishing (PPDP) may not be tied to a specific data mining task since the task may be unknown at the time of the data publishing. Furthermore, most of the PPDP solutions emphasise preserving the data truthfulness at the record level, but often PPDM solutions do not preserve such a property (Aggarwal and Yu, 2008). In recent years, the term 'PPDM' has evolved to cover many other privacy research problems, even though some of them may not directly relate to data mining.

The non-interactive query model in statistical disclosure control (Adam and Worthmann 1989) is another related area of this research, in which the data recipients can submit one query to the system. This type of non-interactive query model may not fully address the information needs of data recipients because, in some cases, it is very difficult for a data recipient to accurately construct a query for a data mining task in one attempt. Consequently, there are a series of studies on the interactive query model (Blum et al., 2008; Dwork, 2006; Dinur and Nissim, 2003), in which the data recipients, unfortunately including attackers, can submit a sequence of queries based on previously received query results. The main limitation of any privacy-preserving query system is that it can only answer a sublinear number of queries in total; otherwise, an attacker (or a group of corrupted data recipients) will be able to reconstruct a large part of the original data (Blum et al., 2008), which is a strong violation of privacy.

This research focuses on a technical problem within the broad domain of privacy protection and de-identification in the data publishing domain. The problem of data protection encompasses many legal, ethical, and technical issues surrounding data ownership, collection, dissemination, and use. Specifically, it investigates the problem of anonymising data with reduced information loss that renders data useful. The problem is currently serious due to the increasing pressure of the data sharing as a result of technology growth. There is much to gain from data sharing, for example healthcare data can be shared with insurance companies, government, researchers, employers, state bureaus of vital statistics, pharmacy benefit managers (companies that track doctors drug prescriptions), local retail pharmacies, attorneys, and others to improve healthcare services (Riedl et al. 2008). However, disclosure of Personal Identifiable Information (PII) may results in privacy breach, prevents each party from sharing data with others. The research focuses on several key issues in Privacy-Preserving Data Publishing (PPDP) such as privacy models, to be discussed in section 3.6.1; anonymisation approaches to be discussed in section 3.6.2; information loss metrics to be discussed in section 5.4; and anonymisation algorithms to be discussed in section 3.6.3.

1.3 Research Motivation

The application of different technologies in different domains results in the collection and storage of large amounts of data. Usefulness of the collected data is reduced due to the presence of Personal Identifiable Information (PII), which has to be protected. Protection of PII usually causes information loss which leaves the data less useful. This can result in a data-rich but information-poor problem.

There is much to gain by allowing access to collected data. The collected data such as medical records may have many reasonable uses serving different purposes inside and outside of the specific domain in which it has initially been collected. The advantages include, doing academic or commercial research, public healthcare, and policy making (Mills et al., 2003). Also, access to sufficient information will enable researchers discover, analyse and predict correct trends and thus can improve all types of decisions by the use of decision support technologies (Goldschmidt, 2005).

There is an increase of data breaches and privacy aweareness which increases the need for privacy. This increases privacy protection and hence decreases sharing of the data. Although privacy and data usefulness are duals of each other, privacy has received much more attention than the sharing of the data. As a result, the concept of data sharing is less considered. Therefore, a technique that insure individual's privacy with lower information loss that renders data useful is important. So, the research motivation for this research is to allow sharing of the data without violating individual's privacy.

1.4 Statement of the problem

In the modern digital world, effective information sharing between individuals and organisations has become a vital requirement. This increases the demand on both data sharing and individual's privacy. The presence of Personal Identifiable Information (PII) such as medical records, financial records and academic records have been identified as a main barrier to data sharing. This limits sharing of the data for different purposes such as academic or commercial research, which are important for supporting various activities in society such as improving public healthcare and policy making.

The problem of how to effectively and efficiently share this data without disclosing PII is still a major challenge. A number of approaches, such as anonymisation, statistical database and encryption, emerged to solve the problem, but this is achieved with substantial information loss. Therefore, there is a problem in sharing microdata while protecting the privacy of the data subjects. The main challenge when disclosing information is to provide as much information as possible while guaranteeing an individual's privacy (Zhang et al., 2007; Zielinski, 2007). This means, limiting disclosure of the shared data requires a careful consideration between the data usefulness and individual's privacy.

Research Questions

This research problem can be represented by using the following main research questions:

1. How can data holders preserve an individual's privacy with reduced information loss that renders data useful?

2. What are the causes of the information loss in the existing algorithms?

3. What approaches can be put in place in order to reduce the amount of information loss while still striving for the individual's privacy?

4. How can anonymisation approaches be designed, developed, and implemented in order to improve individual's privacy and usefulness of the data beyond that provided by a single approach?

This research is based on the idea that: Designing an anonymisation algorithm by using more than one anonymisation approach can be an effective and practical tool for reducing the amount of information loss when ensuring an individual's privacy. The idea was originally presented by the author at the conference of the Healthcare Information Society of Ireland (HISI) and thereafter published in the conference journal as Tinabo et al. (2009b).

1.5 Research Aim and Objectives

The aim of this research is to investigate the problem of anonymising data with minimal information loss that renders data useful. Thus, the research attempts to answer the question of how data holders, such as hospitals, private and government agencies, can release data for different purposes while preserving individual's privacy. Based on these answers, the research proposes an anonymisation algorithm as a solution to the problem. The algorithm is named *kl-redInfo* as it achieves the two main privacy requirements, *k-anonymity* and *l-diversity*, with reduced information loss. To accomplish this aim, the following research objectives will be addressed:

1. *To establish state-of-the-art of the existing techniques*

   A Literature review was conducted in order to establish state-of-the-art of the existing techniques. Most important is to identify characteristics of the existing techniques and identify causes of the information loss. The identified causes of the information loss contributed in proposing approaches that reduce the information loss. Summary of the characteristics of the existing techniques are discussed in Tinabo et al. (2009a).

2. *To analyse and understand the data protection and data usefulness issues*

   Data protection and data usefulness are two conflicting ideas. Therefore, analysis and understanding of these two main issues in this research is important. The problem starts by the need of sharing data for different purposes, such as research, analysis and public education. The presence of Personal Identifiable Information (PII) which its disclosure may result in breach of individual's privacy, makes sharing difficult. Several techniques have been proposed to ensure privacy but this results in information loss which reduce usefulness of the data. Therefore, techniques that ensure individual's privacy with reduced information loss are important. The knowledge obtained from this analysis is used to propose an algorithm which ensures individual's privacy with reduced information loss.

3. *To design the proposed, kl-redInfo, anonymisation algorithm*

   The research designs the proposed algorithm, and names it *kl-redInfo.* The *kl-redInfo* achieves the main privacy requirements, *k-anonymity* and *l-diversity*, with reduced information loss. The algorithm is designed by using the approaches of systematic incorporation of the remaining records, bucketisation and cell-based generalisation together with sorting the records according to quasi-identifier attributes approaches. Using all these approaches significant reduce the amount of information loss.

4. *To evaluate the algorithm*

   To evaluate the proposed *kl-redInfo* algorithm, the research compares the information loss resulting from the application of the *kl-redInfo* algorithm with the widely used algorithms, *l-mondrian* and *g-anatomy* that proved to result in lower information loss. In order to achieve this the analysis and understanding of the existing evaluation metrics is crucial. This results in selection of the three information loss metrics used in this research; Discernibility Penalty (DP) (Bayardo and Agrawal, 2005), Normalised Certainty Penalty (NCP) (Xu et al., 2006) and Kullback_Leibler divergence (KL_divergence) (Kifer and Gehrke, 2006) .

1.6 Research Methodology

This research adopts the design science research paradigm (March and Smith, 1995; Hevner et al., 2004). Design science research relies on the methods used to answer research questions, test research hypotheses and the careful application of these methods (Hevner et al., 2004; Peffers et al., 2007). Therefore, several research methods were conducted in order to achieve the research aim and objectives. The methods used to accomplish this research includes, literature review, data collection, algorithm design and development, implementation, evaluation and validation.

Literature Review

A literature review has been conducted in order to establish what is the state-of-the-art and to draw from the existing theories and knowledge to devise a solution to the defined problem. The literature review is based on the aim and objectives of the research. This leads to the foundation of the detailed description of the existing anonymisation techniques and their limitations as discussed in Chapter 3; and selection of criteria used to measure usefulness of the data as discussed in Chapter 5.

Data Collection

Even though every domain has a problem of ensuring an individual's privacy when sharing data, this research uses healthcare domain as a case study. This is due to the sensitive nature of its data and its importance of sharing the data for secondary uses. Therefore, the process of getting real healthcare data for evaluation purpose was done, but due to data protection issues the use of real data was not possible, as discussed in Section 2.5.2. Thus, this research uses simulated patients' medical dataset, named *PatInfo*, generated by using Data Generator software downloaded from *http://www.generatedata.com/#about.* The schema of the dataset, as discussed in Chatpter 5, is based on the schema of the Muhimbili National Hospital (MNH) in Tanzania where the survey was done.

To show that the *kl-redInfo* algorithm also works in real datasets, as discussed in Chapter 6 this research also used the real-world census dataset, called Adult dataset, downloaded from UCI Machine Learning Repository at *http://archive.ics.uci.edu/ml/datasets/Adult.* This dataset was selected as it is one of the widely used datasets in previous research, and it has most of the information that can be found in any healthcare domain such as age, gender, marital status and address.

Algorithm Design

The design of the proposed *kl-redInfo* algorithm is divided into two phases; the high-level design phase and the detailed design phase. The high-level design phase of the *kl-redInfo* algorithm involves outlining the key components required to form a complete solution. The main outcome in this phase is the solution architecture presented in Section 4.5. The solution architecture represents key components that form the complete solution. These components are a database component, an algorithm engine component, and a user interface component. Generally, these components aim at describing a holistic solution of the problem.

The detailed design phase involves consolidating the *kl-redInfo* algorithm and design of the key components of the solution architecture. The detailed design algorithm contains detailed steps sufficient for implementing an algorithm engine. These details include, clarification of entry and exit points for each approach employed in the algorithm, detailed steps for satisfying both *k-anonymity* and *l-diversity* privacy requirements, and their relationship. The main outcome of this phase is the proposed anonymisation algorithm, named *kl-redInfo*. The detailed *kl-redInfo* algorithm is discussed in Chapter 4.

Implementation

The solution architecture and all consolidated outcomes from the algorithm design phase were used to form a complete testable solution. Tools used during implementation include; Mysql open source relational database management system for back-end and Java programming language for front-end. The *kl-redInfo* algorithm was implemented on the algorithm engine component. A complete working software was evaluated using three different evaluation metrics including Discernibility Penalty (DP), Normalised Certainty Penalty (NCP), and Kullback-Leibler divergence (KL-divergence) as discussed in Chapter 5.

Evaluation

An experimental approach was used to evaluate the *kl-redInfo* algorithm. The comparison was done between the *kl-redInfo* and the widely used algorithms, *l-mondrian* and *g-anatomy* that proved to result in lower information loss. The algorithm was evaluated by calculating the information loss using the three evaluation metrics; Discernibility Penalty (DP), Normalised Certainty Penalty (NCP), and Kullback-Leibler divergence (KL-divergence). The *kl-redInfo* algorithm provides an anonymity table that achieves both *k-anonymity* and *l-diversity* with reduced information loss, as shown in Chapter 6. The lower information loss implies the lower distortation of the original data, therefore the data remains useful.

1.7 Research Contributions

This research contributes the following to the body of knowledge:

- The research proposes an anonymisation algorithm, named *kl-redInfo* that ensures individual's privacy with reduced information loss which renders data useful.
- The research identifies causes of the information loss and proposes modified approaches that can be used to reduce the amount of information loss.
- The research also quantifies the amount of information loss reduced by each of the proposed modified approaches and algorithms.

1.8 Research Dissemination

As part of research dissemination, two conference papers, one journal article and one extended abstract were published in relation to this research. The other three publications are the collaborated work with other colleagues.

1. **Tinabo, R.**; Mtenzi, F.; O'Driscoll, C.; and O'Shea, B. (2010), *"Multiple Anonymisation Technique can Balance Data Usefulness and Protection of Personal Identifiable Information (PII)"*, The International Journal of Web Application (IJWA), Volume 1, Issue 4.
2. **Tinabo, R**, Mtenzi, F.; O'Driscoll, C.; and O'Shea, B. (2010), *"Anonymisation vs. Pseudonymisation: Which one is most Useful for both Privacy Protection and Usefulness of E-healthcare Data"*, The 4th International Conference for Internet Technology and Secured Transactions (ICITST), London, United Kingdom.

3. **Tinabo, R**, Mtenzi, F.; and O'Shea, B. (2009), *"Solving the problem of Balancing Data Usefulness and Protection of Personal Identifiable Information using Multiple Anonymisation Techniques"*, The 1st International Conference on Networked Digital Technologies (NDT), Ostrava, Czech.

4. **Tinabo, R**, Mtenzi, F.; and O'Shea, B. (2009), *"Designing and Developing A New Anonymisation Technique to be Used in E-healthcare"*,The 14th Annual Conference of Healthcare Information Society of Ireland (HISI) , Dublin, Ireland.

Other Publications

5. Lupiana, D.; **Tinabo, R.**; Mtenzi, F.; O'Driscoll, C.; and O'Shea, B. (2011) *Alphanumeric Data: Minimising Privacy Concerns in Smart Environments*, International Journal of Digital Society (IJDS), Volume 2, Issue 3.

6. Doyle, P.; Deegan, M.; **Tinabo, R.**; Masamila, B.; and Tracey, D. (2009), Case Studies in Thin Client", Ubiquitous Computing and Communication Journal (UbiCC), Vol4 Special Issue on ICIT 2009 conference - Applied Computing: pp585-598.

7. Masamila, B.; Mtenzi, F.; Said, J. and **Tinabo, R.** (2010), A Secured Mobile Payment Model for Developing Markets, Networked Digital Technologies, 175182, Springer.

1.9 Thesis Organisation

The remaining Chapters of this thesis are organised as follows:

- Chapter 2: The issue of protecting Personal Identifiable Information (PII) when sharing data for different purposes is a challenge in any domain where data is collected. This research uses the healthcare domain as a case study. Therefore, Chapter 2 provides an overview of the healthcare area mainly focusing on different challenges and characteristics of e-healthcare data.

- Chapter 3: This chapter describes related works on the privacy-enhancing approaches. It discusses strength and weaknesses of several existing privacy models, anonymisation techniques and algorithms.

- Chapter 4: This chapter discusses the proposed algorithm, named *kl-redInfo*, including its detailed design and the high level architecture.

- Chapter 5: This chapter discusses the experimental environment including datasets and parameters used, evaluation metrics, and introduces the implemented algorithms for comparison purposes.

- Chapter 6: This chapter presents an experimental evaluation of the proposed *kl-redInfo* algorithm and the comparison with the widely used algorithms. This chapter clarifies the improvement achieved by the *kl-redInfo* algorithm. Chapter 6 also presents other findings of the research including the impact of the algorithms on the different dataset size and on different k and l parameter values.

- Chapter 7: Conclusions and future work of this research is presented in Chapter 7.

Chapter 2

HEALTHCARE DATA IN INFORMATION SOCIETY

The challenge of issuring individuals' privacy while sharing the data which is still useful is the common challenge. This research uses the healthcare domain as a case study. This chapter discusses different characteristics of the healthcare domain, with the main focus on e-healthcare data. The chapter discusses various characteristics of the healthcare data in section 2.2. Section 2.3 describes drivers for the application of different technologies such as Information and Communication Technology (ICT) in healthcare. Challenges for sharing data for secondary uses are presented in section 2.4. Data accessibility issues for this research and data protection laws are described in section 2.5. Lastly, section 2.6 covers the chapter conclusion and summary.

2.1 Introduction

Application of Information and Communication Technologies (ICT) in the healthcare domain, particularly in provision of the healthcare services is referred to as e-health or e-healthcare (Mukherjee and McGinnis, 2007; Shoniregun et al., 2010). These technologies transform the delivery of the healthcare service from paper-based system to electronic or a hybrid. In comparison to a paper-based system, electronic data is easily stored, retrieved, processed, and transmitted, making it a preferable choice for enhancing the quality of the healthcare service delivery. Examples of the e-healthcare services include e-prescription, telemedicine, healthcare portals, and electronic healthcare record systems.

The advancements of technology in other domains such as social networking, sales and marketing, which lead to the existence of online information databases cause electronic data to be more susceptible to malicious manipulation than paper based data. These databases include demographic and non-demographic information which can easily be linked to identify the identity of individuals. The existence of several online information databases increases the difficulty in protecting personal identifiable information (PII) when disclosing medical records for different purposes. Furthermore, advancement of the storage capacity and processing power of the computing devices can be used maliciously to facilitate linking and mining of the data for the purpose of breaching privacy.

2.2 Characteristics of Healthcare Data

Healthcare data is critical due to prevalence of characteristics which include sensitivity, complexity, volume, and usefulness. These characteristics make the healthcare domain subject to stringent data control compared to other domains. Failure to apply sufficient control on healthcare data can have catastrophic consequences to all healthcare stakeholders. The four identified characteristics of healthcare data are data sensitivity, data diversity, data volume and data usefulness.

2.2.1 Data Sensitivity

Healthcare data is regarded as personal-specific and sensitive. This is because it contains data attributes which when disclosed can affect the data subject. Such effects include irrecoverable social stigma, economic threats, discrimination and mental suffering (Appelbaum, 2003). Therefore, privacy is needed to build public trust in order for people to participate effectively in a particular activity.

There are concerns of privacy in different societies for different application domains. For example, in the healthcare domain, privacy protection is important in order to avoid harm to data subjects and to promote provisions of reliable and accurate data for effective and efficient healthcare services. This is because healthcare information relates to personal aspects of an individual's life (Anderson et al., 2000).

The medical records of an individual may include identifying information, laboratory tests, medical diagnostics and physicians' subjective comments. Also, it may include individual's genetics information, which can be used for inference about the whole family (Mercuri, 2004). Lack of public trust on privacy can cause privacy-sensitive people to avoid healthcare treatment. Also, they may opt to disclose less information to physicians, switch between physicians, or pay service treatment claims from their own pocket. The repercussions for these outcomes are:

Difficult to provide quality care to patients:

Patients have to provide detailed information to their physician during treatment. Also, previous medical records are important for the physician to increase the probability of correct diagnosis and prescription. Lack of complete information can jeopardise the quality of treatment to the patient (Appelbaum, 2003).

Reduce the ability of physicians to diagnose and propose correct treatment:
A Physicians' ability to diagnose and prescribe correct treatment depends on how much information is available. Together with patients' detailed explanations and medical history, information about previous similar cases is important. Unless patients provide detailed and correct information, the reliability of information on previous case will be susceptible. This will affect not only privacy-sensitive patients but also other members in the society.

Increase healthcare costs due to switching healthcare providers and late treatment:
Healthcare costs can be reduced if unnecessary duplications are avoided, such as repeated laboratory tests. In the absence of a shared system such as centralised databases or healthcare information networks, patients who constantly change their healthcare providers are likely to incur additional healthcare costs (Anderson et al., 2000). Also, late diagnosis increases treatment costs. In a narrow view these costs may seem as personal cost, but in a broader view they affect the overall cost of the national healthcare.

Poor outcomes from research, public healthcare, and quality initiatives:
Quality of healthcare depends on continuous quality improvement, effective healthcare administration and public healthcare. An important ingredient to these prime functions is the availability of reliable and accessible healthcare data. When patients avoid care, or give less or false information the whole healthcare system is jeopardised (El Emam et al., 2009).

That said, public trust on privacy is crucial for any healthier society. Building this trust requires effective protection of all information collected from patients that can be used either for primary or secondary use.

2.2.2 Data Diversity

Healthcare data is comprised of different data structures. These data include free form notes, structured and unstructured text and numeric, images, blood sample reports, codes, sounds, and videos (Grimson et al., 2000). This complicates protection of the data, particularly in applying uniform algorithm across different structures. Therefore, this research aimed at anonymising structured/relational data.

2.2.3 Data Volume

Healthcare is an information-intensive domain generating large amounts of data from different areas, including hospitals, primary care surgeries, clinics, and laboratories (Safran et al., 2007). This is due to the prevalence of non interoperable systems and the nomadic nature of patients in seeking treatment from different healthcare providers for different reasons. Also, it is required that patient information be kept for the life time of an individual, that is, from cradle to grave plus retention time (Grimson et al., 2000; Grimson, 2001). Unlike information from other domains, the value of this information does not diminish with time. This property makes its protection challenging.

2.2.4 Data Usefulness

The healthcare domain has several stakeholders, including patients, clinicians, healthcare organisations, public healthcare professionals, policy makers, employers, researchers, and insurers (Appari and Johnson, 2010). Each of these categories is interested in medical records for different reasons as summarised in Figure 2.1.

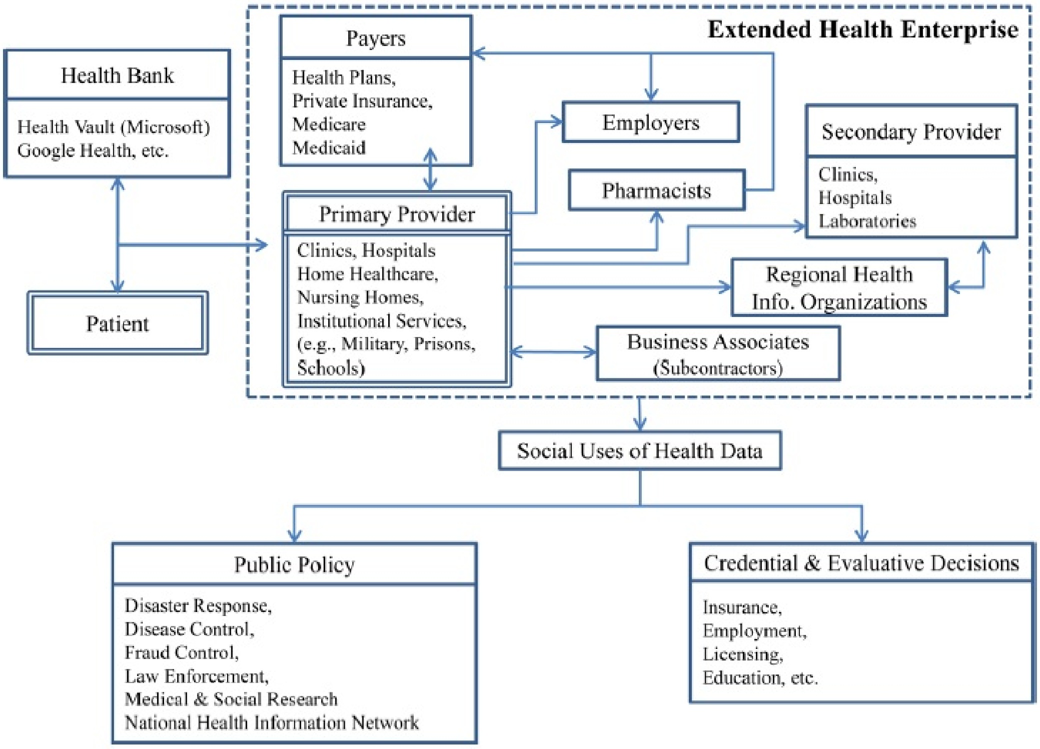


Figure 2.1: Stakeholders of the healthcare data (Appari and Johnson, 2010)

Patients are a primary source of this information. Primary provider need both longitudinal and cross-section information to make evidence-based treatment. Researchers, public healthcare professionals, and policy makers need healthcare data for learning and generating new knowledge and insights for planning and improving healthcare services. Employers and insurers need this information for analysing healthcare costs and to settle the associated bills. However, the same information can be used for personal economic gains such as marketing for drug manufacturers, lending decisions by banks, and employment decisions by employers. The latter two uses can have direct negative impact to the individual patient concerned. In most countries release of medical data to employers or lenders requires explicit patient consent (HIPAA, 1996; statute book, 2003).

The large number of stakeholders with an interest in medical records increases its risk to privacy breaches on disclosed records. This is because medical records can be used for different purposes, thus elevating its usefulness. Thus healthcare data needs stringent control because the number of people who access them is large (Anderson, 1996).

2.3 Benefits of Using Healthcare Data

Despite with all complex characteristics of the healthcare data, the data is needed for different uses; primary and secondary uses. The primary use of the healthcare data is to provide direct health care delivery. Such purposes typically include the provision of adequate and appropriate medical care requested by the patient or deemed necessary for the patient based on the record's contents. These records are necessary on this primary level in order to keep track of the important clinical information that any future medical professional may find useful in encounters with the same patient (Mukherjee and McGinnis, 2007).

In additional of the primary purpose of providing care to patients, there is an increasing demand for the use and sharing of healthcare records. This is due to the adoption of electronic medical and health records throughout the domain and across all sectors of the healthcare system. Any use of the healthcare information for any purpose not directly related to the care of individual patients who are the subject of that information is known as *secondary use.* This includes activities such as analysis, research, quality and safety measurement, public health, payment, provider certification or accreditation, marketing, and other business applications. These activities enhance healthcare services by reducing medical errors, controlling escalating healthcare costs, enhancement of quality care and accessibility of services, and providing timely and relevant information to help physicians during treatment (Safran et al., 2007; Shoniregun et al., 2010). These benefits are described next:

2.3.1 Reducing Medical Error

Medication errors are major concerns in the healthcare domain. Prescription errors and misinterpretation of communications among physicians, nurses and pharmacists due to bad hand writing contribute to the problem (Bates, 2000). According to Kohn et al. (2000), preventable medication errors cost the lives of 44,000 to 98,000 Americans yearly. It is believed that electronic medical records can substantially reduce medical errors (Anderson, 2007; Bates, 2001; Tang et al., 2006). The problem is worldwide, thus it prompted several initiatives to embrace ICT.

2.3.2 Controlling Healthcare Costs

In recent years, there have been concerns on rising healthcare costs (Mukherjee and McGinnis, 2007). Typically, this problem is associated with administration difficulties, lack of reusing the collected information, and shortage of medical professionals. Administration difficulties rise healthcare costs due to complexity in processing medical claims that result in multiple claims for treatment and other fraud. Duplication of medical tests and inability to share the collected data also increases the cost of healthcare (Goldschmidt, 2005).

Furthermore, the number of healthcare professionals with respect to the population they are serving may be relatively small, particularly in developing countries such as Tanzania. This causes ineffective service delivery and thus increases costs. Therefore, the secondary use of the healthcare data is motivated to reduce escalating healthcare costs.

2.3.3 Enhancing Quality and Accessibility of the Care Services

The desire to enhance quality and accessibility of healthcare services motivate the secondary use of the healthcare data. Better management of healthcare information using electronic systems enhances efficiency and communication in the workplace and thus improves healthcare service quality (Mukherjee and McGinnis, 2007; Tang et al., 2006). Also, the use of ICT increases channels through which patients and physicians can interact for example, by using healthcare portals, patients have access to healthcare information.

2.3.4 Provision of Information to Physicians

Physicians need sufficient information to help them to make correct decisions during treatment (Anderson, 2007). This information includes longitudinal and cross-section information for supporting evidence-based care delivery. It is easier and more effective for electronic medical records to serve this demand than its counterpart paper-based medical records. This, generally, improves quality of care.

Therefore, allowing healthcare data to be used for secondary purposes would boost the quality of medical services and overall public health including areas of genetic impacts, disease risk factors, possible interventions, drug side effects, drug safety surveillance, treatment effectiveness, decreased mortality rates, institutional performance tracking and clinical efficiency, could support the identification of disease mechanisms and new discovery areas, accelerate the termination of unsuccessful compounds, decrease patient recruitment cycle times for clinical trials, and improve drug safety surveillance through continuous monitoring (Tang et al., 2006).

In general, the secondary use of health information is a necessity and should be an accepted part of any health system that supports the effectiveness, efficiency and sustainability of the health system and is an integral part of the cycle of research, medical evidence, and accepted knowledge base through to the delivery of care. Therefore, the gains expected from imaginative but responsible uses of healthcare information accrue not only to various interest groups but also to populations in general. Thus, the algorithm to be developed in this research promotes and enables secondary uses while ensuring individual's privacy. It equally ensures there are adequate safeguards to maintain the balance between secondary use of healthcare data and the data protection.

2.4 Challenges in Sharing Healthcare Data

A critical challenge associated with sharing data is the possibility that the data can be disclosed and used for other purposes other than initially stated. Removing PII from these records can not quarantee an individual's privacy as there are possibilities to link the shared data with other data from different databases to identify an individual. Also, removing Personal Identifiable Information (PII) from the data reveals less information and may render it less useful. This affects accuracy and hence quality of knowledge or insight generated from its use which is necessary for different secondary purposes such as planning and improvement of public healthcare in general (Shoniregun et al., 2010).

However, privacy of an individuals whose records are shared must also be protected. Balancing these two conflicting requirements is a challenging problem. Current practice of using contracts and laws (as discussed in section 2.5) cannot guarantee that sensitive data will not be misplaced and end up in the wrong hands.

2.5 Data Protection Laws

This research has considered the use of real patient datasets for evaluation of the developed algorithm. However, the challenge ascribed to patients' privacy hinders its utilisation in this research. This difficulty is due to the difference in data protection laws of the specific countries. While this research targets Tanzania as a case study for supplying real datasets for the evaluation of the implemented algorithm, importing this data to Ireland, is not a straight forward activity. Section 2.5.1 and 2.5.2 discuss in detail the data protection laws of these two countries; Ireland and Tanzania respectively.

2.5.1 Data Protection Law in Ireland

The development and enforcement of European Union Directive 95/46/EC and its interpretetion into laws by its member states, (Ireland in this case) is an indication of how the region is sensitive on personal data privacy (DPC, 1995). In particular and of interest to this research are the directives and protocols that govern trans-border data flow. The Act requires that, the transfer of the personal data to a country or territory outside the European Economic Area (EEA) not to take place unless that country or territory ensures an adequate level of protection for the privacy and the fundamental rights and freedoms of data subjects in relation to the processing of personal data having regard to all the circumstances surrounding the transfer.

The essential concern of the EU Directive, and Ireland Data Protection Laws in particular, is to ensure that their residents' personal information is not transferred to countries that do not adequately protect that information. The Directive says nothing about the information transferred from countries outside EU to EU countries. The main key feature of the Ireland Data Protection Act of 2003 (statute book, 2003) is on the principle that organisations should be held accountable for the personal data that they gather and process. Such accountability is also expected to organisations when they transfer personal data across national border.

The Dublin Institute of Technology (DIT) as an organisation ensures everyone in the institute complies with the eight Data Protection Principles which are set out in the Data Protection Acts. These principles are: information should be obtained and processed fairly; data should be kept for the specified purpose(s) only; the data should be used and disclosed according to the purpose(s); the data should be kept safe and secure; the data should be kept accurate, complete and up-to-date; should not collect excessive data; the data should be retained for a reasonable time; and a copy of an individual's data should be granted when requested.

2.5.2 Data Protection Law in Tanzania

Tanzania is one of the developing countries located in the Eastern part of Africa. As most developing countries, Tanzania has immature data protection laws (Bord et al., 2009). The data holder is the one who has the mandate to grant data for secondary use. From the survey done at The Muhimbili National Hospital (MNH), a commonly used approach of removing PII from medical records is used. This approach is considerably weak for protection of data privacy due to linking attack.

There are several procedures that have to be followed in order for a researcher to access the healthcare data in Tanzania. The procedures includes: getting permission from 1) The Ministry of Health and Social Welfare in Tanzania; 2) The National Institute for Medical Research(NIMR) that oversees all research in healthcare domain in Tanzania; and 3) getting permission from the specific hospital that will provide the data, in this research the hospital is called The Muhimbili National Hospital (MNH). In all three institutes the research proposal that will be assessed by the Ethical Committee of each institute, has to be written. The proposal is assessed by all three institutes in terms of its value of the contributions to the general community, feasibility of the research process, and capability of the researcher to undertake the proposed research.

The researcher followed the procedures and the permission was granted by all three institutes. That allowed the DIT Research Ethics Committee also to grant the researcher DIT Ethical Approval under the condition that the data should be anonymised. The anonymisation technique applied to the data to be given was to remove PII, which is not enough to ensure privacy of individual's. This is due to linking attack that might cause identification of an individual's sensitive information due to growth of data volume and technology.

A consultation with the PhD supervision team and experts from the office of Data

Protection Commissioner in Ireland was then made. The experts from the office of Data Protection Commissioner in Ireland insist the need of getting consent from the patients whose data would be given to researcher. This process is inpractical, therefore the researcher concludes that, it is unethical to transfer the real medical records dataset from Tanzania for research purpose without consent from the patients. Thus, the use of real dataset was not possible and the alternative of simulating data by using schema similar of the Muhimbili National Hospital (MNH) was used.

2.6 Chapter Summary

In the ever increasing online databases and sophistication of technology, ensuring an individual's privacy while sharing data for different purposes becomes difficult in every domain. This research uses the healthcare domain as a case study. This is due to several characteristics of healthcare data such as, data sensitivity, data diversity, data volume and data usefulness, which elevates the risk for data misuses. However, the sharing of this data is necessary for effective and efficient management of healthcare services and improvement of public healthcare. Recognising the need for protecting privacy of data subjects, data protection laws have been enacted. Guidelines and principles are stipulated either to restrict its use or enforce stringent measures. Of themselves, the problem cannot be fully addressed.

Chapter 2 discusses the characteristics, benefits and challenges associated with using healthcare data which set the foundation for analysing the existing techniques for addressing the problem. This investigation was achieved through a literature analysis and by a field study conducted at Muhimbili National Hospital (MNH) in

Tanzania. Chapter 3 discusses different privacy-preserving approaches that can be used to ensure individual's privacy.

Chapter 3

PRIVACY-ENHANCING APPROCHES

The problem of providing useful information while ensuring an individual's privacy is a long standing challenge. Researchers have addressed with limited success and countinue to address the protection of PII when sharing data for different purposes (Sweeney, 1997; Samarati and Sweeney, 1998; Samarati, 2001; Aggarwal et al., 2005; LeFevre et al., 2005; Anderson, 2007; Riedl et al., 2008). Different techniques to address this problem have been proposed, including ethical and legal frameworks, policy and regulatory frameworks, and privacy-enhanced technologies such as statistical techniques, cryptographic techniques, and anonymisation techniques.

This chapter summarises and evaluates the different existing privacy-enhancing approaches inlcuding ethical and legal frameworks (section 3.1), policy and regulatory frameworks (Section 3.2), and privacy-enhancing technologies (section 3.3) including statistical techniques (Section 3.4), cryptographic techniques (Section 3.5) and anonymisation techniques (Section 3.6), and Section 3.7 presents chapter summary and conclusion.

3.1 Ethical and Legal Frameworks

Historically, the Hippocratic Oath by physicians plays a fundamental role in the healthcare domain (Agrawal et al., 2002). Physicians are obliged to maintain the confidentiality of information they see or hear in the course of treatment or outside (U.S. Congress, 1993). In the traditional healthcare system, ethical practices by physicians play a substantial part in building trust on preserving privacy of individuals.

Recognising the importance of secondary use of healthcare information, and increasing use of ICT in the management and delivery of healthcare services implies that the number of people who have legitimate access to medical records increases. The majority of these users are not bound by the Hippocratic Oath. Therefore, when healthcare information leaves healthcare professionals, reliance on ethical frameworks as a means to preserve privacy diminishes. This situation influences legal intervention.

Legal systems establish laws to protect privacy of healthcare information disclosed for secondary use. A typical example is the European Union, which established data protection directives for its member states. Several countries have enacted laws for protection of privacy ($www.informationshield.com/intprivacylaws.html$). The United States represent a significant effort in this trend (HIPAA, 1996). This intervention is important to limit the risk of privacy breaches on the disclosed information. Thus, the legal frameworks play a vital role in limiting illegal information disclosure and processing (statute book, 2003). However, they are not a panacea.

3.2 Policy and Regulatory Frameworks

At organisation level, policy is defined as a set of rules to meet a particular goal (Landwehr, 2001). Privacy policy stipulates rules through which a organisation preserves privacy of information. For example, Anderson (1996) proposed a clinical policy model to help healthcare organisations to maintain the protection of the healthcare information.

Also, regulatory frameworks must unify organisation privacy policies. This is important for ensuring consistent protection across different stakeholders. In modern times, compliance towards regulatory frameworks is considered important for assessing the extent of data protection. Thus, policy and regulatory frameworks form another vital strand for data protection.

3.3 Privacy-enhancing Technologies (PETs)

The complexity of privacy issues requires several protection approaches to effectively preserve the privacy of individuals. The practice of sharing data that relies mainly on policies and guidelines as to what types of data can be shared and on agreements on the use of the shared data, may lead to excessive data distortion or insufficient protection (Schneier, 2000; Fung et al., 2010). Researchers have proposed several technological solutions to address the problem. These solutions are referred to as privacy-enhancing technologies (PETs). The PETs categories include statistical techniques, encryption tools and anonymisation techniques. The summary of the strength and weaknesses of the existing PETs were originally published in Tinabo et al. (2009a) and are further discussed in sections 3.4, 3.5 and 3.6.

3.4 Statistical Techniques

Statistical techniques are the first privacy-enhancing technologies addressing the need of protecting Personal Identifiable Information (PII) when sharing data for secondary use. This is achieved by sharing aggregate/statistical results instead of specific individual information (Muralidhar and Sarathy, 1999). Protecting confidential records and to provide useful information are the general goals of statistical techniques (Benedetti and Franconi, 1998). Database administrators can use statistical techniques to allow users to access aggregate statistical information, rather than information regarding a specific individual.

Even in statistical databases, PII associated with a particular individual can also be infered through a clever choice of queries, leading to disclosure of PII of an individual (Adam and Worthmann, 1989). To solve this problem, statistical databases often use random data perturbation which involves the addition of random noise to confidential numerical attributes. Thus, even if a user manages to compromise data and is able to isolate an individual value of a confidential attribute, the true value is not disclosed (Muralidhar and Sarathy, 1999).

The existing statistical techniques can be classified into three categories: query restriction, input perturbation, and output perturbation (Adam and Worthmann 1989).

3.4.1 Query Restriction

In the query restriction approach, queries are required to comply with a special structure, supposedly to prevent the querying adversary from gaining too much information about specific database entries (Adam and Worthmann, 1989). Query restriction provides exact answers to queries as long as the risk of exact disclosure of individual's PII is small (Nunez et al., 2007). The query restriction techniques works for a relatively small number of queries (Dinur and Nissim, 2003).

Query auditing was introduced to reduce this problem (Chin and Ozsoyoglu, 1982), where a log of the queries is kept, and every new query is checked for possible compromise, allowing or disallowing the query accordingly. But query auditing task is NP-hard (Kleinberg et al., 2000). Also, auditor refusals, in conjunction with the answers to valid queries, may be used by the user who receives the data to achieve a partial or total compromise of the database (Dinur and Nissim, 2003).

3.4.2 Input Perturbation

The Input/Data perturbation is another well known technique for privacy preserving of the data (Kabir et al., 2007). This technique deals with disturbing data before the release. That is, the data is systematically changed to yield answers to queries that are statistically similar to those that would have resulted from the original data (Nunez et al., 2007; Chawla et al., 2005). The Input/Data perturbation approaches are often used to protect confidential, numerical data from unauthorised queries while providing maximum access and accurate information to genuine queries (Muralidhar and Sarathy, 1999). Therefore, in the Input/Data perturbation approach queries are answered according to a disturbed database.

There are several approaches used to achieve the Input/Data perturbation. These approaches include swapping values, where portions of the data are replaced with data taken from the same distribution (Reiss, 1984; Duncan et al., 2001; Samarati, 2001); and fixed perturbations, where a random perturbation is added to every data entry (Agrawal and Srikant, 2000; Agrawal and Aggarwal, 2001).

Even though the Input/Data perturbation approaches guarantee a complete or exact disclosure (i.e., the disclosure of the true value of a confidential attribute), they are vulnerable to partial disclosure. Partial disclosure occurs when the amount of information that a user who manages to compromise data is able to obtain about a confidential attribute through queries and statistical analysis, is more than the amount that the database administrator planned to provide to users (Adam and Worthmann, 1989).

3.4.3 Output Perturbation

The output perturbation approach uses query control mechanism to compute exact answers, but it returns disturbed or noisy answers as a response to the query (Dinur and Nissim, 2003; Beck, 1980; Denning, 1980). Methods of output perturbation include varying output perturbations, where a random perturbation is added to the query answer, with increasing variance as the query is repeated (Beck, 1980); and rounding, either deterministic or probabilistic (Dinur and Nissim, 2003).

3.5 Cryptographic Techniques

Cryptographic approaches are most often associated with the encryption process that can be defined as the ability to convert readable text to unreadable text (Mills et al., 2003). This security mechanism uses mathematical schemes and algorithms to encrypt data into unreadable text. The unreadable text can only be decrypted by the party who possesses the associated key.

There are two types of encryption, known as single key and public key encryption. Single key encryption uses the same key for encrypting and decrypting text while in public key encryption, two keys are used, one for sharing (public) and one is kept secret (private). There is a difference between traditional encryption and potentially truly anonymous procedures such as one-way hashing. By using traditional encryption, data is encrypted but can be decrypted with the use of a key. Theoretically, encrypted data is different from truly anonymous data as the underlying data can be accessed by combining it with the key (and the key might be obtained by applying brute computational force or otherwise) (Clarke, 1999).

It would seem that encryption is the most effective way to preserve privacy of information. Users wishing to access the data could be given keys, and this would summarily solve all privacy issues. Unfortunately, this approach does not work in a data publishing scenario, whose the primary goal is to secure access to confidential information while at the same time sharing useful information.

3.6 Anonymisation Techniques

Anonymisation is the process of making data anonymous. Anonymous data is the data that cannot be manipulated or linked in order to identify an individual. Data can be made anonymous by either suppressing personal-specific data, or by generalising them, or replacing real identifiers with false identifiers. This is done by developing algorithms that fullfill privacy requirements of certain privacy models. The common privacy models are discussed in section 3.6.1 and some of the widely used algorithms are discussed in section 3.6.3.

3.6.1 Privacy Models

There are several privacy models that have to be achieved for the data to be considered protected. Fung et al. (2010) classify them in two categories based on their attack principles, *informative principle* and *uninformative principle*. The first category includes identity disclosure, attribute disclosure and table disclosure. These depend on the privacy threat that occurs when a user who receives the data is able to link an individual's record to a record in the shared data, or to a sensitive attribute in the shared data or to the shared data itself.

The identity disclosure occurs when an individual is linked to a particular record in the shared table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the shared data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data is shared. The identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. The attribute disclosure can occur with or without identity disclosure. In table disclosure an identity disclosure occurs when an individual record is determined to be presence or absence in the shared table. A data table is considered to be privacy-preserving if it can effectively prevent the successfully performing of these disclosures (Xiao and Tao, 2006a; Li et al., 2007).

The second category aims in achieving the *uninformative principle*, which requires that the shared data should provide the user who receives the data with little additional information beyond the background knowledge (Machanavajjhala et al., 2007).*Probabilistic disclosure* occurs when the user who receives the data has a large variation between the prior and posterior knowledge. The two categories may overlap since, many privacy models in the category do not explicitly classify attributes in a data into quasi-identifiers and sensitive attributes, but some of them could also thwart the sensitive disclosure in the first category. Table 3.1 summarises the common used privacy models and the privacy threat that they address.

Table 3.1: Privacy Models (Fung et al., 2010)

Privacy Models	**Privacy Threat**			
	Record Disclosure	*Attribute Disclosure*	*Table Disclosure*	*Probabilistic Disclosure*
k-anonymity	√			
l-diversity	√	√		
t-closeness		√		√
(X,Y)-anonymity	√	√		
(k, e)-anonymity		√		
MultiR k-anonymity	√			
(α, k)-anonymity	√	√		
(ϵ, m)-anonymity		√		
Confidence bounding		√		
Personalised privacy		√		
δ-presence			√	
(c, t)-isolation	√			√
ϵ-differential privacy			√	√
(d, γ)-privacy			√	√
Distributional privacy			√	√

The k-anonymity **Privacy Model**

The k-anonymisation is a privacy model used to provide privacy protection by ensuring that data cannot be traced to an individual with respect to quasi-identifier attributes (Samarati, 2001). The shared data hold *k-anonymity* privacy requirement, if each shared record has at least (*k-1*) other records in the release whose quasi-identifier values are indistinguishable from each other (Sweeney, 2002). The group of records with the same quasi-identifier attributes (QIDs) is known as an *equivalence class*.

Therefore, *k-anonymity* provides privacy protection by guaranteeing that each equivalence class consists of at least k records. Thus, even if the records are directly linked or matched to external information there will be no association between an individual and the record. Also known as identity disclosure. Table 3.3 is an example of the 2-anonymous table of the Patients' Information Table 3.2. Table 3.3 was obtained by generalising QIDs so that there is at least two records with the same QIDs

Table 3.2: Patients' Information Table

No.	*Date of birth*	*Gender*	*P.O. Box*	*Disease*
1	1981/07	M	12386	Cancer
2	1978/02	F	12362	Obesity
3	1962/05	M	12337	Obesity
4	1978/02	F	12395	Malaria
5	1978/10	F	12381	HIV
6	1981/09	M	12352	Obesity
7	1978/10	F	12381	HIV
8	1962/08	F	12394	Cancer
9	1981/04	M	12380	Malaria

Table 3.3: 2-Anonymous Table

No.	*Date of birth*	*Gender*	*P.O. Box*	*Disease*
3	1962	*	123**	Obesity
8	1962	*	123**	Cancer
2	1978/02	F	123**	Obesity
4	1978/02	F	123**	Malaria
5	1978/10	F	123**	HIV
7	1978/10	F	123**	HIV
1	1981	M	123**	Cancer
6	1981	M	123**	Obesity
9	1981	M	123**	Malaria

As it can be seen from Table 3.3 there are at least two records which are indistinguishable from each other. That is why is known as 2-anonymous table.

Since Samarati and Sweeney introduced the *k-anonymity* privacy model, several algorithms have been proposed for implementing the *k-anonymity* privacy model via generalisation and suppression methods. Samarati (2001) proposed the binary search algorithm for full-domain generalisation; Sweeney (2002) proposed a heuristic algorithm for cell generalisation. Meyerson and Williams (2004) and Aggarwal et al. (2005) describe approximation algorithms for the cell-suppression flavor of k-anonymisation. Bayardo and Agrawal (2005) described an optimal search-based algorithm for single-dimensional recoding.

***The l-diversity* Privacy Model**

As recognised by several researchers, *k-anonymity* can only prevent association between individuals and records (identity disclosure), but it cannot prevent the association between individuals and sensitive values (attribute disclosure) (Xiao and Tao, 2006a; Li et al., 2007; Machanavajjhala et al., 2007). Therefore, a user who receives the data can discover the values of sensitive attributes when there is little diversity in those sensitive attributes.

For example, Table 3.3 is a 2-anonymous table of the original Table 3.2 but a user can conclude that a certain woman whose information is in the original table with P.O. Box 12381 and born in October 1978, (1978/10) has HIV disease since all the records with these quasi-identifiable information has HIV disease. This problem is known as homogeneity problem (Machanavajjhala et al., 2007). Machanavajjhala proposes an *l-diversity* privacy model to address this problem of *k-anonymity* privacy model.

The *l-diversity* model requires that each equivalence class has at least l-well-represented values for each sensitive attribute (Machanavajjhala et al., 2007). Machanavajjhala et al. (2007) defines *well represented* as *Distinct l-diversity*, *Entropy l-diversity* and *Recursive (c,l)-diversity.* Since all the definitions of *well represented* result in almost similar results, this research adopts the simple distinct l-diversity definition.

1. *Distinct l-diversity* (also known as p-sensitive k-anonymity (Truta and Vinay, 2006)). This definition ensures that there are at least l distinct values of the sensitive attribute in each equivalence class e_i.

2. *Entropy l-diversity.* A table is said to satisfy entropy l-diversity if for every equivalence class e_i

$$Entropy(e_i) = -\sum_{s \in S} P(e_i, s) * log(P(e_i, s)) \geq log(l) \qquad (3.6.1)$$

   where P (e_i,s) denotes the proportion of each sensitive value s in an equivalence class.

3. *Recursive (c,l)-diversity.* The recursive (c,l)-diversity makes sure that the most frequent value does not appear too frequently, and that the less frequent values do not appear too rarely. A table satisfies recursive *(c,l)-diversity* if every equivalence class satisfies recursive (c,l)-diversity. The equivalence class satisfies recursive *(c,l)-diversity* if $\mathrm{r}_1 < \mathrm{c}(\mathrm{r}_i + \mathrm{r}_{i+1} + \ldots + \mathrm{r}_m)$; where c is the constant and r_i denotes the number of times the i^{th} most frequent sensitive value appears in that equivalence class.

An algorithm for *l-diversity* can be created by changing an algorithm for the *k-anonymity*, and make the algorithm to check for *l-diversity* every time when a table is tested for *k-anonymity.* To make k-anonymous table l-diverse, equivalence classes that are not l-diverse are either suppressed or combined together until they are diverse. This results in unneccesary information loss. By using the 2-anonymous Table 3.3, the 2-diversity table is presented on Table 3.4. Now a user can not identify the diseases of a woman whose information is in the table with P.O. Box 12381 and born in 1978 since there are three different diseases with these quasi-identifiable information. This was achieved by combining together the second and the third equivalence classes and generalising their date of birth by sharing year only.

Table 3.4: 2-diversity Table

No.	*Date of birth*	*Gender*	*P.O Box*	*Disease*
3	1962	*	123**	Obesity
8	1962	*	123**	Cancer
2	1978	F	123**	Obesity
4	1978	F	123**	Malaria
5	1978	F	123**	HIV
7	1978	F	123**	HIV
1	1981	M	123**	Cancer
6	1981	M	123**	Obesity
9	1981	M	123**	Malaria

After the development of this main privacy model; *k-anonymity*, *l-diversity*, several other privacy models have been proposed to address different scenarios that were not considered by the privacy models. Some of the proposed privacy models are discussed next:

The t-closenesss Privacy Model

Li et al. (2007) observed that when the overall distribution of a sensitive attribute is skewed, preventing attribute linkage attacks by using *l-diversity* privacy model results in high information loss. For example, consider a data table containing data of 1000 patients on some quasi-identifier attributes (QIDs) and a single sensitive attribute HIV with two possible values, Yes or No. Assume that there are only 5 patients with HIV = Yes in the table. To achieve k-anonymity with *k=l*, at least one patient with HIV is needed in each equivalence class; therefore, at most 5 equivalence classes can be formed. Enforcing k-anonymity with *k=l* may lead to high information loss in this case.

To prevent skewness attack, Li et al. (2007) proposed a privacy model, called *t-closeness*, which requires the distribution of a sensitive attribute in any group on QID to be close to the distribution of the attribute in the overall table. To measure the closeness between two distributions of sensitive values *t-closeness* uses the Earth Mover Distance (EMD) function. The closeness requires to be lower than *t*.

There are several limitations and weaknesses of *t-closeness* privacy model. First, it lacks the flexibility of specifying different protection levels for different sensitive values. Second, the EMD function is not suitable for preventing attribute linkage on numerical sensitive attributes (Li and Li, 2009). Third, enforcing *t-closeness* would greatly degrade the data usefulness because it requires the distribution of sensitive values to be the same in all equivalence classes. This would significantly damage the correlation between QID and sensitive attributes.

The (X, Y)-anonymity

The *(X, Y)-anonymity* model was proposed to address the assumption that each record represents a distinct individual, assumed by the *k-anonymity* model (Wang and Fung, 2006). Thus if several records in a table represent the same individual, a group of *k* records may represent fewer than *k* individuals, and the individual's information may be identified.

The *(X, Y)-anonymity* specifies that each value on X is linked to at least *k* distinct values on Y, where X and Y are disjoint sets of attributes. The *k-anonymity* is the special case of the *(X, Y)-anonymity* where X is the QID and Y is a sensitive attribute in the table T that uniquely identifies an individual. The *(X, Y)-anonymity* provides a uniform and flexible way to specify different types of privacy requirements. If each value on X describes a group of individuals (e.g., X = Address, Gender, Age) and Y represents the sensitive attribute (e.g., Y = Disease), this means that each group is associated with a diverse set of sensitive values, making it difficult to infer a specific sensitive value.

The (k, e)-anonymity Privacy Model

Most work on *k-anonymity* and its extensions assumes categorical sensitive attributes. Zhang et al. (2007) proposed the notion of *(k, e)-anonymity* to address numerical sensitive attributes such as salary. The general idea is to partition the records into groups so that each group contains at least *k* different sensitive values with a range of at least e.

The MultiRelational k-anonymity Privacy Model.

Instead of anonymising a single data table, the *MultiRelational k-anonymity* was proposed to ensure k-anonymity on multiple relational tables (Nergiz et al., 2007). The *MultiRelational k-anonymity* assumes that a relational database contains a person-specific table T and a set of tables T_1 ,..., T_n, where T contains a person identifier *Pid* and some sensitive attributes, and T_i , for $1 \leq i \leq n$, contains some foreign keys, some attributes in QID, and sensitive attributes.

The general privacy notion is to ensure that for each record *r* contained in the join of all tables T ⋈ T_1 ⋈,..., ⋈ T_n, there exists at least *k-1* other individual's records who share the same QID with *r*. The *MultiRelational k-anonymity* applies the k-anonymisation at the group of individual's record level, not at the record level as in traditional *k-anonymity*. This idea is similar to *(X, Y)-anonymity*, where X = QID and Y = Pid.

The Personalised Privacy Model

In many applications, different subjects have different requirements for privacy. For example, a brokerage customer with a very large account would likely have a much higher level of privacy-protection than a customer with a lower level of privacy protection. In such a case, it is necessary to personalise the privacy-protection algorithm. In personalised privacy-preservation, anonymisations of the data such that different records have a different level of privacy are constructed.

Two examples of personalised privacy-preservation approaches are discussed in Aggarwal and Philip (2005); Xiao and Tao (2006a). The approach in Aggarwal and Philip (2005) uses a condensation approach for personalised anonymisation, while the approach in Xiao and Tao (2006a) uses a more conventional generalisation approach for anonymisation that allows each data subject to specify an individual privacy level. This model assumes that each sensitive attribute has a taxonomy tree and that each data subject specifies a guarding node in this tree. The data subject's privacy is violated if an attacker is able to infer any domain sensitive value within the subtree of the guarding node with a probability, called breach probability, greater than a certain threshold.

In the personalised privacy approach, a guarding node is specified for each record by its owner. The advantage is that each data subject may specify a guarding node according to their own tolerance on sensitivity. Experiments show that this personalised privacy requirement could result in lower information loss than the universal privacy requirement (Xiao and Tao, 2006a). In practice, however, it is unclear how individual data subjects would set their guarding node. Often, a reasonable guarding node depends on the distribution of sensitive values in the whole table or in a group. For example, a woman knowing that her disease is very common, may set a more special (lower privacy protected) guarding node for her record. Nonetheless, the data subjects usually have no access to the distribution of sensitive values in their QID group or in the whole table before the data is published. Without such information, the tendency is to play safe by setting a more general (higher privacy protected) guarding node, which may negatively affect the utility of data.

An anonymised table is considered adequately protected, if it satisfies a privacy model. The privacy models achieve different types of privacy protection; therefore, the choice of a privacy model depends on the needs of the underlying application. The table that does not satisfy the specified privacy requirements must be modified before being shared. The modification is done by applying to the data a sequence of anonymisation approaches. Section 3.6.2 discusses the existing anonymisation approaches.

3.6.2 Anonymisation Approaches

Anonymisation approaches are the approaches used to achieve anonymity. These approaches include, generalisation, suppression, pseudonymisation, and anatomisation. Generalisation approaches replace specific quasi-identifier values with less specific values. Suppression is the highest level of generalisation where values are not shared at all. Pseudonymisation distorts the data by adding noise, aggregating values, swapping values, or generating synthetic data. Anatomisation approach removes the relationship between quasi-identifier and sensitive attributes by grouping and shuffling sensitive values in an equivalence class. These anonymisation approaches are discussed next:

Generalisation

Generalisation approach replaces specific values with less specific values. For example, in Figure 3.1, the parent node “Been-married” is more general than the child nodes “Married”, “Divorced”, and “Widowed”. For a numerical attribute, exact values can be replaced with an interval that covers exact values. The root node, “Any status”, represents the most general value of an attribute, which is also known as suppression. A suppression approach replaces some values with a missing value, indicating that the replaced values are not disclosed. The reverse approach of suppression is called disclosure while the reverse approach of generalisation is called specialisation.

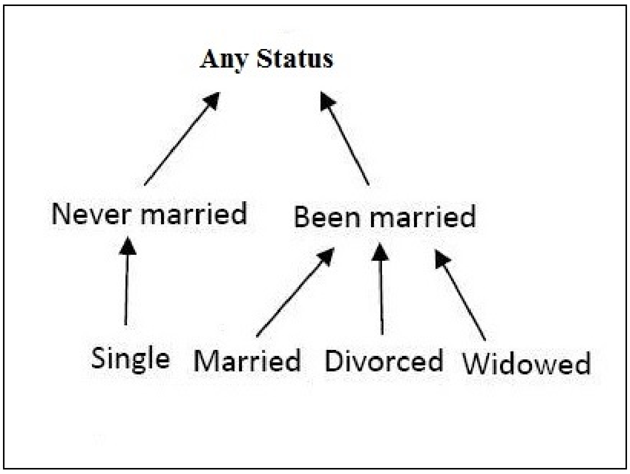


Figure 3.1: Generalisation hierarchy of the Marital status attribute (Samarati, 2001)

The Bottom-up and top-down are the main search strategies used to traverse along the generalisation hierarchies. By bottom-up search strategy, the algorithm starts at the original table, and attribute values are replaced using upper attribute values checking to determine, whether the given anonymity requirement has been achieved. The generalisation process terminates when anonymity requirement has been achieved (Wang et al., 2004). By top-down approach, a table is specialised from the most generalisation state where all attribute values have the most generalised values of their generalisation hierarchies. At each step, the most generalised values are replaced with less general values making checks to determine if anonymity requirement has been violated. The specialisation process terminates when no specialisation can be performed without violating anonymity requirement (Bayardo and Agrawal, 2005).

This research uses a bottom-up search strategy approach as it results in a lower value of the information loss metric compared to top-down search strategy. This is because the top-down search strategy starts from the most general value and stops when anonymity requirement is violated, it may stop at the point where the data is more generalised thus increasing the information loss. This is unlike a bottom-up search strategy that stops at the point where the data is more specific and also achieves the anonymity requirement.

Generalisation can be applied at the cell or attribute level. Most of the solutions proposed in the literature, adopt attribute-based generalisation. This is because the cell-based generalisation produces a table where the values in the cells of the same column may be non homogeneous, since they belong to different domains (e.g., some records report the complete date of birth, while other records only report the year of birth), which cause difficulties in analysis. On the other hand, cell-based generalisation significantly reduces information loss when compared to attribute-based generalisation as it will be discussed further in section 6.2.2, which is the main interest of this research.

Suppression

Any table can be transformed to an anonymised table by using generalisation approach. But sometimes more generalisation may result in more information loss than suppressing the records that are not anonymous. To avoid this weakness, suppression approach is used instead of generalisation.

Suppression is an approach that involves removing data so that it is not disclosed at all. It replaces one or more unique values of an attribute in a record with a missing value. The aim of the approach is to reduce identification of the attributes values. For example, suppose the combination *"Marital status=Widow; Age=20"* is unique in the dataset. If the Age information is suppressed, the combination *"Marital status=Widow; Age=missing "* will not be identifying anymore. Alternatively, if that still identifies an individual, one can suppress the information on Marital status as well.

As in generalisation, suppression can be done at the record or cell level. Record suppression scheme refers to suppressing an entire record (Iyengar, 2002; LeFevre et al., 2005; Samarati, 2001). Therefore, records suppression scheme suppresses every instance of a given cell in a table (Wang et al., 2005, 2007). Cell suppression (or local suppression) refers to suppressing some instances of a given cell in a table (Meyerson and Williams, 2004). Thus, this research uses the cell suppression not the record suppression.

Pseudonymisation

Pseudonymisation is the approach used to replace the true identities of an individual by false-identities that cannot be linked directly to their corresponding identities (De Moor et al., 2003). Unlike previously discussed approaches, pseudonymised data does not corresponds to the real-world individuals represented by the original data. Therefore, even if a user is able to identify an individual it is not possible to perform the sensitive linkages or recover sensitive information from the shared data.

A pseudonymisation approach masks identities of individuals so that information relating to those individuals can be handled without knowing to whom the information relates (Riedl et al., 2008; Claerhout and DeMoor, 2005). Only the statistical properties explicitly selected by the data holder are preserved. In contrast, generalisation and suppression make the data less precise but are semantically consistent with the raw data, and hence preserve the truthfulness of the data.

Generalisation, suppression and pseudonymisation approaches cause information loss that may reduce usefulness of the data for the tasks that require detailed insights. Anatomisation approaches were proposed to reduce the problem.

Anatomisation

Unlike generalisation, suppression and pseudonymisation, anatomisation does not modify the quasi-identifier or the sensitive attribute, but it removes the relationship between the two. This is done by partitioning the records according to distinct sensitive attributes, by the process known as bucketisation and then separates the sensitive attributes from the quasi-idientifiers. Therefore, the data is released in two separate tables; one contains quasi-identifier attributes (QIT), and the other contains sensitive attributes (ST). Both QIT and ST have one common attribute, *GroupID*, for group linking.

The anatomisation approach starts by grouping the records according to their sensitive attributes values. For example, let G_i denote the i^{th} greatest group, S = {G_1 , G_2,..., G_m } denotes the set of groups and l denotes the required number of sensitive values in each group. In each iteration of selection, one record is removed from each of the l largest groups to form a new *bucket*. Note that after every iteration, the size of some groups will be changed. So in the beginning of every iteration, the groups are sorted according to their sizes, this ensures the formed *l-records groups* are as many as possible. Also, by bucketisation the remaining records are sequentially incorporated from the first bucket.

Then, the anatomisation approach creates QIT that contain all records from the original table, but replaces the sensitive values by the GroupIDs, and create ST that contain the count of each sensitive value for each quasi-identifier group. For example, by using Patients' Information Table 3.2, Table 3.6 illustrates the two tables QIT and ST obtained by partitioning the records in the Table 3.2 in groups that satisfy 2-diversity privacy requirement.

Table 3.6: QIT and ST

No.	*DOB*	*Gender*	*P.O. Box*	*GroupID*
3	1962/05	M	12337	1
8	1962/08	F	12394	1
2	1978/02	F	12362	2
4	1978/02	F	12395	2
6	1981/09	M	12352	3
1	1981/07	M	12386	3
9	1981/04	M	12380	4
5	1978/10	F	12381	4
7	1978/10	F	12381	4

GroupID	*Disease*	*Count*
1	Obesity	1
1	Cancer	1
2	Obesity	1
2	Malaria	1
3	Obesity	1
3	Cancer	1
4	Malaria	1
4	HIV	2

This research adopts the bucketisation part of the anatomisation approach and uses a cell-based generalisation approach to achieve *k-anonymity* privacy requirement, rather than separating the table into two parts; QIT and ST tables. The formal approach used in this research is defined as follows; Given a table T, the records are partitioned into buckets (i.e., horizontally partition the table T according to *l*-distinct sensitive attribute). This ensures that each bucket contains exactly *l*-distinct sensitive values. The remaining records are incorporated in the bucket that results in a lower value of the information loss metric when the record is incorporated. Then cell-based generalisation is applied within each bucket to achieve *k-anonymity* privacy requirements. The resulting set of buckets, can then be shared for different purposes.

The two approaches are used together since, by using the bucketisation alone a user can be able to identify an individual if their quasi-identifiers are different. Therefore, cell-based generalisation approach is applied in each bucket to make them indistinguishable from each other. Also, the application of the cell-based generalisation approach depending on the need of the bucket, rather than depending on the need of all attribute values, reduces the amount of information loss, as will be justified in Chapter 6. For example, by using our Patients' information Table 3.2, Table 3.7 depicts a table that is a 2-anonymous and 2-diverse version of the Table 3.2.

Table 3.7: Buckets of the distinct sensitive attributes

No.	*Date of birth*	*Gender*	*P.O. Box*	*Disease*
3	1962	*	123**	Obesity
8	1962	*	123**	Cancer
2	1978	F	123**	Obesity
4	1978	F	123**	Malaria
5	*	*	1238*	HIV
1	*	*	1238*	Cancer
7	*	*	1238*	HIV
6	1981	M	123**	Obesity
9	1981	M	123**	Malaria

3.6.3 Anonymisation Algorithms

There are several anonymisation algorithms that use the approaches discussed in section 3.6.2 to anonymise data. This section discusses the commonly used anonymisation algorithms with their characteristics. The algorithms are summarised in Table 3.8.

Table 3.8: Characteristics of the Existing Algorithms (Author, 2013)

Algorithm	**Characteristics**				
	Methods used	*Type of generalisation*	*Privacy Models*	*Strengths*	*Weaknesesses*
μ-Argus (Hundepool and Willenborg, 1996)	Generalisation and Suppression	Cell-suppression	k-anonymity	Low information loss	The results are not always guaranteed to be k-anonymous (LeFevre et al., 2005)
Datafly (Sweeney, 1997)	Generalisation and Suppression	Full-domain generalisation	k-anonymity	Generalisation is guaranteed to be k-anonymous (LeFevre et al., 2005)	It can over-generalise data (Sweeney, 2002)
Incognito with k-anonymity (LeFevre, 2005)	Generalisation and Suppression	Full-domain generalisation	k-anonymity	Protects against identity disclosure	Cannot resist homogeneity and background attacks (Han and Yu, 2008)
Incognito with l-diversity (Machanavajjhala, 2007)	Generalisation	Full-domain generalisation	k-anonymity and l-diversity	Resist homogeneity and background attacks (LeFevre et al., 2005)	It results in high information loss (Li et al., 2007)
Mondrian (LeFevre, 2006)	Generalisation and Suppression	Multi-dimensional generalisation	k-anonymity	It is more flexible (LeFevre, 2006)	It is less scalable due to the increased search space (Xu et al., 2006)
Anatomy (Xiao and Tao, 2006)	Anatomisation	Not using generalisation	l-diversity	Results in unmodified data (Fung et al., 2010)	It does not achieve *k-anonymity* privacy requirement (Xiao and Tao, 2006b)

μ-argus and DataFly Algorithm

The μ-argus and Datafly are the first algorithms that seek to provide *k-anonymity* protection by using generalisation and suppression approaches (Sweeney, 2002). The μ-argus algorithm, developed by Hundepool and Willenborg, computes the frequency of all 3-value combinations of domain values, then greedily applies generalisations and cell suppressions to achieve *k-anonymity* (Hundepool and Willenborg, 1996). Since the approach limits the size of the attribute combination, the resulting data may not be k-anonymous when more than 3 attributes are considered.

Sweeney's Datafly system was the first k-anonymisation algorithm scalable to handle real-life large datasets (Sweeney, 1997). It achieves k-anonymisation by generating an array of quasi-identifier group sizes and greedily generalising those combinations with less than k occurrences based on a heuristic search metric that selects the attribute having the largest number of distinct values. Datafly employs full-domain generalisation and record suppression schemes. Sweeney (2002) shows that μ-argus can fail to provide adequate protection while Datafly can overdistort the data.

Incognito Algorithm

Samarati (2001) proposed a binary search algorithm that first identifies all minimal generalisations (MinGen), and then finds the optimal generalisation. Enumerating all minimal generalisations is a time consuming operation and, therefore, not scalable for large datasets. LeFevre et al. (2005) observe the problem and propose a suite of bottom-up generalisation algorithms, called Incognito.

The Incognito approach has been proposed for computing a k-minimal generalisation with the use of bottom-up aggregation along domain generalisation hierarchies.

The Incognito approach uses a bottom-up breadth-first search of the domain generalisation hierarchy, in which it generates all the possible minimal k-anonymous tables for an original table. First, it checks *k-anonymity* for each single attribute, and removes all those generalisations which do not satisfy *k-anonymity.* Then, it computes generalisations in pairs, again pruning those pairs which do not satisfy the *k-anonymity* constraints. This approach is continued until, no further pairs can be constructed, or all possible dimensions have been exhausted.

Although Incognito significantly outperforms the binary search in efficiency, the complexity of all three algorithms; MinGen, binary search and Incognito increases exponentially with the size of quasi-identifier (LeFevre, 2006). Also, the Incognito algorithm with *k-anonymity* privacy model cannot ensure diversity of the sensitive attributes, so a generated output table cannot resist homogeneity and background knowledge attacks. Based on this weakness, Incognito with *l-diversity* was proposed by Machanavajjhala to ensure diversity of the sensitive attributes (Machanavajjhala et al., 2007).

Mondrian Multi-dimensional ***k-anonymity*** Algorithm

To address the inflexibility problem of Incognito algorithm, LeFevre (2006) presented a greedy top-down specialisation algorithm for finding a minimal k anonymisation by using the multi-dimensional generalisation approach. The Mondrian performs a specialisation on one quasi-identifier group if each of its specialised quasi-identifier groups contains at least k records. Due to such a relaxed constraint, multi-dimensional generalisation usually results in anonymous data that has a better quality than when using single generalisation.

The trade-off is that multi-dimensional generalisation is less scalable than other types of generalisation due to the increased search space (Fung et al., 2010). Xu et al. (2006) showed that employing cell generalisation could further improve the data quality.

Anatomy Algorithm

The anatomy is a group-based approach addressing the issue of guaranteeing *l-diversity* privacy requirement of the shared data without using generalisation approach (Xiao and Tao, 2006b). It removes the relationship between the quasi-identifier and sensitive attribute by puting the data in two separate tables; one contains quasi-identifier attributes (QIT), and the other contains sensitive attributes (ST).

The anatomy algorithm starts by partitioning the original records into quasi-identifier groups so that, in each group, at most *1/l* of the records contain the same sensitive value. This process involves selecting *l*-records of a different sensitive attribute and sequentially incorporating the remaining records. Then, it creates a QIT table that contains all records from the original table, but replaces the sensitive values by the GroupIDs, and then creates ST table containing the count of each sensitive value for each quasi-identifier group. The anatomy algorithm is further explained in section 4.1.

Even though the anatomy algorithm results in unmodified data in both the QIT and ST tables, it does not achieve the basic *k-anonymity* privacy requirement. So taking into account the importance of both individual's privacy and data usefulness, this research proposed an algorithm, named *kl-redInfo*, which improves the anatomy algorithm. This is done by introducing approaches of systematic incorporation of the remaining records, cell-based generalisation instead of separating the table into two parts, and sorting the records according to their quasi-identifiers in order to reduce the amount of information loss.

3.7 Chapter Summary

The problem of providing useful information while ensuring an individual's privacy is a long standing challenge. Researchers have addressed with limited success and countinue to address the protection of Personal Identifiable Information (PII) when sharing data for different purposes. Different techniques to address this problem have been proposed, including ethical and legal frameworks, policy and regulatory frameworks, and privacy-enhanced technologies such as statistical techniques, cryptographic techniques, and anonymisation techniques, as discussed in Chapter 3. Most of the existing techniques emphasise on ensuring an individual's privacy while the usefulness of such data has not been well-considered.

While identifying the strengths and weaknesses of the existing privacy-enhancing approaches, this research is based on anonymisation techniques. This is due to the fact that unlike other techniques that aim to ensure an individual's privacy from unauthorised user, anonymisation techniques ensures an individual's privacy from both unauthorised and authorised users. The anonymisation algorithms are developed to achieve privacy requirements determined by privacy models by using several approaches including generalisation, suppression, pseudonymisation and bucketisation. Most of the existing anonymisation approaches result in substantial information loss. This is the main motivation of conducting this research.

Chapter 4

THE DESIGN OF THE *kl-redInfo* ALGORITHM

In order to reduce the amount of information loss caused by most of the existing approaches disscussed in Chapter 3, this research proposes anonymisation algorithm, named *kl-redInfo*, that adopts bucket-creation part of the anatomisation approach from an Anatomy algorithm, but instead of sequentially incorporating the remaining records, as is done in the Anatomy algorithm explained in Section 4.1, the *kl-redInfo* algorithm incorporates the remaining records to an equivalence class that results in lower information loss. Also, instead of spliting the table into two parts, the cell-based generalisation approach is added in order to achieve *k-anonymity* privacy requirement, which is not achieved by Anatomy algorithm. Furthermore, a sorting approach is added in order to consider distribution of quasi-identifier attributes.

This chapter presents the proposed *kl-redInfo* algorithm, and discusses key features of the algorithm. The *kl-redInfo* algorithm adopts a bucketisation evolution from the Anatomy algorithm discussed in section 4.1. Section 4.2 describes the *kl-redInfo* algorithm. The algorithm walkthrough is described in section 4.3 and key features of the *kl-redInfo* algorithm are discussed in Section 4.4. Section 4.5 describes the high level architecture of the *kl-redInfo* solution, and lastly a summary of this chapter.

4.1 Anatomy Algorithm

Anatomy is a group-based approach addressing the issue of guaranteeing *l-diversity* privacy requirement of the anonymised dataset without using generalisation approach (Xiao and Tao, 2006b). The Anatomy algorithm is presented in Algorithm 1.

Algorithm 1: Anatomy algorithm (Xiao and Tao, 2006b)

```
Data: Original table T
Result: QIT and ST
QIT = $\emptyset$; ST = $\emptyset$; gcnt = 0 ;
Group the records in $T$ by their sensitive values (each group per sensitive value)
/* The bucket-creation step */
while there are at least l non-empty groups do
    gcnt = gcnt + 1; $QI_{gcnt} = \emptyset$
    $S$ = the set of $l$ largest groups;
    for each group in S do
        remove an arbitrary record $r$ from the group to form a bucket
        $QI_{gcnt} = QI_{gcnt} \cup \{ r \}$
    end
end
/* The remaining-incorporation step */
for each non-empty group do
    $r'$ = the remaining record of the group;
    $S'$ = the set of buckets produced from the previous step;
    sequentially assign $r'$ to a bucket in $S'$ ;
end
/* Populate QIT and ST */
for $j = 1$ to gcnt do
    for each record $r \in QIj$ do
        insert record $(r'_1, ..., r'_d, j)$ into QIT
    end
    for each distinct sensitive value $v$ in QIj do
        cj (v) = the number of records in QIj with As value $v$
        insert record (j, v, cj (v)) into ST
    end
    return QIT and ST
end
```

The algorithm first computes an l-diverse partition of the original table T (Lines 1-13), and then, splits the table into two parts; the table of quasi-identifiers (QIT) and the table of sensitive attributes (ST) (Lines 14-20). The l-diverse partition process involves group-creation and incorporation of the remaining records.

The bucket-creation step is performed in iterations, and continues until when there is less than l non-empty groups (Line 3). Each iteration results in a new quasi-identifier group QI_{gcnt} (Line 4). In order to ensure the formed l-records buckets are as many as possible, the Anatomy first selects a set of groups S consisting of the l groups that currently have the largest number of records (Line 4). Then, from each group in S (Line 5), a record is sequentially selected and added to a bucket QI_{gcnt} (Line 6). Therefore, QI_{gcnt} contains l records with different sensitive attributes values, named *bucket*.

To incorporate each of the remaining records *r*, the Anatomy selects a set *S'* of buckets (produced from the bucket-creation step), which does not have the same records as *r* (Lines 9-11). Then, in line 12, *r* is assigned to an arbitrary bucket in *S'*.

In order to split the table into two parts, the table of quasi-identifier (QIT) and the table of sensitive attributes (ST) (Lines 14-20), each group in *S'* is then associated with a unique group identifier. For each record, both in QIT and ST, notify the identifier of the group to which it belongs. For simplicity, each group in the ST has a record for each sensitive value appearing in the group, and notifies the frequency with which the value is represented in the group. Line 21 returns the formed anonymised QIT and ST tables.

The Anatomy algorithm result in unmodified quasi-identifier and sensitive values in two separate tables; QIT and ST table. The exact quasi-identifiers indicates the presence of a particular individual in the dataset. Therefore, Anatomy does not achieves the *k-anonymity* privacy requirement, which is important for controlling identity disclosure. Also, Anatomy algorithm does not consider the distribution of quasi-identifiers, and the remaining records are sequentially incorporated, which may result in records that are very different to be in the same QI-group, hence making them indistinguishable from each other may results in high information loss.

4.2 The Proposed *kl-redInfo* Algorithm

The *kl-redInfo* algorithm is the set of procedures that ensure an individual's privacy with reduced information loss. The algorithm is named *kl-redInfo* as it achieves *k-anonymity* and *l-diversity* privacy requirements with reduced amount of information loss. The individual's privacy is ensured by achieving the two main privacy requirements, *k-anonymity* and *l-diversity.* The information loss is reduced by

- Incorporating the remaining records to the group that results in a lower value of the information loss metric compared to when the records are incorporated to other groups
- Using both bucketisation and cell-based generalisation approaches
- Sorting the records according to the attributes that can be linked to identify an individual, also known as quasi-identifiers (QIDs).

The problem of ensuring an individual's privacy when sharing data is serious due to the increasing pressure of data sharing as a result of technology growth. In particular, the growth of social networks websites such as Facebook and Tweet that simplify the linking attack, as the information becomes available on a websites can be linked with other information to identify an individual's sensitive information such as disease. In addition the wide use of mobile storage devices such as laptops and external disks, which are easily stolen and misplaced, increases the need for anonymised data. These cause difficulty for data holders to use and share data for various useful purposes such as research, analysis, and public education. Most of the existing techniques may not ensure an individual's privacy or results in substantial information loss.

The *kl-redInfo* will be used by data holders to anonymise data that can be used for different purposes without identifying an individual. The data holders will use the algorithm to anonymise the data before giving them to data recipients such as researchers, analysts and policy makers. The data holders will enter the dataset to be anonymised and the values of parameter k and l, and the algorithm will provide the anonymised dataset.

Specifically, the *kl-redInfo* algorithm adopts bucketisation part of the Anatomy, *l-diversity*-specific algorithm. Then systematically incorporates the remaining records in a group that results in a lower value of the information loss metric instead of sequential incorporation, as is done in the Anatomy algorithm.

Second, instead of splitting the table into two parts, the *kl-redInfo* applies cell-based generalisation approach in every group in order to make the quasi-identifiers indistinguishable from each other. Also, the records are sorted according to the quasi-identifiers in order to consider their distributions. The *kl-redInfo* algorithm can be used by any domain, but its implementation may need to be customised depending on the quasi-identifiers to be anonymised. The available implementation uses the commonly used quasi-identifier attributes including date of birth, address, gender, and marital status. Algorithm 2 presents the *kl-redInfo* algorithm.

Algorithm 2: The proposed *kl-redInfo* algorithm (Author, 2013)

```
Data: Original table T
Result: Anonymised table T*
gcnt = 0 ;
Sort the records in $T$ according to their quasi-identifiers (QIDs)
Group the records in $T$ by their sensitive values (each group per sensitive value)
/* The bucket-creation step */
while there are at least $l$ non-empty groups do
    gcnt = gcnt + 1; $QI_{gcnt} = \emptyset$
    $S$ = the set of $l$ largest groups;
    for each group $G$ in $S$ do
        remove an arbitrary record $r$ from the group to form a bucket B
        B = $QI_{gcnt} \cup \{ r \}$
    end
end
/* Generalisation step */
for each bucket $B$ do
    check if QID values are the same
    if they are not the same then
        generalise the values
    end
end
/* The Incorporation step */
$r$ = the remaining record;
$B$ = the set of buckets produced from the generalisation step;
 while there exists groups G' such that |G'| <$l$ do
    for each remaining record $r$ in $G$ do
        Calculate information loss(B∪r )
    end
    Incorporate r in B with lower information loss
    Insert B into T*
end
return T*
```

The algorithm starts by sorting the original table T according to quasi-identifiers (Line 2) in order to take under consideration the distribution of the quasi-identifiers. Then, the algorithm adopts bucket-creation part from Anatomy algorithm (Line 3-9) to form buckets with l distinct sensitive values. Thereafter, the *kl-redInfo* algorithm applies the cell-based generalisation approach within each bucket (local generalisation) to form equivalence classes (Line 10-14).

Instead of the remaining records being sequentially incorporated, the *kl-redInfo* algorithm calculates the resulting information loss before each remaining record is incorporated in an equivalence class (Line 15-20). Then the remaining record is incorporated into the equivalence class that results in a lower value of the information loss metric (Line 21). Lastly, the bucket is inserted in the anonymised table T* and returned (Line 23).

4.3 Algorithm Walkthrough

The algorithm first sorts the records according to their QID values, then groups the records according to their sensitive attribute values, thereafter recursively selecting l records from l distinct groups to form buckets. Then each bucket is generalised to form equivalence classes. When the number of groups are less than l, the information loss resulting from the application of incorporating each of the remaining records in equivalence classes is calculated. The remaining record is incorporated into an equivalence class that results in lower information loss.

Table 4.1: Patients' Information Table

No.	*Date of birth*	*Gender*	*P.O. Box*	*Disease*
1	1981/07	M	12386	Cancer
2	1978/02	F	12362	Obesity
3	1962/05	M	12337	Obesity
4	1978/02	F	12395	Malaria
5	1978/10	F	12381	HIV
6	1981/09	M	12352	Obesity
7	1978/10	F	12381	HIV
8	1962/08	F	12394	Cancer
9	1981/04	M	12380	Malaria

For example, for the Patients' Information Table 4.1 to satisfy 2-diversity, first, records are sorted according to their QID values; DOB, Gender and Address, as shown in Table 4.2.

Table 4.2: Records sorted according to QID values

Record	*Date of birth*	*Gender*	*P.O. Box*	*Disease*
r_3	1962/05	M	12337	Obesity
r_8	1962/08	F	12394	Cancer
r_2	1978/02	F	12362	Obesity
r_4	1978/02	F	12395	Malaria
r_5	1978/10	F	12381	HIV
r_7	1978/10	F	12381	HIV
r_9	1981/04	M	12370	Malaria
r_1	1981/07	M	12386	Cancer
r_6	1981/09	M	12352	Obesity

Then records are grouped according to the Disease sensitive attribute, and four groups are formed and sorted according to the number of records, highest to smallest: G_1 = {r_3, r_2, r_6 }, G_2 = {r_8, r_1 }, G_3 = { r_4, r_9 }, G_4 = {r_5, r_7 }, where r_i denotes the i^{th} record in the table as shown in Table 4.3.

Table 4.3: Records grouped according to Sensitive attribute

Record	*Date of birth*	*Gender*	*P.O. Box*	*Disease*
r_3	1962/05	M	12337	Obesity
r_2	1978/02	F	12362	Obesity
r_6	1981/09	M	12352	Obesity
r_8	1962/08	F	12394	Cancer
r_1	1981/07	M	12386	Cancer
r_4	1978/02	F	12395	Malaria
r_9	1981/04	M	12370	Malaria
r_5	1978/10	F	12381	HIV
r_7	1978/10	F	12381	HIV

Second, r_3 and r_8 are selected from G_1 and G_2 and bucketised. This forms the first bucket. This process continues until when the number of groups are less than l (in this case l=2). Table 4.4 shows buckets formed with respect to this example. Records are continuously selected from l-distinct groups and bucketised. Then cell-generalisation is applied in each bucket to form an equivalence class.

Table 4.4: The First Bucket

Record	*Date of birth*	*Gender*	*P.O. Box*	*Disease*
r_3	1962/05	M	12337	Obesity
r_8	1962/08	F	12394	Cancer
r_2	1978/02	F	12362	Obesity
r_4	1978/02	F	12395	Malaria
r_5	1978/10	F	12381	HIV
r_1	1981/07	M	12386	Cancer
r_6	1981/09	M	12352	Obesity
r_9	1981/04	M	12370	Malaria

The information loss resulting from the application of incorporating the remaining record r_7 in each bucket is then calculated. Any of the information loss metrics can be used to calculate the information loss, as discussed in Section 5.4. This research uses Normalised Certainty Penalty (NCP) due to the fact that it is a metric that considers the effect of the generalisation process which is the main cause of the information loss when anonymising data. Since incorporating record r_7 in bucket 2 results in a lower value of the information loss metric compared to when it is incorporated in other buckets, the record r_7 is incorporated into bucket 2. The final shared table is created as shown in Table 4.5.

Table 4.5: Anonymised records as a result of *kl-redInfo* algorithm

Record	*Date of birth*	*Gender*	*P.O. Box*	*Disease*
r_3	1962	*	123**	Obesity
r_8	1962	*	123**	Cancer
r_2	1978	F	123**	Obesity
r_4	1978	F	123**	Malaria
r_7	1978	F	123**	HIV
r_5	*	*	1238*	HIV
r_1	*	*	1238*	Cancer
r_6	1981	M	123**	Obesity
r_9	1981	M	123**	Malaria

4.4 Key Features of the *kl-redInfo* algorithm

The proposed *kl-redInfo* algorithm has key unique features compared to the existing algorithms. These features are: systematic incorporation of the remaining records, using both bucketisation and cell-based generalisation approaches, and considering the distribution of the quasi-identifier attributes. These key features cause the *kl-redInfo* to result in significant lower information loss compared to the widely used algorithms. These features are discussed in section 4.4.1, 4.4.2, and 4.4.3, and their significance in reducing the information loss will be evaluated in Chapter 6.

4.4.1 Systematic Incorporation of the Remaining Records

Rather than sequentially incorporating the remaining records, the *kl-redInfo* algorithm incorporates the remaining records to the equivalence class that results in a lower value of the information loss metric (systematically). This helps to reduce the amount of information loss as will be discussed in Chapter 6.

The incorporation process starts if the number of the remaining groups is less than the required *l-value*, therefore the bucket with l-distinct sensitive values cannot be formed. In each iteration, the remaining record is incorporated with the bucket such that the formed bucket has the smallest weighted Normalised Certainty Penalty (NCP). The iteration continues until every remaining record is incorporated in the appropriate bucket.

The weighted NCP was used as it measures the information loss in terms of the generalisation applied instead of the size of equivalence classes measured by the Discernibility Penalty (DP). Therefore, since the size of the equivalence classes is almost equal due to the bucketisation process, DP results in no difference when a remaining record is incorporated. The KL-divergence measures the similarity between the original and the anonymised dataset and not between the groups. The weight was assigned depending on the number of distinct values, the higher the number of distinct values the higher the weight, as that shows the high possibility of identifying an individual. By default, the weighting of each attribute used in the evaluation of information loss is equal to 1/|QID|, where |QID| is the QID size.

4.4.2 Using both bucketisation and cell-based generalisation approaches

Since the implementation of *l-diversity* largely relies on the distribution of sensitive attributes values, a new inspiration is to first, bucketise the records according to their sensitive attributes values, and then recursively selecting l records from l

distinct buckets and groups them into an equivalence class. As for the remaining records, incorporating each of them into an equivalence class results in lower information loss. Cell-based generalisation approach is then applied in each group in order to achieve *k-anonymity* privacy requirement. The resulting table will satisfy both *k-anonymity* and *l-diversity* privacy requirement with lower information loss, as will be justified in Chapter 6.

4.4.3 Considers Distribution of the Quasi-identifiers

The *kl-redInfo* algorithm takes under consideration the distribution of the quasi-identifiers by sorting the records according to quasi-identifier attributes. This approach reduces the possibility of the records that have very different quasi-identifiers being in the same group. This approach seeks to reduce the amount of information loss but its contribution is not significant. This is because the records were again grouped according to the sensitive attribute. These results will be discussed in Chapter 6.

Characteristics of the existing algorithms compared with the proposed *kl-redInfo* are summarised in Table 4.6.

Table 4.6: Characteristics of the Existing Algorithms compared to the Proposed (Author, 2013)

Algorithm	**Characteristics**				
	Methods used	*Type of generalisation*	*Privacy Models*	*Strengths*	*Weaknesesses*
μ-Argus (Hundepool and Willenborg, 1996)	Generalisation and Suppression	Cell-suppression	k-anonymity	Low information loss	The results are not always guaranteed to be k-anonymous (LeFevre et al., 2005)
Datafly (Sweeney, 1997)	Generalisation and Suppression	Full-domain generalisation	k-anonymity	Generalisation is guaranteed to be k-anonymous (LeFevre et al., 2005)	It can over-generalise data (Sweeney, 2002)
Incognito with k-anonymity (LeFevre, 2005)	Generalisation and Suppression	Full-domain generalisation	k-anonymity	Protects against identity disclosure	Cannot resist homogeneity and background attacks (Han and Yu, 2008)
Incognito with l-diversity (Machanavajjhala, 2007)	Generalisation	Full-domain generalisation	k-anonymity and l-diversity	Resist homogeneity and background attacks (LeFevre et al., 2005)	It results in high information loss (Li et al., 2007)
Mondrian (LeFevre, 2006)	Generalisation and Suppression	Multi-dimensional generalisation	k-anonymity	It is more flexible (LeFevre, 2006)	It is less scalable due to the increased search space (Xu et al., 2006)
Anatomy (Xiao and Tao, 2006)	Anatomisation	Not using generalisation	l-diversity	Results in unmodified data (Fung et al., 2010)	It does not achieve *k-anonymity* privacy requirement (Xiao and Tao, 2006b)
kl-redInfo (Author, 2013)	Bucketsation and Generalisation	Cell- generalisation	k-Anonymity and l-Diversity	An adequate level of privacy with reduced information loss	Values might be anonymised in different generalisation levels

4.5 A solution Architecture

In Figure 4.1 the high level representation architecture of the proposed algorithm, *kl-redInfo* is represented. The main components and the relationships between them are identified. These components are databases, algorithm engine, and user interface. The database components are the data storage for original data and anonymised data. The algorithm engine comprises all algorithms designed, and facilitates communication between all components. The user interface component provides interface through which users interact with the system.

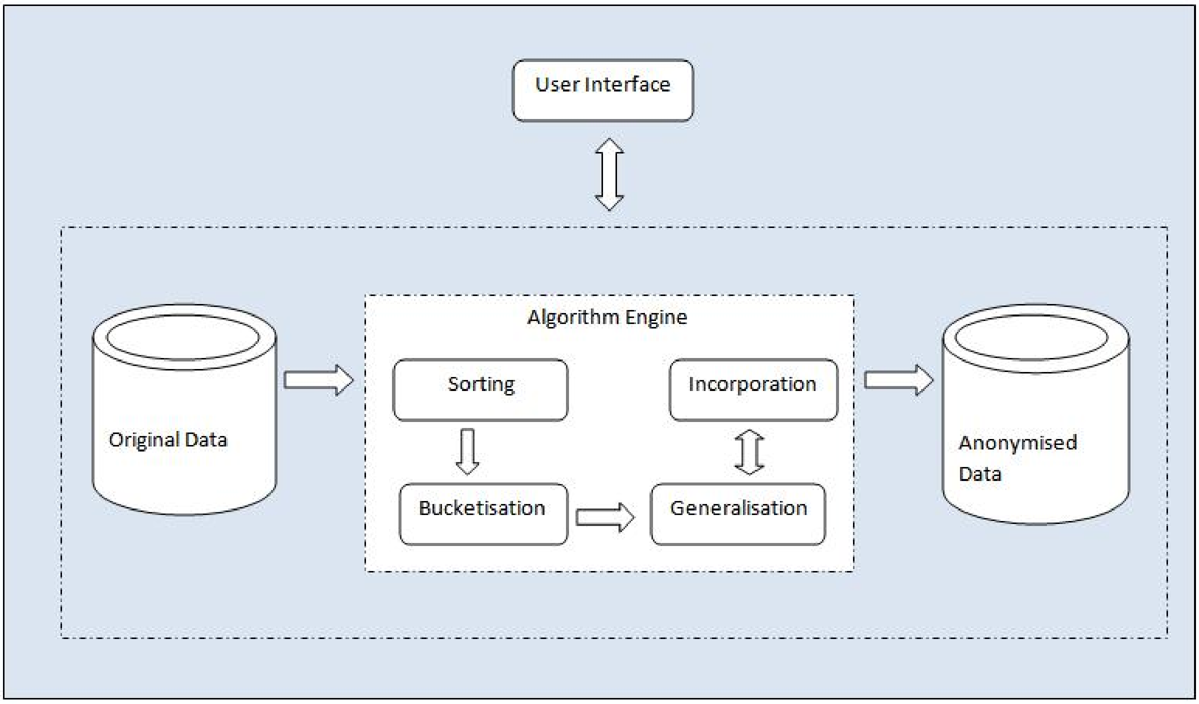


Figure 4.1: A Solution Architecture (Author, 2013)

4.6 Chapter Summary

The *kl-redInfo* algorithm achieves both *k-anonymity* and *l-diversity* privacy requirements. Both privacy requirements are necessary for effective privacy protection. *k-anonymity* ensures that an individual's data cannot be distinguishable by linking the quasi-identifier attributes, also known as identity disclosure. *l-diversity* eliminates a possibility to associate an individual with sensitive attributes. Also known as attribute disclosure.

Most of the existing approaches results in substantial information loss or the anonymisation level achieved may still results in the identification of the individual's sensitive information. Therefore, the proposed *kl-redInfo* algorithm uses systematic incorporation of the remaining records bucketisation and cell-based generalisation approaches, and sorting the records according to quasi-identifier attributes. The combination in this approach generates the anonymisation dataset that satisfies both *k-anonymity* and *l-diversity* privacy requirements with lower information loss. Thus, it maintains the usefulness of the data being shared. The bucketisation approach was used to achieve *l-diversity* privacy requirement while cell-based generalisation approach was used to achieve *k-anonymity* privacy requirement. The significance of each feature will be discussed in Chapter 6.

Chapter 5

EXPERIMENTAL ENVIRONMENT AND SETUP

This chapter discusses the environment used to evaluate and validate the implemented algorithms, *kl-redInfo l-mondrian*, and *g-anatomy*. This environment includes the datasets and the parameters used, and these are discussed in section 5.2. The algorithms used for comparison, *l-mondrian*, *g-anatomy*, are discussed in section 5.3. Three information loss metrics, Discernibility Penalty (DP) (Bayardo and Agrawal, 2005), Normalised Certainty Penalty (NCP) (Xu et al., 2006) and Kullback_Leibler divergence (KL_divergence) (Kifer and Gehrke, 2006), used by this research to calculate the amount of information loss of the implemented algorithms are discussed in Section 5.4 and the Wilcoxon signed-rank test statistics used for analysis is discussed in section 5.6.

5.1 Experiment Setup

The proposed *kl-redInfo* algorithm was experimentally evaluated and compared with the widely used algorithms, *l-mondrian* and *g-anatomy*. The algorithms used are implemented in Java and uses MySQL open source database to store the datasets. All experiments were implemented in Linux (Ubuntu 10.04 LTS- the Lucid Lynx) on a computer with a 2.26 GHz Intel(R) Core(TM)2 Duo CPU and 1 GB RAM.

The Statistical Package for the Social Sciences (SPSS) was used for analysis purpose.

Java is one of the most widely used programming languages and MySQL is the world's most popular open source database management system (Arnold et al., 2000; Flanagan, 2005). In order for the two technologies, Java and MySQL database to work, they have to be connected. MySQL Connector/J driver was used to connect the two technologies. MySQL Connector/J is a native Java driver that converts JDBC (Java Database Connectivity) requests into the network protocol used by the MySQL database. It is the official JDBC driver for MySQL, which can be downloaded from *http://dev.mysql.com/downloads/connector/j/.* The JDBC is an interface for accessing relational databases from Java and is used to maintain the databases connection, issues database queries and updates and receives the results.

A Java Development Kit (JDK) called Eclipse, was installed for compiling and running Java programs. The results from the Java programs were then copied to the SPSS software for graphical represention (histograms) and comparison was done by using a paired difference non-parametric Wilcoxon signed-rank test statistics. This is due to the fact that, the number of the information loss to be evaluate is small and the values are not normally distributed. Therefore, the use of parametric test is not appropriate.

5.2 Datasets and Parameters

The experiments were executed on two different datasets; the generated patient information dataset and the Adult dataset. The generated patient information dataset, (for the purposes ot this research it is named PatInfo), has 30,200 records and eight attributes with seven quasi-identifiers and one sensitive attribute. The research used simulated dataset since the use of real data was not possible, as discussed in section 2.5.2. The schema of the PatInfo dataset is based on the schema of the Muhimbili National Hospital (MNH) in Tanzania where the survey was completed. Table 5.1 provides a description of the PatInfo dataset including the attributes, the number of distinct values for each attribute, and the height of the generalisation hierarchy for each attribute.

Table 5.1: The PatInfo dataset schema

	Attribute	***Domain size***	***Height***
1	Date of Birth	880	3
2	Gender	2	1
3	Address	33	3
4	Marital Status	7	2
5	Admission Date	876	3
6	Discharge Date	879	3
7	Discharge Status	3	1
8	Disease	14	Sensitive attribute

The PatInfo dataset was generated by using Data Generator software downloaded from *http://www.generatedata.com/#about.* The software was installed in the computer where the experiments are implemented. After filling in the possible values of the attributes, the Data Generator software generates Structured Query Language (SQL) syntax for creating a table and randomly inserting the attribute values. The SQL syntax was then copied to the MySQL database where the tables are stored for the experiments.

To show that the *kl-redInfo* algorithm also works in other datasets, this research also used the real-world census dataset, called Adult dataset, downloaded from UCI Machine Learning Repository at *http://archive.ics.uci.edu/ml/datasets/Adult.* The dataset was downloaded from the repository and stored in MySQL database. After removing records with missing values, the dataset remained with 30,162 records and eight attributes with seven quasi-identifiers and one sensitive attribute.

The Adult dataset was selected as it is the most widely used as a benchmark dataset in previous research, therefore it is stable and trusted dataset. In additional to that, Adult dataset has most of the information that can be found in any healthcare domain such as age, gender, marital status and address (Samarati, 2001; Sweeney, 2002; LeFevre et al., 2005; Machanavajjhala et al., 2007). Table 5.2 provides a description of the Adult datasets including the attributes, the number of distinct values for each attribute, and the height of the generalisation hierarchy for each attribute.

Table 5.2: The Adult dataset schema

	Attribute	***Domain size***	***Height***
1	Age	72	4
2	Gender	2	1
3	Marital Status	7	2
4	Race	5	1
5	Education	16	3
6	Native Country	41	2
7	Work class	7	2
8	Occupation	14	Sensitive attribute

This research uses the commonly used generalisation hierarchies such as Date, Gender and Marital status, as presented in Figure 5.1 (Samarati, 2001; Sweeney, 2002; LeFevre et al., 2005).

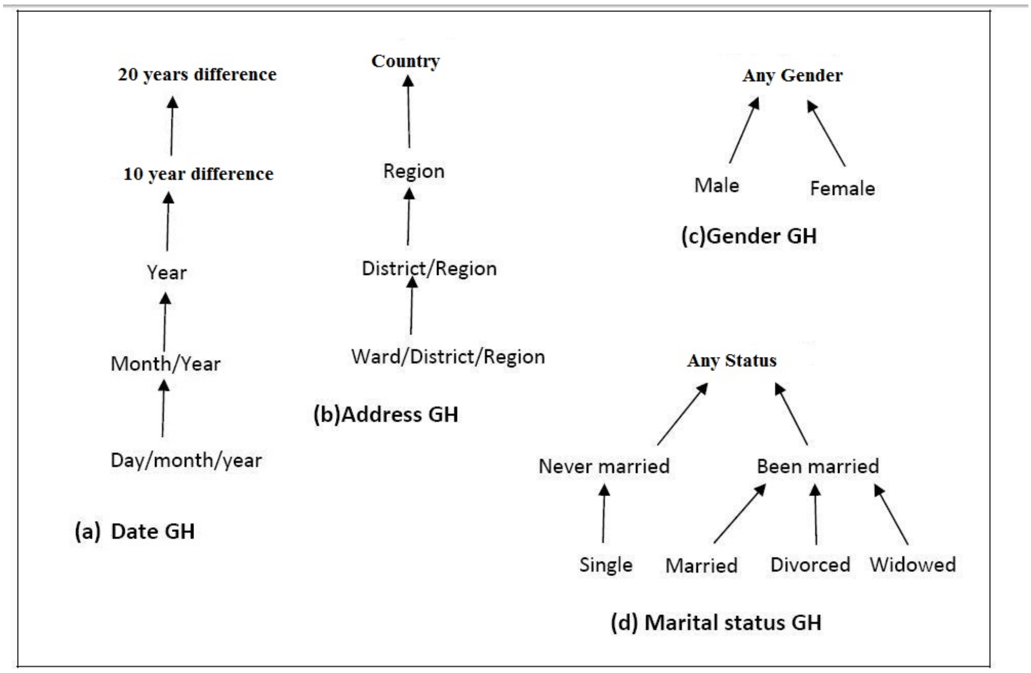


Figure 5.1: Generalisation hierarchies of the QIDs used in this research (LeFevre et al., 2005)

The *k* and *l* are the main parameter values in the experiments. These parameters have two different domains: the *k-value* parameter controls the number of records with the same quasi-identifiers, also known as equivalence class (EC) , while the *l-value* parameter controls the number of sensitive values within each equivalence class. Thus, let n be the total number of records, m be the total number of sensitive values existing in a table, then the *k-value* can vary from 1 to the total number of records $(1 \leq k \leq n)$ while *l-value* varys from 1 to the total number of sensitive values existing in the table $(1 \leq l \leq m)$.

In order to equally consider the identity disclosure, represented by k value, and the attribute disclosure, represented by l value, this research sets the values of $k = l$. The *l-diversity* privacy requirement is defined as for every equivalence class there should be at least l well-represented sensitive values. This indicates that the value of l cannot be greater than the value of k, that is, $l \leq k$.

When $l < k$, the implemented algorithms show no changes on the values of the information loss metrics compared to when $k = l$ was used. This shows that values of l have no effect on the values of the information loss metrics since l-diverse table is automatically l-anonymous. The effect of l values will be on the individual's privacy, since the lower the values of l the lower distinct sensitive values in the equivalence class, thus more possibility of the attribute disclosure.

The k and l values can not be greater than the number of sensitive values in the table ($k \leq m$ and $l \leq m$), 14 in the PatInfo dataset and 14 in the Adult dataset. Hence, there is no table which can be more than 14-diverse for any reasonable definition of *l-diversity* privacy requirement. In practice, a minimal value of k and $l = 3$ is sometimes recommended, but more often a value of k and $l = 5$ is used (Machanavajjhala et al., 2007; Fung et al., 2010). To ensure a reasonable amount of variation in our analysis this research uses all possible values of k and l, that is, all values between 2 and 14 inclusive.

5.3 Implemented Algorithms

When choosing algorithms with which to compare with the proposed *kl-redInfo* algorithm, the following criteria were considered. First, in order to be free from the errors that might be due to the implementation, effort was made to get access to the source code from developers of the algorithms. The source-code of the *l-diversity* version of Incognito was obtained. Second, consideration was not made in comparing against any algorithm that had already been shown to produce lower quality anonymisations than the state-of-the-art *k-anonymity* algorithm, called Mondrian. This eliminated the Incognito algorithm from being used.

Also, since the *kl-redInfo* algorithm achieves *l-diversity* privacy requirement, which is the enhancement of *k-anonymity* privacy requirement, consideration was not made in comparing against algorithms that do not achieve *l-diversity* privacy requirement. The *l-diversity* privacy requirement can be achieved either by extending *k-anonymity* algorithms (Machanavajjhala et al., 2007), or creating a new algorithm that specifically designed to achieve *l-diversity* privacy requirement, such as Anatomy algorithm (Xiao and Tao, 2006a). Even though the Anatomy algorithm is designed to achieve *l-diversity* privacy requirement, it is not achieving the basic *k-anonymity* privacy requirement.

Therefore, this research adds the required criteria in order for the algorithms to achieve both *k-anonymity* and *l-diversity* privacy requirements. Thus, the *l-diversity* criteria was added on the Mondrian algorithm to achieve *l-diversity* privacy requirement, for the purposes of this research this algorithm is named *l-mondrian.* Also, the *k-anonymity* criteria was added on the Anatomy algorithm to achieve *k-anonymity* privacy requirement, this research named the algorithm as *g-anatomy*. Therefore, the *l-mondrian* algorithm extended from Mondrian multidimensional *k-anonymity*, and *g-anatomy* algorithm extended from Anatomy algorithm, were implemented for comparison purposes, and they are further discussed in Section 5.3.1 and 5.3.2.

5.3.1 The *l-mondrian* Algorithm

The *l-mondrian* is an algorithm which achieves *l-diversity* privacy requirement by extending the Mondrian multidimensional *k-anonymity* algorithm (LeFevre, 2006; Xu et al., 2006; Ghinita et al., 2009). The Mondrian algorithm was originally proposed in LeFevre (2006) for *k-anonymity*. The algorithm is extended to achieve *l-diversity* by checking for *l-diversity* in addition, every time when the algorithm is checking for *k-anonymity* privacy requirement (Machanavajjhala et al., 2007).

The Mondrian algorithm uses a greedy top-down approach to recursively partition the (multidimensional) quasi-identifier domain space. It uses a search strategy which recursively splits a group of records at the median value of a chosen attribute, until the partitions created by the split contain at least *k* but no more than *2k-1* records. In order for each group to have approximately uniform partition, the attribute with the largest normalised range of values is used to split the group. This is because the larger the spread/range, the easier the good split point can be found and more likely the data can be further split.

For continuous or ordinal attributes the data is partitioned around the median value of the split attribute. This process is repeated until no allowable split remains, meaning that a particular group cannot be further divided without violating the privacy requirements. Algorithm 3 presents *l-mondrian* algorithm.

Algorithm 3: ***l-mondrian*** **algorithm (LeFevre et al., 2008)**

```
Data: Original table T
Result: Anonymised records
Anonymise(records, attrs)
if no allowable split for records then
    return $\phi$ : r $\in$ records $\rightarrow$ bounding region(records)
end
else
    best $\leftarrow$ Choose Attribute(attrs, records)
    if continuous(best) or ordinal(best) then
        threshold $\leftarrow$ Choose Threshold(best)
        lhs $\leftarrow$ {r $\in$ records : r.best $\leq$ threshold}
        rhs $\leftarrow$ {r $\in$ records : r.best $>$ threshold}
        return Anonymise(rhs,attrs) $\cup$ Anonymise(lhs,attrs)
    end
end
else
    if nominal(best)
    recodings $\leftarrow$ { }
    for each child $v_i$ of root(best.hierarchy) do
        records$_i$ $\leftarrow$ {r $\in$ records : r.best v$_i$ }
        attrs $\leftarrow$ replace root(best.hierarchy) with v$_i$ in attrs
        recodings $\leftarrow$ recodings $\cup$ Anonymise(records$_i$ , attrs )
    end
    return recodings
end
```

5.3.2 The *(g-anatomy)* Algorithm

As the Anatomy algorithm does not prevent identity disclosure, this research updates Anatomy algorithm, presented in Figure 4.1, by adding generalisation approach (Line 14-19) instead of separating the table in two different tables (QIT and ST). This research names this algorithm *g-anatomy.* Algorithm 4 presents the *g-anatomy* algorithm updated from Anatomy algorithm discussed in Section 4.1 presented in Algorithm 1. The added lines 14-19, in Algorithm 4, enables the *g-anatomy* algorithm to achieve the basic *k-anonymity* privacy requirement which prevents identity disclosure.

Algorithm 4: ***g-anatomy*** **algorithm (Xiao and Tao, 2006b)**

```
Data: Original table T
Result: Anonymised table T*
gcnt = 0 ;
Group the records in $T$ by their sensitive values (each group per sensitive value)
/* The bucket-creation step */
while there are at least $l$ non-empty groups do
    gcnt = gcnt + 1; $\mathrm{QI}_{gcnt} = \emptyset$
    $S$ = the set of $l$ largest buckets;
    for each group in $S$ do
        remove an arbitrary record $r$ from the group to form a bucket
        $\mathrm{QI}_{gcnt} = \mathrm{QI}_{gcnt} \cup \{ r \}$
    end
end
/* The remaining-assignment step */
for each non-empty group do
    $r'$ = the remaining record of the group;
    $S'$ = the set of buckets produced from the previous step;
    assign $r'$ to a bucket in $S'$ sequentially;
end
/* Generalisation step */
for $j$ = 1 to gcnt do
    for each bucket $B$ do
        check if QID values are the same
        if they are not the same then
            generalise the values
            else
                Insert QIgcnt into T*
            end
        end
    end
end
return T*
```

Characteristics of the implemented algorithms are summarised in Table 5.3

Table 5.3: Characteristics of the Implemented Algorithms

Algorithm	**Characteristics**			
	Methods used	*Privacy Models*	*Strengths*	*Weaknesesses*
l-mondrian	- Generalisation - Sequential incorporation of ECs	k-Anonymity and l-Diversity	An adequate level of privacy	Substantial information loss
g-anatomy	- Bucketisation and Generalisation -Sequential incorporation of records	k-Anonymity and l-Diversity	An adequate level of privacy	Substantial information loss, the results might be non-homogeneous
kl-redInfo	-Sorting - Bucketisation and Generalisation -Systematic incorporation of records	k-Anonymity and l-Diversity	An adequate level of privacy with reduced information loss	Values might be anonymised in different generalisation levels

The information loss resulting from the application of these two algorithms, *l-mondrian* and *g-anatomy*, were compared to the information loss resulting from the application of the proposed *kl-redInfo* algorithm, to indicate its improvements. The following section discusses the evaluation metrics used in this research.

5.4 Evaluation Metrics

Anonymisation approach has two main aspects; privacy preserving aspect and information retention aspect so that the shared data remains useful. Quantifying the notion of information loss is one of the key challenges in the privacy-preserving data publishing domain (Machanavajjhala et al., 2007; Xu et al., 2006; Kifer and Gehrke, 2006). The information loss of a dataset can be measured based on different characteristics, such as number of records that are indistinguishable from each other, number of generalisation steps, and average group size of the equivalence classes, which results in the existence of several information loss metrics.

This research uses three widely used information loss metrics, Discernibility Penalty (DP) (Bayardo and Agrawal, 2005), Normalised Certainty Penalty (NCP) (Xu et al., 2006) and Kullback-Leibler divergence (KL-divergence) (Kifer and Gehrke, 2006). These three metrics have been acknowledged as appropriate representative metrics in the data anonymisation literature (Ghinita et al., 2007; Machanavajjhala et al., 2007; El Emam et al., 2009). Therefore, these metrics are good indicators of the information loss of the anonymised datasets.

The DP measures the information loss based on the size of the equivalence classes, but it does not measure how much the generalised records approximate the original records. The NCP is used because it takes into account both the size of the equivalence classes and the generalisation process used. Neither the NCP, nor the DP take the data distribution into account, thus this research also uses the KL-divergence which takes into account the data distribution. Therefore, the use of these three metrics has a good spread of the indicators of information loss. These metrics are further discussed in section 5.4.1, 5.4.2, and 5.4.3.

5.4.1 Discernibility Penalty (DP)

The Discernibility Penalty is the measure of information loss based on the number of records that are indistinguishable from each other, also known as equivalence classes. The idea behind the Discernibility Penalty (DP) metric is that, the more records are indistinguishable from each other the more the information loss. This is because more generalisation is required to make the records indistinguishable from each other. Therefore, the ideal algorithm should reduce the Discernibility Penalty (DP) by reducing the size of the equivalence classes. The smaller the size of the equivalence classes results in lower DP which implies the lower information loss.

The Discernibility Penalty (DP) is calculated by assigning a cost to information loss to each record based on how many other records are indistinguishable from it (Bayardo and Agrawal, 2005). If a record is suppressed, its cost to information loss is the number of records in the original dataset |T|. This is due to the fact that a suppressed record cannot be anonymised without anonymising all records in the dataset. If a record is not suppressed, its cost to information loss is the number of records in its anonymised group, also known as equivalence class |E|.

Therefore, the Discernibility Penalty is the sum of the squares of the equivalence class sizes plus the number of records in the original dataset times the number of suppressed records |R|, as shown in equation (5.4.1), where m is the number of the equivalence classes.

$$DP(T) = \sum_{i=1}^{m} \mid E_i \mid^2 + \mid T \mid\mid R \mid \tag{5.4.1}$$

When the record suppression approach is not used, the Discernibility Penalty (DP) is equivalent to the sum of the squares of the sizes of the equivalence classes, mathematically represented as shown in equation (5.4.2), where m is the number of the equivalence classes.

$$DP(T) = \sum_{i=1}^{m} \mid E_i \mid^2 \tag{5.4.2}$$

5.4.2 The Normalised Certainty Penalty (NCP)

The Normalised Certainty Penalty (NCP) is the measure of information loss that measures the importance of the attributes by considering the effect of the generalisation approach (Xu et al., 2006). The Normalised Certainty Penalty for a numeric attribute value measures its normalised interval size after generalisation, while for a categorical attribute value, the NCP measures its normalised number of descendants in the hierarchy tree after generalisation.

By considering the case of numeric attributes, let T be a dataset table with N number of records and n number of quasi-identifiers ($A_1,\ldots,A_n$), where all attributes are numeric. Suppose a record r_1 = ($x_{1i},\ldots,x_{1n}$) is generalised to record $r_1{}^*$ = ([y_{11} , z_{11}],...,[y_{1n} , z_{1n}]) such that $y_{ij} \leq x_{ij} \leq z_{ij}$ (1 ≤ i ≤ n) (1 ≤ j ≤ N). On attribute A_i and weight w_i, the NCP is defined as shown in equation (5.4.3), where $|A_i|$ = $\max_{r\in T}$ {r.A_i } - $\min_{r\in T}$ {r.A_i } is the range of all records on attribute A_i.

$$NCP_{num}(T) = \sum_{j=1}^{N}\sum_{i=1}^{n}(w_i * \frac{z_{ij} - y_{ij}}{\mid A_i \mid}) \tag{5.4.3}$$

A weight is assigned to each attribute to reflect its importance in the analysis on the anonymised data. In this research the weight is assigned depending on the number of the quasi-identifier attributes (QID). That is the weight of each attribute used in the evaluation of information loss is equal to 1/|QID|, where |QID| is the number of quasi-identifier attributes.

For categorical attribute, suppose a record r_1 has the value v_{1j} on a categorical attribute A_1. When it is generalised in anonymisation, the value v_{1j} will be replaced by a set of values $v_{1j},\ldots,v_{nj}$, where $v_{1j},\ldots,v_{nj}$ are the values of records on the attribute that is generalised to the same value u_{ij}. The Normalised Certainty Penalty of T with categorical attributes (NCP_{cat}(T)) is defined as shown in equation (5.4.4), where $|A_i|$ is the number of distinct values on attribute A_i.

$$NCP_{cat}(T) = \sum_{j=1}^{N}\sum_{i=1}^{n}(w_i * \frac{size(u_{ij})}{\mid A_i \mid}) \tag{5.4.4}$$

5.4.3 Kullback-Leibler divergence (KL-divergence)

The Kullback-Leibler divergence (KL-divergence) is the measure of information loss based on the similarity between the values in the original dataset and the values in the anonymised dataset. The KL-divergence is modeled as the difference between two probability distributions (Kifer and Gehrke, 2006; Machanavajjhala et al., 2007). In this research, the two distributions are the original dataset distribution and the anonymised dataset distribution. The KL-divergence is a non-negative metric and is 0 only when the two distributions are identical.

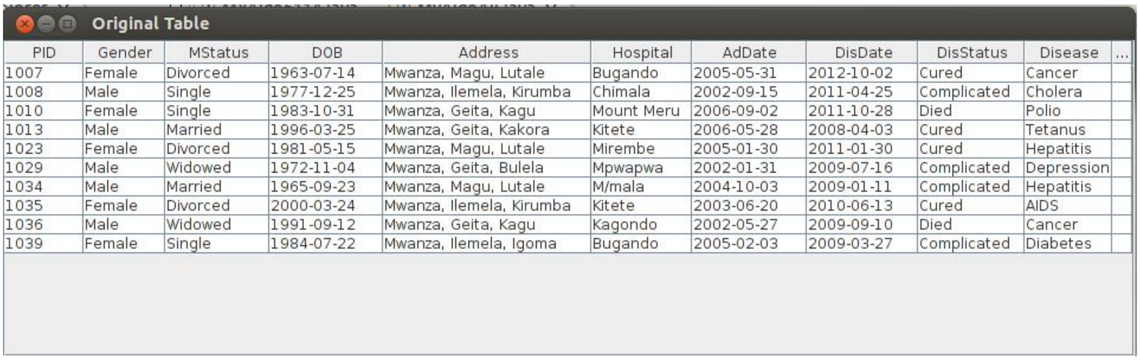
Let ($\mathrm{A}_1,\ldots,\mathrm{A}_n$) be the quasi-identifiers of a table T with values (x_{ij}, *i= 1,. . . ,n; j = 1,. . . ,N*). Let $p_i\ ^{(1)}$ be the probability of A_i according to the distribution F_1 and let $p_i\ ^{(2)}$ be the probability according to distribution F_2, where a probability distribution F_1 associated with the original data, and a probability distribution F_2 associated with the anonymised data. The KL-divergence between F_1 and F_2 is defined as shown in equation (5.4.5) (Kifer and Gehrke, 2006).

$$\text{KL-divergence } = \sum_{i=1}^{n} \mid p_i^{(1)} \log_{10}(\frac{p_i^{(1)}}{p_i^{(2)}}) \mid$$

$$\text{where } p_i = \frac{\text{Number of occurrences of } i}{\text{Total number of values}} \tag{5.4.5}$$

5.5 Software Verification

To ensure the correctness of the software implementation of the algorithms and accuracy in computing the relative information loss metrics, the computational of the three information loss metris, DP, NCP, and KL-divergence of a 10 records dataset were computed by hand. The results computed by hand are the same with the results of the implemented algorithm, as shown in this section. The research uses table shown in Figure 5.2 as the original table that needs to be anonymised.



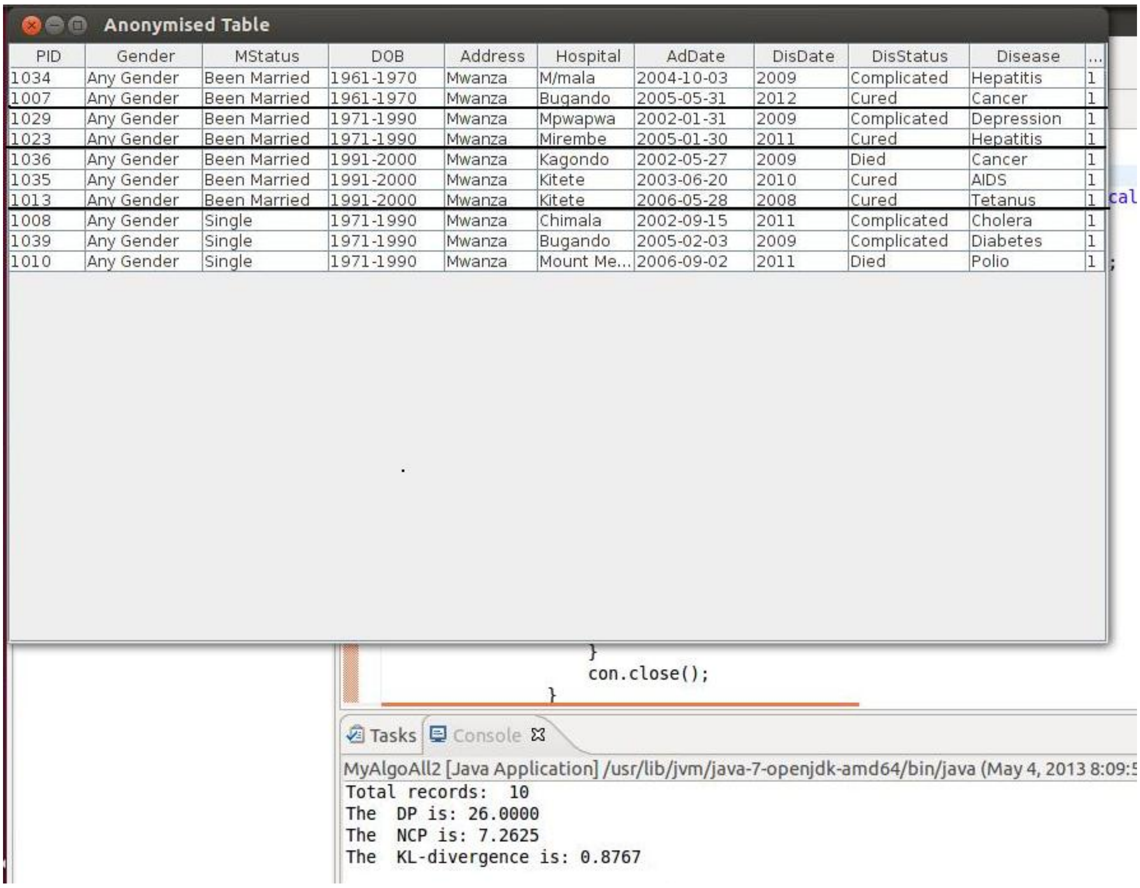
Original Table

PID	Gender	MStatus	DOB	Address	Hospital	AdDate	DisDate	DisStatus	Disease	...
1007	Female	Divorced	1963-07-14	Mwanza, Magu, Lutale	Bugando	2005-05-31	2012-10-02	Cured	Cancer	
1008	Male	Single	1977-12-25	Mwanza, Ilemela, Kirumba	Chimala	2002-09-15	2011-04-25	Complicated	Cholera	
1010	Female	Single	1983-10-31	Mwanza, Geita, Kagu	Mount Meru	2006-09-02	2011-10-28	Died	Polio	
1013	Male	Married	1996-03-25	Mwanza, Geita, Kakora	Kitete	2006-05-28	2008-04-03	Cured	Tetanus	
1023	Female	Divorced	1981-05-15	Mwanza, Magu, Lutale	Mirembe	2005-01-30	2011-01-30	Cured	Hepatitis	
1029	Male	Widowed	1972-11-04	Mwanza, Geita, Bulela	Mpwapwa	2002-01-31	2009-07-16	Complicated	Depression	
1034	Male	Married	1965-09-23	Mwanza, Magu, Lutale	M/mala	2004-10-03	2009-01-11	Complicated	Hepatitis	
1035	Female	Divorced	2000-03-24	Mwanza, Ilemela, Kirumba	Kitete	2003-06-20	2010-06-13	Cured	AIDS	
1036	Male	Widowed	1991-09-12	Mwanza, Geita, Kagu	Kagondo	2002-05-27	2009-09-10	Died	Cancer	
1039	Female	Single	1984-07-22	Mwanza, Ilemela, Igoma	Bugando	2005-02-03	2009-03-27	Complicated	Diabetes	

Figure 5.2: Screenshot of the 10 records PatInfo dataset to be anonymised

5.5.1 Verification of *kl-redInfo*

When the data shown in Figure 5.2 is applied to the software implementation of the *kl-redInfo* algorithm, results to the anonymised table shown at the top of Figure 5.3. The research uses the formulas discussed in Section 5.4 and generalisation hierarchies presented in Figure 5.1. The Discernibility Penalty (DP), Normalised Certainty Penalty (NCP) and Kullback-Leibler divergence (KL-divergence) calculated by the *kl-redInfo* implemented software when *k=l=2* are summarised at the bottom of Figure 5.3 and are calculated next:



Anonymised Table

PID	Gender	MStatus	DOB	Address	Hospital	AdDate	DisDate	DisStatus	Disease	...
1034	Any Gender	Been Married	1961-1970	Mwanza	M/mala	2004-10-03	2009	Complicated	Hepatitis	1
1007	Any Gender	Been Married	1961-1970	Mwanza	Bugando	2005-05-31	2012	Cured	Cancer	1
1029	Any Gender	Been Married	1971-1990	Mwanza	Mpwapwa	2002-01-31	2009	Complicated	Depression	1
1023	Any Gender	Been Married	1971-1990	Mwanza	Mirembe	2005-01-30	2011	Cured	Hepatitis	1
1036	Any Gender	Been Married	1991-2000	Mwanza	Kagondo	2002-05-27	2009	Died	Cancer	1
1035	Any Gender	Been Married	1991-2000	Mwanza	Kitete	2003-06-20	2010	Cured	AIDS	1
1013	Any Gender	Been Married	1991-2000	Mwanza	Kitete	2006-05-28	2008	Cured	Tetanus	1
1008	Any Gender	Single	1971-1990	Mwanza	Chimala	2002-09-15	2011	Complicated	Cholera	1
1039	Any Gender	Single	1971-1990	Mwanza	Bugando	2005-02-03	2009	Complicated	Diabetes	1
1010	Any Gender	Single	1971-1990	Mwanza	Mount Me...	2006-09-02	2011	Died	Polio	1

```
                    }
                    con.close();
                }
```

```
MyAlgoAll2 [Java Application] /usr/lib/jvm/java-7-openjdk-amd64/bin/java (May 4, 2013 8:09:5
Total records:  10
The  DP is: 26.0000
The  NCP is: 7.2625
The  KL-divergence is: 0.8767
```

Figure 5.3: Screenshot of the values of the information loss metric resulting from the application of the *kl-redInfo* algorithm on 10 records PatInfo dataset

- The Discernibility Penalty (DP) is:

$$Dp(T) = \sum_{i=1}^{m} \mid E_i \mid^2$$

$$\begin{aligned} \text{where} \mid E_i \mid &= \text{ Number of records in the equivalence classes} \\ &= 2^2 + 2^2 + 3^2 + 3^2 \\ &= 4 + 4 + 9 + 9 \\ &= 26 \end{aligned} \tag{5.5.1}$$

- The Normalised Certainty Penalty (NCP) using the generalisation hierarchies presented in Figure 5.1.

$$
\begin{aligned}
\text{NCP(T)} &= \sum_{j=1}^{N} NCP(E_j) \text{ where N = Number of equivalence classes} \\
&\text{and } NCP(E_j) = \sum_{i=1}^{n} w_i \frac{size(u_{ij})}{\mid A_i \mid} \\
&\text{where } size(u_{ij}) = \text{The number of leaf nodes that are descendants of } u_{ij} \\
&\text{and } \mid A_i \mid = \text{ The number of distinct values on attribute } A_i \\
&\text{and } w_i = \frac{1}{\mid QIDs \mid} \text{where} \mid QIDs \mid = \text{Number of QIDs} \\
&NCP(E_1) = 2 * \frac{1}{4}(\frac{2}{2} + \frac{3}{4} + \frac{2}{10} + \frac{6}{6}) \\
&= 1.475 \\
&NCP(E_2) = 2 * \frac{1}{4}(\frac{2}{2} + \frac{3}{4} + \frac{5}{10} + \frac{6}{6}) \\
&= 1.625 \\
&NCP(E_3) = 3 * \frac{1}{4}(\frac{2}{2} + \frac{3}{4} + \frac{3}{10} + \frac{6}{6}) \\
&= 2.2875 \\
&NCP(E_4) = 3 * \frac{1}{4}(\frac{2}{2} + \frac{0}{4} + \frac{5}{10} + \frac{6}{6}) \\
&= 1.875 \\
\text{NCP(T)} &= 1.475 + 1.625 + 2.2875 + 1.875 \\
&= 7.2625
\end{aligned}
\tag{5.5.2}
$$

- The Kullback-Leibler divergence (KL-divergence)

$$\text{KL-divergence } = \sum_{i=1}^{n} \mid p_i^{(1)} \log_{10}(\frac{p_i^{(1)}}{p_i^{(2)}}) \mid$$

$$\text{where } p_i = \frac{\text{Number of occurrences of } i}{\text{Total number of occurrences}}$$

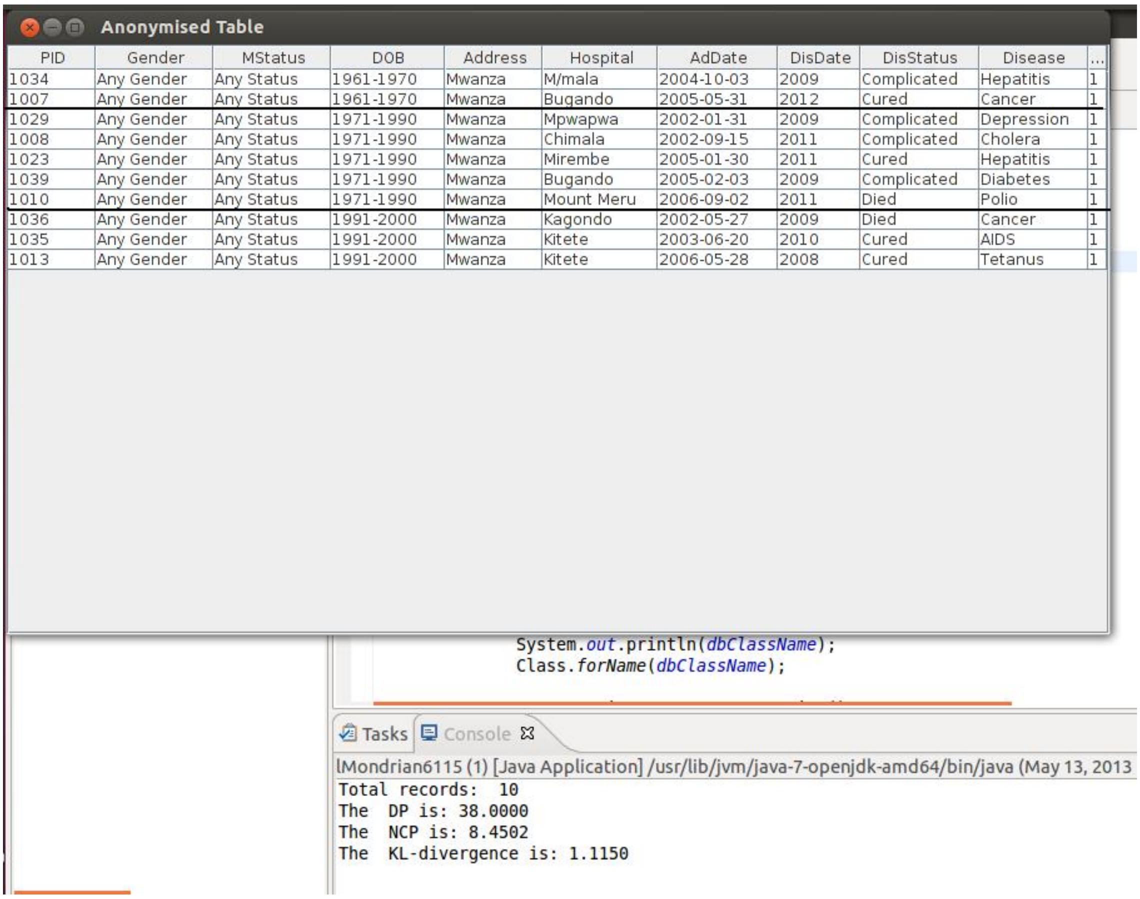
$$\text{For example KL-divergence}_{Female \rightarrow AnyGender} = \mid 0.5 \log_{10}(\frac{0.5}{1}) \mid$$

$$\begin{aligned}\text{KL-divergence}_{Table} &= \mid 0.5 \log_{10}(\frac{0.5}{1}) \mid + \mid 0.5 \log_{10}(\frac{0.5}{1}) \mid + \mid 0.3 \log_{10}(\frac{0.3}{0.7}) \mid + \\ &\mid 0.3 \log_{10}(\frac{0.3}{0.3}) \mid + \mid 0.2 \log_{10}(\frac{0.2}{0.7}) \mid + \mid 0.2 \log_{10}(\frac{0.2}{0.7}) \mid + \\ &\mid 0.1 \log_{10}(\frac{0.1}{0.2}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{0.5}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{0.3}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{1}) \mid \\ &= 0.1505 + 0.1505 + 0.1104 + 0 + 0.1088 + 0.1088 + 0.0301 + 0.0699 + 0.0477 + 0.1 \\ &= 0.8767\end{aligned} \tag{5.5.3}$$

Therefore, the values of the information loss metrics measured by the three information loss metrics, DP, NCP, and KL-divergence, calculated by hand are the same as that calculated by the *kl-redInfo* algorithm as shown at the bottom of Figure 5.3. This shows that the implementation of the *kl-redInfo* algorithm is correct.

5.5.2 Verification of *l-mondrian*

When the data shown in Figure 5.2 is applied to the software implementation of the *l-mondrian* algorithm, results to the anonymised table shown at the top of Figure 5.4. The research uses the formulas discussed in Section 5.4 and generalisation hierarchies presented in Figure 5.1. The Discernibility Penalty (DP), Normalised Certainty Penalty (NCP) and Kullback-Leibler divergence (KL-divergence) calculated by the *l-mondrian* implemented software when *k=l=2* are summarised at the bottom of Figure 5.4 and are calculated next:



PID	Gender	MStatus	DOB	Address	Hospital	AdDate	DisDate	DisStatus	Disease	...
1034	Any Gender	Any Status	1961-1970	Mwanza	M/mala	2004-10-03	2009	Complicated	Hepatitis	1
1007	Any Gender	Any Status	1961-1970	Mwanza	Bugando	2005-05-31	2012	Cured	Cancer	1
1029	Any Gender	Any Status	1971-1990	Mwanza	Mpwapwa	2002-01-31	2009	Complicated	Depression	1
1008	Any Gender	Any Status	1971-1990	Mwanza	Chimala	2002-09-15	2011	Complicated	Cholera	1
1023	Any Gender	Any Status	1971-1990	Mwanza	Mirembe	2005-01-30	2011	Cured	Hepatitis	1
1039	Any Gender	Any Status	1971-1990	Mwanza	Bugando	2005-02-03	2009	Complicated	Diabetes	1
1010	Any Gender	Any Status	1971-1990	Mwanza	Mount Meru	2006-09-02	2011	Died	Polio	1
1036	Any Gender	Any Status	1991-2000	Mwanza	Kagondo	2002-05-27	2009	Died	Cancer	1
1035	Any Gender	Any Status	1991-2000	Mwanza	Kitete	2003-06-20	2010	Cured	AIDS	1
1013	Any Gender	Any Status	1991-2000	Mwanza	Kitete	2006-05-28	2008	Cured	Tetanus	1

```
System.out.println(dbClassName);
Class.forName(dbClassName);
```

```
lMondrian6115 (1) [Java Application] /usr/lib/jvm/java-7-openjdk-amd64/bin/java (May 13, 2013
Total records:  10
The  DP is: 38.0000
The  NCP is: 8.4502
The  KL-divergence is: 1.1150
```

Figure 5.4: Screenshot of the values of the information loss metric resulting from the application of the *l-mondrian* algorithm on 10 records PatInfo dataset

- The Discernibility Penalty (DP) is:

$$Dp(T) = \sum_{i=1}^{m} \mid E_i \mid^2$$

$$\begin{aligned} \text{where} \mid E_i \mid = &\ \text{Number of records in the equivalence classes} \\ &= 2^2 + 5^2 + 3^2 \\ &= 4 + 25 + 9 \\ &= 38 \end{aligned} \tag{5.5.4}$$

- The Normalised Certainty Penalty (NCP) using the generalisation hierarchies presented in Figure 5.1.

$$\begin{aligned}
\text{NCP(T)} &= \sum_{j=1}^{N} NCP(E_j) \text{ where N = Number of equivalence classes} \\
&\text{and } NCP(E_j) = \sum_{i=1}^{n} w_i \frac{size(u_{ij})}{\mid A_i \mid} \\
&\text{where } size(u_{ij}) = \text{The number of leaf nodes that are descendants of } u_{ij} \\
&\text{and } \mid A_i \mid = \text{ The number of distinct values on attribute } A_i \\
&\text{and } w_i = \frac{1}{\mid QIDs \mid} \text{where } \mid QIDs \mid = \text{Number of QIDs} \\
&NCP(E_1) = 2 * \frac{1}{4}(\frac{2}{2} + \frac{4}{4} + \frac{2}{10} + \frac{6}{6}) \\
&= 1.6 \\
&NCP(E_2) = 5 * \frac{1}{4}(\frac{2}{2} + \frac{4}{4} + \frac{5}{10} + \frac{6}{6}) \\
&= 4.375 \\
&NCP(E_3) = 3 * \frac{1}{4}(\frac{2}{2} + \frac{4}{4} + \frac{3}{10} + \frac{6}{6}) \\
&= 2.475 \\
\text{NCP(T)} &= 1.6 + 4.375 + 2.475 \\
&= 8.45
\end{aligned} \tag{5.5.5}$$

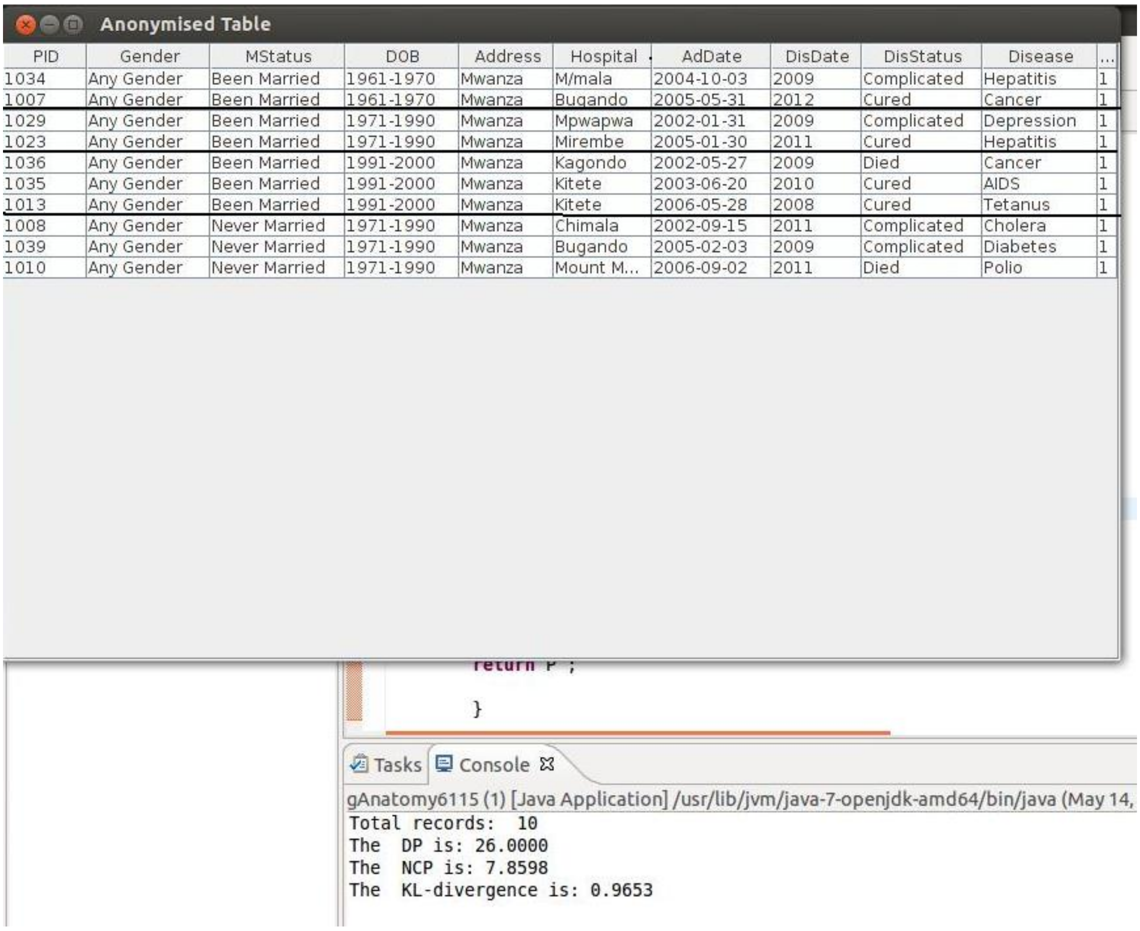
- The Kullback-Leibler divergence (KL-divergence)

$$\begin{aligned}
\text{KL-divergence} &= \sum_{i=1}^{n} \mid p_i^{(1)} \log_{10}(\frac{p_i^{(1)}}{p_i^{(2)}}) \mid \\
&\text{where } p_i = \frac{\text{Number of occurrences of } i}{\text{Total number of occurrences}} \\
\text{KL-divergence}_{Table} &= \mid 0.5 \log_{10}(\frac{0.5}{1}) \mid + \mid 0.5 \log_{10}(\frac{0.5}{1}) \mid + \mid 0.3 \log_{10}(\frac{0.3}{1}) \mid + \\
&\mid 0.3 \log_{10}(\frac{0.3}{1}) \mid + \mid 0.2 \log_{10}(\frac{0.2}{1}) \mid + \mid 0.2 \log_{10}(\frac{0.2}{1}) \mid + \\
&\mid 0.1 \log_{10}(\frac{0.1}{0.2}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{0.5}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{0.3}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{1}) \mid \\
&= 0.1505{+}0.1505{+}0.1569{+}0.1569{+}0.1398{+}0.1398{+}0.0301{+}0.0699{+}0.0206{+}0.1 \\
&= 1.1150
\end{aligned} \tag{5.5.6}$$

Therefore, the values of the information loss metrics measured by the three information loss metrics, DP, NCP, and KL-divergence, calculated by hand are the same as that calculated by the *l-mondrian* algorithm as shown at the bottom of Figure 5.4. This shows that the implementation of *l-mondrian* algorithm is correct.

5.5.3 Verification of *g-anatomy*

When the data shown in Figure 5.2 is applied to the software implementation of the *g-anatomy* algorithm, results to the anonymised table shown at the top of Figure 5.5. The research uses the formulas discussed in Section 5.4 and generalisation hierarchies presented in Figure 5.1. The Discernibility Penalty (DP), Normalised Certainty Penalty (NCP) and Kullback-Leibler divergence (KL-divergence) calculated by the *g-anatomy* implemented software when *k=l=2* are summarised at the bottom of Figure 5.5 and are calculated next:



PID	Gender	MStatus	DOB	Address	Hospital	AdDate	DisDate	DisStatus	Disease	...
1034	Any Gender	Been Married	1961-1970	Mwanza	M/mala	2004-10-03	2009	Complicated	Hepatitis	1
1007	Any Gender	Been Married	1961-1970	Mwanza	Bugando	2005-05-31	2012	Cured	Cancer	1
1029	Any Gender	Been Married	1971-1990	Mwanza	Mpwapwa	2002-01-31	2009	Complicated	Depression	1
1023	Any Gender	Been Married	1971-1990	Mwanza	Mirembe	2005-01-30	2011	Cured	Hepatitis	1
1036	Any Gender	Been Married	1991-2000	Mwanza	Kagondo	2002-05-27	2009	Died	Cancer	1
1035	Any Gender	Been Married	1991-2000	Mwanza	Kitete	2003-06-20	2010	Cured	AIDS	1
1013	Any Gender	Been Married	1991-2000	Mwanza	Kitete	2006-05-28	2008	Cured	Tetanus	1
1008	Any Gender	Never Married	1971-1990	Mwanza	Chimala	2002-09-15	2011	Complicated	Cholera	1
1039	Any Gender	Never Married	1971-1990	Mwanza	Bugando	2005-02-03	2009	Complicated	Diabetes	1
1010	Any Gender	Never Married	1971-1990	Mwanza	Mount M...	2006-09-02	2011	Died	Polio	1

```
        return P ;

        }
```

```
gAnatomy6115 (1) [Java Application] /usr/lib/jvm/java-7-openjdk-amd64/bin/java (May 14,
Total records:  10
The  DP is: 26.0000
The  NCP is: 7.8598
The  KL-divergence is: 0.9653
```

Figure 5.5: Screenshot of the values of the information loss metric resulting from the application of the *g-anatomy* algorithm on 10 records PatInfo dataset

- The Discernibility Penalty (DP) is:

$$Dp(T) = \sum_{i=1}^{m} \mid E_i \mid^2$$

$$\text{where } \mid E_i \mid = \text{ Number of records in the equivalence classes}$$

$$= 2^2 + 2^2 + 3^2 + 3^2$$

$$= 4 + 4 + 9 + 9$$

$$= 26$$

(5.5.7)

- The Normalised Certainty Penalty (NCP) using the generalisation hierarchies presented in Figure 5.1.

$$\text{NCP(T)} \;= \sum_{j=1}^{N} NCP(E_j) \text{ where N = Number of equivalence classes}$$

$$\text{and } NCP(E_j) = \sum_{i=1}^{n} w_i \frac{size(u_{ij})}{\mid A_i \mid}$$

$$\text{where } size(u_{ij}) = \text{The number of leaf nodes that are descendants of } u_{ij}$$

$$\text{and } \mid A_i \mid = \text{ The number of distinct values on attribute } A_i$$

$$\text{and } w_i = \frac{1}{\mid QIDs \mid} \text{where } \mid QIDs \mid = \text{Number of QIDs}$$

$$\begin{aligned}
NCP(E_1) &= 2 * \frac{1}{4}(\frac{2}{2} + \frac{3}{4} + \frac{2}{10} + \frac{6}{6}) \\
&= 1.475 \\
NCP(E_2) &= 2 * \frac{1}{4}(\frac{2}{2} + \frac{3}{4} + \frac{5}{10} + \frac{6}{6}) \\
&= 1.625 \\
NCP(E_3) &= 3 * \frac{1}{4}(\frac{2}{2} + \frac{3}{4} + \frac{3}{10} + \frac{6}{6}) \\
&= 2.2875 \\
NCP(E_4) &= 3 * \frac{1}{4}(\frac{2}{2} + \frac{1}{4} + \frac{5}{10} + \frac{6}{6}) \\
&= 2.4723 \\
\text{NCP(T)} \;&= 1.475 + 1.625 + 2.2875 + 2.4723 \\
&= 7.8598
\end{aligned} \tag{5.5.8}$$

- The Kullback-Leibler divergence (KL-divergence)

$$\begin{aligned}
\text{KL-divergence } &= \sum_{i=1}^{n} \mid p_i^{(1)} \log_{10}(\frac{p_i^{(1)}}{p_i^{(2)}}) \mid \\
&\text{where } p_i = \frac{\text{Number of occurrences of } i}{\text{Total number of occurrences}} \\
\text{KL-divergence}_{\mathit{Table}} &= \mid 0.5 \log_{10}(\frac{0.5}{1}) \mid + \mid 0.5 \log_{10}(\frac{0.5}{1}) \mid + \mid 0.3 \log_{10}(\frac{0.3}{0.7}) \mid + \\
&\mid 0.3 \log_{10}(\frac{0.3}{0.3}) \mid + \mid 0.2 \log_{10}(\frac{0.2}{0.7}) \mid + \mid 0.2 \log_{10}(\frac{0.2}{0.7}) \mid + \\
&\mid 0.1 \log_{10}(\frac{0.1}{0.2}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{0.5}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{0.3}) \mid + \mid 0.1 \log_{10}(\frac{0.1}{1}) \mid \\
&= 0.1505{+}0.1505{+}0.1104{+}0.0886{+}0.1088{+}0.1088{+}0.0301{+}0.0699{+}0.0477{+}0.1 \\
&= 0.9653
\end{aligned} \tag{5.5.9}$$

Therefore, the values of the information loss metrics measured by the three information loss metrics, DP, NCP, and KL-divergence, calculated by hand are the same as that calculated by the *g-anatomy* algorithm as shown at the bottom of Figure 5.5. This shows that the implementation of *g-anatomy* algorithm is correct.

5.6 Wilcoxon signed-rank test

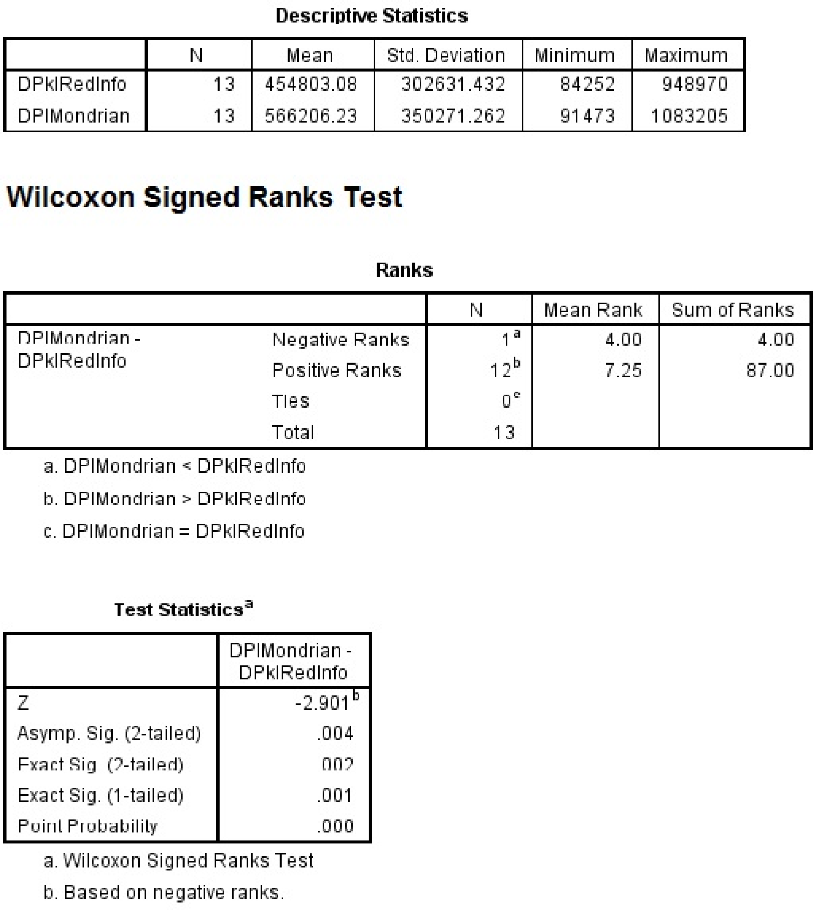
After studying the computed values of the information loss metrics from the implemented algorithms, next step was to quantify the impact of each of the proposed modified approaches and the proposed algorithm. This research uses Wilcoxon signed-rank test statistics to analyse the significance of each of the proposed approaches and algorithm. The Wilcoxon signed-rank test is one of the most commonly used non-parametric statistical test that determines if two datasets differ significantly (Wilcoxon, 1945). The advantage of non-parametric tests over parametric tests is that the parameters are determined from the data and are flexible, not fixed in advance as done by parametric tests (Siegel, 1957). Therefore, non-parametric tests are also called distribution free tests.

The Wilcoxon signed-rank test is the appropriate test statistics due to the reason that, the computed values of the information loss metrics are not normally distributed and the sample sizes are small (13 observation from each algorithm, as a result of 13 values of the information loss metrics resulting from the application of 13 values of k and l ($2 \geq k \leq 14$) where k=l). Therefore, the use of the distribution free test statistic such as Wilcoxon signed-rank test is more appropriate than a parametric test such as t-test statistic (Siegel, 1957).

The logic behind the Wilcoxon test is based solely on the order in which the observations from the two samples fall. The data are ranked to produce two rank totals, one for each condition. If there is a systematic difference between the two conditions, then most of the high ranks will belong to one condition and most of the low ranks will belong to the other one. As a result, the rank totals will be quite different and one of the rank totals will be quite small. On the other hand, if the two conditions are similar, then high and low ranks will be distributed fairly evenly between the two conditions and the rank totals will be fairly similar and quite large.

The comparison results using the Wilcoxon signed-rank test results from the Statistical Package for the Social Sciences (SPSS) are presented in the form of the tables as shown in Table 5.4. The output is divided into three tables. Initially, descriptive data like the number of participants in each group, group averages, their standard deviation, and the minimal and maximal values, appear. Thereafter, the test results appear in two distinct tables. In the first table, between the values of the Ranks, the Mean Rank and the Sum of Ranks given, the N corresponds to the number of observations or participants.

Table 5.4: The example of the Wilcoxon signed-rank test results tables



Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
DPklRedInfo	13	454803.08	302631.432	84252	948970
DPlMondrian	13	566206.23	350271.262	91473	1083205

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
DPlMondrian - DPklRedInfo	Negative Ranks	1[a]	4.00	4.00
	Positive Ranks	12[b]	7.25	87.00
	Ties	0[c]		
	Total	13		

a. DPlMondrian < DPklRedInfo

b. DPlMondrian > DPklRedInfo

c. DPlMondrian = DPklRedInfo

Test Statistics[a]

	DPlMondrian - DPklRedInfo
Z	-2.901[b]
Asymp. Sig. (2-tailed)	.004
Exact Sig. (2-tailed)	.002
Exact Sig. (1-tailed)	.001
Point Probability	.000

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

In addition, in the second table, the tests results appear. The SPSS returns the Z value and the asymptotic significance or the level of significance based on the normal distribution of the statistical test: *Asymp. Sig. (2-tailed).* To evaluate the significance of the differences, the risk level probability (called the alpha level) has to be set (Trochim, 2006). As internationally accepted, the risk level of 0.05

was chosen as the threshold to determine if the difference between two datasets is statistically significant or otherwise. Therefore, the difference is considered to be significant if the *p-value* (labeled as Sig.(2-tailed)) is lower than 0.05. Thus, the results in Table 5.4 show significance difference between the DP of *l-mondrian* and DP of *kl-redInfo* algorithm. This is due to the fact that the *p-value*, $0.004 \leq 0.05$.

The asymptotic significance is based on the assumption that the data sample is large ($N \geq 40$). If the data sample is small or not normally distributed, such as the data used in this research, the asymptotic significance is not in general a good indication of the significance. In this case, the level of significance based on the exact distribution of a statistical test labeled as *Exact Sig. (2-tailed)* corresponds to the statistic of decision. Consequently, studies should use this value when the sample is small, sparse, contains many ties, is badly balanced or does not seem to be normally distributed. Note that *Exact Sig. (2-tailed):* represents level of significance for a two-tailed test, and *Exact Sig. (1-tailed):* represents level of significance for a one-tailed test.

5.7 Chapter Summary

This chapter presents the experimental environment and setup including datasets, algorithms used for comparison and information loss metrics. The research uses simulated dataset named PatInfo and real-world Adult dataset downloaded from UCI repository. The two widely used algorithms, *l-mondrian* and *g-anatomy*, are used for comparison purpose. The two algorithms are appropriate as each of the algorithm is well established and proved to result in a lower value of the information loss metric from the two ways of achieving *l-diversity* privacy requirement. The *l-mondrian* is the *l-diversity* algorithm extended from Mondrian *k-anonymisation* algorithm while *g-anatomy* is a modified version of the Anatomy algorithm, which is the *l-diversity* specific algorithm.

Ensuring an individual's privacy is one aspect of the anonymisation approach, the other aspect is the retention of information so that the shared data remains useful. There are several metrics that can be used to measure the information loss as discussed in Section 5.4. This research uses three well established metrics, Discernibility Penalty (DP), Normalised Certainty Penalty (NCP), and Kullback-Leibler divergence (KL-divergence). These three metrics have been used because each of them has different characteristics that are useful to the research. In order to investigate if the impact of the proposed modified approaches was statistically significant, the Wilcoxon signed-rank non-parametric test is used, as discussed in Section 5.6. This test was used as the sample datasets are small (13 values of the information loss metrics as a result of the values of ($2 \geq k \leq 14$) where $k=l$) and the values are not normally distributed.

Chapter 6

EVALUATION of the *kl-redInfo*

This research performs a set of experiments in order to evaluate the proposed *kl-redInfo* algorithm. The algorithm is evaluated based on it's ability to preserve the quality of the data, by reducing the amount of information loss, while achieving both *k-anonymity* and *l-diversity* privacy requirements. The information loss is indicated by the three information loss metrics, Discernibility Penalty (DP), Normalised Certainty Penalty (NCP), and Kullback-Leibler divergence (KL-divergence), discussed in Section 5.4. The metrics are unitless, therefore there is no general accepted absolute benchmarks for their interpretation (El Emam et al., 2009).

The research first quantifies the impact of each of the proposed modified approaches in reducing the values of the information loss metrics that indicate the reduction in information loss. The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm is compared with the modified versions of the widely used *Mondrian* and *Anatomy* algorithms, that achieved both *k-anonymity* and *l-diversity* privacy requirements, as discussed in Section 5.3.1 and 5.3.2.

The values of the information loss metrics resulting from the application of the implemented algorithms are presented as *histograms*. The *y-axis* of the histograms represents the information loss measured in terms of the three metrics, DP, NCP, and KL-divergence. The Discernibility Penalty (DP) indicates the information loss based on how many other records are indistinguishable from each other with respect to quasi-identifiers. The NCP indicates the values of the information loss resulting from the application of the generalisation approach. The KL-divergence indicates the probability deviation of the anonymised dataset from the original dataset. The smaller the value of the information loss metric, the better it approximates the original dataset, therefore the lower the information loss.

The *x-axis* of the histograms represent the values of parameter k and l, which control the level of the privacy provided by the algorithm. The higher the values of k and l indicates the higher level of the privacy. The value of k controls the identity disclosure while the value of l controls the attribute disclosure. In order to ensure equal protection of both identity disclosure and attribute disclosure, this research uses the values of $k = l$.

When the values of $l<k$ was used, the values of the information loss metrics was the same as when values of $k = l$ was used. This shows that values of l have no effect on the values of the information loss metrics. The effect of l values are on the individual's privacy, since the lower the value of l the lower distinct sensitive values are in the equivalence class thus more possibility of the attribute disclosure.

6.1 The Information loss Metrics of the Algorithms without the Proposed Modifications

In order to investigate the impact of the proposed modified approaches, the values of the information loss metrics resulting from the application of the proposed *kl-redInfo* algorithm with the existing approaches, and the *l-mondrian* and *g-anatomy* algorithms, was calculated by using the PatInfo and Adult datasets. The simulated PatInfo dataset has 30,200 records and eight attributes with four quasi-identifiers and disease as a sensitive attribute. The real-world Adult dataset was downloaded from UCI repository and it has 30,162 records and eight attributes with seven quasi-identifiers and occupation as a sensitive attribute, as disscussed in Section 5.2.

The research identifies three existing approaches that increase the values of the information loss metrics and proposes modified approaches that can be used to reduce the values of the information loss metrics. These causes are;

- sequential incorporation of the remaining records after forming the groups of records with distinct sensitive attributes.
- the use of either bucketisation or cell-based generalisation approaches.
- not taking under consideration the distribution of quasi-identifier attributes.

The values of the information loss metrics resulting from the application of the proposed *kl-redInfo* algorithm without the proposed modified approaches, the *l-mondrian*, and *g-anatomy* algorithms are presented in Figure 6.1, 6.2 and 6.3. The information loss is based on the three information loss metrics; Discernibility Penalty (DP), Normalised Certainty Penalty (NCP), and Kullback-Leibler divergence (KL-divergence), further disscussed in section 5.4.

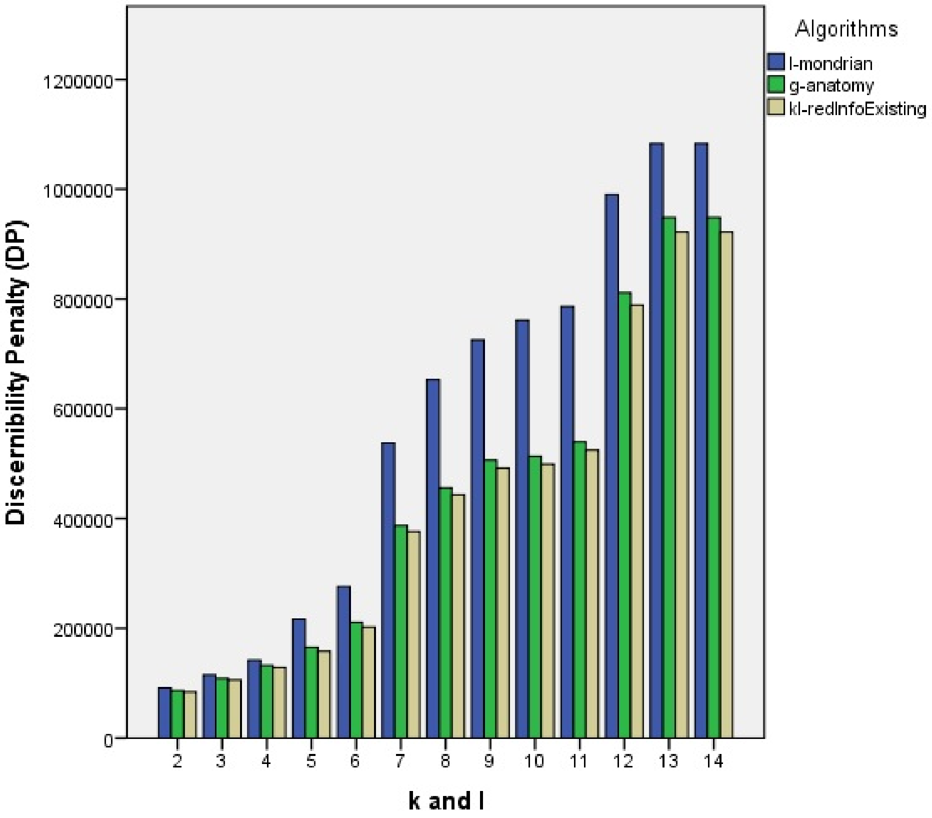


Figure 6.1: DP resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* algorithm without the proposed modifications on the PatInfo dataset

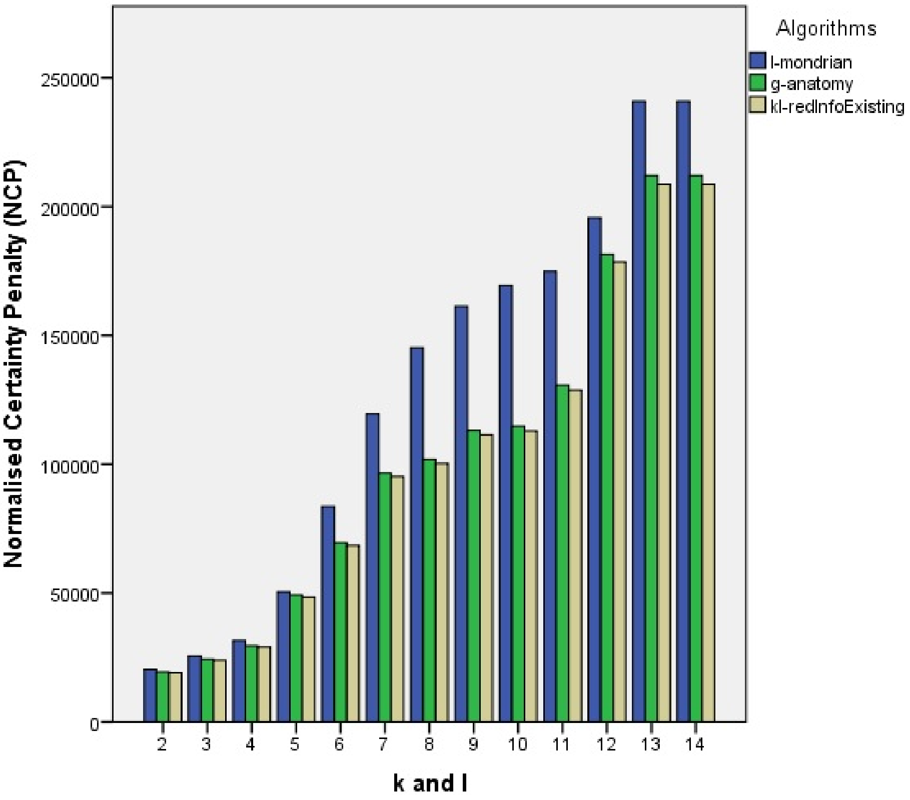


Figure 6.2: NCP resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* algorithm without the proposed modifications on the PatInfo dataset

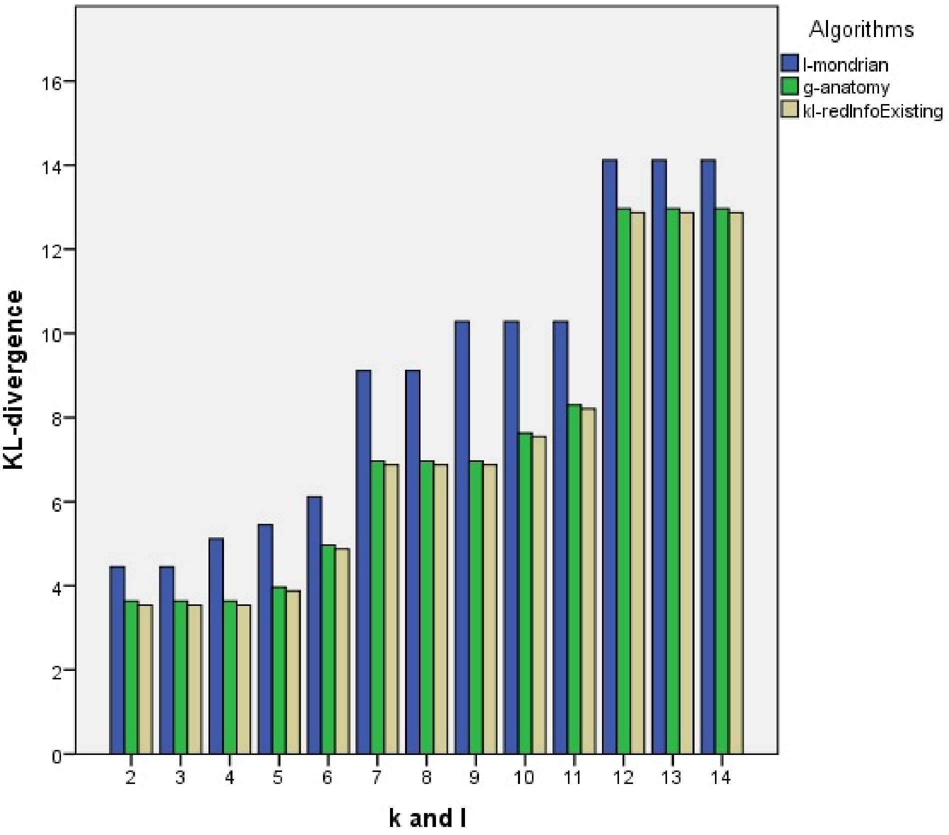


Figure 6.3: KL-divergence resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* algorithm without the proposed modifications on the PatInfo dataset

As it can be seen in Figure 6.1, 6.2 and 6.3, the *l-mondrian* algorithm results in higher values of the information loss metrics compared to *g-anatomy* and *kl-redInfo* algorithm without the proposed modified approaches in all three information loss metrics. This is due to the fact that the *l-mondrian* algorithm incorporates an equivalence class, instead of a record as *g-anatomy* and *kl-redInfo* do.

The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modified approaches is relatively lower compared to the *g-anatomy.* This is due to the order in which the approaches are applied. The *kl-*

redInfo does generalisation approach before incorporation approach and more generalisation is applied after incorporation approach only if the incorporated record needs to be generalised to make it indistinguishable from other records. The *g-anatomy* algorithm does the generalisation approach after incorporation approach. This causes the *g-anatomy* algorithm to has large number of records that need to be indistinguishable from each other, thus increases the value of the information loss metrics which indicates the increase of information loss. Therefore, the order in which the approaches are applied contributes in the reduction of information loss.

As discussed in Section 5.6, identifying significance of the differences between the algorithms cannot be achieved without examining the statistical significance of the results. The Wilcoxon signed-rank non-parametric test statistic is an appropriate statistical test as the collected data are of small size (13 observations in each algorithm, as a result of the values of the information loss metrics resulting from the application of 13 values of k and l ($2 \leq k \leq 14$) where $k=l$) and the data are not normally distributed. This also shows that the level of significance should be based on the exact distribution of a statistical test (labeled as *Exact Sig.(2-tailed)* on the table of Wilcoxon signed-rank test) rather than asmptotic significance which assumes the data samples are large and are normally distributed.

Even though the proposed *kl-redInfo* algorithm without the proposed modified approaches results in reduced values of the information loss metrics compared to *l-mondrian* and *g-anatomy* algorithms, the difference is significant when the *kl-redInfo* is compared to *l-mondrian*, but the difference is not significant when *kl-redInfo* is compared to *g-anatomy* algorithm, as shown by the Wilcoxon signed-rank test statistics. This is due to the feature of *l-mondrian* algorithm to incorporate equivalence classes rather than individual records as *g-anatomy* and *kl-redInfo* algorithms do.

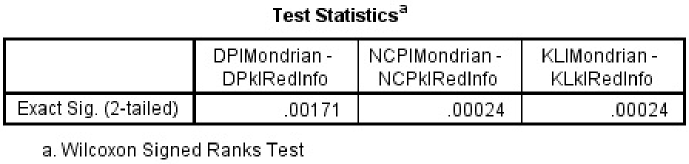
The difference between the algorithms can also be evaluated by using the difference percentages of their values of the information loss metrics. The higher the percentage difference the higher the difference between the compared algorithms, hence significance difference. Table 6.1 summarises the values of the information loss metrics, the difference (presented as **Diff**), and the percentage of the difference of the values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed approaches with respect to the values of the information loss metrics resulting from the application of the existing *l-mondrian* algorithm (presented in the brackets). The comparison of the results by using the Wilcoxon signed-rank test statistics are presented in Table 6.2.

As it can be seen in Table 6.1, the values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modified approaches are reduced by an average of 21% of DP, 18% of NCP and 25% of KL-divergence when compared to the values of the information loss metrics resulting from the application of the *l-mondrian* algorithm. This indicates that there is a reduction in information loss resulting from the application of the *kl-redInfo* without the proposed modified approaches compared to the *l-mondrian* algorithm, as indicated by the three information loss metrics.

Table 6.1: The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and *l-mondrian* on the PatInfo dataset

k,l	Information loss Metrics								
	DP			**NCP**			**KL-divergence**		
	l-mondrian	*kl-without*	**Diff (%)**	*l-mondrian*	*kl-without*	**Diff (%)**	*l-mondrian*	*kl-without*	**Diff (%)**
2	91473	84252	7221(8)	20341	19070	1271(6)	4.4483	3.4643	0.984(22)
3	114821	105756	9065(8)	25533	23937	1596(6)	4.4483	3.4643	0.984(22)
4	141843	128527	13316(9)	31542	29091	2451(8)	5.115	3.4643	1.6507(32)
5	216855	157932	58923(27)	50461	48382	2079(4)	5.4483	3.5893	1.859(34)
6	275847	202354	73493(27)	83579	68436	15143(18)	6.115	4.7143	1.4007(23)
7	537525	376379	161146(30)	119532	85191	34341(29)	9.115	6.5893	2.5257(28)
8	653152	443040	210112(32)	145244	100280	44964(31)	9.115	6.5893	2.5257(28)
9	725257	492152	233105(32)	161279	111396	49883(31)	10.2817	6.5893	3.6924(36)
10	761364	499002	262362(34)	169307	112946	56361(33)	10.2817	7.2143	3.0674(30)
11	786277	524681	261596(33)	174848	118759	56089(32)	10.2817	7.8393	2.4424(24)
12	789857	788735	1122(0)	195644	178526	17118(9)	14.115	12.2143	1.9007(13)
13	1083205	921970	161235(15)	240877	208683	32194(13)	14.115	12.2143	1.9007(13)
14	1083205	921970	161235(15)	240877	208683	32194(13)	14.115	12.2143	1.9007(13)
		Average	21%			18%			25%

Table 6.2: The comparison results for the values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and *l-mondrian* on the PatInfo dataset



Test Statistics[a]

	DPlMondrian - DPklRedInfo	NCPlMondrian - NCPklRedInfo	KLlMondrian - KLklRedInfo
Exact Sig. (2-tailed)	.00171	.00024	.00024

a. Wilcoxon Signed Ranks Test

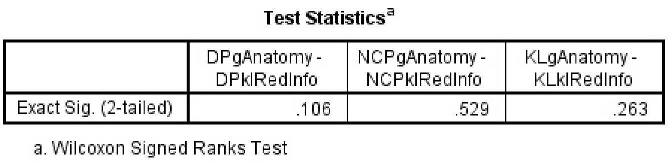
The comparison results shown in Table 6.2 shows that, there is a significant difference when comparing *kl-redInfo* algorithm without the proposed modified approaches by the *l-mondrian* algorithm. This is due to the *p-values*, 0.00171 for DP, 0.00024 for NCP, and 0.00024 for KL-divergence, being lower than the acceptance risk level of 0.05.

Table 6.3: The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and the *g-anatomy* on the PatInfo dataset

k,l	Information loss Metrics								
	DP			**NCP**			**KL**		
	g-anatomy	*kl-without*	**Diff (%)**	*g-anatomy*	*kl-without*	**Diff (%)**	*g-anatomy*	*kl-without*	**Diff (%)**
2	86659	84252	2407(3)	19371	19070	301(2)	3.63	3.4643	0.1677(5)
3	108778	105756	3022(3)	24323	23937	386(2)	3.632	3.4643	0.1677(5)
4	132199	128527	3672(3)	29560	29091	469(2)	3.632	3.4643	0.1677(5)
5	165302	157932	7370(4)	49321	48382	939(2)	3.9653	3.5893	0.376(9)
6	210993	202354	8639(4)	69538	68436	1102(2)	4.9653	4.7143	0.251(5)
7	387133	376379	10754(3)	86563	85191	1372(2)	6.9653	6.5893	0.376(5)
8	455699	443040	12659(3)	101894	100280	1614(2)	6.9653	6.5893	0.376(5)
9	506214	492152	14062(3)	113189	111396	1793(2)	6.9653	6.5893	0.376(5)
10	513259	499002	14257(3)	114765	112946	1819(2)	7.632	7.2143	0.4177(5)
11	539672	524681	14991(3)	120671	118759	1912(2)	8.2986	7.8393	0.4593(6)
12	811271	788735	22536(3)	181400	178526	2874(2)	12.9653	12.2143	0.751(6)
13	948312	921970	26342(3)	212042	208683	3359(2)	12.9653	12.2143	0.751(6)
14	948312	921970	26342(3)	212042	208683	3359(2)	12.9653	12.2143	0.751(6)
		Average	3%			2%			6%

The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm without the proposed modified approaches are reduced by an average of 3% of DP, 2% of NCP and 6% of KL-divergence when compared to the values of the information loss metrics resulting from the application of the *g-anatomy*, as shown in Table 6.3. Table 6.4 shows that, even though the information loss metrics, shown in Figure 6.1, 6.2 and 6.3, show reduction in the values of the information loss metris resulting from the application of the *kl-redInfo* algorithm without the proposed modified approaches when compared to *g-anatomy* algorithm, the comparison results show that the difference is not significant. This is due to the *p-values*, 0.106 for DP, 0.529 for NCP, and 0.263 for KL-divergence, being greater than the acceptance risk level of 0. 05.

Table 6.4: The comparison results for the values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modified approaches and *g-anatomy* on the PatInfo dataset



Test Statistics[a]

	DPgAnatomy - DPklRedInfo	NCPgAnatomy - NCPklRedInfo	KLgAnatomy - KLklRedInfo
Exact Sig. (2-tailed)	.106	.529	.263

a. Wilcoxon Signed Ranks Test

To study the impact of the implemented algorithms on different and independent datasets, the three implemented algorithms, the proposed *kl-redInfo* algorithm without the proposed modified approaches, *l-mondrian*, and *g-anatomy*, were applied to the real-world Adult dataset downloaded from UCI repository (Asuncion and Newman, 2007). The values of the three information loss metrics; DP, NCP, and KL-divergence, were calculated and the results are summarised in Appendix C.

The results show that, even in this real-world Adult dataset, the proposed *kl-redInfo* algorithm without the proposed modified approaches, results in reduced values of the information loss metrics compared to the existing, *l-mondrian* and *g-anatomy* algorithms. The difference is significant when *kl-redInfo* is compared to *l-mondrian* algorithm, but the difference is not significant when *kl-redInfo* is compared to *g-anatomy* algorithm.

Section 6.2 discusses the causes of the information loss and the proposed modified approaches that can be used to reduce the values of the information loss metrics that indicate reduction in information loss. The proposed modified approaches were then used to design the proposed *kl-redInfo* algorithm.

6.2 Proposed Modifications

This research proposes three approaches that can be used to reduce the values of the information loss metrics that indicate reduction in information loss. These approaches are:

- Systematic incorporation of the remaining records in the equivalence class that results in lower value of the information loss metric.
- Using both bucketisation and cell-based generalisation approaches.
- Sorting the records according to the quasi-identifier attributes in order to take under consideration their distribution.

These approaches and their significance in reducing the values of the information loss metrics are discussed in detail in Section 6.2.1, 6.2.2 and 6.2.3 respectively.

6.2.1 Systematic Incorporation of the Remaining records in the equivalence classes

Achieving the *l-diversity* privacy requirement involves incorporation of the remaining records that do not achieve the privacy requirement to the group of records that satisfy the privacy requirements. This research studies the effects resulting from the way the remaining records are incorporated on the created equivalence classes. Sequential incorporation of the remaining records, as done by the existing algorithms, may result in increase of the values of the information loss metrics. This is because a record may be incorporated in an equivalence class that has very different QIDs, thus more generalisation is needed to make them indistinguishable from each other, which results in high values of the information loss metrics.

This research uses the approach of incorporating the remaining records in the equivalence classes that results in lower value of information loss metric compared to other equivalence classes. This is done by calculating the value of information loss metric before the record is incorporated in an equivalence class. This research named the approach as *systematic incorporation approach*.

Figure 6.4, 6.5 and 6.6 show the values of the information loss metrics resulting from the application of the *kl-redInfo* when remaining records are sequential incorporated, represented as *kl-redInfosequencial*, and when remaining records are systematically incorporated, represented as *kl-redInfosystematic*, on simulated PatInfo dataset. The values of the information loss metrics resulting from the application of the algorithms on the real-world Adult dataset are presented in Appendix D.

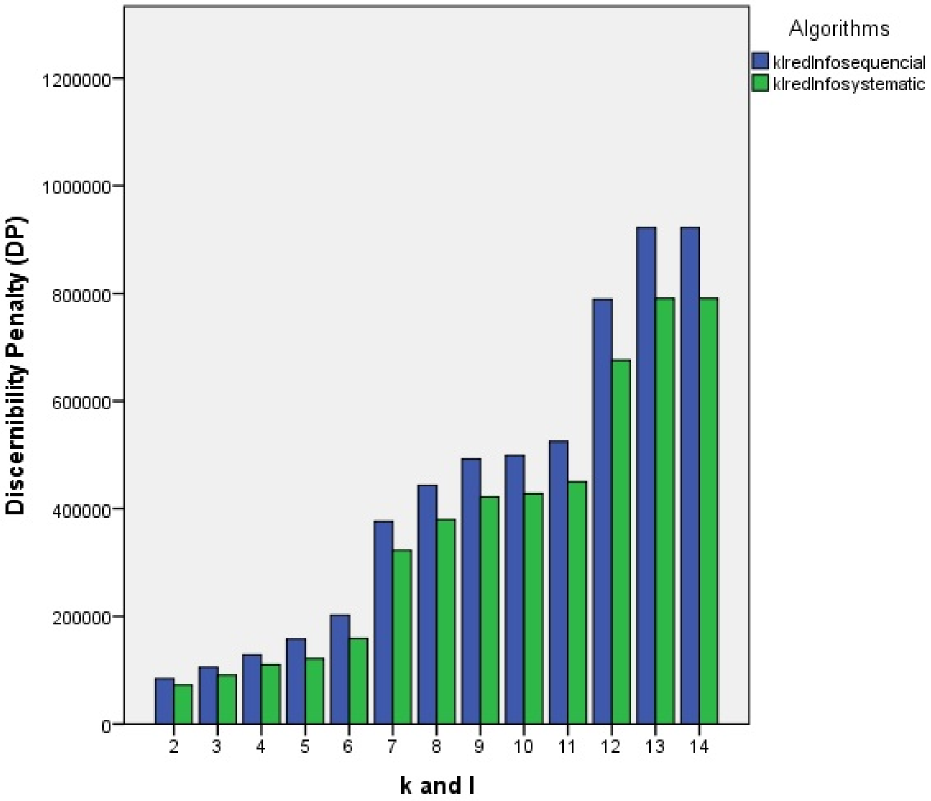


Figure 6.4: DP resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the PatInfo dataset

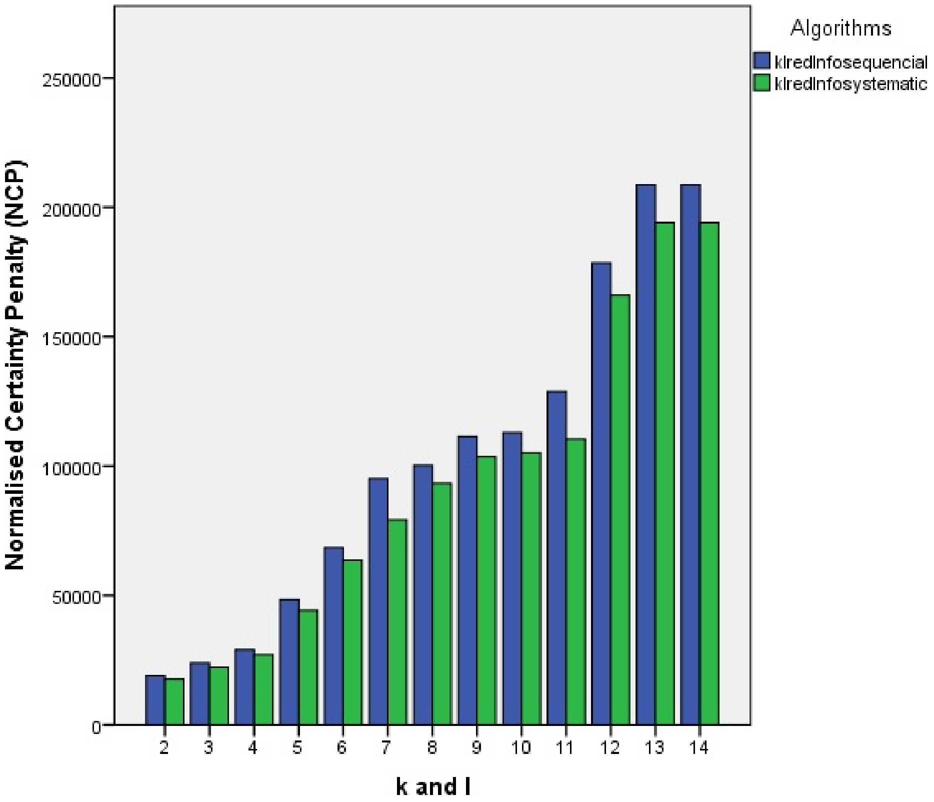


Figure 6.5: NCP resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the PatInfo dataset

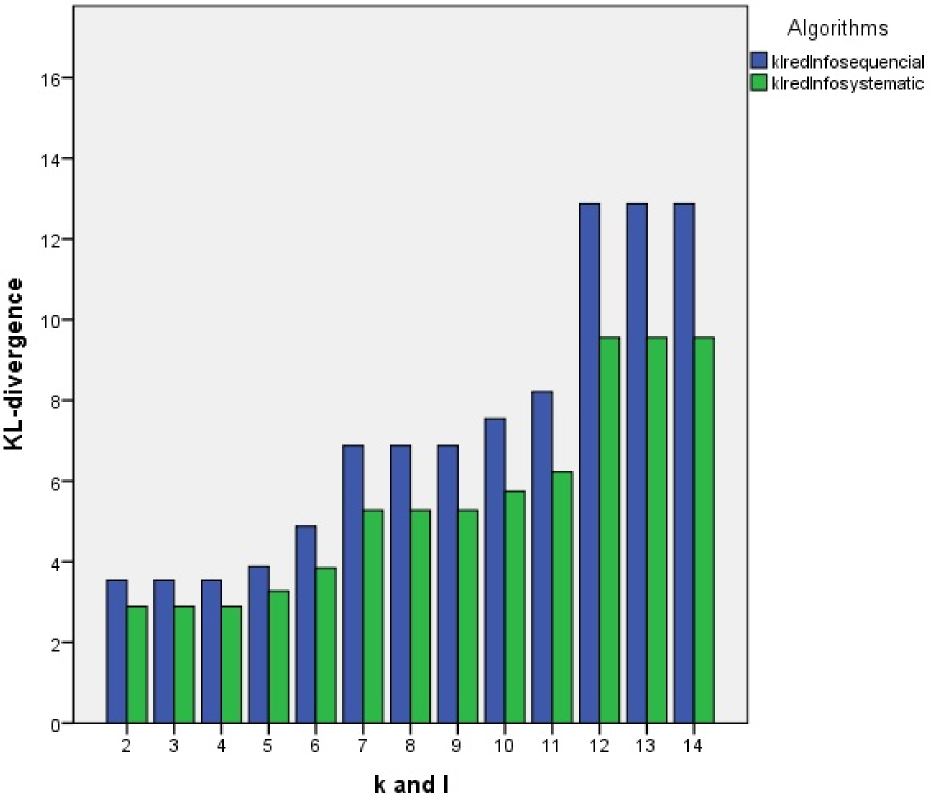
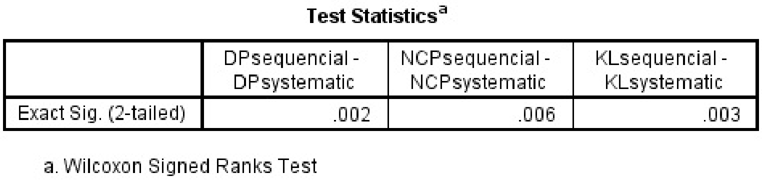


Figure 6.6: KL-divergence resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the PatInfo dataset

As it can be seen from Figure 6.4, 6.5 and 6.6, the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm when records are systematically incorporated is lower than the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm when records are sequentially incorporated in all three information loss metrics, DP, NCP, and KL-divergence. Therefore, this indicates that the approach of systematic incorporation of the remaining records reduces the values of the information loss metrics that indicate reduction in the information loss. Similar results were drawn when Adult dataset was used, refer to Appendix D for the Adult dataset results.

Table 6.5 shows the comparison results of the *kl-redInfo* when records are sequentially incorporated and when records are systematically incorporated, measured by the three information loss metrics, DP, NCP, and KL-divergence. Observing the comparison test results table, the *p-values* are 0.002 for DP, 0.006 for NCP and 0.003 for KL-divergence.

Table 6.5: The comparison results for the values of the information loss metrics resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the PatInfo dataset



Test Statistics[a]

	DPsequencial - DPsystematic	NCPsequencial - NCPsystematic	KLsequencial - KLsystematic
Exact Sig. (2-tailed)	.002	.006	.003

a. Wilcoxon Signed Ranks Test

Since the *p-values* are lower than the acceptable risk level of 0.05, this indicates that there is a significant difference on the values of the information loss metrics resulting from the application of the *kl-redInfo* when remaining records are sequentially incorporated and when they are systematically incorporated. Therefore, this research concludes that there is a significant evidence on the reduction in the values of information loss metrics when the remaining records are systematically incorporated compared to when they are sequentially incorporated at the risk level of 0.05.

The values of the information loss metrics when Adult dataset was used also shows that there is a significant evidence on the reduction in the values of information loss when remaining records are systematically incorporated compared to when they are sequentially incorporated at the risk level of 0.05. This is due to the fact that the *p-values* are lower than the acceptable risk level of 0.05. The *p-values* are 0.033 for

DP, 0.033 for NCP, and 0.019 for KL-divergence, as shown in Appendix D Table D.2.

6.2.2 Using both bucketisation and cell-based generalisation approaches

Most of the solutions proposed in the literature use bucketisation or generalisation approach to achieve privacy requirements. For example, Mondrian algorithm uses generalisation approach but not bucketisation approach, which results in substantial information loss, and Anatomy algorithm uses bucketisation but not generalisation approach, which result in violation of an individual's privacy.

Therefore, this research proposes the use of both bucketisation and cell-based generalisation approaches. These approaches reduce the values of the information loss metrics while ensuring an individual's privacy. The values of the information loss metrics are reduced by the use of generalisation approach depending on the need of the group and not all values of the attribute, also known as cell-generalisation.

On the other hand, these approaches produce results that may be non homogeneous as generalisation is applied depending on the need of the group of records and not all records. Therefore, records of each group might be anonymised in different levels of the generalisation hierarchy. This makes the analysis process difficult (e.g., some records report the complete date of birth, while other records only report the year of birth). But this approach reduces the values of the information loss metrics when compared to bucketisation and generalisation approaches when are used separate, which is the aim of this research. Thus, this research uses both bucketisation and cell-based generalisation approaches in order to reduce the values of the information loss metric that indicate reduction in information loss and still ensures an individual's privacy.

Figure 6.7, 6.8 and 6.9 show the values of the information loss metrics resulting from the application of the *kl-redInfo* when only generalisation approach is used, named *kl-redInfoCell*, and the *kl-redInfo* when bucketisation and cell-based generalisation approaches are used, named *kl-redInfoBucketCell*. The information loss is measured in terms of DP, NCP, and KL-divergence metrics.

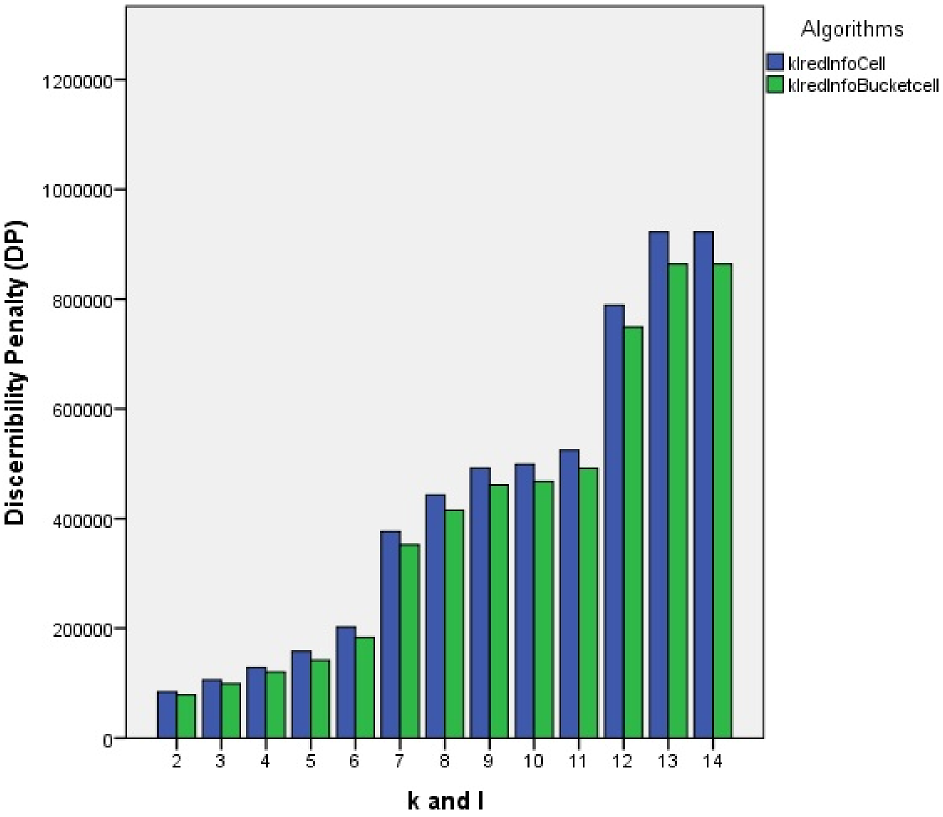


Figure 6.7: DP resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the PatInfo dataset

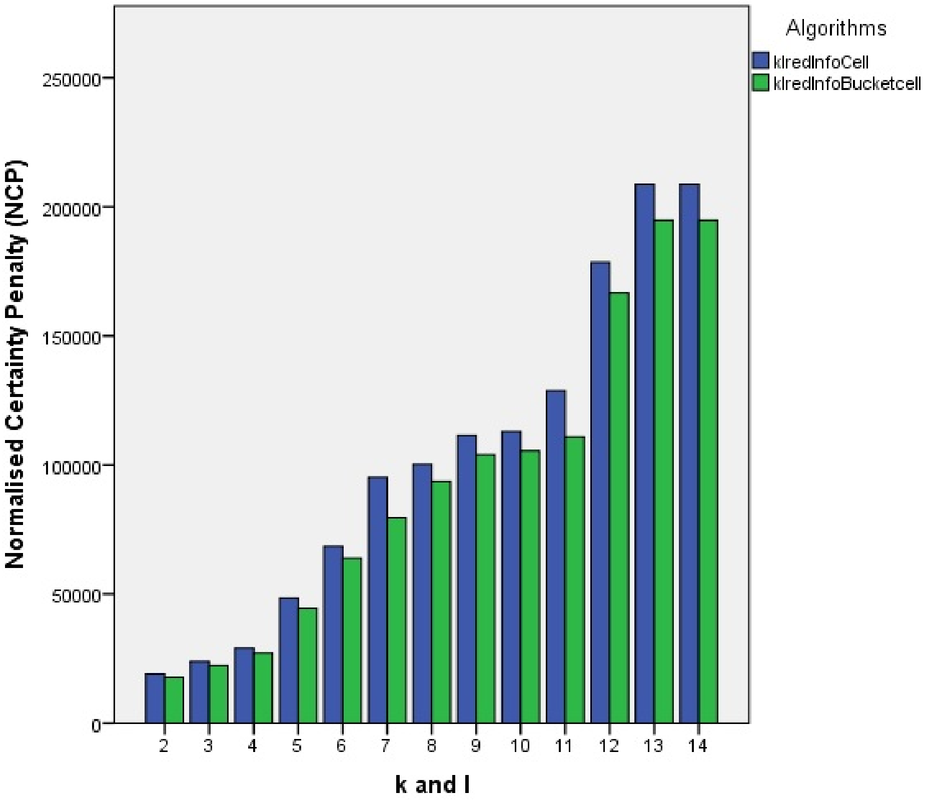


Figure 6.8: NCP resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the PatInfo dataset

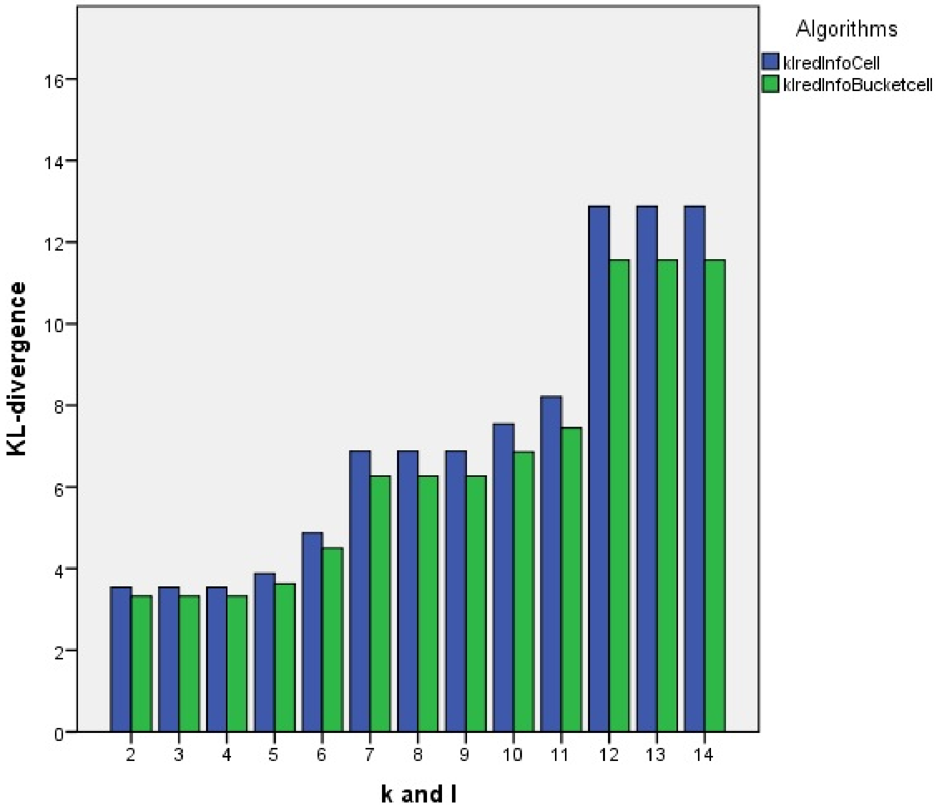
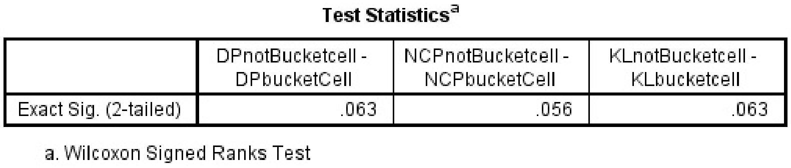


Figure 6.9: KL-divergence resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the PatInfo dataset

Observing Figure 6.7, 6.8 and 6.9, the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm when using both bucketisation and cell-based generalisation approaches are lower than the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm when using generalisation approach in all three information loss metrics. Therefore, this research concludes that the approach of using both bucketisation and cell-based generalisation approaches reduces the values of the information loss metrics that indicate reduction in the information loss.

Table 6.6 shows the Wilcoxon signed-rank test statistics results between the *kl-redInfo* with generalisation and the *kl-redInfo* with both bucketisation and cell-based generalisation approaches, measured by the three information loss metrics, DP, NCP, and KL-divergence. Observing the comparison results, the *p-values* are 0.063 for DP, 0.056 for NCP, and 0.063 for KL-divergence. These results indicate that there is not a significant difference between the values of the information loss metrics, since the *p-values* are greater than the acceptable risk level of 0.05. Therefore this indicates that, there is not a significant difference on the values of the information loss metrics resulting from the application of the *kl-redInfo* with generalisation and *kl-redInfo* with both bucketisation and cell-based generalisation approaches.

Table 6.6: The comparison results from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation approaches on the PatInfo dataset



Test Statistics[a]

	DPnotBucketcell - DPbucketCell	NCPnotBucketcell - NCPbucketCell	KLnotBucketcell - KLbucketcell
Exact Sig. (2-tailed)	.063	.056	.063

a. Wilcoxon Signed Ranks Test

When the Adult dataset is used, the results also show there is not significant evidence of the reduction in the values of information loss metrics when using both the bucketisation and cell-based generalisation approaches compared to when using generalisation approach only at the risk level of 0.05. This is due to the *p-value* shown in Appendix E Table E.2 being 0.127 for DP, 0.244 for NCP, and 0.213 for KL-divergence, which are greater than the risk level of 0.05. Refer to Appendix E for the results when Adult dataset is used.

6.2.3 Sorting records according to Quasi-identifiers

The research also studies the effect of the distribution of quasi-identifier attributes (QIDs). This was done by sorting the records according to quasi-identifier attributes before anonymising them. The sorting approach reduces the possibility of the record values that are very different to be in the same equivalence class, which forces the need for high generalisation level to make them indistinguishable from each other in order to achieve *k-anonymity* privacy requirement.

The values of the information loss metrics resulting from the application of the algorithm when the records are sorted according to quasi-identifier attributes were calculated using the three information loss metrics; DP, NCP, and KL-divergence. The results were compared with the values of the information loss metrics resulting from the application of the algorithm without sorting the records. The values of the information loss metrics resulting from the application of the *kl-redInfo* without sorting the records, named *kl-redInfoNotSorted*, and when the records are sorted according to quasi-identifier attributes, named *kl-redInfoSorted*, are presented in Figure 6.10, 6.11, and 6.12. Refer to Appendix F for the results when Adult dataset is used.

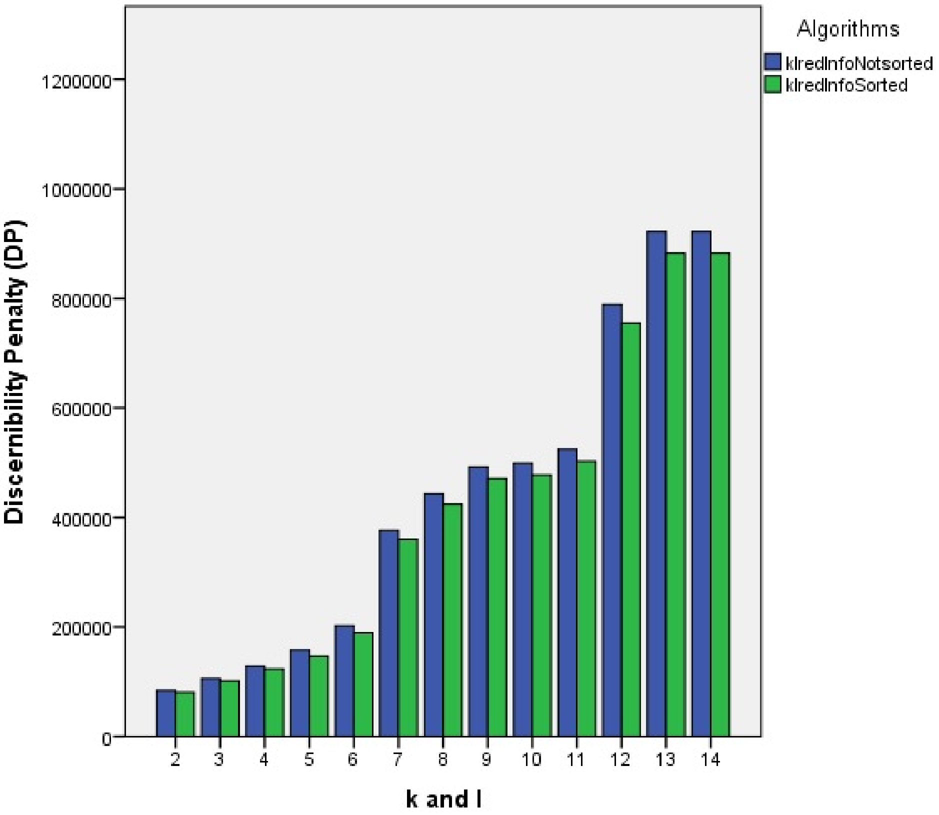


Figure 6.10: DP resulting from the application of the *kl-redInfo* without and with sorting approach on the PatInfo dataset

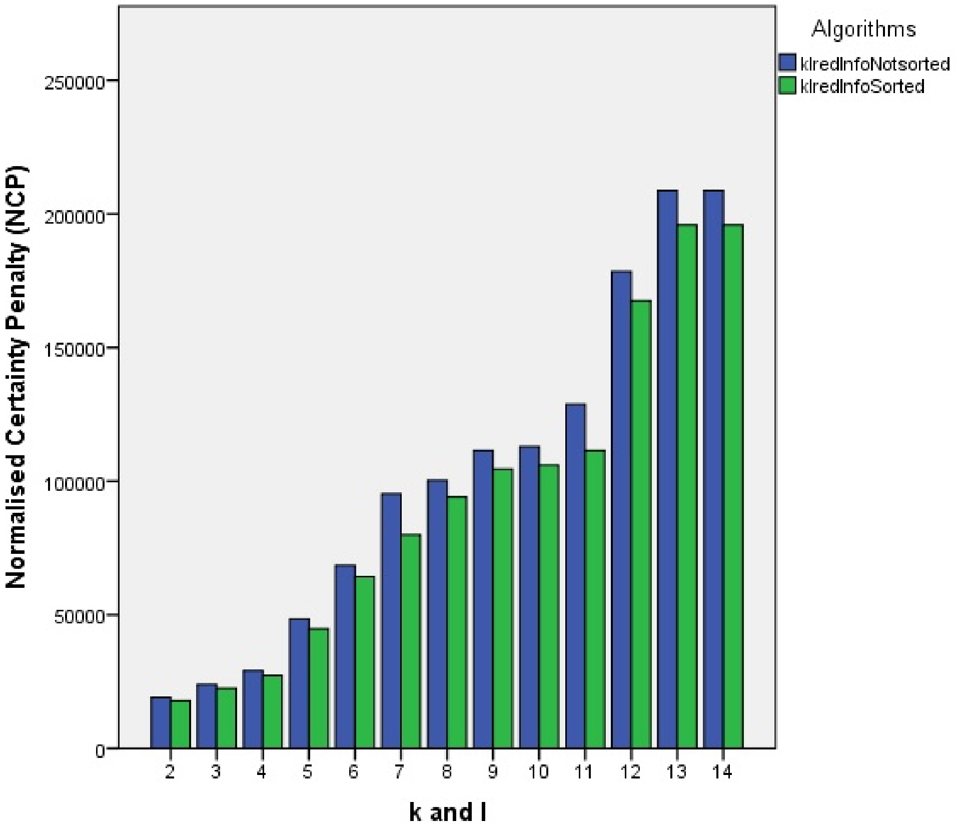


Figure 6.11: NCP resulting from the application of the *kl-redInfo* without and with sorting approach on the PatInfo dataset

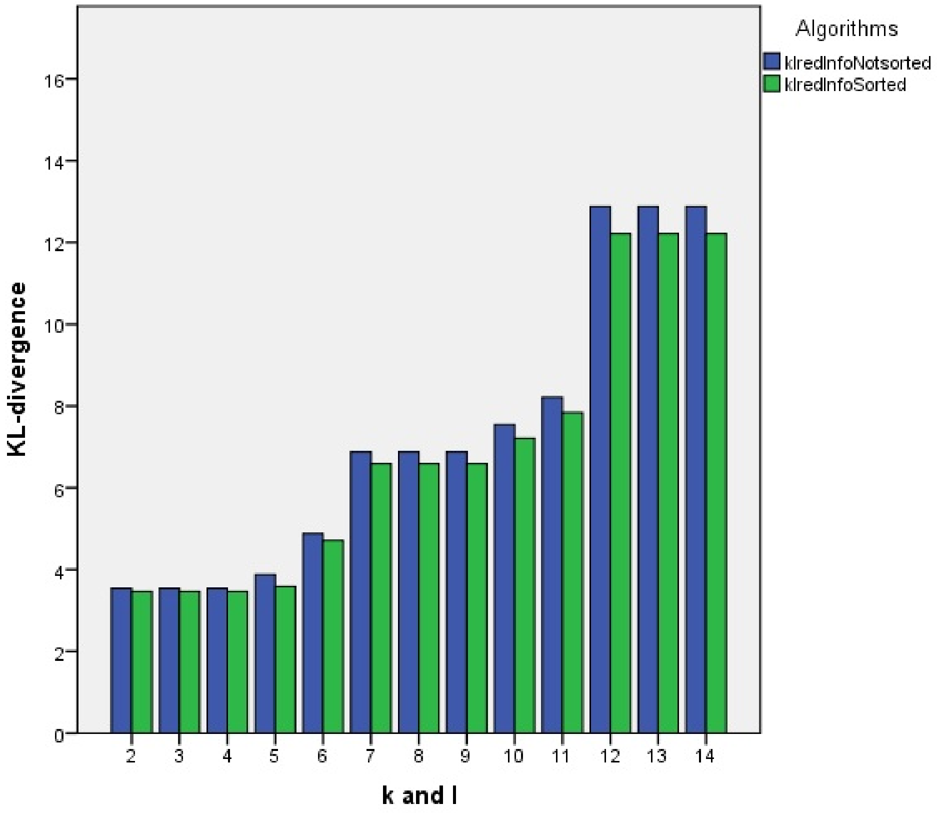
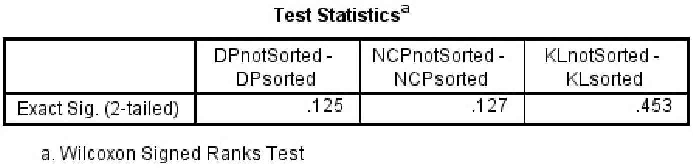


Figure 6.12: KL-divergence resulting from the application of the *kl-redInfo* without and with sorting approach on the PatInfo dataset

As shown in Figure 6.10, 6.11 and 6.12, the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm when records are sorted according to QIDs is lower than the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm when records are not sorted in all three information loss metrics. Therefore, this research concludes that the approach of sorting the records according to QIDs reduces the values of the information loss metrics that indicate reduction in information loss.

The significance of the difference of the values of the information loss metrics when the records are not sorted and when the records are sorted according to the quasi-identifiers attributes, were calculated by the Wilcoxon signed-rank test statistic. Table 6.7 shows the comparison results between the *kl-redInfo* algorithm when records are not sorted according to QIDs and when records are sorted according to QIDs. The table shows that the p-values are 0.125 for DP, 0.127 for NCP, and 0.453 for KL-divergence metric.

Table 6.7: The comparison results from the application of the *kl-redInfo* without and with sorting approach on PatInfo dataset



Test Statistics[a]

	DPnotSorted - DPsorted	NCPnotSorted - NCPsorted	KLnotSorted - KLsorted
Exact Sig. (2-tailed)	.125	.127	.453

a. Wilcoxon Signed Ranks Test

Thus, the *p-values* are greater than the acceptable risk level of 0.05. This indicates that, even though by looking to the histograms the values of the information loss metrics was reduced, the Wilcoxon signed-rank test statistics indicate that there is not enough evidence to conclude that there is a significant difference in the values of the information loss metrics. The difference is not significant due to the fact that the records were later grouped according to distinct sensitive attributes to achieve *l-diversity* privacy requirement, thus decreasing the usefulness of sorting the records according to QIDs. Therefore, this research concludes that there is not significant evidence on the reduction in the values of information loss metrics between the *kl-redInfo* algorithm when records are not sorted according to QIDs and when records are sorted at the risk level of 0.05.

Refer to Appendix F for the results of the values of the information loss metrics of the algorithm on the Adult dataset. The p-values are also higher that 0.05, 0.497 for DP, 0.455 for NCP, and 0.364 for KL-divergence, as shown in Appendix F Table F.2. Therefore, even when the Adult dataset is used the conclusion of not having significant evidence on the reduction in the values of information loss metrics when records are sorted and when records are not sorted in *kl-redInfo* algorithm is drawn.

The values of the information loss metrics resulting from the *kl-redInfo* algorithm when a pair of two approaches is used was also studied. The pairs are: systematic incorporation and the use of both bucketisation and cell-based generalisation, systematic incorporation and sorting, and the use of both bucketisation and cell-based generalisation approach and sorting. The results show that, the values of the information loss metrics resulting from the application of the algorithms with a pair of two approaches are further reduced compared to when the approaches are used separate.

Also, the results show that there is more reduction in the values of the information loss metrics every time when the approach of systematic incorporation of the remaining records was used. This indicates that the approach of systematic incorporation of the remaining records has more effect in reducing the values of the information loss metrics that indicate reduction in information loss. Section 6.3 discusses further the impact of the proposed approaches.

6.3 The Significance of the Proposed Modifications

Each of the proposed approaches contributes in reducing the value of information loss metrics that indicate reduction in information loss, but their impacts are not the same. This is indicated by not only the *p-values* of the Wilcoxon signed-rank test statistics, presented in their respectively sections, but also based on the percentage difference of the value of information loss metrics when the approaches are not used compared to when the approaches are used. The higher the percentage difference of the information loss metric indicates the more impact of the approach in reducing the values of the information loss metrics.

Table 6.8, 6.9 and 6.10 summarise the values of the information loss metrics, the difference of the information loss metrics (Diff), and the percentage difference of the information loss metrics with respect to the values of the information loss metrics resulting from the application of the existing approach on the PatInfo dataset.

Table 6.8: The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm when records sequentially and systematically incorporated on the PatInfo dataset

k,l	Information loss Metrics								
	DP			**NCP**			**KL**		
	kl-sequential	*kl-systematic*	**Diff (%)**	*kl-sequential*	*kl-systematic*	**Diff (%)**	*kl-sequential*	*kl-systematic*	**Diff (%)**
2	84252	72216	12036(14)	19070	17735	1335(7)	3.4643	2.8891	0.5752(17)
3	105756	90648	15108(14)	23937	22262	1675(7)	3.4643	2.8891	0.5752(17)
4	128527	110166	18361(14)	29091	27055	2036(7)	3.4643	2.8891	0.5752(17)
5	157932	121085	36847(23)	48382	44295	4087(8)	3.5893	3.27	0.3193(9)
6	202354	159161	43193(21)	68436	63646	4790(7)	4.7143	3.8414	0.8729(19)
7	376379	322611	53768(14)	95191	79229	15962(17)	6.5893	5.27	1.3193(20)
8	443040	379749	63291(14)	100280	93261	7019(7)	6.5893	5.27	1.3193(20)
9	492152	421845	70307(14)	111396	103599	7797(7)	6.5893	5.27	1.3193(20)
10	499002	427716	71286(14)	112946	105041	7905(7)	7.2143	5.7462	1.4681(20)
11	524681	449727	74954(14)	128759	110446	18313(14)	7.8393	6.2224	1.6169(21)
12	788735	676059	112676(14)	178526	166030	12496(7)	12.2143	9.557	2.6573(22)
13	921970	790260	131710(14)	208683	194076	14607(7)	12.2143	9.557	2.6573(22)
14	921970	790260	131710(14)	208683	194076	14607(7)	12.2143	9.557	2.6573(22)
		Average	16%			8%			19%

Table 6.8 shows that, the values of the information loss metrics resulting from the application of the *kl-redInfo* with the systematic incorporation of the remaining records approach is reduced by an average of 16% of DP, 8% of NCP, and 19% of KL-divergence metrics when compared with the values of the information loss metrics resulting from the application of the *kl-redInfo* with sequential incorporation of the remaining records approach. This indicates that, there is a reduction in the amount of information loss resulting from the application of the approach of systematic incorporation of the remaining records when compared with the approach

of sequential incorporation of the remaining records .

Table 6.9: The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm with generalisation and with both bucketisation and cell-generalisation approaches on the PatInfo dataset

k,l	**Information loss Metrics**								
	DP			**NCP**			**KL**		
	kl-cell	*kl-BucketCell*	**Diff (%)**	*kl-cell*	*kl-BucketCell*	**Diff (%)**	*kl-cell*	*kl-BucketCell*	**Diff (%)**
2	84252	78956	5296(6)	19070	17800	1270(7)	3.4643	3.3272	0.1371(3)
3	105756	99108	6648(6)	23937	22344	1593(7)	3.4643	3.3272	0.1371(3)
4	128527	120448	8079(6)	29091	27155	1936(7)	3.4643	3.3272	0.1371(3)
5	157932	141720	16212(10)	48382	44496	3886(8)	3.5893	3.5284	0.0609(3)
6	202354	193349	9005(4)	68436	63882	4554(7)	4.7143	4.5037	0.2106(4)
7	376379	352721	23658(6)	95191	79522	15669(16)	6.5893	6.2684	0.3209(6)
8	443040	415192	27848(6)	100280	93606	6674(7)	6.5893	6.2684	0.3209(6)
9	492152	461217	30935(6)	111396	103982	7414(7)	6.5893	6.2684	0.3209(6)
10	499002	467636	31366(6)	112946	105429	7517(7)	7.2143	6.8567	0.3576(6)
11	524681	491702	32979(6)	128759	120855	7904(6)	7.8393	7.4449	0.3944(7)
12	788735	749157	39578(5)	178526	166644	11882(7)	12.2143	11.5625	0.6518(11)
13	921970	864018	57952(6)	208683	194794	13889(7)	12.2143	11.5625	0.6518(11)
14	921970	864018	57952(6)	208683	194794	13889(7)	12.2143	11.5625	0.6518(11)
		Average	6%			7%			6%

As shown in Table 6.9, the values of the information loss resulting from the application of the *kl-redInfo* with the bucketisation and cell-based generalisation approach is reduced by an average of 6% of DP, 7% of NCP, and 6% of KL-divergence metrics when compared with the values of the information loss metrics resulting from the application of the approach of generalisation only. This indicates that, there is reduction in the amount of the information loss resulting from the application of the bucketisation and cell-based generalisation approach when compared with the generalisation approach only.

Table 6.10: The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm without and with sorting approach on the PatInfo dataset

k,l	**Information loss Metrics**								
	DP			**NCP**			**KL**		
	kl-NoSort	*kl-sort*	**Diff (%)**	*kl-NoSort*	*kl-sort*	**Diff (%)**	*kl-NoSort*	*kl-sort*	**Diff (%)**
2	84252	80641	3611(4)	19070	17893	1177(6)	3.4643	3.3885	0.0758(2)
3	105756	101223	4533(4)	23937	22460	1477(6)	3.4643	3.3885	0.0758(2)
4	128527	123019	5508(4)	29091	27297	1794(6)	3.4643	3.3885	0.0758(2)
5	157932	146878	11054(7)	48382	44780	3602(7)	3.5893	3.4188	0.1705(5)
6	202354	189396	12958(6)	68436	64214	4222(6)	4.7143	4.6007	0.1136(2)
7	376379	360249	16130(4)	95191	89935	5256(6)	6.5893	6.4188	0.1705(3)
8	443040	424053	18987(4)	100280	94093	6187(6)	6.5893	6.4188	0.1705(3)
9	492152	471060	21092(4)	111396	104523	6873(6)	6.5893	6.4188	0.1705(3)
10	499002	477616	21386(4)	112946	105978	6968(6)	7.2143	7.0249	0.1894(3)
11	524681	502195	22486(4)	128759	121432	7327(6)	7.8393	7.631	0.2083(3)
12	788735	754932	33803(4)	178526	167511	11015(6)	12.2143	11.8734	0.3409(3)
13	921970	882457	39513(4)	208683	195808	12875(6)	12.2143	11.8734	0.3409(3)
14	921970	882457	39513(4)	208683	195808	12875(6)	12.2143	11.8734	0.3409(3)
		Average	5%			6%			3%

The values of the information loss metrics resulting from the application of the *kl-redInfo* with sorting approach is redured by an average of 5% of DP, 6% of NCP, and 3% of KL-divergence, when compared with the values of the information loss metrics resulting from the application of the *kl-redInfo* without sorting the records, as shown in Table 6.10. This indicates that, there is reduction in the amount of the information loss resulting from the application of the sorting approach when compared with the amount of the information loss when the records are not sorted.

As it can be seen in Table 6.8, 6.9 and 6.10, the approach of systematic incorporation of the remaining records has higher percentage difference of the values of the information loss metrics compared with the approach of using bucketisation and cell-based generalisation, and the approach of sorting the records according to their quasi-identifier attributes (QIDs). This is due to the reason that the approach of systematic incorporation of the remaining records ensures that records are incorporated in an equivalence class that results to lower values of the information loss metrics.

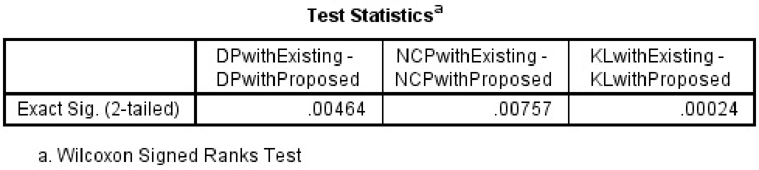
The research find that the values of the information loss metrics resulting from the application of an individual approach does not results in significant reduction in the values of information loss metrics, as shown in their respective sections. Therefore, the research proposed the use of all modified approaches that significantly reduce the values of the information loss metrics that indicate reduction in information loss, as shown in Table 6.11 and confirmed by Wilcoxon signed-rank test results presented in Table 6.12.

The average of the percentage difference of the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm with the proposed modified approaches is 23% lower than the DP of the *kl-redInfo* algorithm without the proposed modified approaches, and is 10% lower than the NCP of the *kl-redInfo* algorithm without the proposed modified approaches, and 40% lower than the KL-divergence of the *kl-redInfo* algorithm without the proposed modified approaches, as shown by **Diff(%)** column. This indicates reduction in the information loss.

Table 6.11: The values of the information loss metrics resulting from the application of the *kl-redInfo* without and with the proposed modified approaches on the PatInfo dataset

k,l	Information loss Metrics								
	DP			NCP			KL		
	kl-without	*kl-with*	Diff (%)	*kl-without*	*kl-with*	Diff (%)	*kl-without*	*kl-with*	Diff (%)
2	84252	62587	21665(26)	19070	17482	1588(8)	3.4643	2.0888	1.3755(40)
3	105756	78562	27194(26)	23937	21944	1993(8)	3.4643	2.0888	1.3755(40)
4	128527	95477	33050(26)	29091	26669	2422(8)	3.4643	2.0888	1.3755(40)
5	157932	151607	6325(4)	48382	43521	4861(10)	3.5893	3.0604	0.5289(15)
6	202354	184606	17748(9)	68436	62738	5698(8)	4.7143	3.4969	1.2174(26)
7	376379	279596	96783(26)	95191	78099	17092(18)	6.5893	3.604	2.9853(45)
8	443040	329116	113924(26)	100280	91931	8349(8)	6.5893	3.604	2.9853(45)
9	492152	365599	126553(26)	111396	102121	9275(8)	6.5893	3.604	2.9853(45)
10	499002	370687	128315(26)	112946	103543	9403(8)	7.2143	3.907	3.3073(46)
11	524681	389763	134918(26)	128759	108871	19888(15)	7.8393	5.21	2.6293(34)
12	788735	585918	202817(26)	178526	153662	24864(14)	12.2143	6.3312	5.8831(48)
13	921970	684892	237078(26)	208683	191308	17375(8)	12.2143	6.3312	5.8831(48)
14	921970	684892	237078(26)	208683	191308	17375(8)	12.2143	6.3312	5.8831(48)
		Average	23%			10%			40%

Table 6.12: The comparison results from the application of the *kl-redInfo* without and with the proposed modified approaches on the PatInfo dataset



Test Statistics[a]

	DPwithExisting - DPwithProposed	NCPwithExisting - NCPwithProposed	KLwithExisting - KLwithProposed
Exact Sig. (2-tailed)	.00464	.00757	.00024

a. Wilcoxon Signed Ranks Test

These conclusions are also drawn when Adult dataset is used, refer to Appendix D, E and F for the results on the Adult dataset. The values of the information loss metrics resulting from the application of the approach of systematic incorporation of the remaining records is reduced by an average of 17% of DP, 18% of NCP, and 24% of KL-divergence metrics when compared with the values of the information loss resulting from the application of the approach of sequential incorporation of the remaining records. The values of the information loss metrics resulting from the application of the approach of bucketisation and cell-based generalisation is reduced by an average of 8% of DP, 10% of NCP, and 9% of KL-divergence metrics when compared with the values of the information loss metrics resulting from the application of the approach of generalisation. The values of the information loss metrics resulting from the application of the sorting approach is reduced by an average of 5% of DP, 4% of NCP, and 8% of KL-divergence metrics when compared with the values of the information loss metrics when the records are not sorted. Therefore, this indicates that the approach of systematic incorporation of the remaining records has most impact in reducing the values of the information loss metrics compared to other approaches.

The results that compare the values of the information loss metrics of *kl-redInfo* with the proposed modified approaches and without proposed modified approaches on the Adult dataset are presented in Appendix G. The values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm with the proposed modified approaches is reduced by an average of 34% of DP, 31% of NCP, and 40% of KL-divergence when compared with the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm without the proposed modified approaches. The results show a significance difference of the values of the information loss metrics when the proposed modified approaches are used compared to when they are not used. This is due to the values of *p-value* being 0.00342 for DP, 0.03271 for NCP, and 0.00024 for KL-divergence, as shown in Appendix G Table G.2, which are lower than the acceptable risk level of 0.05.

6.4 Comparison of *kl-redInfo* with the Existing Algorithms

In order to evaluate the improvement achieved by the proposed *kl-redInfo* algorithm, a comparison with the widely used algorithms, *l-mondrian* and *g-anatomy*, was studied. This was done by comparing the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm with the three proposed approaches and the values of the information loss metrics resulting from the application of the *l-mondrian* and *g-anatomy* algorithms. Figure 6.13, 6.14 and 6.15 show the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modifications, *l-mondrian* and *g-anatomy* algorithms.

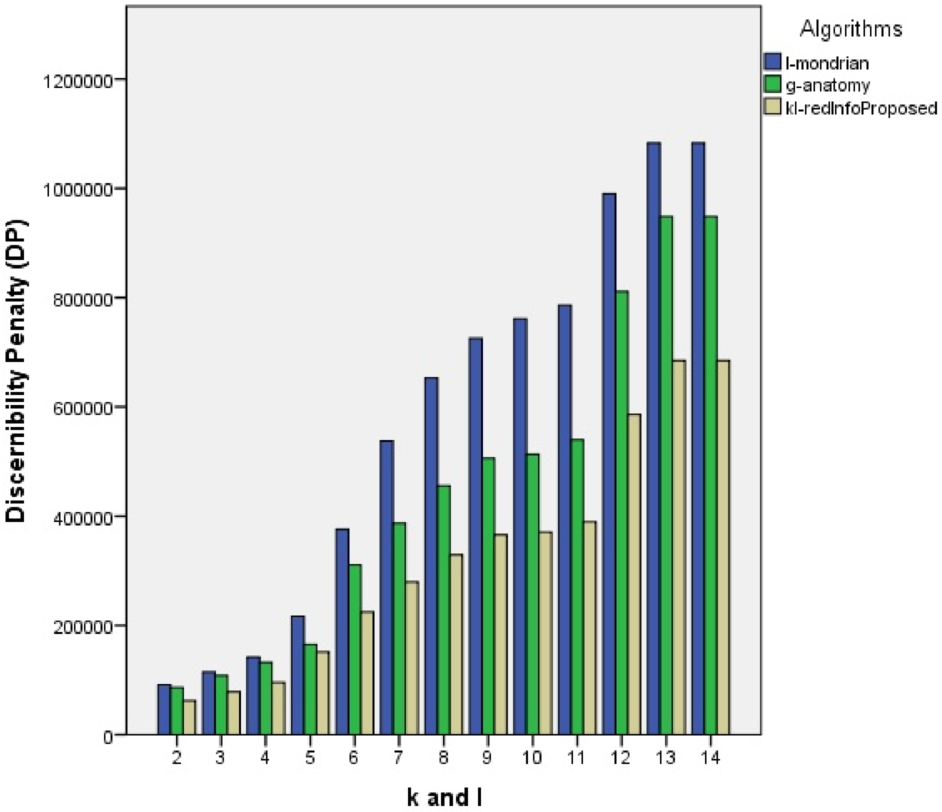


Figure 6.13: DP of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modified approaches on the PatInfo dataset

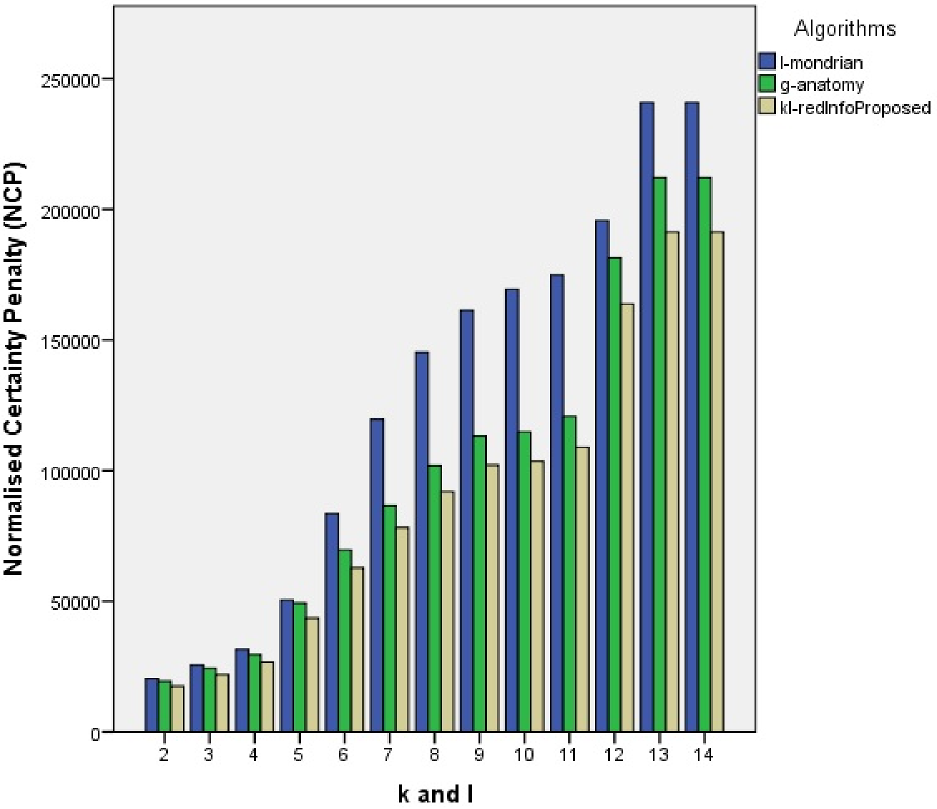


Figure 6.14: NCP of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modified approaches on the PatInfo dataset

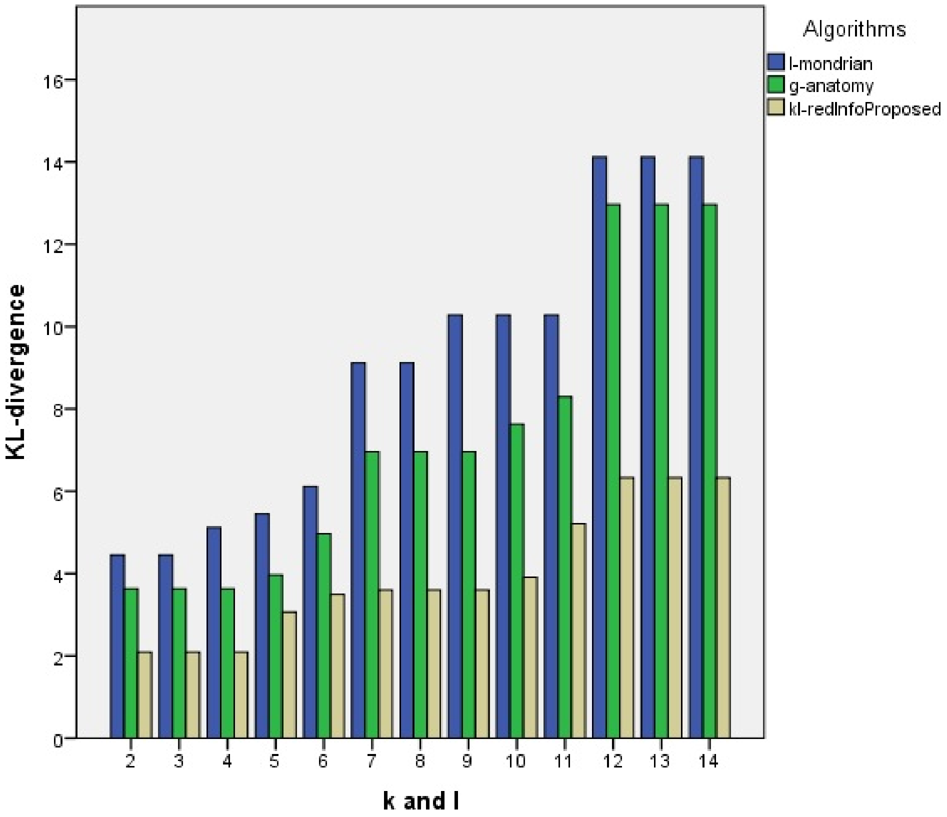


Figure 6.15: KL-divergence of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modified approaches on the PatInfo dataset

As shown in Figure 6.13, 6.14 and 6.15 the proposed *kl-redInfo* algorithm results in lower values of the information loss compared to *l-mondrian* and *g-anatomy* algorithms. The values of the information loss metrics resulting from the application of the *l-mondrian* is high compared with the values of the information loss metrics resulting from the application of the *g-anatomy* and *kl-redInfo* algorithms. This is due to the fact that the *l-mondrian* algorithm incorporates equivalence classes that do not achieve the privacy requirement rather than a record, as *g-anatomy* and *kl-redInfo* do. This increases the size of the equivalence class and hence increases the values of the information loss metrics when transforming them to be indistinguishable from each other.

The differences between the algorithms were also evaluated by using percentages difference of their values of the information loss metrics. The percentage difference of the values of the information loss metrics between the *kl-redInfo* algorithm and *g-anatomy*, and *kl-redInfo* and *l-mondrian*, are summarised in Table 6.13 and 6.14.

As it can be seen in Table 6.13, the results show that the average of the percentage difference of the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm with the proposed modified approaches is 39% lower than the DP of the *l-mondrian* algorithm, and it is 24% lower than the NCP of the *l-mondrian* algorithm, and 55% lower than the KL-divergence of the *l-mondrian* algorithm, as shown by **Diff (%)** column. This indicates the reduction in the information loss resulting from the application of the *kl-redInfo* when compared with *l-mondrian* measured in terms of DP, NCP, and KL-divergence metrics.

Table 6.13: The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modified approaches and *l-mondrian* on the PatInfo dataset

k,l	**Information loss Metrics**								
	DP			**NCP**			**KL**		
	l-mondrian	*kl-redInfo*	**Diff (%)**	*l-mondrian*	*kl-redInfo*	**Diff (%)**	*l-mondrian*	*kl-redInfo*	**Diff (%)**
2	91473	62587	28886(32)	20341	17482	2859(14)	4.4483	2.0888	2.3595(53)
3	114821	78562	36259(32)	25533	21944	3589(14)	4.4483	2.0888	2.3595(53)
4	141843	95477	46366(33)	31542	26669	4873(15)	5.115	2.0888	3.0262(59)
5	216855	151607	65248(30)	50461	43521	6940(14)	5.4483	3.0604	2.3879(44)
6	275847	184606	46366(33)	83579	62738	20841(25)	6.115	3.4969	2.6181(43)
7	537525	279596	257929(48)	119532	78099	41433(35)	9.115	3.604	5.511(60)
8	653152	329116	324036(50)	145244	91931	53313(37)	9.115	3.604	5.511(60)
9	725257	365599	359658(50)	161279	102121	59158(37)	10.2817	3.604	6.6777(65)
10	761364	370687	390677(51)	169307	103543	65764(39)	10.2817	3.907	6.3747(62)
11	786277	389763	396514(50)	174848	108871	65977(38)	10.2817	5.21	5.0717(49)
12	789857	585918	203939(26)	175644	163662	11982(7)	14.115	6.3312	7.7838(55)
13	1083205	684892	398313(37)	240877	191308	49569(21)	14.115	6.3312	7.7838(55)
14	1083205	684892	398313(37)	240877	191308	49569(21)	14.115	6.3312	7.7838(55)
		Average	39%			24%			55%

The results in Table 6.14, show that the average of the percentage difference of the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm with the proposed modified approaches is 25% lower than the DP of the *g-anatomy* algorithm, and it is 10% lower than the NCP of the *g-anatomy* algorithm, and 43% lower than the KL-divergence of the *g-anatomy* algorithm, as shown by **Diff (%)** column. This indicates the reduction in the information loss resulting from the application of the *kl-redInfo* when compared with *g-anatomy* measured in terms of DP, NCP, and KL-divergence metrics.

Table 6.14: The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modified approaches and *g-anatomy* on the PatInfo dataset

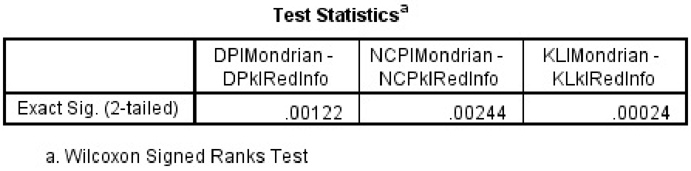
k,l	**Information loss Metrics**								
	DP			**NCP**			**KL**		
	g-anatomy	*kl-redInfo*	**Diff (%)**	*g-anatomy*	*kl-redInfo*	**Diff (%)**	*g-anatomy*	*kl-redInfo*	**Diff (%)**
2	86659	62587	24072(28)	19371	17482	1889(10)	3.632	2.0888	1.5432(42)
3	108778	78562	30216(28)	24323	21944	2379(10)	3.632	2.0888	1.5432(42)
4	132199	95477	36722(28)	29560	26669	2891(10)	3.632	2.0888	1.5432(42)
5	165302	151607	13695(8)	49321	43521	5800(12)	3.9653	3.0604	0.9049(23)
6	210993	184606	26387(13)	69538	62738	6800(10)	4.9653	3.4969	1.4684(30)
7	387133	279596	107537(28)	86563	78099	8464(10)	6.9653	3.604	3.3613(48)
8	455699	329116	126583(28)	101894	91931	9963(10)	6.9653	3.604	3.3613(48)
9	506214	365599	140615(28)	113189	102121	11068(10)	6.9653	3.604	3.3613(48)
10	513259	370687	142572(28)	114765	103543	11222(10)	7.632	3.907	3.725(49)
11	539672	389763	149909(28)	120671	108871	11800(10)	8.2986	5.21	3.0886(37)
12	811271	585918	225353(28)	181400	163662	17738(10)	12.9653	6.3312	6.6341(51)
13	948312	684892	263420(28)	212042	191308	20734(10)	12.9653	6.3312	6.6341(51)
14	948312	684892	263420(28)	212042	191308	20734(10)	12.9653	6.3312	6.6341(51)
		Average	25%			10%			43%

Table 6.13 and 6.14 show significant reduction in the values of the information loss metrics resulting from the application of the proposed *kl-redInfo* algorithm with the proposed modified approaches when compared with the *l-mondrian* and *g-anatomy* algorithms. This is due to the use of the proposed modified approaches. The values of the information loss metrics are much higher when *kl-redInfo* algorithm with the proposed modified approaches is compared with *l-mondrian* algorithm. This is due to the fact that *l-mondrian* incorporates equivalence class rather than a record as

done by *g-anatomy* and *kl-redInfo* algorithms. This forces *l-mondrian* to have large number of the records that have to be indistinguishable from each other in order to form an equivalence class.

The significance of the difference of information loss metrics between the *kl-redInfo* algorithm with the proposed modified approaches and the widely used *l-mondrian*, *g-anatomy* algorithms, were evaluated by the Wilcoxon signed-rank test statistics. The results are presented in Table 6.15, and 6.16.

Table 6.15: The comparison results from the application of the *l-mondrian* and *kl-redInfo* with the proposed modified approaches on the PatInfo dataset



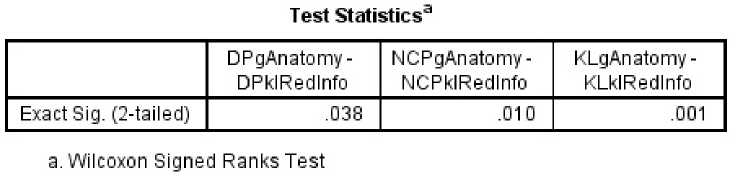
Test Statistics[a]

	DPlMondrian - DPklRedInfo	NCPlMondrian - NCPklRedInfo	KLlMondrian - KLklRedInfo
Exact Sig. (2-tailed)	.00122	.00244	.00024

a. Wilcoxon Signed Ranks Test

The *p-values* in Table 6.15 are 0.00122 for DP, 0.00244 for NCP, and 0.00024 for KL-divergence. Since the *p-values* are lower than the acceptable risk level of 0.05, this implies that there is enough evidence to conclude that there is significant difference between the values of the information loss metrics resulting from the application of the *kl-redInfo* and the values of the information loss metrics resulting from the application of the *l-mondrian* algorithm. This indicates that there is significant reduction in information loss resulting from the application of the *kl-redInfo* when compared with the *l-mondrian* algorithm, measured in terms of DP, NCP, and KL-divergence metrics.

Table 6.16: The comparison results from the application of the *g-anatomy* and *kl-redInfo* with the proposed modified approaches on the PatInfo dataset



Test Statistics[a]

	DPgAnatomy - DPklRedInfo	NCPgAnatomy - NCPklRedInfo	KLgAnatomy - KLklRedInfo
Exact Sig. (2-tailed)	.038	.010	.001

a. Wilcoxon Signed Ranks Test

Table 6.16 shows the comparison results for the values of the information loss metrics resulting from the application of the *kl-redInfo* and *g-anatomy* algorithms on the PatInfo dataset. The *p-values* are 0.038 for DP, 0.010 for NCP, and 0.001 for KL-divergence. This shows that, there is enough evidence to conclude that there is statistically significant difference between the values of the information loss metrics caused *kl-redInfo* and *g-anatomy* algorithm at the risk level of 0.05. This is indicated by the *p-values* which are lower than the acceptable risk level of 0.05. This also indicates that there is significant reduction in information loss resulting from the application of the *kl-redInfo* when compared with the *g-anatomy* algorithms, measured in terms of DP, NCP, and KL-divergence metrics.

The results on real-world Adult dataset also show enough evidence in the reduction of the values of the information loss metrics resulting from the application of the *kl-redInfo* algorithm with the proposed modified approaches compared with the *l-mondrian* and *g-anatomy* algorithms. The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modified approaches is reduced by an average of 43% of DP, 39% of NCP and 50% of KL-divergence when compared with the values of the information loss metrics resulting from the application of *l-mondrian.* The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modified approaches is reduced by an average of 35% of DP, 28% of NCP and 46% of KL-divergence when compared with the values of the information loss metrics resulting from the application of the *g-anatomy.* Refer to Appendix H for detailed comparison results of the *kl-redInfo* algorithm with the proposed modified approaches compared with the *l-mondrian* and *g-anatomy* algorithms on Adult dataset.

6.5 Other Experiments and Findings

In addition to the evaluation experiments, this research also investigates the impact of different characteristics on the implemented algorithms, *kl-redInfo*, *l-mondrian* and *g-anatomy.* These characteristics include; different dataset size considered in Section 6.5.1, different parameter values k and l considered in Section 6.5.2, and the performance speed of the algorithms considered in Section 6.5.3.

6.5.1 Impact of the different datasets size

The research investigates the impact of the different size of the datasets on the implemented algorithms, *kl-redInfo*, *l-mondrian* and *g-anatomy.* This was done by calculating the three information loss metrics on randomly selected sets of 5,000, 10,000, 15,000, 20,000, 25,000, and 30,000 records of both the PatInfo and Adult datasets. Figure 6.16, 6.17, and 6.18 represent the information loss metrics resulting from the application of the proposed *kl-redInfo* algorithm, on the 5,000, 10,000, 15,000, 20,000, 25,000, and 30,000 records of the PatInfo dataset when $k = l = 6$. The information loss is measured by the three information loss metrics, Discernibility Penalty (DP), Normalised Certainty Penalty (NCP), and Kullback-Leibler divergence (KL-divergence) respectively. Refer to Appendix I for the values of the information loss metrics resulting from the application of the proposed *kl-redInfo* algorithm on different size of the Adult dataset.

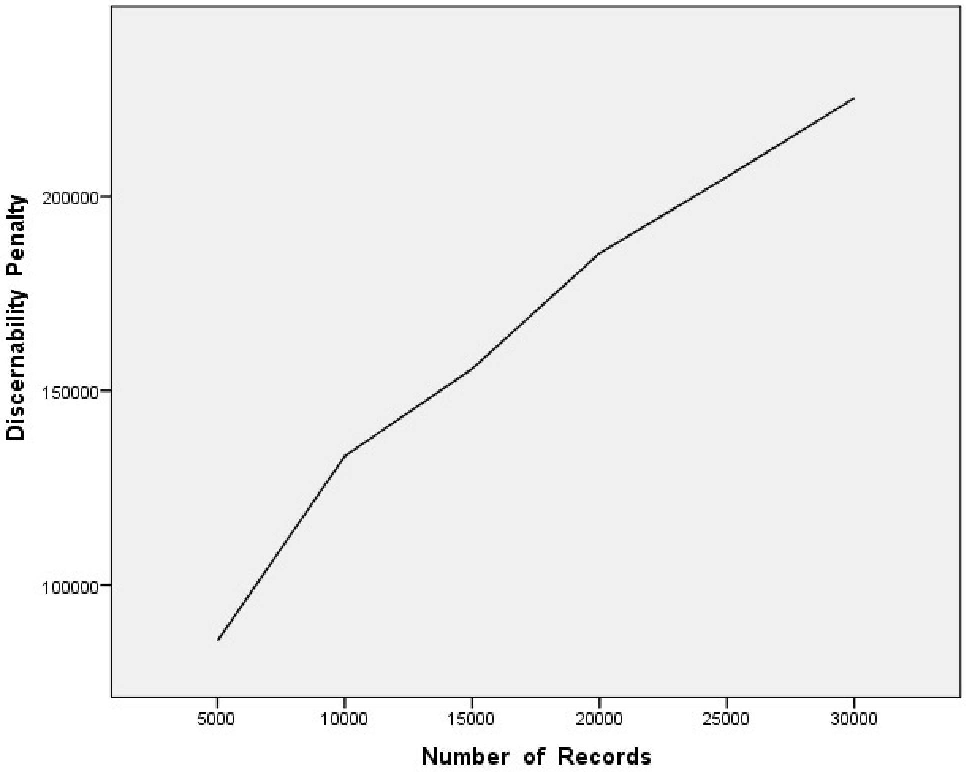


Figure 6.16: DP resulting from the application of the *kl-redInfo* algorithm on different size of the PatInfo dataset when $k = l = 6$

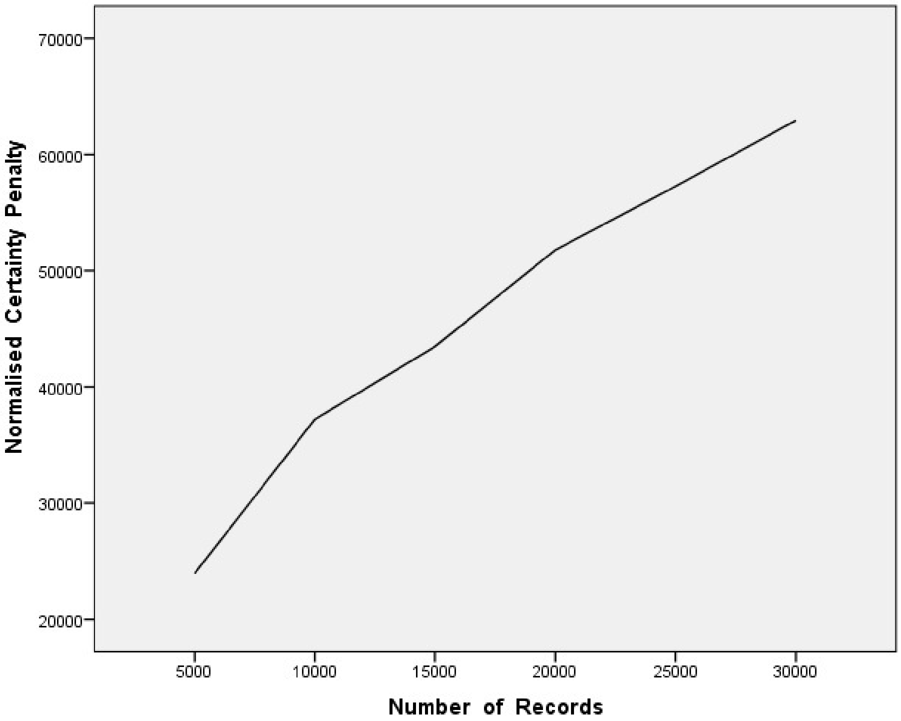


Figure 6.17: NCP resulting from the application of the *kl-redInfo* algorithm on different size of the PatInfo dataset when $k = l = 6$

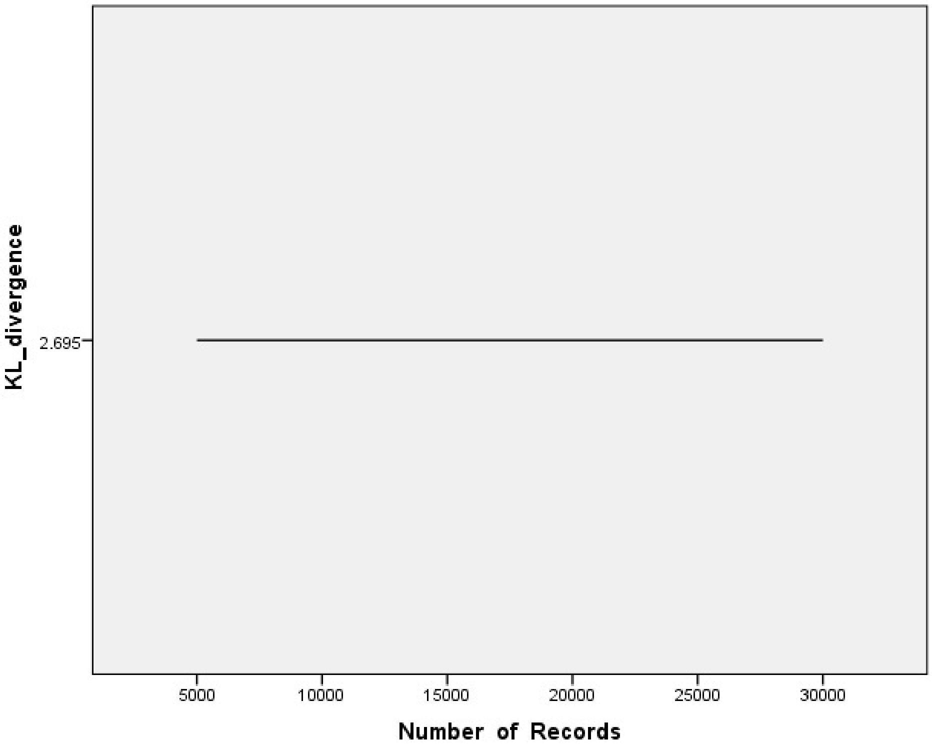


Figure 6.18: KL-divergence resulting from the application of the *kl-redInfo* algorithm on different size of the PatInfo dataset when $k = l = 6$

Figure 6.16 and 6.17 show that there is a linear relationship between the values of the information loss metrics and the size of the dataset. The KL-divergence metrics shown in Figure 6.18 shows similarity of the information loss when different size of the datasets are used. This shows that the probability deviation of the anonymised datasets from the original dataset mostly depends on the values of the parameter k and l, not on the dataset size. The results was also shown when different size of the Adult dataset are used, refer to Appendix I for the results.

In general, as in the literature, when the number of records increases, the values of the information loss metrics also increase, which indicates a increase in the amount of information loss. This is due to the fact that when the number of records increases the size or number of the equivalence classes also increases and hence more generalisation needs to be applied in order to make them indistinguishable from each other.

6.5.2 Impact of the different Parameter values

This research also studied the tradeoff between the level of privacy and the values of the information loss metrics. The level of privacy is indicated by the parameter values of the k and l, when the values increase the level of privacy also increases. The results show that when the values of the parameter k and l increase, the values of the information loss metrics also increases. This is due to the fact that when the values of the parameter k and l increase, the number of records required to be indistinguishable from each other also increases, thus more generalisation should be applied to make them indistinguishable from each other which increase the values of the information loss metric.

As shown in most of the histograms in previous sections, when the value of k and l are greater than 6 there is a steep increase in the values of the information loss metrics. Therefore, even though there is no one value of the parameter k and l that fits all requirements, this research proposes k and l= 6 to be the most appropriate value of the parameter k and l when anonymising the data. Thus, data holders can start from this value on deciding the appropriate values depending on the other characteristics such as the amount of data required, purposes and the users of the anonymised dataset. This comes at the expense of increasing the values of the information loss metrics when the value of $k = l > 6$ or decreasing an individual's privacy when the value of $k = l < 6$.

6.5.3 Algorithms Perfomance

Even though the main interest of this research is to reduce the amount of information loss, the research also evaluated the perfomance speed of each of the implemented algorithms based on the same hardware configurations and datasets. The results show that the *l-mondrian* algorithm performs faster (Average of 55 seconds on the 30200 records of the PatInfo dataset) than *g-anatomy* (Average of 62 seconds on the 30200 records of the PatInfo dataset) and the *kl-redInfo* algorithm (Average of 65 seconds on the 30200 records of the PatInfo dataset).

This is due to the fact that the *l-mondrian* incorporates the group of the records (equivalence class) that does not achieve *l-diversity* privacy requirement instead of incorporating each remaining record as it is done by *g-anatomy* and the *kl-redInfo* algorithms. The *g-anatomy* also performs slightly better than the *kl-redInfo* algorithm due to the fact that *g-anatomy* incorporates records sequentially rather than systematically as *kl-redInfo* does.

In general, most of the characteristics show that the *kl-redInfo* algorithm reduces the values of the information loss metrics that indicate a reduction in information loss. This shows that the *kl-redInfo* with the proposed modifications is better than the widely used *l-mondrian* and *g-anatomy* algorithms.

6.6 Chapter Summary

There are several approaches that can be used to achieve the privacy models, including generalisation, suppression, pseudonymisation and bucketisation. Most of the existing anonymisation approaches result in substantial information loss. To reduce this weakness, this research proposes a systematic incorporation of the remaining records, the use of both bucketisation and cell-based generalisation approaches, and sorting the records according to their quasi-identifiers. The use of these three approaches results in lower values of the information loss metrics that indicates a reduction in the information loss. The approach of systematic incorporation of the remaining records shows most impact in reducing the amount of the information loss compared to other approaches.

The research also found that, the proposed *kl-redInfo* algorithm results in significant reduction in the values of the information loss metrics compared with the widely used *l-mondrian* and *g-anatomy* algorithms. The values of the information loss metrics were more reduced when *kl-redInfo* was compared with *l-mondrian* than when *kl-redInfo* is compared with *g-anatomy* algorithm. This is due to the fact that the *l-mondrian* incorporates equivalence classes that do not achieve *l-diversity* instead of individual records as *g-anatomy* and *kl-redInfo* do.

Chapter 7

CONCLUSIONS AND FUTURE WORK

The use and sharing of the collected data is limited due to the presence of Personal Identifiable Information (PII), whose sharing may breach an individual's privacy. The difficulties in sharing the data arise mainly from the fact that ensuring an individual's privacy results in information loss that renders data less useful. The challenge of ensuring an individual's privacy while providing useful information makes Privacy-Preserving Data Publishing (PPDP) a challenging domain. Most of the existing solutions result in a substantial information loss that renders data less useful.

7.1 Conclusions

This research designed an anonymisation algorithm, named *kl-redInfo*, which results in a reduced amount of information loss compared to the widely used and well-established *l-mondrian* and *g-anatomy* algorithms. The reduction in the information loss is indicated by the lower values of the information loss metrics, disscussed in Chapter 6. This is due to the fact that the three information loss metrics, Discernibility Penalty (DP), Normalised Certainty Penalty (NCP), and Kullback-Leibler divergence (KL-divergence), have useful characteristics that indicate a reduction in information loss.

The DP measures the information loss based on the size of the equivalence classes. The larger the size of the equivalence classes, the higher the level of the generalisation hierarchy that is required to make the records indistinguishable from each other. Not only the size of the equivalence classes that indicates the amount of the information loss, but also the generalisation process used. The NCP metric was used as it takes into account both the size of the equivalence classes and the generalisation process used. Neither the DP, nor the NCP take the data distribution into account, thus this research also used the KL-divergence which takes into account the data distribution. Therefore, the use of these three metrics has a good spread of the indicators of the information loss.

The *kl-redInfo* algorithm was evaluated by comparing the values of the three information loss metrics resulting from the application of the *kl-redInfo* with the *l-diversity* versions of the well-established and widely used *Mondrian* and *Anatomy* algorithms. The results shows that, there is a significant reduction in the values of all three information loss metrics, DP, NCP, and KL-divergence, resulting from the application of the *kl-redInfo* compared to *l-mondrian* and *g-anatomy* algorithms. This implies that there is a significant reduction in information loss when *kl-redInfo* was used compared to when *l-mondrian* and *g-anatomy* algorithms were used. The reduction in information loss is due to the use of the three proposed modifications approaches:

1. Systematic incorporation of the remaining records in the equivalence class that results in a lower value of the information loss metric

2. Using both bucketisation and cell-based generalisation approaches

3. Sorting the records according to the quasi-identifier attributes in order to take

under consideration their distribution

Each of the proposed modified approach contributes in reducing the values of the information loss metrics that indicate a reduction in information loss. The approach of systematic incorporation of the remaining records in the equivalence class that results in a lower value of the information loss metric has more impact in reducing the amount of the information loss compared to the approach of using both bucketisation and cell-based generalisation approach, and the approach of sorting the records according to the quasi-identifier attributes.

The research also found that, the combined use of all three proposed modified approaches results in a significant reduction in the values of the information loss metrics compared to when an individual approach is used. Therefore, this research uses all three approaches to design the proposed *kl-redInfo* algorithm that significantly reduces the values of the information loss metrics, that indicate a reduction in the information loss, compared to the well-established and widely used *l-mondrian* and *g-anatomy* algorithms.

7.2 Research Contributions

The main contribution of this research is the designed *kl-redInfo* anonymisation algorithm that ensure's an individual's privacy with a reduced amount of the information loss. An individual's privacy is ensured by achieving the two main privacy requirements, *k-anonymity* and *l-diversity.* The two privacy requirements ensure individual's privacy against the two main disclosures, identity and attribute disclosures. The information loss is reduced by the use of the proposed modified approaches that are used to design the proposed *kl-redInfo* algorithm. These approaches are:

- Systematic incorporation of the remaining records in the equivalence classes that result in a lower value of the information loss metric
- Using both bucketisation and cell-based generalisation approaches
- Sorting the records according to the quasi-identifier attributes in order to take under consideration their distribution

The proposed modifications are further discussed next:

- **Systematic Incorporation approach**

  A key challenge when anonymising data is what do we do with the records that do not achieve privacy requirements, named *remaining records*. Suppressing the records ensures an individual's privacy but the data cannot be used for various useful purposes such as research, analysis, quality and safety measurement, public health, and marketing, that can enhance quality of services and minimise cost. The approach of sequential incorporation of the remaining records to the equivalence classes, as done by the existing algorithms, results in substantial information loss that renders the data less useful.

  This research proposed the approach that systematically incorporates the remaining records to the equivalence classes, named *systematic incorporation approach*. This is done by calculating the values of the information loss metrics before the record is incorporated in an equivalence class that results in a lower information loss. Unlike the sequential incorporation approach, systematic incorporation approach ensures that the record is incorporated in an equivalence class that results in a lower information loss metric. The research shows that, the values of the information loss metrics resulting from the application of the systematic incorporation approach is reduced when compared to the values of the information loss metrics resulting from the application of the sequential incorporation approach.

- **Bucketisation and Cell-based generalisation approaches**

  Another challenge in anonymising data is on deciding the approaches that can be used in order to ensure an individual's privacy while still striving to reduce the amount of the information loss. Most of the existing solutions use either bucketisation or generalisation approach to achieve privacy requirements. For example, *Mondrian* algorithm uses a generalisation approach but not a bucketisation approach, which results in a substantial information loss, and *Anatomy* algorithm uses a bucketisation but not a generalisation approach, which results in violating an individual's privacy.

  Therefore, this research proposed the use of both bucketisation and cell-based generalisation approaches. These approaches reduce the values of the information loss metrics while ensuring an individual's privacy. The research found that, the values of the information loss metrics, that indicate a reduction in the information loss, resulting from the application of both a bucketisation and cell-generalisation approaches are reduced when compared to the values of the information loss metrics resulting from the application of the generalisation approach only.

- **Sorting the records according to quasi-identifiers**

  The distribution of quasi-identifier attributes (QIDs) is another factor that contributes to the increase of the information loss. This research proposed the approach of sorting not only sensitive attributes, as done by the existing algorithms, but also QIDs in order to consider their distribution. This was done by introducing the approach of sorting the records according to quasi-identifier attributes before anonymising them.

The sorting approach reduces the possibility of the record values that are very different to be in the same equivalence class, which forces the need for a high level of the generalisation hierarchy in order to make them indistinguishable from each other. That increases the values of the information loss metrics that indicate an increase in the amount of the information loss. The results show that, the sorting approach results in a reduction of the values of the information loss metrics when compared with the values of the information loss metrics when the records are not sorted.

7.3 Benefits from this Research

There is much to gain from sharing information which is under pressure due to the growth of different technologies in the modern digital world. This will enhance real time services by providing real-time decision aid such as alerts and reminders. Also, access to enough information and interchange will enable researchers to discover, analyse and predict correct trends of services. Moreover, electronic accessibility to the collected data and knowledge can improve all types of decisions by the use of decision support technology and will transform services. In general, sharing the information will reduce service costs and improve quality by relating the outcomes to service processes. Therefore, the proposed *kl-redInfo* algorithm anonymises the data with a reduced information loss that renders it useful for many purposes.

The beneficiaries of this research include:

- Data holders such as research and healthcare institutions, social networking websites, and search engines, seeking to share their data without violating an individual's privacy. This is because the *kl-redInfo* algorithm is designed to ensure an individual's privacy with reduced information loss that renders data useful. The data holder will enter the dataset to be anonymised and the values of k and l depending on the user of the data and the purpose of use. This is due to the fact that the level of privacy/anonymisation differs depending on the user of the data and the purpose of use. For example, if the data is needed by the doctor for treatment purpose, its level of anonymisation should be lower compared to the data needed by the researcher for a research purpose.
- Privacy-enhancing researchers seeking to expand knowledge and understanding how they can reduce the values of the information loss metrics on the privacy-enhancing approaches.

The *kl-redInfo* algorithm can be used by any domain, but its software implementations may need to be customised depending on the quasi-identifiers to be anonymised. The available software implementation uses the commonly used quasi-identifier attributes including date of birth, address, gender, and marital status. Application of the *kl-redInfo* algorithm on the real healthcare environment, mainly Muhimbili National Hospital in Tanzania where the survey was done, will be of benefit. This will be followed by the prototype system to be registered under open source general license in order to facilitate its adoption for the benefits of different data holders.

7.4 Future Work

Several important issues regarding the designing of privacy-enhancing algorithms have been addressed by this research. The research provides both theoretical and empirical research on the domain. The main contribution made is the proposed anonymisation algorithm for devising and implementing privacy-enhancing algorithms with reduced amount of the information loss. In the process, several challenges were encountered and addressed. This section highlights some of the remaining challenges as future research directions.

This research can be extended in different research directions including:

- **Extending the Algorithm to achieve other Privacy Models**

  The proposed *kl-redInfo* algorithm ensures an individual's privacy by achieving *k-anonymisation* and *l-diversity* privacy requirements. Therefore, the algorithm inherites characteristics of the achieved privacy requirements, as discussed in Section 3.6.1. Thus, expanding the algorithm to achieve other privacy models such as *t-closeness* and *personalised privacy*, will be of benefit to the body of knowledge.

- **Extending the Algorithm to anonymise other types of data**

  The proposed algorithm anonymises structured/relational type of data, which is the subset of the healthcare dataset. Due to diversity nature of the healthcare data, extending the algorithm so that it can anonymise other types of data including, unstructured text and numeric, images, blood sample reports, codes, sounds, and videos, will be of great importance.

- **Extending the Implementation to other Domains**

  The *kl-redInfo* algorithm can be used not only on the healthcare domain but also to other domains such as financial and education domain. This can be achieved by customising its implementation depending on the quasi-identifiers to be anonymised. The implemented algorithms uses the commonly used quasi-identifier attributes including date of birth, address, gender, and marital status. The generalisation hierarchies for these quasi-identifiers were manually formed, automating this process or doing it intelligentlly may improve the process significantly.

- **Developing a Unified Metric for Quantification of Information loss**

  Quantifying the amount of the information loss is still a challenging problem. Each of the existing metric uses different aspects that indicates a reduction in the amount of the information loss. For example, the DP measures the information loss based on the size of the equivalence classes, the NCP takes into account the size of the equivalence classes and the generalisation process, while the KL-divergence takes into account the data distribution. Therefore, there is no single metric that fully measures the amount of the information loss. Developing a metric that takes into consideration all aspects of the information loss will simplify the process of quantifying the amount of the information loss.

- **Developing Parameter Benchmarks for different Data Recipients**

  Deciding the values of parameter k and l to appropriate anonymise the dataset is still a challenging problem. This is because the selection of the parameters is subjective depending on the recipient of the data and the purpose of use. The implemented algorithms can be improved in order to automatically provide an anonymised dataset that takes under consideration the data recipient and the purpose of use. Developing the parameter benchmarks for different data recipients will be the solution of the parameter selection problem. To achieve this, a comprehensive study has to be done in order to take into consideration characteristics of different data recipient and the purpose of use.

Bibliography

Adam, Nabil R. and Worthmann, John C. Security-control methods for statistical databases: a comparative study. *ACM Comput. Surv.*, 21(4):515–556, 1989.

Aggarwal, C.C. and Philip, S.Y. On variable constraints in privacy preserving data mining. In *Proc. of the 5th SIAM International Conference on Data Mining*, pages 115–125, 2005.

Aggarwal, C.C. and Yu, P.S. On static and dynamic methods for condensation-based privacy-preserving data mining. *ACM Transactions on Database Systems (TODS)*, 33(1):1–39, 2008.

Aggarwal, G.; Feder, T.; Kenthapadi, K.; Motwani, R.; Panigrahy, R.; Thomas, D., and Zhu, A. Anonymizing tables. *Database Theory-ICDT 2005*, pages 246–258, 2005.

Agrawal, Dakshi and Aggarwal, Charu C. On the design and quantification of privacy preserving data mining algorithms. In *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255, New York, NY, USA, 2001. ACM.

Agrawal, Rakesh and Srikant, Ramakrishnan. Privacy-preserving data mining. In *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 439–450, New York, NY, USA, 2000. ACM.

Agrawal, Rakesh; Kiernan, Jerry; Srikant, Ramakrishnan, and Xu, Yirong. Hippocratic databases. In *VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases*, pages 143–154. VLDB Endowment, 2002.

Al-Fedaghi, S and Al-Azmi, A. Experimentation with personal identifiable information. *Intelligent Information Management*, Vol. 4 No. 4:123–133, 2012.

Anderson, J.G. Social, ethical and legal barriers to e-health. *International Journal of Medical Informatics*, 76(5-6):480–483, 2007.

Anderson, J.G. and others, . Security of the distributed electronic patient record: a case-based approach to identifying policy issues. *International Journal of Medical Informatics*, 60(2):111, 2000.

Anderson, R.J. A security policy model for clinical information systems. In *Proc. IEEE Symposium on Security and Privacy*, pages 30–43, 1996. doi: 10.1109/SECPRI.1996.502667.

Appari, A. and Johnson, M.E. Information security and privacy in healthcare: current state of research. *International journal of Internet and enterprise management*, 6(4):279–314, 2010.

Appelbaum, P.S. Privacy in psychiatric treatment: threats and responses. *Focus*, 1(4):396, 2003.

Arnold, Ken; Gosling, James, and Holmes, David. *The Java programming language*, volume 2. Addison-wesley Reading, MA, 2000.

Asuncion, A. and Newman, D.J. Uci machine learning repository, 2007.

Bates, D.W. Using information technology to reduce rates of medication errors in hospitals. *British Medical Journal*, 320(7237):788, 2000.

Bates, W. Reducing the frequency of errors in medicine using information technology. *Journal of the American Medical Informatics Association*, 8(4):299, 2001.

Bayardo, Roberto J. and Agrawal, Rakesh. Data privacy through optimal k-anonymization. In *ICDE '05: Proceedings of the 21st International Conference on Data Engineering*, pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.

Beck, Leland L. A security mechanism for statistical databases. Volume 5 , Issue 3:316–338, 1980.

Benedetti, R. and Franconi, L. Statistical and technological solutions for controlled data dissemination. In *Pre-proceedings of New Techniques and Technologies for Statistics*, volume 1, pages 225–232, 1998.

Blum, Avrim; Ligett, Katrina, and Roth, Aaron. A learning theory approach to non-interactive database privacy. *In Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 609–618, 2008.

Bord, A.; Fromm, C.; Kapadia, F.; Molla, D.S.; Sherwood, E.; Sørensen, J.B., and Mahdi, A.R. Ict in health for development. 2009.

Chaum, David L. Untraceable electronic mail, return addresses, and digital pseudonym. *Commun. ACM*, 24(2):84–90, 1981. doi: http://0-doi.acm.org.ditlib.dit.ie:80/10.1145/358549.358563.

Chawla, Shuchi; Dwork, Cynthia; McSherry, Frank; Smith, Adam, and Wee, Hoeteck. Toward privacy in public databases. 2005.

Chin, F.Y. and Ozsoyoglu, G. Auditing and inference control in statistical databases. SE-8(6):574–582, 1982.

Claerhout, B. and DeMoor, GJE. Privacy protection for clinical and genomic data:: The use of privacy-enhancing techniques in medicine. *International Journal of Medical Informatics*, 74(2-4):257–265, 2005.

Clarke, Roger. Introduction to dataveillance and information privacy, and definitions of terms. 1999.

Dalenius, T. and Reiss, S. P. Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6:73–85, 1982.

De Moor, G.; Claerhout, B., and De Meyer, F. Privacy enhancing techniques: The key to secure communication and management of clinical and genomic data. *Meth Info Med*, 42:148–153, 2003.

Denning, Dorothy E. Secure statistical databases with random sample queries. *ACM Trans. Database Syst.*, 5(3):291–315, 1980.

Dinur, Irit and Nissim, Kobbi. Revealing information while preserving privacy. pages 202 – 210, 2003.

DPC, . Eu data protection directive 95/46/e. http://www.dataprotection.ie/viewdoc.asp? m=&fn=/documents/legal/6aii-3.htm25 accessed 30Sept, 2010, 1995.

Duncan, George T.; Fienberg, Stephen E.; Krishnan, Ramayya; Padman, Rema, and Roehrig, Stephen F. Disclosure limitation methods and information loss tabular data. 2001.

Dwork, C. Differential privacy. *Automata, languages and programming*, pages 1–12, 2006.

El Emam, K. Heuristics for de-identifying health data. *IEEE Security & Privacy*, 6(4):58–61, 2008.

El Emam, K.; Dankar, F.K.; Issa, R.; Jonker, E.; Amyot, D.; Cogo, E.; Corriveau, J.P.; Walker, M.; Chowdhury, S.; Vaillancourt, R., and others, . A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

Flanagan, David. *Java in a Nutshell.* O'Reilly Media, Incorporated, 2005.

Fung, B.C.M.; Wang, K.; Chen, R., and Yu, P.S. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Surveys*, 2010.

Gavish, B. and Gerdes Jr, JH. Anonymous mechanisms in group decision support systems communication. *Decision Support Systems*, 23(4):297–328, 1998.

Gavison, R. Privacy and the limits of law. *Yale LJ*, 89:421, 1979.

Gehrke, Johannes. Models and methods for privacy-preserving data analysis and publishing. *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, page 105, 2006.

Ghinita, G.; Karras, P.; Kalnis, P., and Mamoulis, N. Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*, pages 758–769. VLDB Endowment, 2007.

Ghinita, G.; Karras, P.; Kalnis, P., and Mamoulis, N. A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Transactions on Database Systems (TODS)*, 34(2):9, 2009.

Goldschmidt, Peter G. Hit and mis: implications of health information technology and medical information systems. *Commun. ACM*, 48(10):68–74, 2005.

Grimson, Jane. Delivering the electronic healthcare record for the 21st century. *International Journal of Medical Informatics,*, 64:111–127, 2001.

Grimson, Jane; Grimson, William, and Hasselbring, Wilhelm. The si challenge in health care. *Communications of the ACM*, 43:48– 55, 2000.

Han, Jianmin; Huiqun Yu; and Yu, Juan. An improved l-diversity model for numerical sensitive attributes. *Communications and Networking in China, 2008. ChinaCom 2008. Third International Conference*, pages 938–943, 2008.

Herranz, J. and Nin, J. Privacy and anonymity in information management systems. *Privacy and Anonymity in Information Management Systems*, pages 3–6, 2010.

Hevner, A.R.; March, S.T.; Park, J., and Ram, S. Design science in information systems research. *Mis Quarterly*, pages 75–105, 2004.

HIPAA, . Health insurance portability and accountability act of 1996 (hipaa). www.hhs.gov/ocr/privacy, 1996.

Hou, Y.C. and Tu, S.F. A visual cryptographic technique for chromatic images using multi-pixel encoding method. *Journal of Research and Practice in Information Technology*, 37(2):179–192, 2005.

Hundepool, A. and Willenborg, L. -and -argus: software for statistical disclosure control. *Third International Seminar on Statistical Confidentiality*, 1996.

Iyengar, Vijay S. Transforming data to satisfy privacy constraints. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288, New York, NY, USA, 2002. ACM.

Kabir, S.M.A.; Youssef, A.M., and Elhakeem, A.K. On data distortion for privacy preserving data mining. In *Proc. Canadian Conference on Electrical and Computer Engineering CCECE 2007*, pages 308–311, 2007.

Kifer, Daniel and Gehrke, Johannes. Injecting utility into anonymized datasets. pages 217–228, 2006.

Kim, J.J. A method for limiting disclosure in microdata based on random noise and transformation. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 303–308, 1986.

Kleinberg, Jon; Papadimitriou, Christos, and Raghavan, Prabhakar. Auditing boolean attributes. pages 86 – 91, 2000.

Klimavicz, Joseph F. Compliance with dr. sampsons november 6, 2006 memo. 2007.

Kohn, L.T.; Corrigan, J.; Donaldson, M.S., and others, . *To err is human: Building a safer health system.* National Academy Press Washington, DC, 2000.

Lambert, D. Measures of disclosure risk and harm. *Journal of Official Statistics*, 9:313–331, 1993.

Landwehr, C.E. Computer security. *International Journal of Information Security*, 1(1):3–13, 2001.

K.; DeWittLeFevre, D.J.; Ramakrishnan R.;. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference*, 2006.

LeFevre, Kristen; DeWitt, David J., and Ramakrishnan, Raghu. Incognito: Efficient full-domain k-anonymity. In *SIGMOD '05: Proceedings of the 2005 ACM*

SIGMOD international conference on Management of data, pages 49–60, New York, NY, USA, 2005. ACM.

LeFevre, Kristen; DeWitt, David J., and Ramakrishnan, Raghu. Workload-aware anonymization techniques for large-scale datasets. *ACM Trans. Database Syst.*, 33(3):1–47, 2008.

Li, Ninghui; Li, Tiancheng, and Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. IEEE 23rd International Conference on Data Engineering ICDE 2007*, pages 106–115, 2007.

Li, Tiancheng and Li, Ninghui. On the tradeoff between privacy and utility in data publishing. *KDD09*, pages 517– 525, 2009.

Machanavajjhala, Ashwin; Kifer, Daniel; Gehrke, Johannes, and Venkitasubramaniam, Muthuramakrishnan. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3, 2007.

March, S.T. and Smith, G.F. Design and natural science research on information technology. *Decision Support Systems*, 15(4):251–266, 1995.

Mercuri, R.T. The hipaa-potamus in health care data security. *Communications of the ACM*, 47(7):25–28, 2004.

Meyerson, Adam and Williams, Ryan. On the complexity of optimal k-anonymity. In *PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228, New York, NY, USA, 2004. ACM.

Mills, S. Keith; Yao, Rico S., and Chan, Yolande E. Privacy in canadian health networks: challenges and opportunities. *Leadership in Health Services*, 16, Issue 1:1–10, 2003.

Mukherjee, Avinandan and McGinnis, John. E-healthcare: an analysis of key themes in researcher. *International Journal of Pharmaceutical and Healthcare Marketing*, 1; Issue: 4:349–363, 2007.

Muralidhar, K. and Sarathy, R. Security of random data perturbation methods. *ACM Transactions on Database Systems (TODS)*, 24(4):487–493, 1999.

Muralidhar, K.; Parsa, R., and Sarathy, R. A general additive data perturbation method for database security. *Management Science*, 45(10):1399–1415, 1999.

Nergiz, M.E.; Atzori, M., and Clifton, C. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, page 676. ACM, 2007.

Nunez, Manuel A.; Garfinkel, Robert S., and D., Ram. Data perturbation and query restriction combination suggested to tackle problems facing confidential databases. *Operations Research Journal of Biomedical Informatics*, 2007.

Peffers, K.; Tuunanen, T.; Rothenberger, M.A., and Chatterjee, S. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77, 2007.

Pfitzmann, A. and Hansen, M. A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. *URL: http://dud. inf. tu-dresden. de/literatur/Anon_Terminology_v0*, 34, 2010.

Pfitzmann, Andreas and Hansen, Marit. Anonymity, unlinkability, unobservability, pseudonymity, and identity management a consolidated proposal for terminology. 2008.

Pommerening, K. and Michael, R. Secondary use of the ehr via pseudonymisation. *Medical and Care Compunetics 1*, page 441, 2004.

Quantin, C.; Allaert, F.A., and Dusserre, L. Anonymous statistical methods versus cryptographic methods in epidemiology. *International journal of medical informatics*, 60(2):177–183, 2000.

Reiss, Steven P. Practical data-swapping: The first steps. Volume 9 , Issue 1:20–37, 1984.

Riedl, B.; Grascher, V., and S.and NeubauerFenz, T. Pseudonymization for improving the privacy in e-health applications. pages 255–255, 2008.

Safran, C.; Bloomrosen, M.; Hammond, W.E.; Labkoff, S.; Markel-Fox, S.; Tang, P.C., and Detmer, D.E. Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *Journal of the American Medical Informatics Association*, 14(1):1–9, 2007.

Samarati, Pierangela. Protecting respondents' identities in microdata release. 13: 1010–1027, 2001.

Samarati, Pierangela and Sweeney, Latanya. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

Schneier, B. *Secrets & lies: digital security in a networked world.* John Wiley & Sons, Inc. New York, NY, USA, 2000.

Shoniregun, C.A.; Dube, K., and Mtenzi, F. *Electronic healthcare information security.* Springer, 2010.

Siegel, Sidney. Nonparametric statistics. *The American Statistician*, 11(3):13–19, 1957.

Stallings, William and Brown, Lawrence V. *Computer security: Principles and Practice.* Prentice-Hall, 2012.

statute book, Irish. Data protection (amendment) act 2003. http://www.irishstatutebook.ie/2003/en/act/pub/0006/print.html, 2003.

Sweeney, Latanya. Guaranteeing anonymity when sharing medical data, the datafly system. *Uncertainty Fuzziness and Knowledge Based Systems*, pages 51– 55, 1997.

Sweeney, Latanya. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, pages 571–588, 2002.

Sweeney, Latanya. k- anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Volume 10, Issue 5:557–570, 2002a.

Tang, P.C.; Ash, J.S.; Bates, D.W.; Overhage, J.M., and Sands, D.Z. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. *Journal of the American Medical Informatics Association*, 13(2):121–126, 2006.

Tinabo, Rose; Mtenzi, Fredrick, and Brendan, O'Shea. Solving the problem of balancing data usefulness and protection of personal identifiable information using multiple anonymisation techniques. *The 1st International Conference on Networked Digital Technologies (NDT), Ostrava, Czech.*, pages 37–42, 2009a.

Tinabo, Rose; Mtenzi, Fredrick, and O'Shea, Brendan. Designing and developing a new anonymisation technique to be used in e-healthcare. *The 14th Annual Conference of Healthcare Information Society of Ireland (HISI) , Dublin, Ireland.*, 2009b.

Trochim, William M. The research methods knowledge base, 2nd edition. Internet WWW page, october 2006. URL `http://www.socialresearchmethods.net/kb/`.

Truta, T.M. and Vinay, B. Privacy protection: p-sensitive k-anonymity property. In *Proceedings of the 22nd International Conference on Data Engineering Workshops, the Second Intenational Workshop on Privacy Data Management (PDM06)*, page 94, 2006.

U.S. Congress, Office of Technology Assessment. Protecting privacy in computerized medical information. Technical report, Washington, DC:U.S. Government Printing Office, 1993.

Walters, G.J. *Human Rights in an Information Age: A Philosophical Analysis.* Univ of Toronto Pr, 2002.

Wang, K.; Yu, P.S., and Chakraborty, S. Bottom-up generalization: A data mining solution to privacy protection. In *Fourth IEEE International Conference on Data Mining, 2004. ICDM'04*, pages 249–256, 2004.

Wang, K.; Fung, B.C.M., and Yu, P.S. Template-based privacy preservation in classification problems. In *Fifth IEEE International Conference on Data Mining*, page 8, 2005.

Wang, K.; Fung, B.C.M., and Yu, P.S. Handicapping attacker's confidence: an alternative to k-anonymization. *Knowledge and Information Systems*, 11(3):345–368, 2007.

Westin, A.F. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

Wilcoxon, Frank. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

Wright, T. Security, privacy, and anonymity. *Crossroads*, 11(2):5, 2004.

Xiao, X. and Tao, Y. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 229–240. ACM New York, NY, USA, 2006a.

Xiao, X. and Tao, Y. Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd international conference on Very large data bases*, page 150. VLDB Endowment, 2006b.

Xu, J.; Wang, W.; Pei, J.; Wang, X.; Shi, B., and Fu, A.W.C. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 790. ACM, 2006.

Zhang, Lei; Jajodia, Sushil, and Brodsky, Alexander. Information disclosure under realistic assumptions: Privacy versus optimality. pages 573–583, 2007.

Zielinski, Marek Piotr. Balancing privacy and information utility in data anonymisation. 2007.

Appendix A

Information Classification

Information on the area of Privacy-preserving Data Publishing (PII) can be classfied as shown in Figure A.1 and explained in Table A.1. Note that, the privacy significance increases when moving inward.

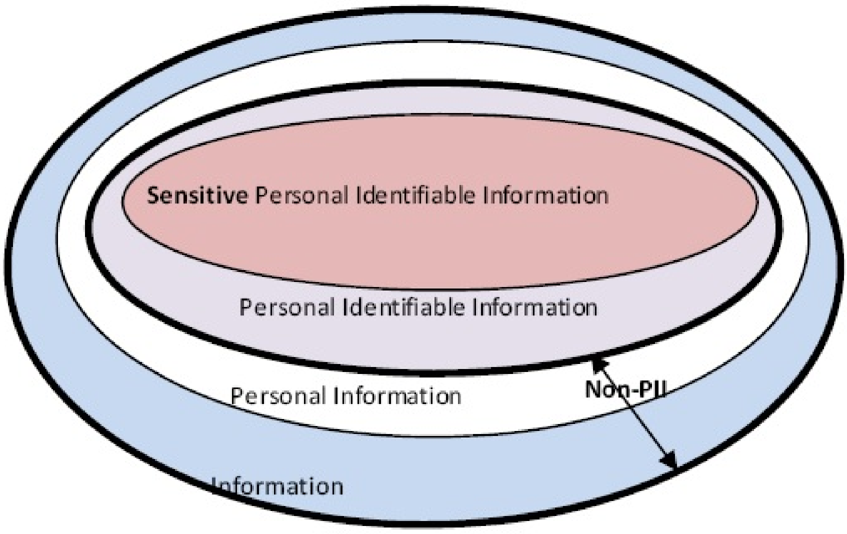

Sensitive Personal Identifiable Information
Personal Identifiable Information
Personal Information
Non-PII
Information

Figure A.1: Information Classification (Al-Fedaghi and Al-Azmi, 2012)

Table A.1: Information Classification

Term	*Explanation*	*Examples*	*Protection Measures*
Information	Information is data that has been processed or analysed to produce something useful.	Census records, criminal records, and voter registrations	Not considered confidential
Personal Information	Information belonging to the private life of an individual that can uniquely identify that individual when are linked together and not on their own.	Gender, Marital status, DOB, Address, Country, and Race	Restricted Access.
Personal Identifiable Information (PII)	Any information that permits the identity of an individual to be directly and indirectly inferred	Name, date and place of birth, biometric records, medical, education, financial, and employment information	Rules and Regulations
Sensitive Personal Identifiable Information	PII that embeds sensitive information	inherently sensitive intimate (e.g., medical or sexual information), judgmental data, and biographical data	More stringent controls

Appendix B

Glossary of Terms

Term	*Explanation*
Anonymisation	Anonymisation is the process that ensures individual information remain un-identified within the set of data.
Data	The building blocks for information. These can be described as numbers, symbols, words, images and graphics that have been validated but yet to be organised or analysed.
Database	A collection of data that is organised so that its contents can easily be accessed, managed, and updated.
Data holder	A data holder is the individual or the legal person who either alone or with others, controls and is responsible for the keeping and use of personal information on computer or in structured manual files.

Data recipient	Any user of data or information produced by the data holder. The data is used for a number of purposes, including planning, decision making and research.
Data subject	An individual who is the subject of data, for example, a patient admitted to a hospital.
Discernibility Penalty (DP)	The information loss metrics that measures the information loss based on size of the equivalence classes.
Equivalence Class	The number of records that have the same quasi-identifiable attributes.
Information	Information is data that has been processed or analysed to produce something useful.
Information Loss	Information loss due to un-identification process
Information and communication technology (ICT)	The tools and resources used to communicate, create, disseminate, store, and manage information electronically.
Input Perturbation	Privacy preserving technique that deals with disturbing data before the release
Kullback-Leibler divergence (KL-divergence)	Information loss metric that measures the information loss based on the data distribution difference of the anonymised dataset compared to the original dataset.
Muhimbili National Hospital (MNH)	The national hospital in Tanzania where the survey was done

Normalised Certainty Penalty (NCP)	The information loss metric that measures the information loss based on effect of generalisation
National Institute for Medical Research (NIMR)	The overseer of all healthcare research in Tanzania
Output Perturbation	Privacy preserving technique that uses query control mechanism to compute exact answers, but it returns disturbed or noisy answers as a response to the query
Personal Identifiable Information (PII)	Personal information is data relating to an individual who is or can be identified either from the data or from the data in conjunction with other information that is in, or is likely to come into, the possession of the data holder
Privacy	The right of an individual to remain un-identifiable within a set of data
Privacy-Preserving Data Publishing (PPDP)	One of the broad areas of privacy-preserving that deals with un-identifying the data so that individuals privacy remains preserved when shared for different purposes
Quasi-identifier Attribute (QID)	The Quasi-Identifier (QID) is a set of attributes that could potentially identify the data subject such as gender, race and marital status.

Sensitive attributes	The Sensitive Attributes consists of sensitive person-specific information such as disease, salary, and disability status.

Table B.1: Glossary of Terms

Appendix C

The Information loss metrics of the algorithms without the proposed modifications

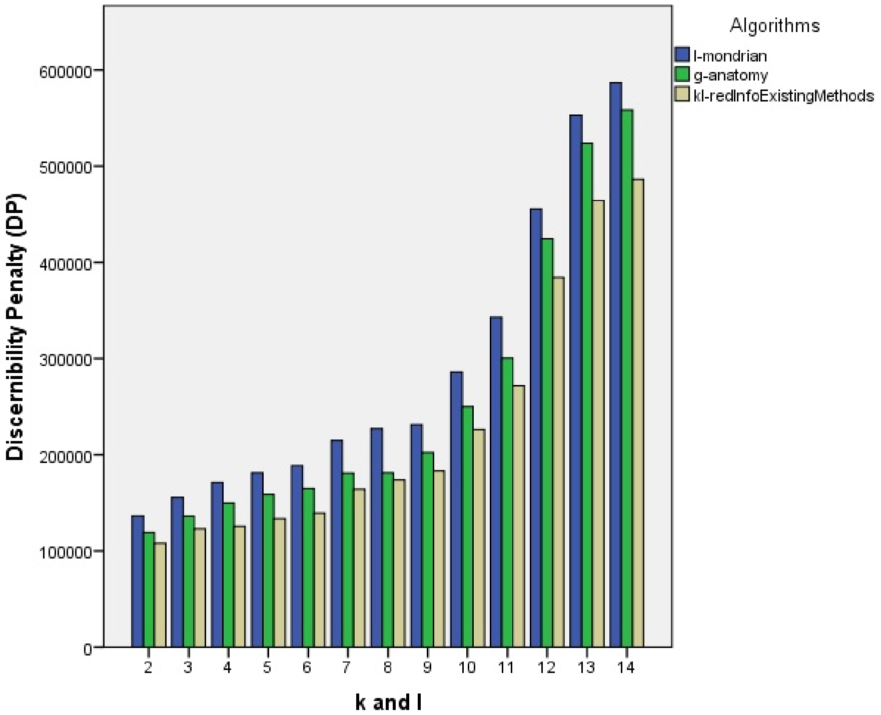


Figure C.1: DP resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* algorithm without the proposed modifications on the Adult dataset

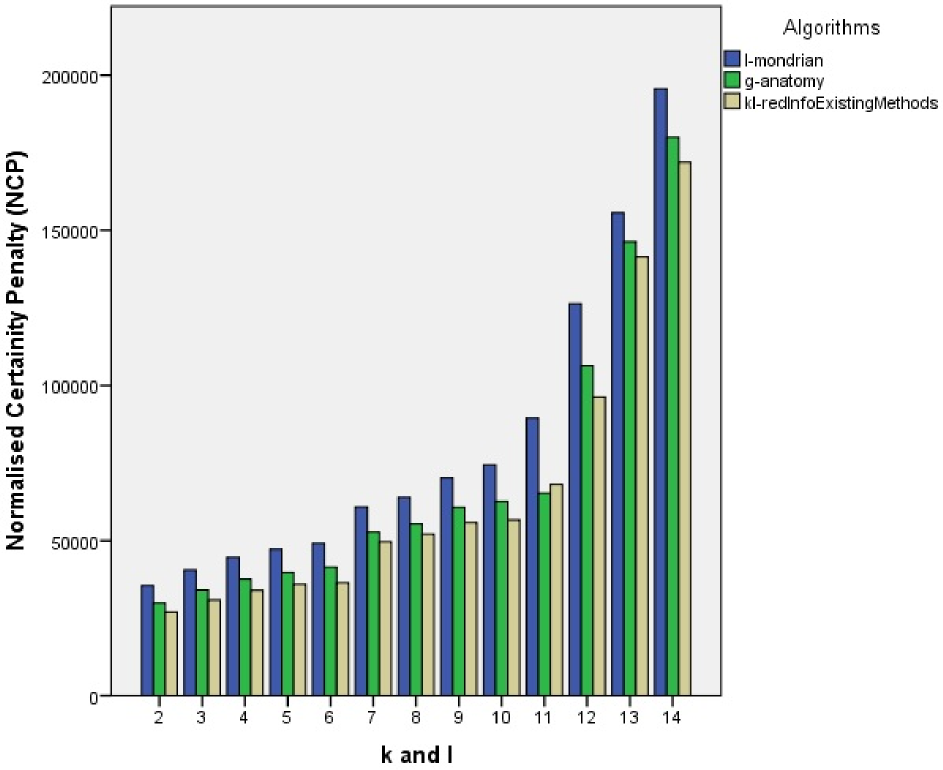


Figure C.2: NCP resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* without the proposed modifications algorithms on the Adult dataset

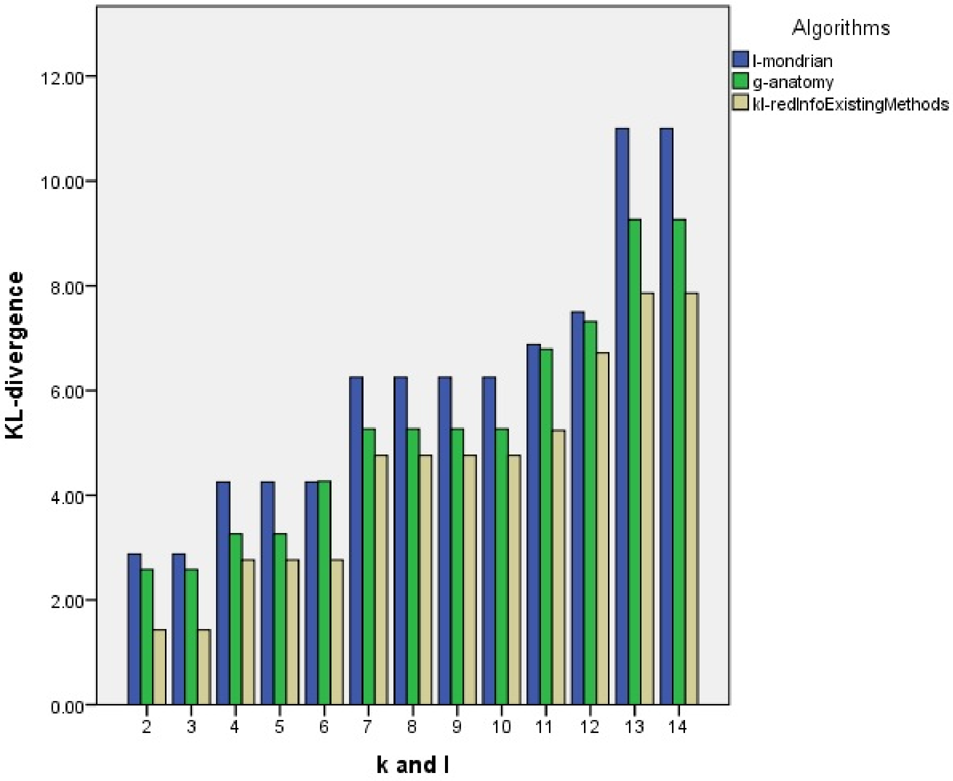
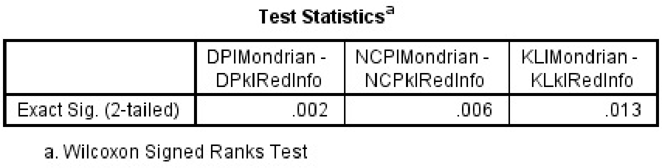


Figure C.3: KL-divergence resulting from the application of the *l-mondrian*, *g-anatomy*, and *kl-redInfo* without the proposed modifications algorithms on the Adult dataset

Table C.1: The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and *l-mondrian* algorithm on the Adult dataset

k,l	Information loss Metrics								
	DP			NCP			KL		
	l-mondrian	*kl-without*	**Diff (%)**	*l-mondrian*	*kl-without*	**Diff (%)**	*l-mondrian*	*kl-without*	**Diff (%)**
2	136183	107812	28371(21)	35464	27020	8444(24)	1.875	1.429	0.446(24)
3	155527	123126	32401(21)	40502	30859	9643(24)	1.875	1.429	0.446(24)
4	171242	125567	45675(27)	44594	33977	10617(24)	3.25	2.762	0.488(15)
5	181186	133439	47747(26)	47184	35950	11234(24)	3.25	2.762	0.488(15)
6	188753	139429	49324(26)	49154	36451	12703(26)	3.25	2.762	0.488(15)
7	215012	164384	50628(24)	60784	49693	11091(18)	5.25	4.762	0.488(9)
8	227067	173928	53139(23)	63924	52085	11839(19)	5.25	4.762	0.488(9)
9	231156	192998	38158(17)	70197	55864	14333(20)	5.25	4.762	0.488(9)
10	285732	226204	59528(21)	74409	56693	17716(24)	5.25	4.762	0.488(9)
11	343044	291861	51183(15)	89428	68136	21292(24)	5.875	5.238	0.637(11)
12	455246	414153	41093(9)	126366	96279	30087(24)	7.5	6.714	0.786(10)
13	552877	464361	88516(16)	155645	141444	14201(9)	11	7.857	3.143(29)
14	586458	485946	100512(17)	195640	171916	23724(12)	11	7.857	3.143(29)
		Average	20%			21%			16%

Table C.2: The comparison results from the application of the *kl-redInfo* algorithm without the proposed modifications and *l-mondrian* on the Adult dataset



Test Statistics[a]

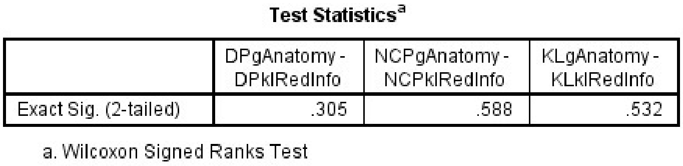
	DPlMondrian - DPklRedInfo	NCPlMondrian - NCPklRedInfo	KLlMondrian - KLklRedInfo
Exact Sig. (2-tailed)	.002	.006	.013

a. Wilcoxon Signed Ranks Test

Table C.3: The values of the information loss metrics resulting from the application of the *kl-redInfo* without the proposed modifications and *g-anatomy* algorithm on the Adult dataset

k,l	Information loss Metrics								
	DP			NCP			KL		
	g-anatomy	*kl-without*	Diff (%)	*g-anatomy*	*kl-without*	Diff (%)	*g-anatomy*	*kl-without*	Diff (%)
2	119160	107812	11348(10)	29865	27020	2845(10)	1.579	1.429	0.15(9)
3	136086	123126	12960(10)	34107	30859	3248(10)	1.579	1.429	0.15(9)
4	149837	125567	24270(16)	37553	33977	3576(10)	3.263	2.762	0.501(15)
5	158537	133439	25098(16)	39734	35950	3784(10)	3.263	2.762	0.501(15)
6	165159	149429	15730(10)	41393	36451	4942(12)	3.263	2.762	0.501(15)
7	180635	164384	16251(9)	52766	49693	3073(6)	5.263	4.762	0.501(10)
8	181184	173928	7256(4)	55409	52085	3324(6)	5.263	4.762	0.501(10)
9	202261	192998	9263(5)	60692	55864	4828(8)	5.263	4.762	0.501(10)
10	250015	226204	23811(10)	62661	56693	5968(10)	5.263	4.762	0.501(10)
11	300478	291861	8617(3)	68308	68136	172(0)	5.789	5.238	0.551(10)
12	424591	414153	10438(2)	106414	96279	10135(10)	7.316	6.714	0.602(8)
13	523767	464361	59406(11)	146333	141444	4889(3)	8.263	7.857	0.406(5)
14	558150	485946	72204(13)	180013	171916	8097(4)	8.263	7.857	0.406(5)
		Average	9%			7%			10%

Table C.4: The comparison results from the application of the *kl-redInfo* algorithm without the proposed modifications and *g-anatomy* on the Adult dataset



Test Statistics[a]

	DPgAnatomy - DPklRedInfo	NCPgAnatomy - NCPklRedInfo	KLgAnatomy - KLklRedInfo
Exact Sig. (2-tailed)	.305	.588	.532

a. Wilcoxon Signed Ranks Test

Appendix D

The information loss metrics of the *kl-redInfo* with sequential and systematic incorporation approaches

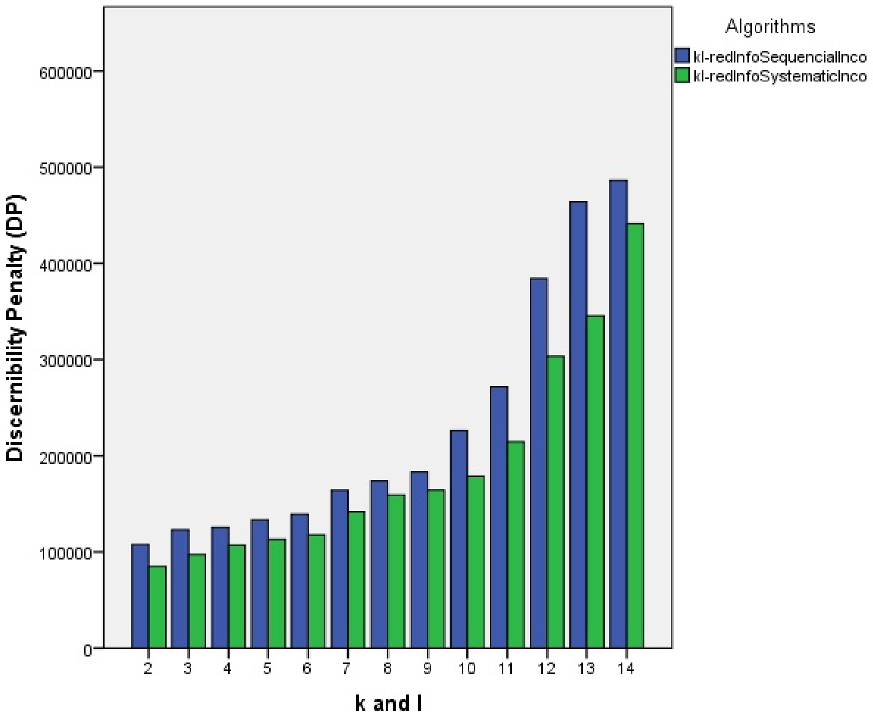


Figure D.1: DP resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset

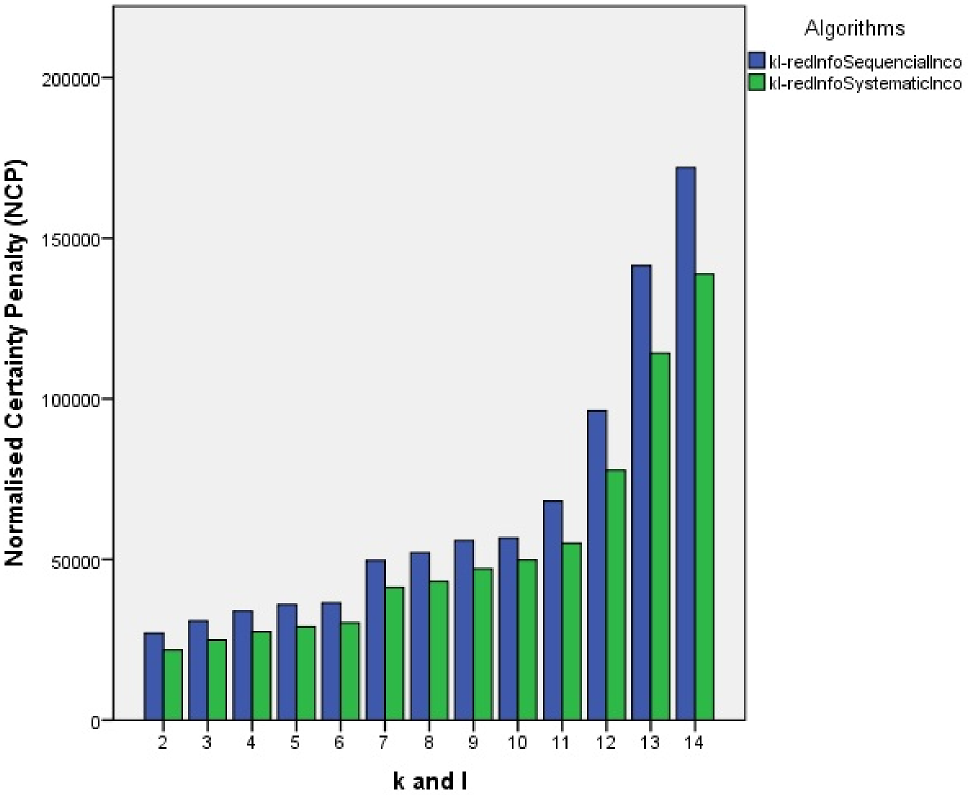


Figure D.2: NCP resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset

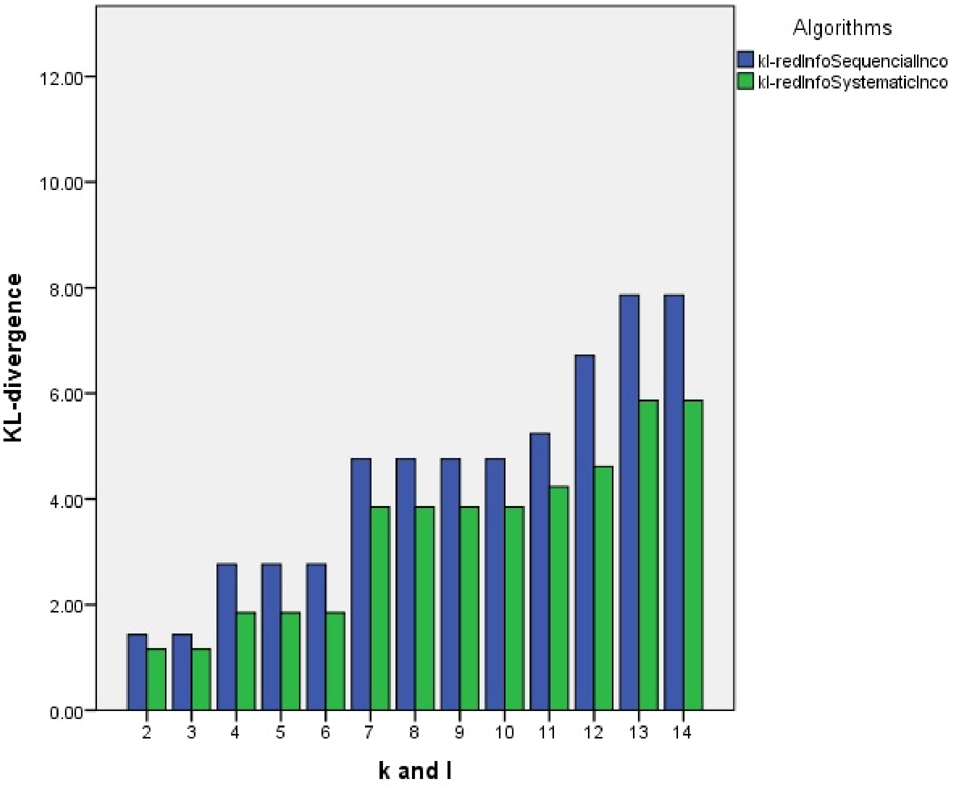
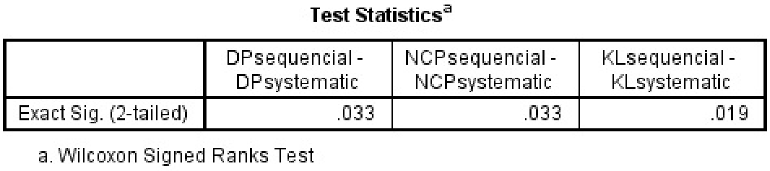


Figure D.3: KL-divergence resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset

Table D.1: The values of the information loss metrics resulting from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset

k,l	Information loss Metrics								
	DP			NCP			KL		
	kl-sequential	*kl-systematic*	Diff (%)	*kl-sequential*	*kl-systematic*	Diff (%)	*kl-sequential*	*kl-systematic*	Diff (%)
2	107812	85114	22698(21)	27020	21824	5196(19)	1.429	1.154	0.275(19)
3	123126	97204	25922(21)	30859	24924	5935(19)	1.429	1.154	0.275(19)
4	125567	107026	18541(15)	33977	27443	6534(19)	2.762	1.846	0.916(33)
5	133439	113241	20198(15)	35950	29036	6914(19)	2.762	1.846	0.916(33)
6	139429	117970	21459(15)	36451	30249	6202(17)	2.762	1.846	0.916(33)
7	164384	141882	22502(14)	49693	41252	8441(17)	4.762	3.846	0.916(19)
8	173928	159417	14511(8)	52085	43184	8901(17)	4.762	3.846	0.916(19)
9	182998	164472	18526(10)	55864	47044	8820(16)	4.762	3.846	0.916(19)
10	226204	178582	47622(21)	56693	49790	6903(12)	4.762	3.846	0.916(19)
11	271861	214627	57234(21)	68136	55033	13103(19)	5.238	4.231	1.007(19)
12	384153	303279	80874(21)	96279	77764	18515(19)	6.714	4.615	2.099(31)
13	464361	345548	118813(26)	141444	114243	27201(19)	7.857	5.862	1.995(25)
14	485946	441536	44410(9)	171916	138855	33061(19)	7.857	5.862	1.995(25)
		Average	17%			18%			24%

Table D.2: The comparison results from the application of the *kl-redInfo* with sequential and systematic incorporation approaches on the Adult dataset



Test Statistics[a]

	DPsequencial - DPsystematic	NCPsequencial - NCPsystematic	KLsequencial - KLsystematic
Exact Sig. (2-tailed)	.033	.033	.019

a. Wilcoxon Signed Ranks Test

Appendix E

The information loss metrics of the *kl-redInfo* with generalisation and with both bucketisation and generalisation

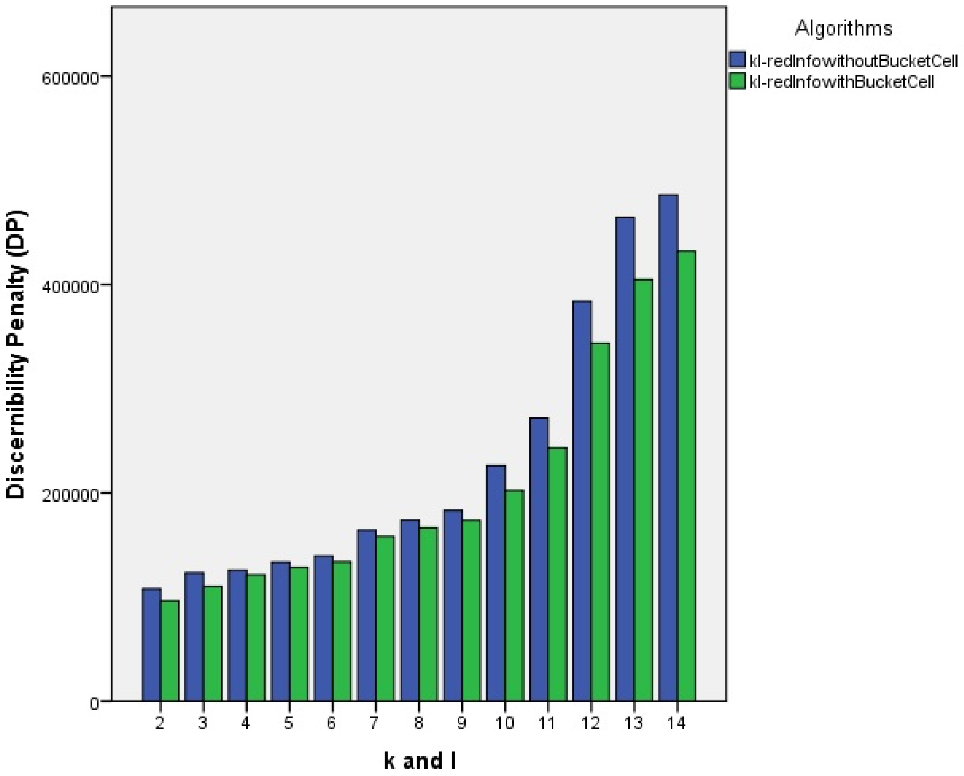


Figure E.1: DP resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the Adult dataset

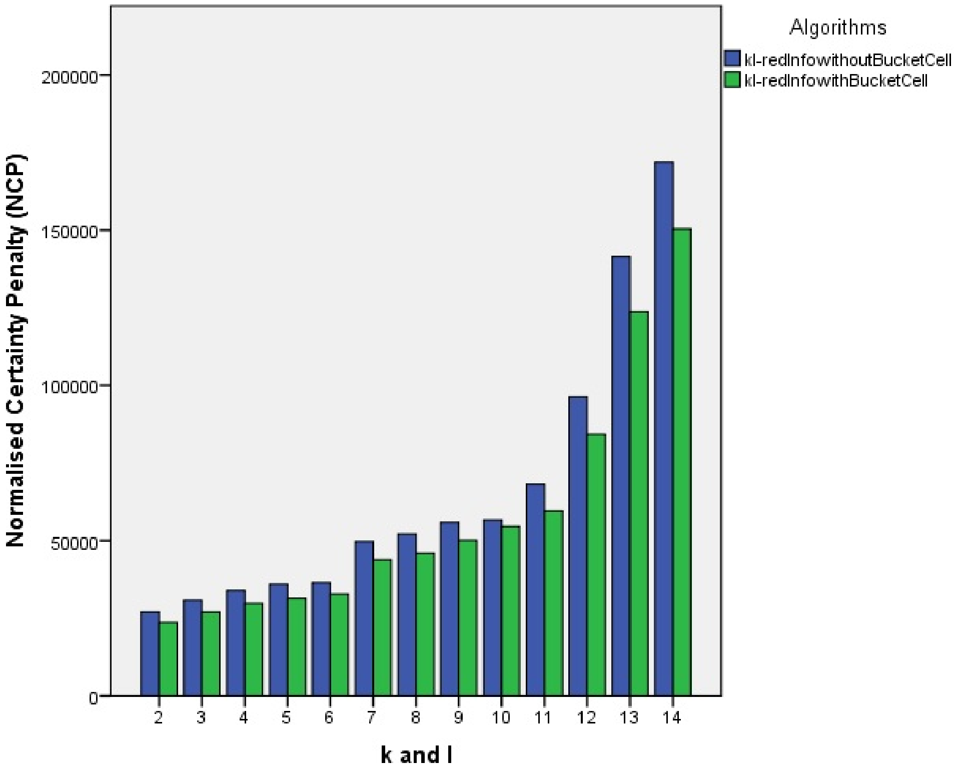


Figure E.2: NCP resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the Adult dataset

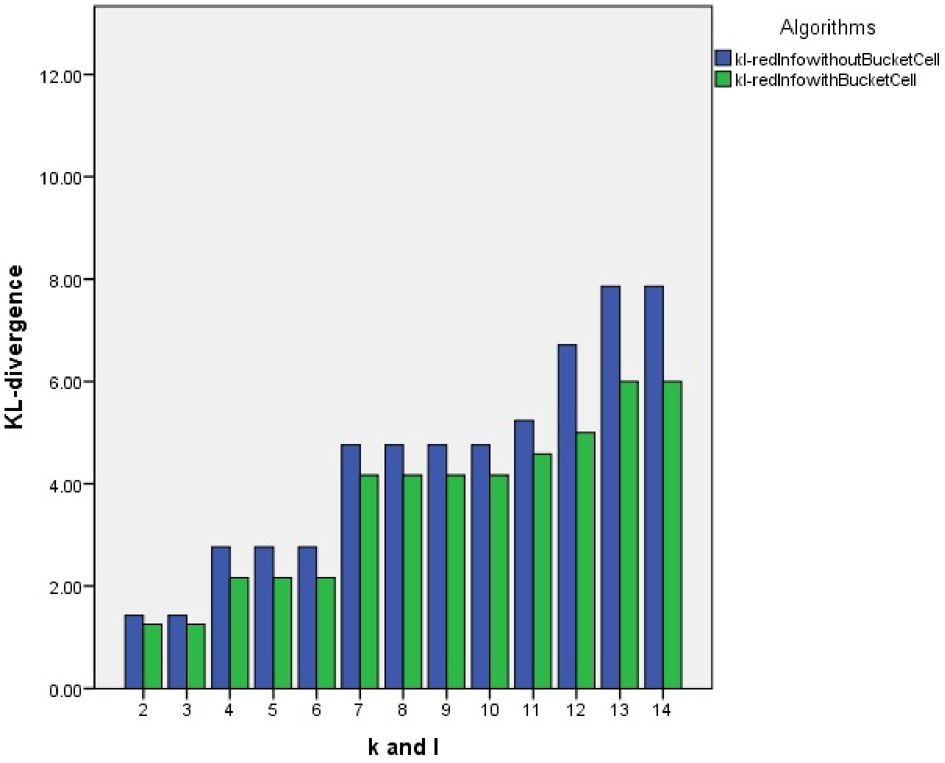
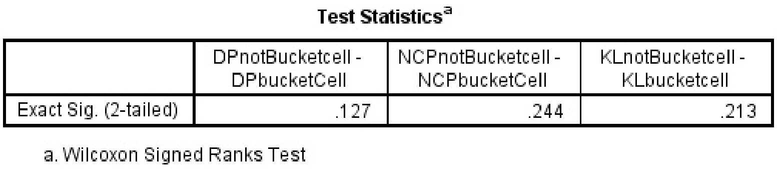


Figure E.3: KL-divergence resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the Adult dataset

Table E.1: The values of the information loss metrics resulting from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-generalisation approaches on Adult dataset

k,l	Information loss Metrics								
	DP			**NCP**			**KL**		
	kl-cell	*kl-BucketCell*	**Diff (%)**	*kl-cell*	*kl-BucketCell*	**Diff (%)**	*kl-cell*	*kl-BucketCell*	**Diff (%)**
2	107812	96463	11349(11)	27020	23643	3377(12)	1.429	1.25	0.179(13)
3	123126	110165	12961(11)	30859	27001	3858(13)	1.429	1.25	0.179(13)
4	125567	121297	4270(3)	33977	29730	4247(12)	2.762	2.467	0.295(11)
5	133439	128340	5099(4)	35950	31456	4494(13)	2.762	2.467	0.295(11)
6	139429	133700	5729(4)	36451	32770	3681(10)	2.762	2.567	0.195(7)
7	164384	158133	6251(4)	49693	43856	5837(12)	4.762	4.167	0.595(12)
8	173928	166673	7255(4)	52085	45949	6136 (12)	4.762	4.167	0.595(12)
9	182998	173735	9263(5)	55864	50131	5733(10)	4.762	4.167	0.595(12)
10	226204	202393	23811(11)	56693	54606	2087(4)	4.762	4.167	0.595(12)
11	271861	243244	28617(11)	68136	59619	8517(13)	5.238	4.983	0.255(5)
12	384153	343716	40437(11)	96279	94244	2035(2)	6.714	6.5	0.214(3)
13	464361	404954	59407(13)	141444	133763	7681(5)	7.857	7.468	0.389(5)
14	485946	431741	54205(11)	171916	150427	21489(12)	7.857	7.468	0.389(5)
		Average	8%			10%			9%

Table E.2: The comparison results from the application of the *kl-redInfo* with generalisation and with both bucketisation and cell-based generalisation on the Adult dataset



Test Statistics[a]

	DPnotBucketcell - DPbucketCell	NCPnotBucketcell - NCPbucketCell	KLnotBucketcell - KLbucketcell
Exact Sig. (2-tailed)	.127	.244	.213

a. Wilcoxon Signed Ranks Test

Appendix F

The information loss metrics of the *kl-redInfo* without and with sorting approach on Adult dataset

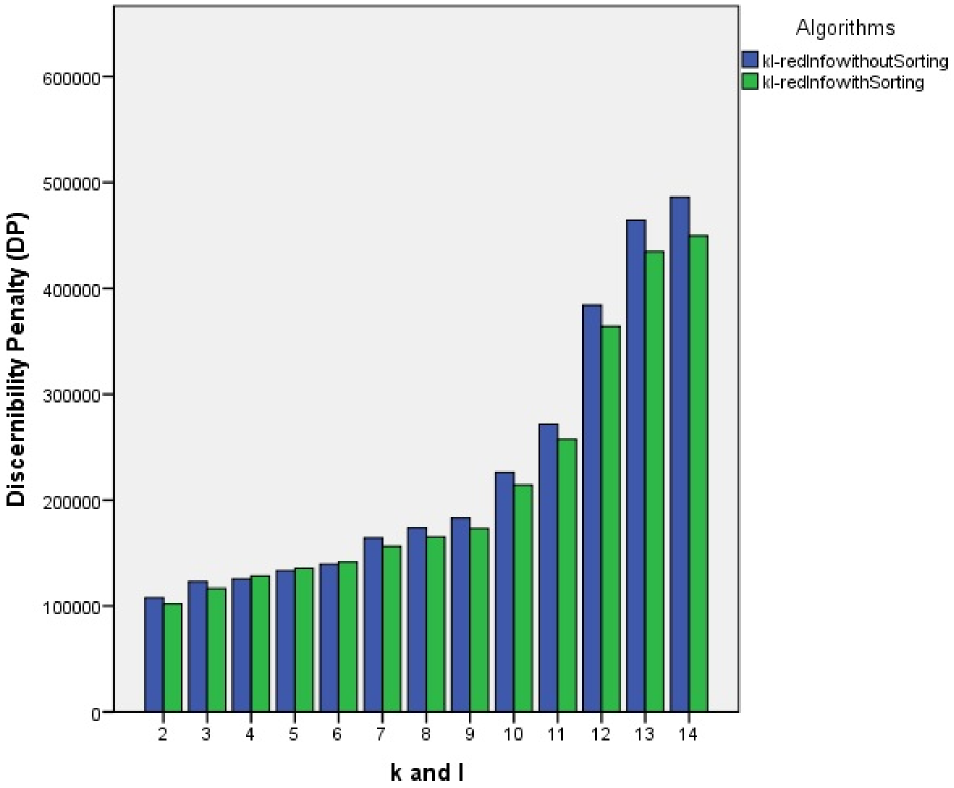


Figure F.1: DP resulting from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset

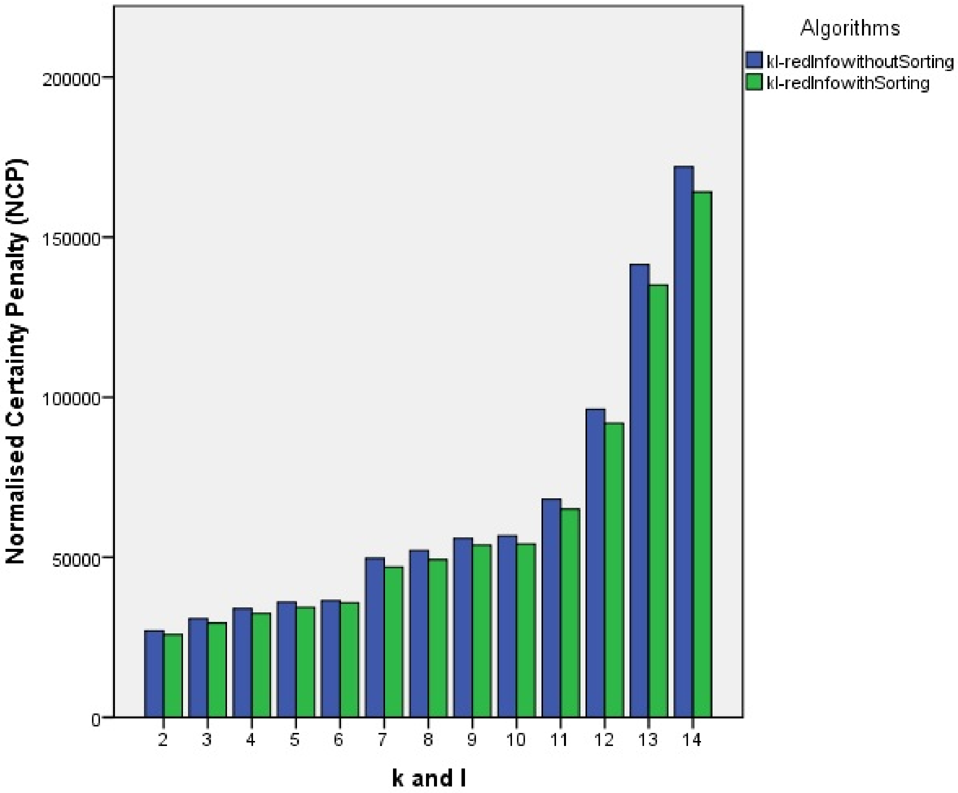


Figure F.2: NCP resulting from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset

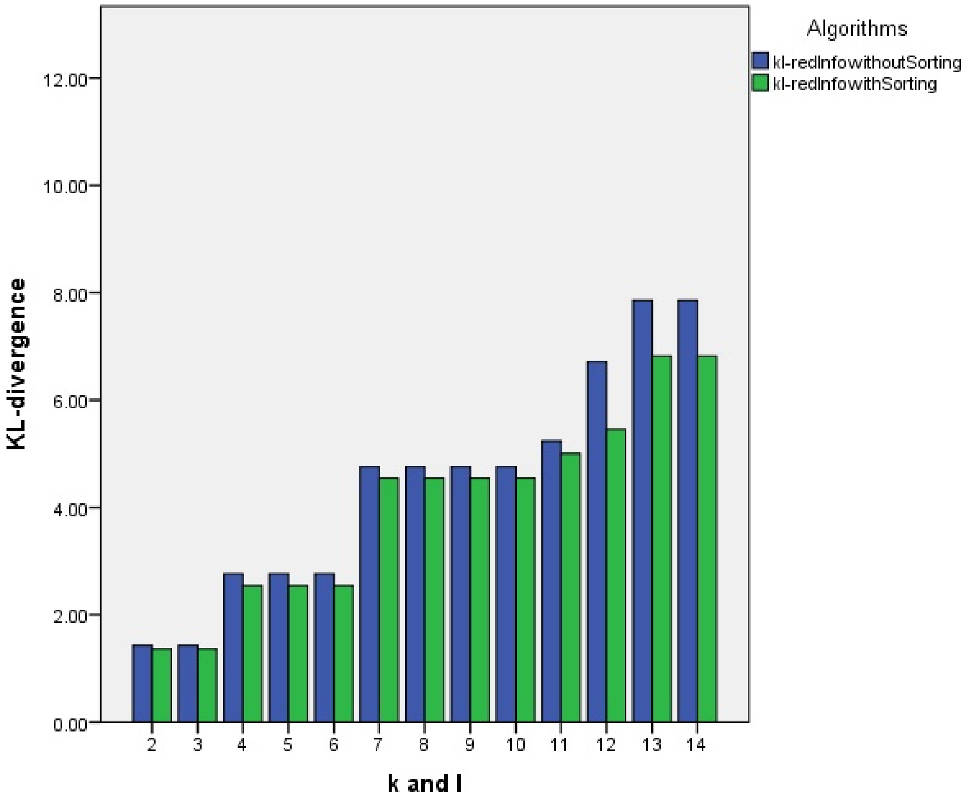
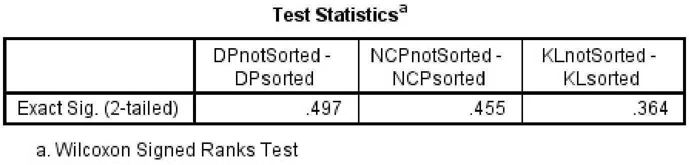


Figure F.3: KL-divergence resulting from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset

Table F.1: **The values of the information loss metrics resulting from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset**

k,l	Information loss Metrics								
	DP			NCP			KL		
	kl-NoSort	*kl-sort*	**Diff (%)**	*kl-NoSort*	*kl-sort*	**Diff (%)**	*kl-NoSort*	*kl-sort*	**Diff (%)**
2	107812	102137	5675(5)	27020	25792	1228(5)	1.429	1.364	0.065(5)
3	123126	116645	6481(5)	30859	29456	1403(5)	1.429	1.364	0.065(5)
4	128567	125432	3135(2)	33977	32432	1545(5)	2.762	2.545	0.217(8)
5	135439	133889	1550(1)	35950	34315	1635(5)	2.762	2.545	0.217(8)
6	149429	141565	7864(5)	36451	35749	702(2)	2.762	2.545	0.217(8)
7	164384	156259	8125(5)	49693	46934	2759(6)	4.762	4.545	0.217(5)
8	173928	165300	8628(5)	52085	49217	2868(6)	4.762	4.545	0.217(5)
9	182998	173367	9631(5)	55864	53780	2084(4)	4.762	4.545	0.217(5)
10	226204	214299	11905(5)	56693	54116	2577(5)	4.762	4.545	0.217(5)
11	271861	257553	14308(5)	68136	65039	3097(5)	5.238	5	0.238(5)
12	384153	363935	20218(5)	96279	91903	4376(5)	6.714	5.455	1.259(19)
13	464361	434658	29703(6)	141444	135015	6429(5)	7.857	6.818	1.039(13)
14	485946	449843	36103(7)	171916	164102	7814(5)	7.857	6.818	1.039(13)
		Average	5%			4%			8%

Table F.2: **The comparison results from the application of the *kl-redInfo* without and with sorting approach on the Adult dataset**



Test Statistics[a]

	DPnotSorted - DPsorted	NCPnotSorted - NCPsorted	KLnotSorted - KLsorted
Exact Sig. (2-tailed)	.497	.455	.364

a. Wilcoxon Signed Ranks Test

Appendix G

The information loss metrics of the *kl-redInfo* without and with the proposed modifications on Adult dataset

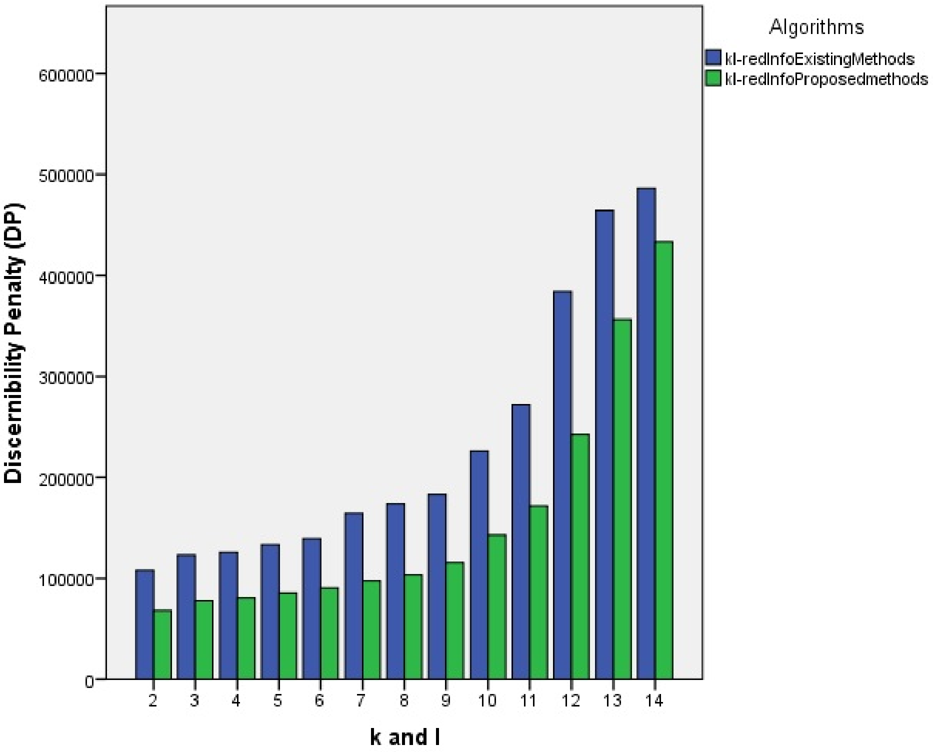


Figure G.1: DP resulting from the application of the *kl-redInfo* without and with the proposed modifications on the Adult dataset

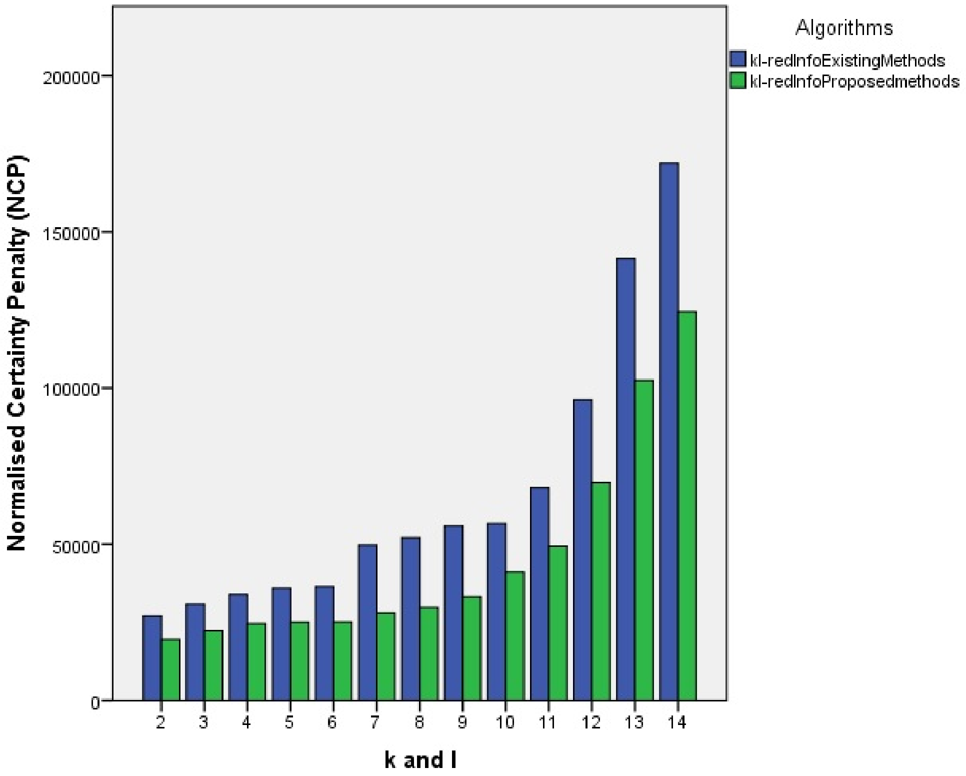


Figure G.2: NCP resulting from the application of the *kl-redInfo* without and with the proposed modifications on the Adult dataset

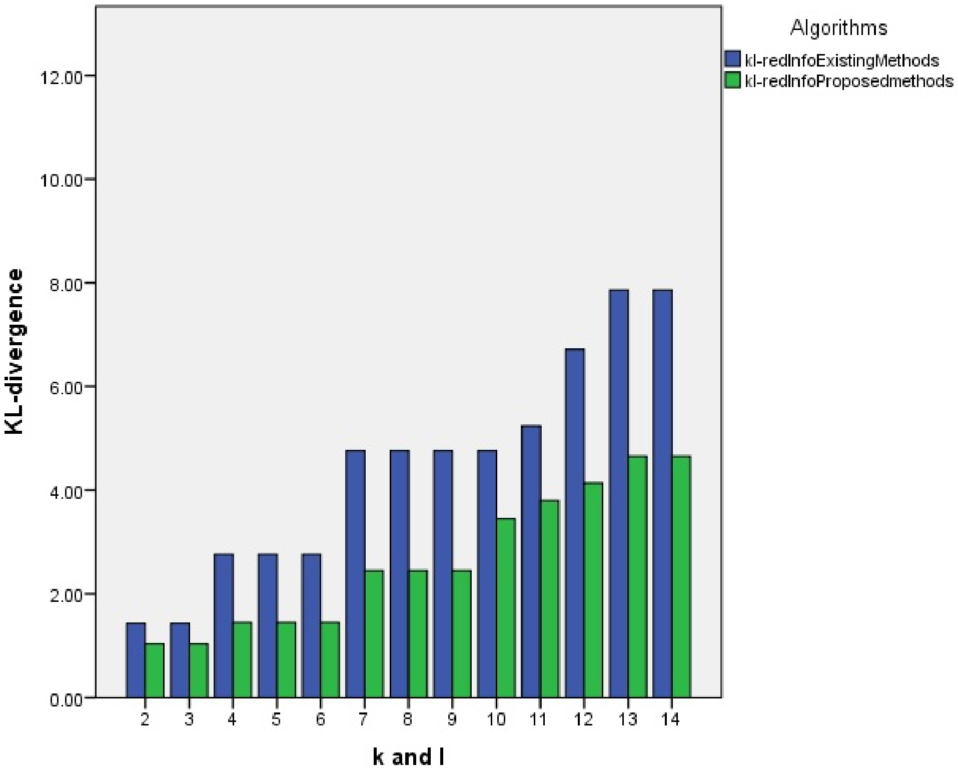
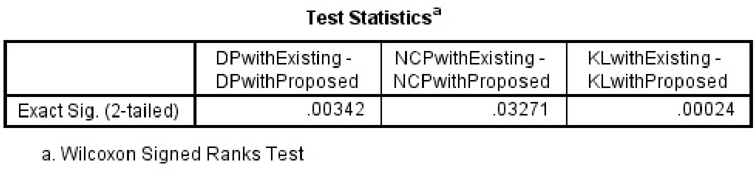


Figure G.3: KL-divergence resulting from the application of the *kl-redInfo* without and with the proposed modifications on the Adult dataset

Table G.1: The values of the information loss metrics resulting from the application of the *kl-redInfo* without and with the proposed modifications on the Adult dataset

k,l	Information loss Metrics								
	DP			**NCP**			**KL**		
	kl-without	*kl-with*	**Diff (%)**	*kl-without*	*kl-with*	**Diff (%)**	*kl-without*	*kl-with*	**Diff (%)**
2	107812	68092	39720(37)	27020	19567	7453(28)	1.429	1.034	0.395(28)
3	123126	77764	45362(37)	30859	22346	8513(28)	1.429	1.034	0.395(28)
4	125567	80621	44946(36)	33977	24604	9373(28)	2.762	1.448	1.314(48)
5	133439	85593	47846(36)	35950	25032	10918(30)	2.762	1.448	1.314(48)
6	139429	90376	49053(35)	36451	25120	11331(31)	2.762	1.448	1.314(48)
7	164384	97506	66878(41)	49693	28019	21674(44)	4.762	2.448	2.314(49)
8	173928	103534	70394(40)	52085	29751	22334(43)	4.762	2.448	2.314(49)
9	182998	115578	67420(37)	55864	33212	22652(41)	4.762	2.448	2.314(49)
10	226204	142866	83338(37)	56693	41053	15640(28)	4.762	3.448	1.314(28)
11	271861	171702	100159(37)	68136	49340	18796(28)	5.238	3.798	1.44(27)
12	384153	242623	141530(37)	96279	69719	26560(28)	6.714	4.138	2.576(38)
13	464361	356438	107923(23)	141444	102425	39019(28)	7.857	4.652	3.205(41)
14	485946	433229	52717(11)	171916	124491	47425(28)	7.857	4.652	3.205(41)
		Average	34%			31%			40%

Table G.2: The comparison results from the application of the *kl-redInfo* algorithm without and with the proposed modifications on the Adult dataset



Test Statistics[a]

	DPwithExisting - DPwithProposed	NCPwithExisting - NCPwithProposed	KLwithExisting - KLwithProposed
Exact Sig. (2-tailed)	.00342	.03271	.00024

a. Wilcoxon Signed Ranks Test

Appendix H

The information loss metrics of the *kl-redInfo* with proposed modifications, *l-mondrian*, and *g-anatomy*

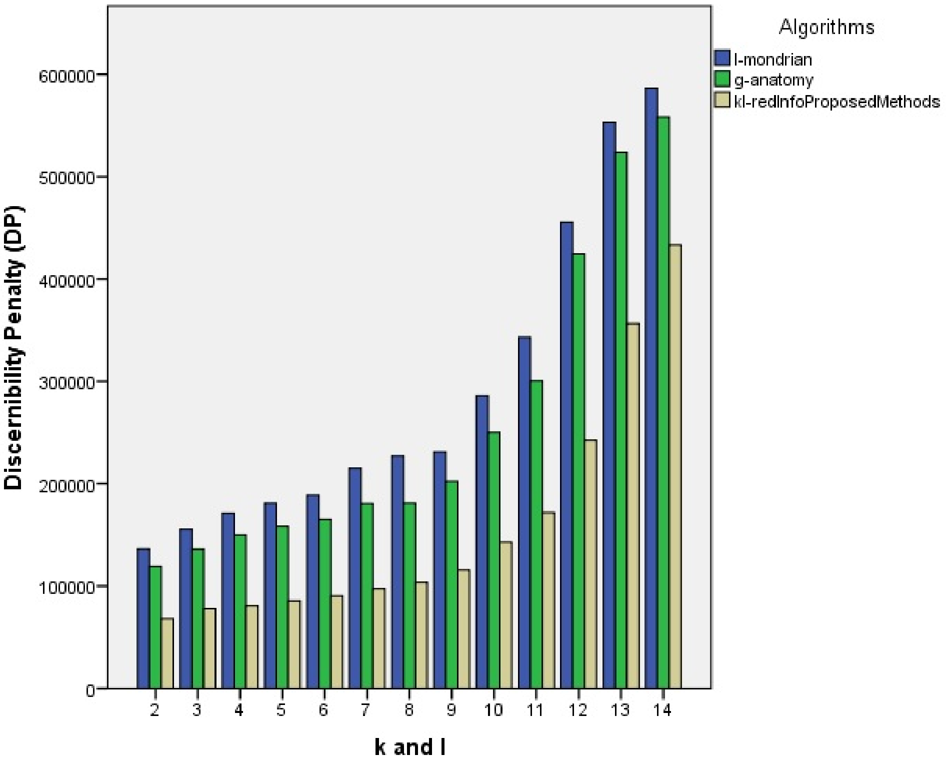


Figure H.1: DP resulting from the application of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modifications on the Adult dataset

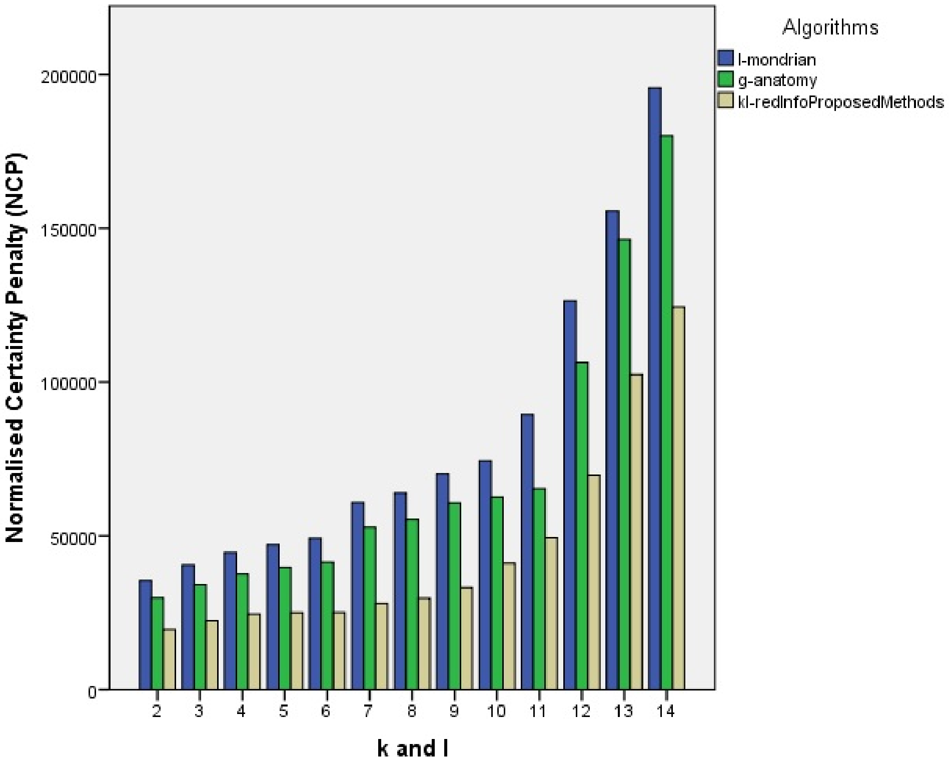


Figure H.2: NCP resulting from the application of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modifications on the Adult dataset

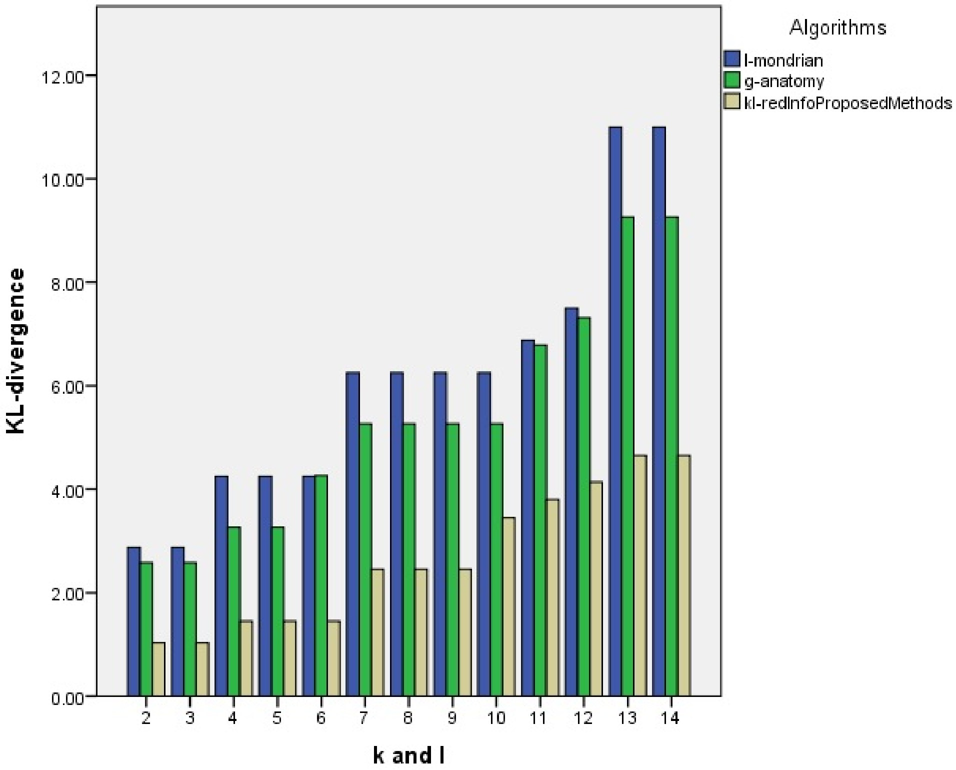
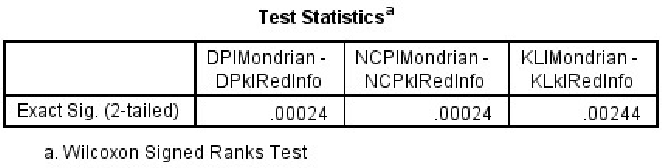


Figure H.3: KL-divergence resulting from the application of the *l-mondrian*, *g-anatomy*, and the *kl-redInfo* algorithm with the proposed modifications on the Adult dataset

Table H.1: The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modifications and *l-mondrian* on the Adult dataset

k,l	Information loss Metrics								
	DP			NCP			KL		
	l-mondrian	*kl-redInfo*	Diff (%)	*l-mondrian*	*kl-redInfo*	Diff (%)	*l-mondrian*	*kl-redInfo*	Diff (%)
2	136183	68092	68091(50)	35464	19567	15897(45)	1.875	1.034	0.841(45)
3	155527	77764	77763(50)	40502	22346	18156(45)	1.875	1.034	0.841(45)
4	171242	100621	70621(41)	44594	24604	19990(45)	3.25	1.448	1.802(55)
5	181186	105593	75593(42)	47184	25032	22152(47)	3.25	1.448	1.802(55)
6	188753	116376	72377(38)	49154	29120	20034(41)	3.25	1.448	2.802(55)
7	215012	127506	87506(41)	60784	36019	24765(41)	5.25	2.448	2.802(53)
8	227067	129534	97533(43)	63924	39751	24173(38)	5.25	2.448	2.802(53)
9	231156	135578	95578(41)	70197	43212	26985(38)	5.25	2.448	2.802(53)
10	285732	142866	142866(50)	74409	51053	23356(31)	5.25	3.448	1.802(53)
11	343044	171702	171342(50)	89428	64340	25088(28)	5.875	3.798	2.077(34)
12	455246	242623	212623(47)	126366	79719	46647(37)	7.5	4.138	3.362(35)
13	552877	356438	196439(36)	155645	102425	53220(34)	11	4.652	6.348(58)
14	586458	433229	153229(26)	195640	124491	71149(36)	11	4.652	6.348(58)
		Average	43%			39%			50%

Table H.2: The comparison results from the application of the *kl-redInfo* algorithm with the proposed modifications and *l-mondrian* on the Adult dataset



Test Statistics[a]

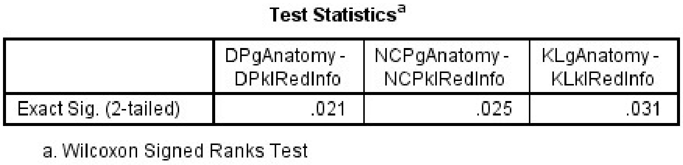
	DPlMondrian - DPklRedInfo	NCPlMondrian - NCPklRedInfo	KLlMondrian - KLklRedInfo
Exact Sig. (2-tailed)	.00024	.00024	.00244

a. Wilcoxon Signed Ranks Test

Table H.3: The values of the information loss metrics resulting from the application of the *kl-redInfo* with the proposed modifications and *g-anatomy* on the Adult dataset

k,l	Information loss Metrics								
	DP			NCP			KL		
	g-anatomy	*kl-redInfo*	Diff (%)	*g-anatomy*	*kl-redInfo*	Diff (%)	*g-anatomy*	*kl-redInfo*	Diff (%)
2	119160	68092	51068(43)	29865	19567	10298(34)	1.579	1.034	1.545(35)
3	136086	77764	58322(43)	34107	22346	11761(34)	1.579	1.034	1.545(35)
4	149837	100621	49216(33)	37553	24604	12949(34)	3.263	1.448	1.815(56)
5	158537	105593	52944(33)	39734	25032	14702(37)	3.263	1.448	1.815(56)
6	165159	116376	48783(30)	41393	29120	12273(30)	3.263	1.448	1.815(56)
7	180635	127506	53129(29)	52766	36019	16747(32)	5.263	2.448	2.815(53)
8	181184	129534	51650(29)	55409	39751	15658(28)	5.263	2.448	2.815(53)
9	202261	135578	66683(33)	60692	43212	17480(29)	5.263	2.448	2.815(53)
10	2590015	142866	107149(43)	62661	51053	11608(19)	5.263	3.448	1.815(34)
11	300478	171702	128776(43)	65308	64340	968(1)	5.789	3.798	1.991(34)
12	424591	242623	181968(43)	106414	79719	26695(25)	7.316	4.138	3.178(43)
13	523767	356438	167329(32)	146333	102425	43908(30)	8.263	4.652	3.611(44)
14	558150	433229	124921(22)	180013	124491	55522(31)	8.263	4.652	3.611(44)
		Average	35%			28%			46%

Table H.4: The comparison results from the application of the *kl-redInfo* algorithm with the proposed modifications and *g-anatomy* on the Adult dataset



Test Statistics[a]

	DPgAnatomy - DPklRedInfo	NCPgAnatomy - NCPklRedInfo	KLgAnatomy - KLklRedInfo
Exact Sig. (2-tailed)	.021	.025	.031

a. Wilcoxon Signed Ranks Test

Appendix I

The information loss metrics of the *kl-redInfo* algorithm, on the different size of the Adult dataset

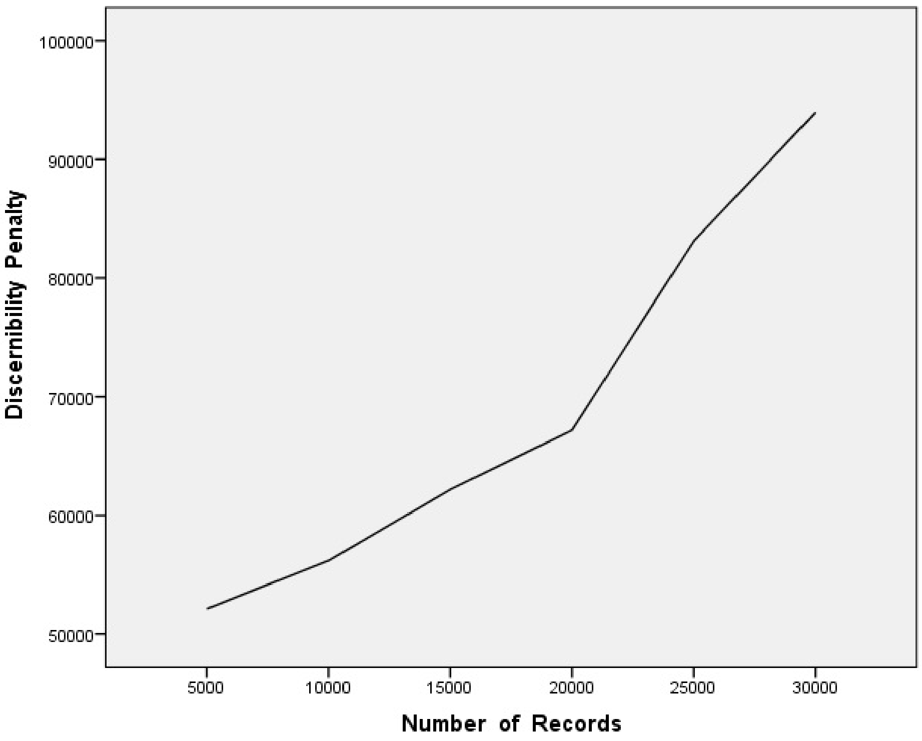


Figure I.1: DP resulting from the application of the *kl-redInfo* algorithm on different size of the Adult dataset when $k = l = 6$

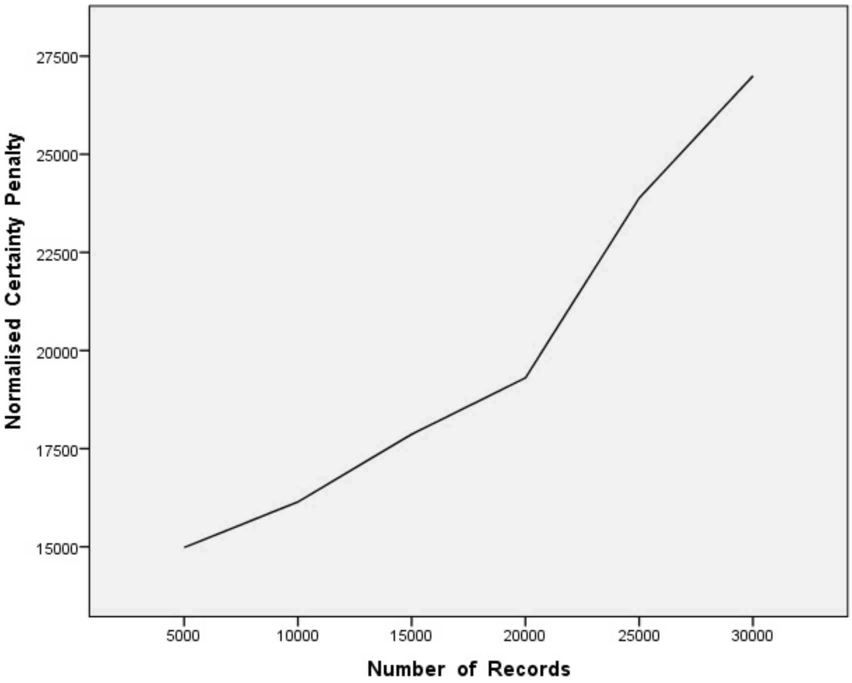


Figure I.2: NCP resulting from the application of the *kl-redInfo* algorithm on different size of the Adult dataset when $k = l = 6$

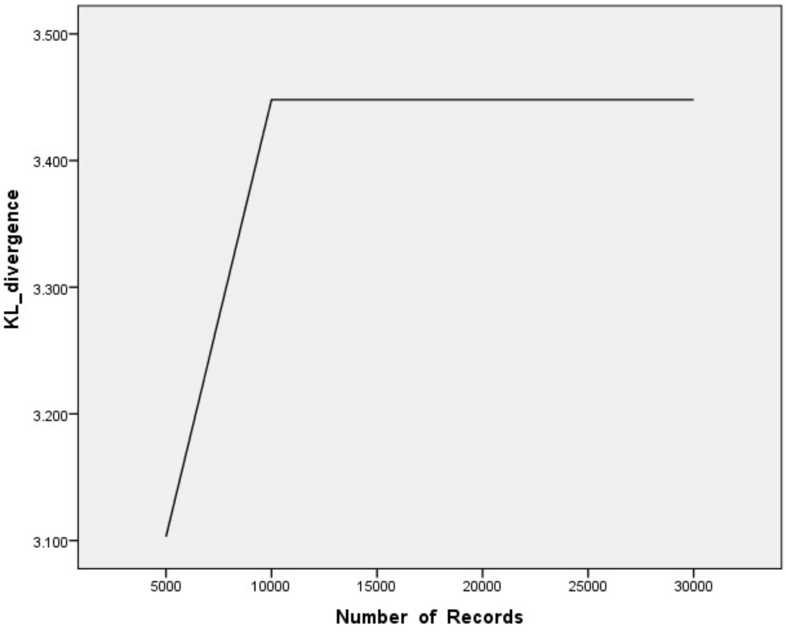


Figure I.3: KL-divergence resulting from the application of the *kl-redInfo* algorithm on different size of the Adult dataset when $k = l = 6$