Privacy Preserving Data Sharing in Data Mining Environment

PH.D DISSERTATION

BY

SUN, XIAOXUN

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF SOUTHERN QUEENSLAND IN FULLFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

PRINCIPAL SUPERVISOR: DR. HUA WANG
ASSOCIATE SUPERVISOR: DR. ASHLEY PLANK

JUNE, 2010
DEDICATION

Dedicated to my parents Jianlu Sun and Yanping Liu

and

my beloved wife Min Li
STATEMENT

I hereby declare that the work presented in this dissertation is in my own and is, to the best of my knowledge and belief, original except as acknowledgement in the text. It has not previously been submitted either in whole or in part for a degree at this or any other university.

Xiaoxun Sun

Signature of Candidate Date

ENDORSEMENT

Signature of Supervisor Date
ACKNOWLEDGEMENT

This dissertation would not be possible without the support and help from many professors, friends, and my family members over many years.

First, I would like to thank my advisor Dr. Hua Wang. I feel very fortunate to have such a great advisor for my Ph.D study. Thank you for your patience, insightful suggestions, financial support and unending encouragement during my Ph.D research. Additionally, I would like to thank Dr. Ashley Plank, for your guidance and suggestions to my research. I would also like to thank Dr. Jiuyong Li from University of South Australia for your valuable feedback and comments on my research.

I would like to thank Dr. Karsten Schulz from SAP Research Brisbane. It has been an honor to work with you, and I have learned so much from our collaborations. The time I spent as an intern at SAP Research Brisbane in 2009 will always be a precious memory in my life.

I sincerely thank the Centre for Systems Biology (CSBi), Department of Mathematics & Computing, Faculty of Science and Research and Higher Degree office of The University of Southern Queensland for providing the excellent study environment and financial support. It is a great pleasure to study at the Department of Mathematics & Computing.

I also acknowledge Dr. Henk Huijser from Learning and Teaching Support Unit at University of Southern Queensland for his help on proof-reading the dissertation.

Last, but not the least, I would like to give my special thanks to my parents Jianlu Sun and Yanping Liu, and my beloved wife Min Li, for their continued support and encouragement to me.
Abstract

Numerous organizations collect and distribute non-aggregate personal data for a variety of different purposes, including demographic and public health research. In these situations, the data distributor is often faced with a quandary: on one hand, it is important to protect the anonymity and personal information of individuals. While on the other hand, it is also important to preserve the utility of the data for research.

This thesis presents an extensive study of this problem. We focus primarily on notions of anonymity that are defined with respect to individual identity, or with respect to the value of a sensitive attribute. We discuss the anonymization techniques over relational data and large survey rating data. For relational data, we propose a variety of techniques that use generalization (also called recoding) and microaggregation to produce a sanitized view, while preserving the utility of the input data. Specifically, we provide a new structure called “Privacy Hash Table”; propose three enhanced privacy models to limit the privacy leakage; we inject the purpose and trust into the data anonymization process to increase the utility of the anonymized data, and we enhance the microaggregation method by using concepts from Information Theory. For survey rating data, we investigate two important problems (satisfaction and publication problems) in anonymizing survey rating data. By utilizing the characteristics of sparseness and high dimensionality, we develop a slicing technique for satisfaction problems. By using graphical representation, we provide a comprehensive analysis of graphical modification strategies. For all the techniques developed in this thesis, we include a set of extensive evaluations to indicate that the techniques are possible to distribute high-quality data that respect several meaningful notions of privacy.
# Table of Contents

1  **Introduction**  
   1.1 Privacy Preserving Data Sharing ........................................ 10  
   1.2 Scope of The Research ..................................................... 13  
   1.2.1 Data model .............................................................. 14  
   1.2.2 Publishing model ....................................................... 14  
   1.2.3 Privacy model ........................................................... 15  
   1.2.4 Attack model ............................................................ 15  
   1.3 Contributions ............................................................... 16  
   1.4 Dissertation Outline ...................................................... 18  

2  **Privacy Hash Table**  
   2.1 Motivation ................................................................. 19  
   2.2 Preliminaries ............................................................... 20  
   2.2.1 $K$-Anonymity ......................................................... 20  
   2.2.2 Generalization Relationship .......................................... 23  
   2.2.3 Generalized table and minimal generalization ..................... 28  
   2.3 Privacy hash table ....................................................... 32  
   2.3.1 The hash-based algorithm ............................................ 34  
   2.4 Extended privacy hash table ............................................. 37  
   2.5 An example ................................................................. 42  
   2.6 Summary ................................................................. 45  

3  **Enhanced $k$-Anonymity Models** ........................................ 46  
   3.1 Motivation ................................................................. 46
3.2 Preliminaries .................................................. 48
3.3 New Privacy Protection Models .......................... 50
3.4 NP-Hardness .................................................. 53
3.5 Utility Measurements ....................................... 55
3.6 The Anonymization Algorithms ......................... 57
3.7 Proof-of-concept Experiments .......................... 59
3.7.1 First Set of Experiments ............................... 60
3.7.2 Second Set of Experiments ......................... 64
3.8 Summary ...................................................... 66

4 INJECTING PURPOSE AND TRUST INTO DATA ANONYMISATION 67
4.1 Motivation ....................................................... 67
4.2 Attribute priority ........................................... 70
4.2.1 Mutual information measure ......................... 71
4.3 Degree of data anonymisation .......................... 78
4.3.1 Data anonymisation model ............................. 79
4.3.2 Degree of data anonymisation ....................... 80
4.4 The decomposition algorithm ............................ 83
4.5 Proof-of-concept experiments .......................... 87
4.5.1 Experiment Setup ....................................... 87
4.5.2 First set of experiments ............................... 89
4.5.3 Second set of experiments ......................... 91
4.6 Summary ...................................................... 94

5 PRIVACY PROTECTION THROUGH APPROXIMATE MICROAGGREGATION 95
5.1 Motivation ....................................................... 95

TABLE OF CONTENTS
5.2 Preliminary ................................................................. 97
  5.2.1 Microaggregation with its algorithms ...................... 98
5.3 Approximate Microaggregation ...................................... 102
  5.3.1 Dependency Tree ................................................. 102
  5.3.2 Application to $K$-Anonymity ................................. 107
5.4 Proof-of-concept Experiments ...................................... 108
  5.4.1 Experiment setup .................................................. 108
  5.4.2 Experimental results ............................................. 109
5.5 Summary ............................................................... 111

6 ANONYMIZING LARGE SURVEY RATING DATA ....................... 113
  6.1 Motivation ............................................................. 114
  6.2 Problem Definition .................................................. 117
    6.2.1 Background knowledge ...................................... 118
    6.2.2 New privacy principles .................................... 119
    6.2.3 Hamming groups .............................................. 122
  6.3 Publishing Anonymous Survey Rating Data ..................... 123
    6.3.1 Distortion Metrics ............................................ 123
    6.3.2 Graphical Representation .................................. 124
    6.3.3 Graphical modification .................................... 132
    6.3.4 Data modification ............................................ 135
  6.4 Proof-of-concept experiments .................................... 141
    6.4.1 Data sets ..................................................... 141
    6.4.2 Efficiency ..................................................... 141
    6.4.3 Data utility .................................................. 143
    6.4.4 Statistical properties ...................................... 144

TABLE OF CONTENTS
6.5 Summary ................................................................. 146

7 SATISFYING PRIVACY REQUIREMENTS IN SURVEY RATING DATA 147
7.1 Characteristics of \((k, \epsilon, l)\)-anonymity .................................................. 147
7.2 The Satisfaction algorithm .......................................................... 152
   7.2.1 Search by slicing ................................................................. 153
   7.2.2 To determine \(k\) and \(l\) when \(\epsilon\) is given ............................... 154
   7.2.3 To determine \(\epsilon\) and \(l\) when \(k\) is given ............................. 157
   7.2.4 To determine \(k\) and \(\epsilon\) when \(l\) is given ............................. 158
   7.2.5 Pruning and adjusting .......................................................... 160
7.3 Algorithm complexity ............................................................. 161
7.4 Experimental study ............................................................... 165
   7.4.1 Data sets ............................................................................ 166
   7.4.2 Efficiency ........................................................................... 166
   7.4.3 Space complexity ................................................................. 170
7.5 Summary ................................................................................. 170

8 DISCUSSION ............................................................................. 171
8.1 Summary of contributions .......................................................... 171
8.2 Related work ........................................................................... 174
   8.2.1 Policy-based Privacy Enforcement ........................................... 174
   8.2.2 Privacy-Preserving Data Mining .............................................. 175
   8.2.3 Macrodata/Microdata Protection ............................................. 176
8.3 Future work .............................................................................. 181
LIST OF FIGURES

2.1 Domain and value generalization hierarchies for Zip code, Age and Gender 25
2.2 The hierarchy of $DGH_{<G_0,Z_0>}$ .................................................. 27
2.3 Domain and value generalization strategies ................................. 29
2.4 Generalized table for $PT$ ................................................................. 30
2.5 Hierarchy $DGH_{<G_0,Z_0>}$ and corresponding lattice on distance vectors . 31
2.6 Extended domain generalization $EDGH_{<G_0,Z_0>}$ .......................... 40
2.7 Extended domain generalization $EDGH_{<G_0,Z_0>}$ with entropy ............... 41
2.8 $DGH$ and $VGH$ for Age and Zip of the example ............................ 43
2.9 The hierarchy of $DGH_{<A_0,Z_0>}$ ..................................................... 44
2.10 2-anonymous (2-diverse) data ....................................................... 44
2.10 One (extended) domain and value generalization strategy from Figure 2.9 44
3.1 Algorithm illustration for QI=$\{Zip\ Code\}$ ........................................ 58
3.2 Execution time vs. three privacy measures ........................................ 61
3.3 Distortion ratio vs. two enhanced privacy measures ............................. 62
3.4 Performance comparisons I ................................................................. 64
3.5 Performance comparisons II ............................................................... 65
4.1 The architecture of data anonymisation by injecting purposes and trust . . 69
4.2 Generalization hierarchy (taxonomy tree) for attributes Gender and Postcode 79
4.3 Correctness of the anonymisation degree decomposition ..................... 86
4.4 Performance of different methods with variant $t$ ............................... 90
4.5 Performance of different methods with variant $k$ ............................... 91
4.6 Performance vs. attribute priority ...................................................... 92
4.7 Performance vs. classification and predication accuracy I . . . . . . . . . . 93
4.8 Performance vs. classification and predication accuracy II . . . . . . . . 94

5.1 Example of microaggregation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99
5.2 The graph with its minimum spanning tree . . . . . . . . . . . . . . . . . . . .103
5.3 Proof of Theorem 5.1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .106
5.4 Running time comparison between different methods . . . . . . . . . . . . 110
5.5 Number of key attributes and information loss comparisons . . . . . . . . 111

6.1 Hardness proof of Problem 6.1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . .127
6.2 Two possible modifications of the rating data set $T$ with $k = 6, \epsilon = 1$ . . 127
6.3 An example of domino effects . . . . . . . . . . . . . . . . . . . . . . . . . . . . .129
6.4 Graphical representation example . . . . . . . . . . . . . . . . . . . . . . . . . . .129
6.5 Two possible 2-decompositions of $G_1$ . . . . . . . . . . . . . . . . . . . . . . . 130
6.6 A counter example . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .130
6.7 Merging and modification process for subcase 2.1 . . . . . . . . . . . . . . . 133
6.8 Borrowing nodes from other connected graphs . . . . . . . . . . . . . . . . . . 134
6.9 Combining two 2-cliques . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .135
6.10 The modification of graphical representation $G$ for Case 2.1.1 . . . . . . . 136
6.11 The modification of graphical representation $G$ for Case 2.1.2 . . . . . . . 137
6.12 The modification of graphical representation $G$ for Case 2.2.1 . . . . . . . 138
6.13 The modification of graphical representation $G$ for Case 2.2.2 . . . . . . . 139
6.14 Running time on MovieLens and Netflix data . . . . . . . . . . . . . . . . . . 142
6.15 Performance comparisons on MovieLens and Netflix data . . . . . . . . . . 143
6.16 Statistical properties analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . .145

7.1 The slicing technique . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .154
7.2 2-D illustration ................................................................. 162
7.3 Running time comparison I .................................................. 165
7.4 Running time comparison II .................................................. 165
7.5 Running time comparison III ............................................... 167
7.6 Running time comparison IV ............................................... 167
7.7 Space Complexity comparison I ......................................... 168
7.8 Space Complexity comparison II ........................................ 169
LIST OF TABLES

2.1 An example of microdata ........................................ 23
2.2 A 3-anonymous microdata ..................................... 23
2.3 An example of hash table ...................................... 33
2.4 Hash table with COUNT ....................................... 33
2.5 Hash table of generalization strategy 1 in Figure 2.3 .......... 36
2.6 External available information .................................. 37
2.7 Extended privacy hash table with sensitive attributes .......... 40
2.8 An example data set ............................................ 42
2.9 Hash table of generalization strategy in Figure 2.10 ........... 45

3.1 Raw microdata ................................................... 47
3.2 2-sensitive 4-anonymous microdata .............................. 48
3.3 Categories of Disease ........................................... 49
3.4 $2^+$-sensitive 4-anonymous microdata ......................... 50
3.5 $(3, 1)$-sensitive 4-anonymous microdata ...................... 51
3.6 Sample data ..................................................... 58
3.7 Features of QI attributes ....................................... 60
3.8 Attribute disclosures ............................................ 60
3.9 Categories of Income ............................................ 64

4.1 Sample data with global and local recoding .................... 71
4.2 Features of two real-world databases ............................ 88

5.1 Sample data ..................................................... 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>A raw microdata</td>
<td>99</td>
</tr>
<tr>
<td>5.3</td>
<td>A 2-anonymous microdata</td>
<td>99</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary of attributes in CENSUS</td>
<td>109</td>
</tr>
<tr>
<td>6.1</td>
<td>Sample survey rating data</td>
<td>115</td>
</tr>
<tr>
<td>6.2</td>
<td>Sample survey rating data (I)</td>
<td>124</td>
</tr>
<tr>
<td>6.3</td>
<td>Sample survey rating data (II)</td>
<td>124</td>
</tr>
<tr>
<td>7.1</td>
<td>Sample rating data</td>
<td>152</td>
</tr>
</tbody>
</table>