# Preserving Privacy in Data Publishing and Analysis

By

**Yidong Li**

Thesis submitted for the degree of

**Doctor of Philosophy**

In

**Computer Science**


THE UNIVERSITY *of* ADELAIDE

**The University of Adelaide, Australia**

**December 1st, 2010**

© Copyright by
Yidong Li
2010

*To Lan,*

*who made all of this possible*

*with her endless encouragement, patience and love.*

# Contents

# Abstract

As data collection and storage techniques being greatly improved, data analysis is becoming an increasingly important issue in many business and academic collaborations that enhances their productivity and competitiveness. Multiple techniques for data analysis, such as data mining, business intelligence, statistical analysis and predictive analytics, have been developed in different science, commerce and social science domains. To ensure quality data analysis, effective information sharing between organizations becomes a vital requirement in today's society. However, the shared data often contains person-specific and sensitive information like medical records. As more and more real-world datasets are released publicly, there is a growing concern about privacy breaches for the entities involved. To respond to this challenge, this thesis discusses the problem of eliminating privacy threats while, at the same time, preserving useful information in the released database for data analysis.

The first part of this thesis discuss the problem of privacy preservation on relational data. Due to the inherent drawbacks of applying equi-depth data swapping in distance-based data analysis, we study efficient swapping algorithms based on equi-width partitioning for relational data publishing. We develop effective methods for both univariate and multivariate data swapping. With extensive theoretical analysis and experimental validation, we show that, Equi-Width Swapping (EWS) can achieve a similar performance in privacy preservation to that of Equi-Depth Swapping (EDS) if the number of partitions is sufficiently large (e.g. $\geq \sqrt{n}$, where $n$ is the size of dataset). In addition, our analysis shows that the multivariate EWS algorithm has much lower computational complexity $O(n)$ than that of the multivariate EDS (which is $O(n^3)$ basically), while it still provides good protection for sensitive information.

The second part of this thesis focuses on solving the problem of privacy preservation on graphs, which has increasing significance as more and more real-world graphs modelling complex systems such as social networks are released publicly, . We point out that the real labels of a large portion of nodes can be easily re-identified with some weight-related attacks in a weighted graph, even the graph is perturbed with weight-independent invariants like degree. Two concrete attacks have been identified based on the following elementary weight invariants: 1) volume: the sum of adjacent weights for a vertex; and 2) histogram: the neighborhood weight distribution of a vertex. In order to protect a graph from these attacks, we formalize a general model for weighted graph anonymization and provide efficient methods with respect to a two-step framework including property anonymization and graph reconstruction. Moreover, we theoretically prove the histogram anonymization problem is NP-hard in the general case, and present an efficient heuristic algorithm for this problem running in near-quadratic time on graph size.

The final part of this thesis turns to exploring efficient privacy preserving techniques for hypergraphs, meanwhile, maintaining the quality of community detection. We first model a background knowledge attack based on so-called rank, which is one of the important properties of hyperedges. Then, we show empirically how high the disclosure risk is with the attack to breach the real-world data. We formalize a general model for rank-based hypergraph anonymization, and justify its hardness. As a solution, we extend the two-step framework for graph anonymization into our new problem and propose efficient algorithms that perform well on preserving data privacy. Also, we explore the issue of constructing a hypergraph with a specified rank set in the first place so far as we know. The proposed construction algorithm also has the characteristics of minimizing the bias of community detection on the original and the perturbed hypergraphs. In addition, we consider two de-anonymizing schemes that may be used to attack an anonymizied hypergraph and verify that both schemes fail in breaching the privacy of a hypergraph with rank anonymity in the real-world case.

# Statement of Originality

This work contains no material which has been accepted for the award of any other degree or diploma in any University or other tertiary institution to Yidong Li and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature:

Yidong Li

December 1st, 2010

# Acknowledgments

I am deeply grateful to my principal supervisor, Prof. *Hong Shen*, for guiding my research during the PhD candidateship since year 2007. From the discussion of my proposal to the last corrections of this document, he has been part of my research in all the stages. And it is from him that I learned critical thinking and skills in technical writing, which laid solid foundation for my future research. Furthermore, I would like to thank my co-supervisor, Dr. *Michael Sheng*, for his support on the three and a half years study at the University of Adelaide.

I owe great thanks to the School of Computer Science at the University of Adelaide for granting me scholarships, which make it possible for me to pursue the degree. Thank the staff in the institution for providing their selfless help and professional assistance in my work and life.

I would like to thank all the people who either directly or indirectly provide me knowledge, experience and support. I would like to thank my friends *Yingpeng Sang, Haibo Zhang and Yawen Chen* for sharing their experience in research. Especially, I would like to thank my colleagues: *Shihong Xu, Longkun Guo and Bo Chen*, who shared with me the office room and those stressful times when struggling on a research problem; and *Donglai Zhang*, who initiated the regular tea time in our group and shared his precious insights into technologies. I will always remember that it is these brilliant people that accompanied me intellectually and personally in these years.

I would like to thank my family for their support and love which make my life enjoyable even in those stressful times. I thank my parents for making it possible for me to receive high-level education, and my wife Lan Sun for supporting my research faithfully all the time. They have been and will always be the reasons that I strive for

excellence.

Last but not the least, I would like to express my gratitude to the anonymous examiners of this thesis. Review of a PhD thesis contains considerable amount of work which requires reading the full document and looking for problems and potential enhancements with a critical mind. I would like to thank them for their precious time and constructive comments on this thesis.

# Papers Published

[1] Yidong Li and Hong Shen: Anonymizing graphs against weight-based attacks. Workshop on The *10th IEEE International Conference on Data Mining (ICDMW'10)*. December 14-17, 2010. Sydney, Australia.

[2] Yidong Li and Hong Shen: On identity disclosure in weighted graphs. The *11th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'10)*, December 8-11, 2010, Wuhan, China.

[3] Yidong Li and Hong Shen: Multivariate Equi-width data swapping for private data publication. The *14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'10)*, pp 209-215. June 21-24, 2010. Hyderabad, India. .

[4] Yidong Li and Hong Shen: Equi-width data swapping for private data publication. The *10th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'09)*, pp. 231-238. December 8-11, 2009, Hiroshima, Japan.

# Papers under Review and in Preparation

[1] Yidong Li and Hong Shen: Privacy preserving on weighted graphs. Submitted to the *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

[2] Yidong Li and Hong Shen: Preventing identity disclosure with community preservation in hypergraphs. Submitted to the *SIAM International Conference on Data Mining (SDM'11)*.

[2] Yidong Li and Hong Shen: Anonymizing graphs against weight-based attacks: An extension. *Journal of Computing Science and Engineering (JCSE).* Invited paper.

[3] Yidong Li and Hong Shen: Hypergraph anonymization. Journal paper 16 pages (double columns), forthcoming.

[4] Yidong Li and Hong Shen: An attack to de-anonymize hypergraphs. Conference paper 10 pages (double columns), forthcoming.

# List of Figures

# List of Tables

# List of Algorithms