

# Towards Effective Spatial Data Mining: Uncertainty, Condensity and Privacy



Bozhong Liu

Faculty of Engineering and Information Technology  
University of Technology, Sydney

A thesis submitted for the degree of

*Doctor of Philosophy*

April 2017



## **Certificate of Original Authorship**

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Student: Bozhong Liu

Date: 04/21/2017



## Acknowledgements

Firstly, I would like to express my sincere gratitude to my principal supervisor, Dr. Ling Chen, for her continuous support of my Ph.D study and related research, for her patience, kindly support and inspiring motivation. She is always supportive when I feel frustrated or despairing. Her guidance helps me in all the time of my research life. Also, I am very thankful to Prof. Zhu, for his immense knowledge and wonderful advices. Without their guidance, my research life would be much more difficult. I am also grateful to my supervisor in Shanghai Jiao Tong University, Prof. Qiu, for his grate support and encouragement. Without his effort, I might have lost this precious opportunity of studying in UTS.

Secondly, I would like to thank my fellow students for their suggestion, discussion, cooperation and of course friendship. Their patience and support help me in overcoming numerous obstacles I have been facing through my research. Especially, I am grateful to Chunyang Liu, Meng Fang, Zhe Xu, Zhibin Hong and Shirui Pan for their kindness and sincerity. They not only help me a lot when I came to Sydney, but also provide many nice advices for my research work. They make my Ph.D study more colorful and wonderful.

Last but not the least, I would like to thank my parents for supporting me spiritually throughout writing this thesis and my life in general.



## Abstract

Spatial data mining (SDM) is a process of knowledge discovery that the observing data is related to geographical information. It has become an important data mining task due to the explosive growth and pervasive use of spatial data. It is more difficult to extract interesting and useful patterns from spatial datasets due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. Although existing methods can handle the spatial mining task properly, as the arrival of the big data era, new challenges for SDM are arising.

Firstly, traditional SDM methods usually focus on deterministic datasets, where spatial events occur affirmatively at precise locations. However, the inherent uncertainty of spatial data makes the mining process more difficult. Classical spatial data mining algorithms are no longer applicable or need delicate modification. Secondly, traditional SDM frameworks produce an exponential number of patterns, which makes it hard for users to understand or apply. To solve the condensity issue, novel techniques such as summarization or representation must be carefully investigated. Thirdly, spatial data usually involves an individual's location information, which incurs location privacy problem. It would be a challenge to protect location privacy with enhanced data security and improved resulting accuracy.

To address the uncertainty issue, we study the problem of discovering co-location patterns in the context of continuously distributed uncertain data, namely Probabilistic Co-location Patterns Mining (PCPM). We develop an effective probabilistic co-location mining framework integrated with optimization strategies to address the challenges.

To address the condensity issue, we investigate the problem of Representative Co-location Patterns Mining (RCPM). We define a new measure to quantify the distance between co-location patterns, and develop two efficient algorithms for summarization.

---

To address the privacy issue, we solve the problem of protecting Location Privacy in Spatial Crowdsourcing (LPSC). We propose a secure spatial crowdsourcing framework based on encryption, and devise a novel secure indexing technique for efficient querying.

The experimental results demonstrate the effectiveness and efficiency of our proposed solutions. The methods and techniques used in solving concrete SDM tasks can also be applied or extended to other SDM scenarios.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Mining Co-location Patterns from Uncertain Data</b>	<b>7</b>
2.1	Introduction . . . . .	8
2.2	Related Works . . . . .	10
2.3	Problem Definitions . . . . .	11
2.3.1	Co-location Patterns in Deterministic Data . . . . .	11
2.3.2	Co-location Patterns in Gaussian-based Data . . . . .	13
2.4	Probabilistic Participation Ratio Computation . . . . .	15
2.5	Probabilistic Co-location Mining Framework . . . . .	18
2.6	Finding Probabilistic Neighbors . . . . .	20
2.6.1	Minimum Bounding Sphere . . . . .	21
2.6.2	The filtering . . . . .	22
2.7	Performance Study . . . . .	23
2.7.1	Experiment Setup . . . . .	23
2.7.2	Comparisons with other methods . . . . .	24
2.7.3	Efficiency of Filtering . . . . .	26
2.7.4	Parameter Evaluation . . . . .	27
2.8	Conclusion . . . . .	28
<b>3</b>	<b>Summarizing Spatial Co-location Patterns</b>	<b>29</b>
3.1	Introduction . . . . .	30
3.2	Related Works . . . . .	33
3.3	Preliminary . . . . .	35
3.3.1	Co-location Patterns . . . . .	35
3.3.2	Co-location Distance Measure . . . . .	36

---

3.3.3	Problem Statement . . . . .	38
3.4	The <i>RCPFast</i> Algorithm . . . . .	40
3.5	The <i>RCPMS</i> Algorithm . . . . .	42
3.5.1	Optimization Strategy . . . . .	45
3.5.2	Approximation Strategy . . . . .	49
3.5.3	The <i>gen_cover_set()</i> Function . . . . .	51
3.6	Experimental Study . . . . .	52
3.6.1	Experiments on Synthetic Data . . . . .	52
3.6.2	Experiments on Real Data . . . . .	59
3.7	Conclusions . . . . .	62
<b>4</b>	<b>Protecting Location Privacy in Spatial Crowdsourcing</b>	<b>63</b>
4.1	Introduction . . . . .	64
4.2	Related Works . . . . .	66
4.2.1	Location Privacy . . . . .	66
4.2.2	Secure Index . . . . .	67
4.2.3	Task Assignment in SC . . . . .	68
4.3	Preliminary . . . . .	69
4.3.1	Spatial Crowdsourcing Model . . . . .	69
4.3.2	Threat Model . . . . .	70
4.3.3	Paillier Cryptosystem . . . . .	70
4.4	The HESI Framework . . . . .	71
4.4.1	The Dual-Server Architecture . . . . .	71
4.4.2	The System Workflow . . . . .	72
4.5	Secure Distance Computation . . . . .	75
4.6	Secure Indexing . . . . .	78
4.6.1	SKD-tree . . . . .	78
4.6.2	Fast Pruning . . . . .	82
4.7	Secure Task Assignment . . . . .	86
4.7.1	Assignment Strategy . . . . .	86
4.7.2	Secure Assignment . . . . .	89
4.8	Analysis . . . . .	90
4.8.1	Security Analysis . . . . .	91
4.8.2	Complexity Analysis . . . . .	93

## Contents

---

4.9	Performance Evaluation . . . . .	95
4.9.1	Benchmark Data . . . . .	95
4.9.2	Experimental Results . . . . .	96
4.10	Conclusions . . . . .	101
<b>5</b>	<b>Conclusion</b>	<b>103</b>
	<b>References</b>	<b>105</b>



# List of Tables

1.1	Relationships among non-spatial and spatial data [1]. . . . .	2
2.1	<i>Determinization</i> method vs. our method on EPA data. . . . .	26
2.2	<i>Discretization</i> method vs. our method on ITF data. . . . .	27
3.1	Prevalent patterns in the example. . . . .	32
3.2	Parameters used in synthetic data generation. . . . .	52
4.1	The outline of the secure protocols. . . . .	75
4.2	Potential assignments for each task. . . . .	88
4.3	Complexity summary. . . . .	94
4.4	Performance of distance computation for different location distribution. . . . .	99
4.5	Communication cost. . . . .	100
4.6	Task assignment evaluation on Yelp. . . . .	101
4.7	Task assignment evaluation on Gowalla. . . . .	101



# List of Figures

2.1	An example of deterministic spatial data. . . . .	12
2.2	An example of an uncertain spatial data. . . . .	16
2.3	Using Minimum Bounding Spheres to bound $\rho$ -regions. . . . .	22
2.4	Evaluation of filtering technique. . . . .	27
2.5	Parameter evaluation. . . . .	28
3.1	A motivating example. . . . .	31
3.2	An example illustrating <i>RCPFast</i> algorithm. . . . .	42
3.3	An illustration of the optimization strategy based on Theorem 3.3. . . . .	48
3.4	Examples of the approximation strategy. . . . .	49
3.5	Compression rate tests on synthetic data sets. . . . .	54
3.6	Framework comparison on synthetic data sets. . . . .	56
3.7	Performance tests with <i>minpi</i> and $\epsilon$ on synthetic data sets. . . . .	58
3.8	Co-location distance computation analysis on synthetic data sets. . . . .	59
3.9	Compression rate differences between <i>RCPMS</i> and <i>RCPFast</i> on synthetic data sets. . . . .	60
3.10	Compression rate tests on EPA and POI data sets. . . . .	60
3.11	Performance on EPA and POI data sets. . . . .	61
4.1	The HESI framework. . . . .	73
4.2	A small spatial dataset. . . . .	76
4.3	An example of a normal KD-tree <i>vs.</i> an SKD-tree with reference to worker locations in Figure 4.2. . . . .	80
4.4	Evaluation of tree construction. . . . .	96

4.5	Tree operation evaluation. . . . .	97
4.6	Overall Performance. . . . .	98
4.7	Performance Improvement. . . . .	99