

Faculty of Engineering and Information Technology
University of Technology, Sydney

**Nonparametric Bayesian Models for
Learning Network Coupling
Relationships**

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Xuhui Fan

May 2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

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.

Signature of Candidate

*To my parents and my wife
for your love and support*

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisors Prof. Longbing Cao for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I also would like to appreciate my co-supervisor Dr. Richard Yi Da Xu, for leading into the wonderful world of nonparametric Bayesian methods and for the sleepless nights we were working together before the deadlines. Without his professional guidance and persistent help, this thesis would not have been possible.

I thank my fellow labmates in Advanced Analytics Institute: Lin Zhu, Xiaodong Yue, Yin Song, Wei Wei, Can Wang, Zhong She, etc., for our stimulating discussions and for all the fun we have had in the last four years.

Also, I would like to thank the magic of machine learning. This fantastic world has made me lucky enough to have the fun while exploring it during all of my Ph.D period.

Last but not the least, I would like to thank my family: my wife, my parents and my parents in law, for their unconditional support, both financially and emotionally throughout the whole PhD studying.

Xuhui Fan

December 2014 @ UTS

Contents

Certificate	i
Acknowledgment	v
List of Figures	xiii
List of Tables	xv
List of Publications	xvii
Abstract	xix
Chapter 1 Introduction	1
1.1 Problem Statement	1
1.2 Research Methods	6
1.2.1 Relational Models	6
1.2.2 Nonparametric Bayesian Method	8
1.3 Aims and objectives	9
1.4 Research Challenges	9
1.4.1 Coupling relations between the mixed-membership dis- tributions across time	10
1.4.2 Coupling relations within the communities of networks	10
1.4.3 Coupling relations between the node information and mixed-membership distribution	11
1.5 Research Contributions	11
1.6 Thesis Structure	13
Chapter 2 Preliminaries and Literature Review	16
2.1 Preliminaries of Nonparametric Bayesian methods	16

2.1.1	Dirichlet Process	17
2.1.2	Hierarchical Dirichlet Process	20
2.1.3	Beta Process	23
2.1.4	Indian Buffet Process	24
2.2	Preliminaries of Markov Chain Monte Carlo (MCMC) methods	25
2.2.1	Metropolis-Hastings Sampling	26
2.2.2	Gibbs Sampling	26
2.3	Preliminaries & Literature Review of Relational Models	28
2.3.1	Latent Class model	28
2.3.2	Latent Feature Model	30
2.3.3	Literature Review of the relational models	30
2.4	Literature Review of Coupling Relations Learning	35
2.4.1	Limitation of the coupling Relational Learning	36
Chapter 3 Dynamic Infinite Mixed-Membership Stochastic		
Blockmodel 37		
3.1	Introduction	37
3.2	The dynamic infinite mixed-membership stochastic blockmodel (DIM3)	40
3.2.1	The general settings	40
3.2.2	The mixture time variant (MTV) Model	41
3.2.3	The mixture time invariant (MTI) Model	42
3.2.4	Discussion and comparison	44
3.3	Inference	44
3.3.1	Gibbs Sampling for the MTV model	44
3.3.2	Adapted Slice-Efficient Sampling for the MTV model .	48
3.3.3	Hyper-parameter Sampling	50
3.3.4	Gibbs Sampling for the MTI model	50
3.3.5	Inference discussions	52
3.4	Experiments	53
3.4.1	Synthetic Dataset	53
3.4.2	Real World Datasets Performance	62

3.4.3	Kapferer Tailor Shop	64
3.4.4	Sampson Monastery Dataset	65
3.4.5	Hypertext 2009 dynamic contact network	66
3.5	Summary	68
3.6	Limitation & future work	69

Chapter 4 Copula Mixed-Membership Stochastic Blockmodel for Intra-group Correlation 71

4.1	Introduction	71
4.2	Copula Model	73
4.3	Graphical Model Description	74
4.4	Inference & Further Discussion	78
4.4.1	Marginal Conditional on π only: cMMSB $^\pi$	79
4.4.2	Marginal Conditional on u and v only: cMMSB uv	82
4.4.3	Relations with Classical MMSB	85
4.4.4	Relations with Dependent Dirichlet Process	86
4.4.5	Computational Complexity Analysis	86
4.5	Experiments	86
4.5.1	Synthetic Data	87
4.5.2	Real-world Datasets for Link Prediction	92
4.6	Summary	97
4.7	Limitation & Future Work	97

Chapter 5 Learning Relational Models by Efficiently Involving Node Information in a Network 99

5.1	Introduction	99
5.2	Literature Review of Stick-breaking Process	102
5.3	Generative Model	102
5.3.1	Node-information Involved Mixed-Membership Model	103
5.3.2	Node-information Involved Latent Feature Model	106
5.4	Inference	107
5.4.1	Informative Mixed Membership Model	107

CONTENTS

5.4.2	Informative Latent Feature model	109
5.4.3	π_i -Collapsed Sampling for the niMM Model	112
5.4.4	Computational Complexity	112
5.5	Experiments	113
5.5.1	Performance on the Lazega Dataset	116
5.5.2	MIT Reality Mining	118
5.5.3	Convergence Behaviour	119
5.6	Summary	120
5.7	Limitations & Future Work	121
Chapter 6 Conclusions and Future Work		122
6.1	Conclusions	122
6.2	Future Work	124
6.2.1	Future Work on Large-scale Bayesian inference for non- parametric Bayesian methods	124
6.2.2	Future work from the relational models perspective	125
6.2.3	Future work on the inconsistent estimators of the com- ponent number	127
Appendix A Derivation equations for the model of MTV-g		128
A.1	Sample β	128
A.2	Sample $Z = \{s_{ij}^t, r_{ij}^t\}_{n \times n}^{1:T}$	128
A.3	Sampling \hat{m}	132
Appendix B Several fundamental distributions used in the thesis		133
B.1	Bernoulli distribution	133
B.2	Multinomial distribution	133
B.3	Beta distribution	134
B.4	Dirichlet distribution	134
B.5	Gamma distribution	134
Appendix C List of Symbols		135

Bibliography 138

List of Figures

1.1	The network structure.	3
1.2	The structure of this thesis	15
2.1	The mixed-membership stochastic blockmodel (MMSB) Model	29
2.2	The graphical model for latent feature model (LFM).	31
3.1	The mixture time variant (MTV) Model	41
3.2	The mixture time invariant (MTI) Model	43
3.3	Four Cases of the Compatibility Matrix.	52
3.4	Top: the training log-likelihood trace plot on the MTV-g model. Bottom: the training log-likelihood trace plot on the MTI-g model.	57
3.5	Log-likelihood Performance on all the four cases	57
3.6	Training Log-likelihood Performance (95% Confidence Interval = Mean \mp \times Standard Deviation)	60
3.7	AUC Performance (95% Confidence Interval = Mean \mp \times Standard Deviation)	61
3.8	Larger dataset's role-compatibility matrix	62
3.9	The MTI model's Performance on Kapferer Tailor Shop Dataset.	64
3.10	The nodes' mixed-membership distribution of the MTI model on Sampson Monastery Dataset (from Left to Right: Time 1-3.) Blue, Loyal Opposition; Red, Outcasts; Green, Young Turks; Magenta, interstitial group.	66
3.11	Role Compatibility Matrix (Left: MTV-g; Right: MTI)	66

LIST OF FIGURES

3.12	The MTI model's performance on the Hypertext 2009 dynamic contact network.	67
4.1	Graphical model of Copula MMSB	75
4.2	Mixed-membership Distribution	89
4.3	Role-compatibility Matrix	89
4.4	Comparison of Models' Posterior Predictive Distribution on the Training data.	91
5.1	The generative model for the niMM and niLF models.	103
5.2	Trace plot of the AUC value versus iteration time in the Lazega dataset	119
5.3	Trace plot of the AUC value versus iteration time the Reality dataset	119

List of Tables

1.1	Persons' Information	3
3.1	Integrated Autocorrelation Times Estimator $\hat{\tau}$ for K	55
3.2	Integrated Autocorrelation Times Estimator $\hat{\tau}$ for D	56
3.3	Average l_2 Distance to the Ground-truth	58
3.4	Running Time (Seconds per iteration)	59
3.5	Dataset Information	63
4.1	Computational Complexity for Different Models	87
4.2	Model Performance (Mean \mp Standard Deviation) on Synthetic Data of Full Correlation.	88
4.3	θ 's 95% Confidence Interval	91
4.4	Model Performance on NIPS Co-author dataset(Mean \mp Standard Deviation)	93
4.5	Model Performance on MIT Reality dataset(Mean \mp Standard Deviation)	94
4.6	Model Performance on Lazega dataset(Mean \mp Standard Deviation)	95
5.1	Computational Complexity for Different Models	113
5.2	Performance on Lazega Dataset (Mean \mp Standard Deviation)	114
5.3	Performance on Reality Dataset (Mean \mp Standard Deviation)	115
5.4	Attribute-community importance learning for $\boldsymbol{\eta}$	117

LIST OF TABLES

5.5 Mixing rate (Mean \mp Standard Deviation) for different models, with the bold type denoting the best ones within each row. 118

List of Publications

Papers Published

- **Xuhui Fan**, Longbing Cao, Richard Yi Da Xu. Dynamic Infinite Mixed-Membership Stochastic Blockmodel. Accepted in IEEE Transaction on Neural Networks and Learning Systems.
- **Xuhui Fan** and Longbing Cao. A Convergence Theorem for the Graph Shift-type Algorithms. Accept in Pattern Recognition.
- **Xuhui Fan**, Lin Zhu, Longbing Cao, Xia Cui and Yew Soon Ong. Maximum Margin Clustering on Evolutionary Data. Proceedings of the 21th ACM international conference on Information and knowledge management(ACM CIKM). October 2012.
- Wei Wei, **Xuhui Fan**, Jinyan Li and Longbing Cao. Model the Complex Dependence Structures of Financial Variables by Using Canonical Vine. Proceedings of the 21th ACM international Conference on Information and knowledge management(ACM CIKM). October 2012.
- **Xuhui Fan** and Longbing Cao. A theoretical framework of the Graph Shift Algorithm. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence(AAAI-12 poster). July 2012.

Papers to be Submitted/Under Review

- **Xuhui Fan**, Eric Gaussier, Longbing Cao, Richard Yi Da Xu. Supervised Categorical Metric Learning with Schattern p -norm. Submitted

LIST OF PUBLICATIONS

to SDM 2015.

- **Xuhui Fan**, Richard Yi Da Xu, Longbing Cao, Yin Song. Learning Hidden Structures with Relational Models by Adequately Involving Rich Information in A Network. CoRR abs/1310.1545(2013) Submitted to AISTATS 2015.
- **Xuhui Fan**, Longbing Cao, Richard Yi Da Xu. Copula Mixed-Membership Stochastic Blockmodel for Intra-Subgroup Correlations. CoRR abs/1306.2733(2013) Submitted to IEEE Transaction on Pattern Analysis and Machine Intelligence.
- **Xuhui Fan**, Yiling Zeng, Longbing Cao. Non-parametric Power-law Data Clustering. CoRR abs/1306.3003(2013).

Abstract

As the traditional machine learning setting assumes that the data are identically and independently distributed (i.i.d), this is quite like a perfect conditioned vacuum and seldom a real case in practical applications. Thus, the non-i.i.d learning (Cao, Ou, Yu & Wei 2010)(Cao, Ou & Yu 2012)(Cao 2014) has emerged as a powerful tool in describing the fundamental phenomena in the real world, as more factors to be well catered in this modelling. One critical factor in the non-i.i.d. learning is the relations among the data, ranging from the feature information, node partitioning to the correlation of the outcome, which is referred to as the coupling relation in the non-i.i.d. learning. In our work, we aim at uncovering this coupling relation with the nonparametric Bayesian relational models, that is, the data points in our work are supposed to be coupled with each other, and it is this coupling relation we are interested in for further investigation. The coupling relation is widely seen and motivated in real world applications, for example, the hidden structure learning in social networks for link prediction and structure understanding, the fraud detection in the transactional stock market, the protein interaction modelling in biology.

In this thesis, we are particularly interested in the learning and inferencing on the relational data, which is to further discover the coupling relation between the corresponding points. For the detail modelling perspective, we have focused on the framework of *mixed-membership stochastic blockmodel*, in which *membership indicator* and *mixed-membership distribution* are noted to represent the nodes' belonging community for one relation and the his-

togram of all the belonging communities for one node. More specifically, we are trying to model the coupling relation through three different aspects: 1) the mixed-membership distributions' coupling relation across the time. In this work, the coupling relation is reflected in the sticky phenomenon between the mixed-membership distributions in two consecutive time; 2) the membership indicators' coupling relation, in which the Copula function is utilized to depict the coupling relation; 3) the node information and mixed-membership distribution's coupling relation. This is achieved by the new proposal transform for the node information's integration. As these three aspects describe the critical parts of the nodes' interaction with the communities, we are hoping the complex hidden structures can thus be well studied.

In all of the above extensions, we set the number of the communities in a nonparametric Bayesian prior (mainly Hierarchical Dirichlet Process), instead of fixing it as in the previous classical models. In such a way, the complexity of our model can grow along with the data size. That is to say, while we have more data, our model can have a larger amount of communities to account for them. This appealing property enables our models to fit the data better. Moreover, the nice formalization of the Hierarchical Dirichlet Process facilitates us to some benefits, such as the conjugate prior. Thus, this nonparametric Bayesian prior has introduced new elements to the coupling relations' learning.

Under this varying backgrounds and scenarios, we have shown our proposed models and frameworks for learning the coupling relations are evidenced to outperform the state-of-the-art methods via literature explanation and empirical results. The outcomes are sequentially accepted by top journals. Therefore, the nonparametric Bayesian models in learning the coupling relations presents high research value and would still be attractive opportunities for further exploration and exploit.