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 Analystics 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

# List of Figures

# List of Tables

# 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 Arti.cial 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 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 histogram 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.