

Heterogeneous Network Analysis on Academic Collaboration Networks

A Thesis Submitted for the Degree of
Doctor of Philosophy

By

Qinxue Meng

in

School of Software
UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA
JULY 2014

© Copyright by Qinxue Meng, 2014

UNIVERSITY OF TECHNOLOGY, SYDNEY
SCHOOL OF SOFTWARE

The undersigned hereby certify that they have read this thesis entitled “**Heterogeneous Network Analysis on Academic Collaboration Networks**” by **Qinxue Meng** and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of **Doctor of Philosophy**.

Dated: July 2014

Research Supervisors: _____
Paul J. Kennedy

CERTIFICATE

Date: **July 2014**

Author: **Qinxue Meng**

Title: **Heterogeneous Network Analysis on Academic
Collaboration Networks**

Degree: **Ph.D.**

I certify that this thesis has not already been submitted for any degree and is not being submitted as part of candidature for any other degree.

I also certify that the thesis has been written by me and that any help that I have received in preparing this thesis, and all sources used, have been acknowledged in this thesis.

Signature of Author

Acknowledgements

I would like to express my gratitude to my supervisor, Paul J. Kennedy for his continuous encouragement, advice, help and invaluable suggestions to my study and my life. He is such a nice, generous, helpful and kindhearted person. At the beginning of my study, it is he who held a series of lectures in lab meetings covering the fundamental knowledge of research such as common methods and tools in data mining, writing in Latex and explaining doctoral framework. During my study at UTS, he builds a relaxing, comfortable and active environment and I owe my research achievements to his experienced supervision.

Many thanks go to my lab mates Ahmad Al-oqaily, Hamid Ghous, Siamak Tafavogh, Hooman Homayoonfard and Ali Anaissi. The discussions with them in lab meetings are extremely useful to my research and inspire my research. I appreciate the travel support for attending the international conferences which I received from the School of Software and QCIS Lab.

I also would like to thank my wife, Wang Wenjun, for her understanding, assistance and company during my study in Australia. I also thank my parents for the support of my overseas study. This thesis could not have been completed without their supports.

Last but not least, special thanks are given to UTS Research & Innovation Office for providing raw datasets for my research.

Wish you all every success in the future.

Table of Contents

Table of Contents	vii
List of Figures	1
List of Tables	4
Abstract	6
Table of Symbols	11
1 Introduction	14
1.1 What are heterogeneous networks?	16
1.2 Significance of mining heterogeneous networks	20
1.3 Why study academic collaboration?	21
1.4 Research questions	23
1.5 Contributions to knowledge	25
1.6 Organisation of contents	27
2 Literature review	30
2.1 Similarity measures in networks	33
2.1.1 Distance-based similarity measures	33
2.1.2 Neighborhood-based similarity measures	35
2.1.3 Probability-based similarity measures	38
2.2 Community detection	39
2.2.1 Similarity-based community detection	41
2.2.2 Hierarchical clustering	44
2.2.3 Spectral-based clustering algorithms	46
2.2.4 Modularity partitioning	51
2.2.5 Other community detection methods on heterogeneous networks	53

2.2.6	Community detection validation	54
2.3	Determining the number of clusters	58
2.3.1	Clustering result-based methods	58
2.3.2	Topological feature-based methods	59
2.3.3	Support Vector Machine (SVM)	63
2.4	Link prediction	64
2.5	Ranking	67
2.6	Research gaps	69
3	Community detection on heterogeneous networks	71
3.1	Methodology	72
3.1.1	Multiple semantic-path clustering	72
3.1.2	Semantic path assessment	75
3.2	Clustering evaluation	76
3.2.1	Cluster comparison	76
3.2.2	Cluster validation	78
3.3	Experimental dataset	78
3.4	Experimental results	81
3.4.1	Collective similarity calculation	81
3.4.2	Path assessment	83
3.4.3	Community detection and validation	87
3.5	Contribution and discussion	89
4	Determine the number of clusters by leaders	91
4.1	Leader detection and grouping clustering	93
4.1.1	Leader identification	93
4.1.2	Leader group formation	96
4.1.3	Community detection	99
4.2	Clustering validation	100
4.3	Experimental dataset	101
4.4	Experiment	102
4.4.1	Leader identification	102
4.4.2	Community detection	105
4.5	Contribution and discussion	107
5	Network evolution-based link prediction	110
5.1	Methodology	112
5.1.1	Modeling vertex activeness	112
5.1.2	Network Evolution-based link prediction	115

5.1.3	Evaluation	117
5.2	Experimental dataset	118
5.3	Experimental results	119
5.3.1	Modeling vertex activeness evolution	119
5.3.2	Determining vertex evolving patterns	120
5.3.3	Link prediction	120
5.4	Contribution and discussion	122
6	Co-ranking on complex bipartite heterogeneous networks	124
6.1	Data model	125
6.1.1	Network model	125
6.1.2	Matrix model	126
6.2	Methodology	127
6.2.1	Ranking based on rules	127
6.2.2	The co-ranking framework	128
6.2.3	Evaluation	130
6.3	Experimental dataset	132
6.4	Experiment	134
6.4.1	Extracting ranking rules	134
6.4.2	Co-ranking authors and publications	138
6.4.3	Evaluation	141
6.4.4	Divergence analysis	142
6.5	Contribution and discussion	143
7	Conclusions	146
7.1	Future research directions	151

List of Figures

1.1	An example of how to decompose complex heterogeneous networks . . .	17
2.1	A simple example of community detection to show that three communities have been found.	40
2.2	An example of hierarchical tree. Horizontal cuts correspond to partitions of a network in communities from Newman & Girvan (2004). . .	44
2.3	An example of graph partitioning.	47
2.4	Scree graph	60
2.5	SVM aims to draw a boundary among objects and those objects in the same side are classified into one cluster (Cortes & Vapnik 1995). . . .	63
3.1	Examples of semantic paths.	73
3.2	The constitution of Field of Research (FoR) codes	78
3.3	The schema of the network presenting the academic collaboration at UTS	80
3.4	Semantic paths derived from the academic collaboration heterogeneous networks.	82
3.5	The similarity graph of researchers in Scenario I	84
3.6	The similarity graph of researchers in Scenario II	85

3.7	Some laboratories are labeled in the researcher similarity network of Scenario III.	86
4.1	An example of why leaders should be grouped	97
4.2	The working process of community detection by Leader Detection and Grouping Clustering. Circles are vertices and LG_i refers to leader groups. The similarity between vertices and leader groups are calculated and vertices are allocated to those leader groups with highest similarity.	98
4.3	The schema of the experimental heterogeneous network.	102
4.4	The distribution of researcher degree centrality	103
4.5	The distribution of researcher betweenness centrality	103
4.6	The result of leader detection by SVM in R. Triangles and circles are data points from two clusters. Circles stand for leaders and triangles are community members. Solid triangles and circles are support vectors of these two clusters respectively.	104
4.7	The eigenvalues of the random-walk Laplacian matrix.	106
5.1	The schema of the heterogeneous academic collaboration network at UTS.	118
6.1	An example of a complex bipartite heterogeneous network is used for co-ranking.	126
6.2	The working flow between different rules in the co-ranking method.	129
6.3	The data sources of the experimental dataset.	132

6.4	The schema of the experimental complex bipartite heterogeneous network built from the DBLP website.	135
6.5	The academic collaboration network is represented by a matrix for further computation.	136
6.6	Mutual improvement of co-ranking publications and authors through iterations. Each of the six pairs of graphs shows the distributions of publication and author ranks. The x axis in each diagram is ranking score and the y axis is the frequency of objects.	138
6.7	Convergence analysis: the rates of convergence from the proposed co-ranking method, PageRank and HITS are illustrated: (a) describes the process of publication ranking by co-ranking and PageRank; (b) compares the converge rate between coranking and HITS for author ranking.	143

List of Tables

2.1	Summarization of Neighborhood-based similarity measures	35
2.2	Laplacian matrices	48
3.1	The number of records in files from 2009 to 2011	80
3.2	Similarity calculation of all semantic paths related to researchers. . .	82
3.3	The paths and their corresponding scalars	83
3.4	Clustering validation	88
3.5	Clustering quality validation	89
4.1	A sample of new built dataset	95
4.2	The number of records in files from 2009 to 2011	101
4.3	Clustering results by spectral clustering	106
4.4	Clustering results by spectral clustering and LDGC	107
5.1	The experimental dataset from UTS	118
5.2	Statistics of the academic collaboration network from 2006 to 2011 . .	119
5.3	Categories of vertices	120
5.4	Link prediction accuracy comparison.	121
6.1	Statistics of the academic network used to validate the proposed co- ranking approach	133

6.2	Top 10 authors and publications by co-ranking	140
6.3	Top 10 authors by co-ranking and their H-index scores by CiteSeer .	141
6.4	Evaluation of ranking results by co-ranking, PageRank and HITS . .	142

Abstract

Heterogeneous networks are a type of complex network model which can have multi-type objects and relationships. Nowadays, research on heterogeneous networks has been increasingly attracting interest because these networks are more advantageous in modeling real-world situations than traditional networks, that is homogenous networks, that can only have one type of object and relationship. For example, the network of Facebook has vertices including photographs, companies, movies, news and messages and different relationships among these objects. Besides that, heterogeneous networks are especially useful for representing complex abstract concepts, such as friendship and academic collaboration. Because these concepts are hard to measure directly, heterogeneous networks are able to represent these abstract concepts by concrete and measurable objects and relationships. Because of these features, heterogeneous networks are applied in many areas including social networks, the World Wide Web, research publication networks and so on. This motivates the thesis to work on network analysis in the context of heterogeneous networks.

In the past, homogeneous networks were the research focus of network analysis and therefore many methods proposed by previous studies for social network analysis were designed for homogenous networks. Although heterogeneous networks can be considered as an extension of homogenous networks, most of these methods are

not applicable on heterogeneous networks because these methods can only address one type of object and relationships instead of dealing with multi-type ones. In network analysis, there are three basic problems including community detection, link prediction and object ranking. These three questions are the basis of many practical questions, such as network structure extraction, recommendation systems and search engines. Community detection, also called clustering, aims to find the community structure of a network including subgroups of vertices that are closely related, which can facilitate people to understand the structure of networks. Link prediction is a task for finding links which are currently non-existent in networks but may appear in the future. Object ranking can be viewed as an object evaluation task which aims to order a set of objects based on their importance, relevance, or other user defined criteria. In addition to these three research issues, approaches for determining the number of clusters *a priori* is also important because it can improve the quality of community detection significantly. This thesis works on heterogeneous network and proposes a set of methods to address the four main research problems in network analysis including community detection, determining the number of clusters, link prediction and object ranking.

There are four contributions in this thesis. Contribution 1 proposes a Multiple Semantic-path Clustering method which can facilitate users to achieve a desired clustering in heterogeneous networks. Contribution 2 develops a Leader Detection and Grouping Clustering method which can determine the number of clusters *a priori*, thereby improving the quality of clustering. Contribution 3 introduces a Network Evolution-based Link Prediction method which can improve link prediction accuracy by modeling evolution patterns of objects. Contribution 4 proposes a co-ranking

method which can work on complex bipartite heterogeneous networks where one type of vertex can connect to themselves directly and indirectly.

The performance of all developed methods in the thesis in terms of clustering quality, link prediction accuracy and ranking effectiveness, is evaluated in the context of a research management dataset of University of Technology, Sydney (UTS) and public bibliographic DBLP (DataBase systems and Logic Programming) dataset. Moreover, all the results of the proposed methods in this thesis are compared with state-of-the-art methods and these experimental results suggest that the proposed methods outperform these state-of-the-art methods in quantitative and qualitative analysis.

Publications

Below is the list of the journal and conference papers associated with my PhD research:

1. Meng, Q., Tafavogh, S. & Kennedy, P. J. (2014), ‘Community Detection on Heterogeneous Networks by Multiple Semantic-Path Clustering’, *in* ‘Proceedings of the 6th International Conference on Computational Aspects of Social Networks (CASoN)’, IEEE.
2. Tafavogh, S., Meng, Q., Catchpoole, D. R., Kennedy, P. J. (2014), ‘Automated quantitative and qualitative analysis of the whole slide images of neuroblastoma tumor for making a prognosis decision’, *in* ‘Proceedings of The 11th IASTED International Conference on Biomedical Engineering (BioMed 2014)’, IEEE.
3. Asabere, N. Y., Xia, F., Meng, Q., Li, F. & Liua, H. (2014), ‘Scholarly paper recommendation based on social awareness and folksonomy’, *International Journal of Parallel, Emergent and Distributed Systems*.
4. Meng, Q. & Kennedy, P. J. (2013b), ‘Survey on spectral clustering and its applications in social networks’, *Computer Engineering and Applications* **49**(3), 213–221.
5. Meng, Q. & Kennedy, P. J. (2013a), ‘Discovering influential authors in heterogeneous academic networks by a co-ranking method’, *in* ‘Proceedings of the 22nd ACM International Conference on Information & Knowledge Management’, ACM, pp. 1029–1036.

6. Meng, Q. & Kennedy, P. J. (2012c), ‘Using network evolution theory and singular value decomposition method to improve accuracy of link prediction in social networks’, *in* ‘Proceedings of the Tenth Australasian Data Mining Conference’, Volume 134, Australian Computer Society, Inc., pp. 175–181.
7. Meng, Q. & Kennedy, P. J. (2012b), ‘Using field of research codes to discover research groups from co-authorship networks’, *in* ‘Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)’, IEEE Computer Society, pp. 289–293.
8. Meng, Q. & Kennedy, P. J. (2012a), ‘Determining the number of clusters in co-authorship networks using social network theory’, *in* ‘2012 Second International Conference on Social Computing and Its Applications (SCA 2012)’, IEEE, pp. 337–343.

Table of Symbols

Symbols	Description
G	Networks or graphs
V	Vertex set
$ V $	the number of vertices
E	Edge set
$ E $	the number of edges
P	Semantic path set
$ P $	the number of semantic paths
V_n	The set of vertices in type n
E_m	The set of edges in type m
v, u	vertices
e_{uv}	The edge from vertex u to v
A	Adjacency matrix
a_{uv}	An element of adjacency matrix A . If $a_{uv} = 1$, there is an edge from u and v ; If $a_{uv} = 0$, vertex u and v are not connected.

Symbols	Description
W	Weighted adjacency matrix
w_{uv}	the weight of edge e_{uv}
d_v	Degree of vertex v , $d_v = \sum_{i=1}^{ V } w_{vi}$
D	Degree matrix which is a diagonal matrix with the degrees $d_1, \dots, d_{ V }$
L	Laplacian matrix
l_i	i th eigenvalue of Laplacian matrix
I	Identify matrix
S	Similarity matrix
s_{uv}	The similarity between vertex u and v
C	Cluster indicator matrix
k	Number of clusters
$W^{V_i V_j}$	The weight adjacency matrix between object type V_i and V_j
$w_{uv}^{V_i V_j}$	$w_{uv}^{V_i V_j} = w_{uv}$ where $u \in V_i$ and $v \in V_j$
C_D	Vertex degree centrality
C_B	Vertex betweenness centrality
LG_i	i th leader group

Symbols	Description
$N_{uv}(i)$	The number of paths between vertex u and v and that belong to semantic path i
$len(i)$	The length of semantic path i
X, Y	Network partitions
$T(u, v)$	Time of randomly moving agent from starting vertex u to the end vertex v
Ω	Network evolution
$neighbor(v, t)$	The neighborhood set of vertex v in timeslot t
$rank(v, i)$	The ranking scores of vertex v in i th iteration
$Diff(i, i + 1)$	The difference of vertex ranking scores between i th iteration and $(i + 1)$ th iteration