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

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

List of Figures

List of Tables

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 $d_1, \ldots, d_{\|V\|}$
L	Laplacian matrix
l_i	ith eigenvalue of Laplacian matrix
I	Identify matrix
S	Similarity matrix
s_{uv}	The similarity beween vertex u and v
C	Cluster indicator matrix
k	Number of clusters
$W^{V_iV_j}$	The weight adjacency matrix
	between object type V_i and V_j
$w_{uv}^{V_iV_j}$	$w_{uv}^{V_iV_j} = w_{uv}$ where $u \in V_i$ and $v \in V_j$
C_D	Vertex degree centrality
C_B	Vertex betweenness centrality
LG_i	ith 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
$neighor(v, t)$	The neighborhood set of vertex v in timeslot t
$rank(v, i)$	The ranking scores of vertex v in ith iteration
$Diff(i, i+1)$	The difference of vertex ranking scores between ith iteration and $(i+1)$th iteration