Structural Community Detection in Big Graphs

by

DONG WEN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Centre for Artificial Intelligence (CAI)
Faculty of Engineering and Information Technology (FEIT)
University of Technology Sydney (UTS)

June, 2018

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.

Production Note:
Signature removed prior to publication.

Signature of Candidate

ACKNOWLEDGEMENTS

First, I would like to deliver my sincere gratitude to my supervisor Dr. Lu Qin for his continuous support and guidance of my PhD career. He is professional, efficient, patient and diligent. His valuable ideas always guided me and extended my knowledge not only in the database area, but also about the practical research skills. Additionally, Lu is a good mentor and friend for me. He is selfless and glad to share his own experience with me. Thanks to his encouragement, I am always positive and brave when experiencing challenges and even failures. This thesis could not reach its present form without his illuminating instructions.

Secondly, I would like to express my great gratitude to my co-supervisor Prof. Ying Zhang for his guidance and advice, especially for his strong confidence in me and support for my academic career. He gave me many wonderful ideas and inspirations. His guidance significantly extended my knowledge and helped me pursue a correct research direction all the time. He is also visionary, and often encouraged and guided me to optimize the plan for the career development.

Thirdly, I would like to thank Prof. Xuemin Lin and Dr. Lijun Chang for supporting the works in this thesis, as most of the works were conducted in collaboration with them. I thank Prof. Lin for offering an interesting but rigorous

research environment. I learned the characteristics of an excellent researcher from Prof. Lin — passion, preciseness and earnest. I thank Dr. Chang for his brilliant ideas and confidence in me. His wonderful research works always inspired me, and he gave me many accurate and valuable suggestions to improve the quality of the works in this thesis.

I would also like to thank the following people: Prof. Jeffrey Xu Yu, Dr. Wenjie Zhang, Dr. Xin Cao, Dr. Zengfeng Huang, Dr. Ling Chen, Dr. Xiaoyang Wang, Dr. Shiyu Yang, for sharing valuable ideas and experiences. Thanks to Dr. Fan Zhang, Dr. Long Yuan, Dr. Longbin Lai, Dr. Xing Feng, Dr. Xiang Wang, Dr. Jianye Yang, Dr. Fei Bi, Mr. Xubo Wang, Mr. Wei Li, Mr. Haida Zhang, Ms. Chen Zhang, Dr. Qing Bing, Dr. Shenlu Wang, Mr. Yang Yang, Mr. Kai Wang, Mr. You Peng, Mr. Hanchen Wang, Mr. Xuefeng Chen, Mr. Boge Liu, Mr. Jiahui Yang, Dr. Wei Wu, Dr. Qian Zhang, Dr. Zhaofeng Su, Mr. Dian Ouyang, Mr. Wentao Li, Mr. Bohua Yang, Mr. Mingjie Li, Ms. Conggai Li, Mr. Wei Song, for sharing the happiness with me in my PhD career.

Last but not least, I would like to thank my father Mr. Rongxiao Wen and my mother Mrs. Tiefang Liu, for bringing me a wonderful life, and other relatives for their understanding, encouragement and love. Thanks to my wife Mrs. Qing Lin, for her company and support in my PhD study.

ABSTRACT

Community detection in graphs is a fundamental problem widely experienced across industries. Given a graph structure, one popular method to identify communities is classifying the vertices, which is formally named graph clustering. Additionally, community structures are always dense and highly connected in graphs. There are also a large number of research works focusing on mining cohesive subgraphs for community detection. Even though the community detection problem is extensively studied, challenges still exist. With the development of social media, graphs are highly dynamic, and the size of graphs is sharply increasing. The large time and space cost of traditional solutions may hardly be endured in big and dynamic graphs.

In this thesis, we propose an index-based algorithm for the structural graph clustering (SCAN). Based on the proposed index structure, the time expended to compute structural clustering depends only on the result size, not on the size of the original graph. The space complexity of the index is bounded by $O(m)$, where m is the number of edges in the graph. We also propose algorithms and several optimization techniques for maintaining our index in dynamic graphs.

For the cohesive subgraph detection, we study both degree-constrained (k-core) and connectivity-constrained (k-VCC) cohesive subgraph metrics. A k-core is a maximal connected subgraph in which each vertex has degree at least k. We study I/O efficient core decomposition following a semi-external model, which only allows vertex information to be loaded in memory. We propose an optimized

I/O efficient algorithm for both core decomposition and core maintenance. In addition, we extend our algorithm to compute the graph degeneracy order, which is an important graph problem that is highly related to core decomposition.

A k-vertex connected component (k-VCC) is a connected subgraph in which the removal of any $k-1$ vertices will not disconnect the subgraph. A k-VCC has many outstanding structural properties, such as high cohesiveness, high robustness, and subgraph overlapping. The state-of-the-art solution enumerates all k-VCCs following a partition-based framework. It requires high computational cost in connectivity tests. We prove the upper bound of the number of partitions, which implies the polynomial running time of this framework. We propose two effective optimization strategies, namely neighbor sweep and group sweep, to largely reduce the number of local connectivity tests.

We conducted extensive performance studies using several large real-world datasets to show the efficiency and effectiveness of all our approaches.

PUBLICATIONS

- ***Dong Wen***, *Lu Qin, Xuemin Lin, Ying Zhang, and Lijun Chang, Enumerating k-Vertex Connected Components in Large Graphs, in submission.*
- ***Dong Wen***, *Lu Qin, Ying Zhang, Xuemin Lin, and Jeffrey Xu Yu, I/O Efficient Core Graph Decomposition: Application to Degeneracy Ordering, TKDE, revision submitted.*
- ***Dong Wen***, *Lu Qin, Ying Zhang, Lijun Chang, and Xuemin Lin, Efficient Structural Graph Clustering: an Index-Based Approach, to appear in PVLDB 2018.*
- *Lingxi Yue,* ***Dong Wen***, *Lizhen Cui, Lu Qin, and Yongqing Zheng, K-Connected Cores Computation in Large Dual Networks, to appear in DASFAA 2018.*
- ***Dong Wen***, *Lu Qin, Ying Zhang, Xuemin Lin, and Jeffrey Xu Yu, I/O Efficient Core Graph Decomposition at Web Scale, ICDE2016, Best Paper Award.*

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES