Cohesive Subgraph Detection on Large Graphs

by

Bohua Yang

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

Australian Artificial Intelligence Institute (AAII)
Faculty of Engineering and Information Technology (FEIT)
University of Technology Sydney (UTS)

April, 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Bohua Yang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 09/04/2021

ACKNOWLEDGEMENTS

First of all, I would like to deliver my sincere gratitude to my principal supervisor A/Prof. Lu Qin. Thanks for his selfless help during my PhD life. He is a good mentor and friend to me. He is very professional; he is a perfect guide who leads me to the beautiful world of the graph. He is very patient; he is always ready to help me whenever I meet problems in research or life. I could not finish my thesis without his illuminating instructions.

Secondly, I would like to thank my co-supervisor Prof. Ying Zhang. He gives me much accurate and valuable advice on my research topics. Besides, he often encourages and guides me to optimize my career development plan.

Thirdly, I would like to thank Dr. Dong Wen. All the research works presented in this thesis are conducted together with him. He patiently guides me in acquiring research skills, such as thinking about problems, writing papers, and preparing presentations. I learn a lot from the collaboration with him.

Fourthly, I would like to thank Prof. Xuemin Lin and Dr. Lijun Chang for supporting the works in this thesis. I would like to thank Prof. Lin for offering an interesting but rigorous research environment. I would like to thank Dr. Chang for his insightful advice on my research works.

Fifthly, I am very grateful to A/Prof. Xin Zhao, my mentor at the Renmin

University of China, for stimulating my research interest and providing helpful advice for my academic career. Furthermore, I would like to thank Prof. Jirong Wen, A/Prof. Hui Sun, Prof. Qing Zhu, and A/Prof. Yan Yu at the Renmin University of China for their kind help.

I would also like to appreciate Prof. Jeffrey Xu Yu, Prof. Wenjie Zhang, Prof. Ronghua Li, A/Prof. Ling Chen, Dr. Yulei Sui, Dr. Xin Cao, A/Prof. Zengfeng Huang, Dr. Shiyu Yang, A/Prof. Yixiang Fang, Prof. Weiwei Liu, and Prof. Xiaoyang Wang for sharing brilliant thoughts and experiences. Thanks to Dr. Dian Ouyang, Dr. Wentao Li, Dr. Conggai Li, Dr. Mingjie Li, Dr. Adi Lin, Dr. Yuan Liang, Mr. Yuxuan Qiu, Mr. Junhua Zhang, Dr. Fan Zhang, Dr. Wanqi Liu, Mr. Hanchen Wang, Mr. Yilun Huang, Dr. Xubo Wang, Dr. Longbin Lai, Dr. Long Yuan, Dr. Xing Feng, Dr. Haida Zhang, Dr. Wei Li, Dr. Chen Zhang, Dr. Yang Yang, Dr. Kai Wang, Dr. You Peng, Dr. Boge Liu, Mr. Yuanhang Yu, Mr. Peilun Yang, Ms. Xiaoshuang Chen, Mr. Xuefeng Chen, Mr. Yuren Mao, Mr. Yixing Yang, Mr. Zhengyi Yang, Mr. Qingyuan Linghu, Mr. Michael Ruisi Yu, Mr. Chenji Huang, Mr. Yu Hao, Mr. Kongzhang Hao, Dr. Peng Zhang, Mr. Xunxiang Yao, Dr. Binbin Huang, Dr. Yarui Chen, Ms. Kun Wang, Mr. Yijun Li, and Ms. Ruilin Liu. The days we spent together bear unforgettable memories.

Finally, I would like to thank my grandfather Mr. Zhencai Yang, and my grandmother Ms. Yuling Zhao, who give me selfless love and endless encouragement. I would like to thank the support from other relatives and friends.

ABSTRACT

Graphs have been widely used to model sophisticated relationships between different entities due to their strong representative properties. Social networks, traffic networks, and biological networks are among the applications that benefit from being expressed as graphs. The cohesive subgraph is an essential structure for understanding the organization of many real-world networks, and cohesive subgraph detection is a crucial problem in network analysis. There are many cohesive subgraph models, such as k-core, strongly connected component, and maximum density subgraph.

Uncertain graph management and analysis have attracted much research attention. Among them, computing k-cores in uncertain graphs (aka, (k, η)-cores) is an important problem and has emerged in many applications. However, the existing algorithms for computing (k, η)-cores heavily depend on the two input parameters k and η. In addition, computing and updating the η-degree for each vertex is the costliest component in the algorithm, and the cost is high.

To overcome these drawbacks, we propose an index-based solution for computing (k, η)-cores. The index size is well-bounded by $O(m)$, where m is the number of edges in the graph. Based on the index, queries for any k and η can be answered in optimal time. We propose an algorithm for index construction with several different optimizations.

We also discuss the (k, η)-core computation when graphs cannot be entirely stored in memory. We adopt the semi-external setting, which allows $O(n)$ mem-

ory usage, where n is the number of vertices in the graph. This assumption is reasonable in practice, and it has been widely adopted in massive graph analysis. We design an index-based solution for I/O efficient (k, η)-core computation.

Given the frequent updates in many real-world graphs, detecting strongly connected components (SCC) in dynamic graphs is a very complicated problem. In the thesis, we study the fully dynamic depth-first search (DFS) problem in directed graphs, which is a crucial basis of dynamic SCC detection. In the literature, most works focus on the dynamic DFS problem in undirected graphs and directed acyclic graphs. However, their methods cannot easily be applied in the case of general directed graphs. Motivated by this, we propose a framework and corresponding algorithms for both edge insertion and deletion in general directed graphs. We further give several optimizations to speed up the algorithms.

We conduct extensive experiments on several large real-world graphs to practically evaluate the performance of all proposed algorithms.

PUBLICATIONS

- ***Bohua Yang***, *Dong Wen, Lu Qin, Ying Zhang, Lijun Chang, and Rong-Hua Li. Index-Based Optimal Algorithm for Computing K-Cores in Large Uncertain Graphs. In 2019 IEEE 35th International Conference on Data Engineering (ICDE), pages 64-75. IEEE, 2019.*

- ***Bohua Yang***, *Dong Wen, Lu Qin, Ying Zhang, Xubo Wang, and Xuemin Lin. Fully Dynamic Depth-First Search in Directed Graphs. Proceedings of the VLDB Endowment, 13(2):142-154, 2019.*

- *Dong Wen,* ***Bohua Yang***, *Lu Qin, Ying Zhang, Lijun Chang, and Rong-Hua Li. Computing K-Cores in Large Uncertain Graphs: An Index-based Optimal Approach. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2020.*

- *Dong Wen,* ***Bohua Yang***, *Ying Zhang, Lu Qin, Dawei Cheng, and Wenjie Zhang. Span-Reachability Querying in Large Temporal Graphs. The VLDB Journal, Revision Submitted.*

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES