Cohesive Subgraph Mining on Social Networks

by

Fan Zhang

B.E. Zhejiang University, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

the Centre for Artificial Intelligence (CAI)
the Faculty of Engineering and Information Technology (FEIT)
the University of Technology Sydney (UTS)
August, 2017

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

UNIVERSITY OF
TECHNOLOGY SYDNEY

ACKNOWLEDGEMENTS

First and foremost, I would like to deliver my sincere gratitude to my supervisor Prof. Ying Zhang for his continuous support of my PhD study and research, especially for his professionalism, patience, passion and diligence. His guidance extends my knowledge in computer science, improves my capacity in scientific research and elevates my love in exploring the fields which are significant, undiscovered and challenging. Besides the character of supervisor, Ying has also been a friend and mentor. Thanks to his confidence in me, I never lost hope when experiencing failures and was always positive during my PhD study. Without his consistent and illuminating instruction, this thesis could not have reached its present form.

Secondly, I would like to express my great gratitude to my co-supervisor Dr. Lu Qin for his constant encouragement and guidance, especially for his brilliant ideas and inspirations. His work efficiency gave me the hope to conduct good research without abandoning too many of other interests, which prevented me to be negative in the early stage of my PhD study. Lu always has confidence in solving research problems, regardless of their complexities, which encourages me to keep thinking and challenging myself.

Thirdly, parts of the work in this thesis were conducted in collaboration with Prof. Xuemin Lin and Dr. Wenjie Zhang. I thank them for supporting the work presented in this thesis. A special thanks to Prof. Lin for providing a rigorous and self-motivated research environment, and the support for my academic career. I learnt the characteristics of an excellent researcher from Prof. Lin - passion, diligence and earnestness in the research.

I would also like to thank the following people at UNSW and UTS, Australia: Dr. Xiaoyang Wang, Dr. Shiyu Yang, Dr. Lijun Chang, Dr. Zengfeng Huang, Dr. Xin Cao, and Dr. Gaoping Zhu, for sharing your brilliant ideas and experiences. Thanks to Ms. Haiyan Hua, Ms. Chen Zhang, Dr. Xing Feng, Ms. Shan Xue, Dr. Long Yuan, Dr. Longbin Lai, Dr. Xiang Wang, Dr. Jianye Yang, Dr. Chengyuan Zhang, Mr. Xubo Wang, Mr. Fei Bi, Mr. Wei Li, Mr. Dong Wen, Mr. Haida Zhang, Ms. Qing Bing, Ms. Wen Li, Ms. Lu Liu, Ms. Shenlu Wang, Ms. Conggai Li, Mr. Yang Yang, Mr. Kai Wang, Mr. Mingjie Li, Mr. Wentao Li and Mr. Hanchen Wang, for sharing the happiness and bitterness with me during my PhD study. The time we spent together will be memorized forever.

Last but not least, thank my family: my father Mr. Lingchi Zhang and my mother Mrs. Liqin Zhou, for bringing me a happy and wonderful life in the world, and other relatives for their support, encouragement and love.

PUBLICATIONS

- **Fan Zhang**, Ying Zhang, Lu Qin, Wenjie Zhang, Xuemin Lin. When Engagement Meets Similarity: Efficient (k,r)-Core Computation on Social Networks. PVLDB 2017. (Chapter 3)

- **Fan Zhang**, Wenjie Zhang, Ying Zhang, Lu Qin, Xuemin Lin. OLAK: An Efficient Algorithm to Prevent Unraveling in Social Networks. PVLDB 2017. (Chapter 4)

- **Fan Zhang**, Ying Zhang, Lu Qin, Wenjie Zhang, Xuemin Lin. Efficiently Reinforcing Social Networks over User Engagement and Tie Strength. Under submission. (Chapter 4)

- **Fan Zhang**, Ying Zhang, Lu Qin, Wenjie Zhang, Xuemin Lin. Finding Critical Users for Social Network Engagement: The Collapsed k-Core Problem. AAAI 2017. (Chapter 5)

TABLE OF CONTENT

CERTIFICATE OF AUTHORSHIP/ORGINALITY iii

ACKNOWLEDGEMENTS iv

PUBLICATIONS vi

TABLE OF CONTENT vii

LIST OF FIGURES x

LIST OF TABLES xii

ABSTRACT xiii

Chapter 1 Introduction 1
- 1.1 Motivation . . . 2
  - 1.1.1 Cohesive Subgraph Discovery . . . 2
  - 1.1.2 Prevent Network Unraveling . . . 4
  - 1.1.3 Finding Critical Users . . . 6
- 1.2 Contribution . . . 7
  - 1.2.1 Cohesive Subgraph Discovery . . . 7
  - 1.2.2 Prevent Network Unraveling . . . 8
  - 1.2.3 Finding Critical Users . . . 10
- 1.3 Organization . . . 11

Chapter 2 Literature Review 12
- 2.1 Cohesive Subgraph Models . . . 12
  - 2.1.1 Clique . . . 12
  - 2.1.2 k-Core . . . 14
  - 2.1.3 k-Truss . . . 16
  - 2.1.4 Other Models . . . 18
- 2.2 Social Network Components . . . 19
  - 2.2.1 User Engagement . . . 19

2.2.2 User Similarity 20
2.2.3 Tie Strength 21
2.3 Mining Attributed Graphs 22

Chapter 3 Cohesive Subgraph Discovery 24
3.1 Introduction 24
3.2 Preliminaries 28
3.2.1 Problem Definition 29
3.2.2 Problem Complexity 31
3.3 The Clique-based Approach 33
3.4 Warming Up for Our Approach 34
3.5 Finding All Maximal (k,r)-Cores 36
3.5.1 Reducing Candidate Size 36
3.5.2 Early Termination 39
3.5.3 Checking Maximal 40
3.5.4 Advanced Enumeration Method 42
3.6 Finding the Maximum (k,r)-core 45
3.6.1 Algorithm for Finding the Maximum One 45
3.6.2 Size Upper Bound of (k,r)-Core 46
3.6.3 Algorithm for (k,k')-Core Upper Bound 49
3.6.4 Finding the Top-m Maximal (k,r)-Cores 50
3.7 Search Order 51
3.7.1 Important Measurements 51
3.7.2 Finding the Maximum (k,r)-Core 52
3.7.3 Enumerating All Maximal (k,r)-Core 53
3.7.4 Checking Maximal 54
3.8 Performance Evaluation 54
3.8.1 Experimental Setting 54
3.8.2 Effectiveness 57
3.8.3 Efficiency 60
3.9 Conclusion 66

Chapter 4 Prevent Network Unraveling 67
4.1 Introduction 67
4.2 Preliminaries 73
4.2.1 Problem Definition 73
4.2.2 Problem Complexity 76
4.3 Our Approach 77
4.3.1 Motivation 77
4.3.2 Reducing the Number of Candidate Anchors 79
4.3.3 Efficiently Finding Followers 80

4.3.4 The OLAK Algorithm 90
4.4 Prevent Collapse of Strong Tie Communities 94
4.4.1 The Problem of Anchored k-Truss 95
4.4.2 The Problem Hardness 97
4.4.3 The Edge Onion Layers 101
4.4.4 Candidate Anchors and Followers 103
4.4.5 Efficiently Finding Candidate Followers 104
4.4.6 Finding Followers with Early Termination 107
4.4.7 The AKT Algorithm 111
4.5 Performance Evaluation 113
4.5.1 Experimental Setting for Anchored k-Core 113
4.5.2 Effectiveness of OLAK 115
4.5.3 Efficiency of OLAK 120
4.5.4 Experimental Setting for Anchored k-Truss 124
4.5.5 Effectiveness of AKT 126
4.5.6 Efficiency of AKT 129
4.6 Conclusion 132

Chapter 5 Finding Critical Users 133
5.1 Introduction 133
5.2 Preliminaries 136
5.2.1 Problem Definition 136
5.2.2 Problem Complexity 138
5.3 Our Approach 142
5.3.1 Motivation 142
5.3.2 Reducing Candidate Collapsers 143
5.3.3 CKC Algorithm 145
5.4 Performance Evaluation 146
5.4.1 Experimental Setting 146
5.4.2 Effectiveness 147
5.4.3 Efficiency 150
5.5 Conclusion 152

Chapter 6 Epilogue 153
6.1 Conclusions 153
6.2 Future Work 155

REFERENCES 156

LIST OF FIGURES

3.1 Group by Friendship and Interests 26
3.2 Group by Friendship and Locations 27
3.3 Example of the Search Tree 35
3.4 Pruning and Retaining Candidates 39
3.5 Early Termination and Check Maximal 41
3.6 Upper Bound Examples 48
3.7 Case Study on DBLP (k=15, r=top 3‰) 58
3.8 Case Study on Gowalla (k=10, r=10km) 59
3.9 (k,r)-core Statistics 60
3.10 Evaluate Clique-based Method 61
3.11 Evaluate Pruning Techniques 62
3.12 Evaluate Upper Bounds 62
3.13 Evaluate Search Orders 63
3.14 Performance on Four Datasets 64
3.15 Effect of k and r for Enumeration 65
3.16 Effect of k and r for Maximum 65

4.1 Motivating Example for Anchored k-Core 68
4.2 Motivating Example for Anchored k-Truss 70
4.3 Bounded Anchors and Followers 80
4.4 Onion Layer Structure (L_0^3) 83
4.5 Early Termination 87
4.6 Followers Pruning and Upper Bound Pruning 91
4.7 Construction Example for NP-hardness Proof, $k = 4$ 98
4.8 Examples for Non-submodular 100
4.9 Examples for Edge Onion Layers $\mathcal{L}$, $k = 4$ 104
4.10 Number of Followers 116
4.11 Greedy vs Exact 117
4.12 Effect of k and b 118
4.13 Case Studies 119
4.14 Reducing Candidate Anchors 120
4.15 Pruning Candidate Followers 121

4.16 Effectiveness of Early Termination 122
4.17 Further Pruning Candidate Anchors 123
4.18 Running Time on Different Datasets 124
4.19 Effect of k and b 125
4.20 Number of Followers 127
4.21 Greedy vs Exact 128
4.22 Case Studies 129
4.23 Reducing Candidate Anchors and Followers 130
4.24 Effect of Algorithms with Different k and b 131
4.25 Running Time on Different Datasets 131

5.1 Motivating Example 134
5.2 Examples for NP-hardness Proof 139
5.3 Examples for Non-submodular 142
5.4 Number of the Followers 148
5.5 Greedy vs Optimal 149
5.6 Case Study on DBLP, k=20, b=1 149
5.7 Effectiveness of Reducing Candidate Collapsers 151
5.8 Performance of Baseline and CKC 151

LIST OF TABLES

3.1 The summary of notations 29
3.2 Statistics of Datasets 55
3.3 Summary of Algorithms 56

4.1 Summary of Notations 74
4.2 Summary of New Notations 96
4.3 Statistics of Datasets 113
4.4 Summary of Algorithms for Anchored k-Core 114
4.5 Summary of Algorithms for Anchored k-Truss 126

5.1 Summary of Notations 137
5.2 Statistics of Datasets 147

ABSTRACT

Graphs are widely used to represent the abundant information in social networks for discovering promising communities, reinforcing network stability, and finding critical users, to name a few. Cohesive subgraph mining, as one of the most fundamental problems in graphs, gains increasing popularity in social network study for its effectiveness. In this thesis, some basic social components are considered in cohesive subgraphs to better accommodate various real-life applications.

Firstly, we investigate the problem of (k,r)-core which intends to find cohesive subgraphs on social networks considering both user engagement and similarity. Efficient algorithms are proposed to enumerate all *maximal* (k,r)-cores and find the *maximum* (k,r)-core, where both problems are shown to be NP-hard. Effective pruning techniques and search orders substantially reduce the search space of two algorithms. A novel upper bound enhances performance of the maximum (k,r)-core computation. Comprehensive experiments on real-life data demonstrate that the algorithms efficiently find interesting communities.

Secondly, we study the problem of the anchored k-core, which was introduced by Bhawalkar and Kleinberg *et* al. in the context of user engagement in social networks. The problem has been shown to be NP-hard and inapproximable. We propose an efficient algorithm, namely `OLAK`, as the first to solve the problem on general graphs. An *onion layer* structure is designed together with efficient candidates exploration, early termination and pruning techniques to significantly simplify computation and greatly reduce the search space.

Besides considering user engagement, we further explore the unraveling phenomenon with tie strength, which leads us to the model of k-truss. We then investigate the anchored k-truss problem which is also NP-hard and propose an *edge onion layer* structure based algorithm, namely `AKT`. Efficient candidate exploration and pruning techniques are designed based on the *edge onion layers*. Comprehensive experiments on real-life graphs for the above two problems demonstrate the effectiveness and efficiency of our proposed methods.

Finally, we study the leave of critical users, which may greatly break network engagement. Accordingly, we propose the collapsed k-core problem to find the vertices whose leave can lead to the smallest k-core. We prove the problem is NP-hard. Then, an efficient algorithm is proposed, which significantly reduces the number of candidate vertices to speed up computation. Comprehensive experiments on real-life social networks demonstrate effectiveness of the model and efficiency of the proposed techniques.