

Measuring Graphs with Shortest Distances

by

WENTAO LI

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

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

January, 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Wentao Li 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: 12/01/2021

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my principal supervisor A/Prof. Lu Qin. Lu is a responsible teacher, a kind friend, and a supportive brother. As a teacher, he guides me to acquire the necessary research skills. I have benefited greatly from discussions with him. As a friend, he is always ready to help me. Whenever I feel frustrated, he is ready to offer advice and practical support. As a brother, he teaches me how to balance research and life. His encouragement has helped me to have a positive attitude towards life.

Secondly, I would like to thank my co-supervisor Prof. Ying Zhang. He always provides enlightening advice on my research topics. In addition, he often encourages and guides me to optimize my career development plan.

Thirdly, I would like to thank my co-supervisor Dr. Miao Qiao. She always encourages me to explore the essence of a problem, from which I experience the enjoyment of conducting research. She patiently guides me in acquiring research skills, such as writing papers and preparing presentations. I learn a lot from her way of thinking about problems and her rigorous attitude to research.

Fourthly, I would like to acknowledge Prof. Xuemin Lin and Dr. Lijun Chang for their support — most of the works presented in this thesis are conducted together with them. I would like to thank Prof. Lin for serving as an example of an excellent researcher for me. I would like to thank Dr. Chang for his insightful

suggestions to improve the quality of the works.

I would also like to appreciate Prof. Wenjie Zhang, Dr. Xin Cao, Dr. Zengfeng Huang, Dr. Shiyu Yang, Dr. Yixiang Fang, Dr. Weiwei Liu, for sharing brilliant thoughts and experiences. Thanks to Dr. Dong Wen, Dr. Xubo Wang, Dr. Dian Ouyang, Dr. Fan Zhang, Dr. Longbin Lai, Dr. Long Yuan, Dr. Xing Feng, Dr. Haida Zhang, Dr. Wei Li, Dr. Chen Zhang, Dr. Yang Yang, Dr. Wanqi Liu, Dr. Kai Wang, Dr. You Peng, Dr. Boge Liu, Mr. Bohua Yang, Ms. Conggai Li, Mr. Mingjie Li, Mr. Junhua Zhang, Mr. Yuxuan Qiu, Mr. Hanchen Wang, Mr. Yilun Huang, 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. Peng Zhang, Mr. Tao Shen, Mr. Xunxiang Yao, Mr. Xiaolin Zhang, Ms. Muwen Yang, Ms. Qian Li, Mr. Chengzhang Zhu. The days we spent together bear unforgettable memories.

I am also very grateful to my supervisors at Chongqing University: A/Prof. Hua Li and A/Prof. Min Gao, for stimulating my research interest and providing useful advice for my study. Furthermore, I would like to express my gratitude to Prof. Qingyu Xiong, Prof. Junhao Wen, A/Prof. Jun Zeng and A/Prof. Wei Zhou of Chongqing University for their kind help.

Finally, I would also like to acknowledge my father Mr. Dongyun Li, and my mother Ms. Linren Wang, who provide me with selfless love and endless encouragement. I would like to thank my grandparents, my sister, and other relatives and friends for their love and support.

ABSTRACT

Shortest distances characterize the pair-wise relationships among nodes in a graph. Given a graph with a node set and an edge set, the shortest distance between two nodes is defined as the minimum path length between them. Computing the shortest distance between two nodes is a fundamental operation of graphs, which can be used both as a primary function and as a building block for applications. Given the important roles of shortest distances, this thesis focuses on the efficient computation of shortest-distance-based measures on graphs.

Firstly, we investigate how to accelerate the index time of 2-hop labeling. 2-hop labeling approaches are widely adopted to speed up the online performance of shortest distance queries. The construction of the 2-hop labeling, however, can be exhaustive especially on big graphs. For a major category of large graphs, small-world networks, the state-of-the-art approach is Pruned Landmark Labeling (PLL). However, PLL's strong sequential nature hinders it from being parallelized. It becomes an urgent issue on massive small-world networks whose index can hardly be constructed by a single thread within a reasonable time. Based on the dependency analysis of PLL, we propose a Parallelized Shortest-distance Labeling (PSL) scheme to exploit parallelism to shorten the index time.

Secondly, we study how to reduce the index size of 2-hop labeling. While the index time can be shortened by parallelized labeling, the index size becomes the bottleneck for a massive real graph with a relatively large treewidth — 2-hop labeling can hardly be constructed due to the oversized index. We disclose

the theoretical relationships between the graph treewidth and 2-hop labeling’s index size and query time. To scale up distance labeling, we propose Core-Tree (CT) Index to dramatically reduce the index size, thereby enabling CT-Index to handle massive graphs that no existing approaches can process.

Thirdly, we compute and maintain the eccentricities of all nodes. Given a graph, eccentricity measures the shortest distance from each node to its farthest node. Existing eccentricity computation algorithms are not scalable enough to handle real large networks. Our solution optimizes existing eccentricity computation algorithms on their bottlenecks — one node eccentricity computation and the upper/lower bounds update — based on a line of original insights; it also provides the first algorithm on maintaining the eccentricities of a dynamic graph without recomputing the eccentricity distribution upon each edge update.

Extensive empirical studies validate the efficiency of our techniques.

PUBLICATIONS

- **Wentao Li**, Miao Qiao, Lu Qin, Ying Zhang, Lijun Chang, and Xuemin Lin. “Scaling Distance Labeling on Small-World Networks.” In *Proceedings of the 2019 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pp. 1060-1077. 2019. (**Chapter 3**)
- **Wentao Li**, Miao Qiao, Lu Qin, Ying Zhang, Lijun Chang, and Xuemin Lin. “Scaling Up Distance Labeling on Graphs with Core-Periphery Properties.” In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pp. 1367-1381. 2020. (**Chapter 4**)
- **Wentao Li**, Miao Qiao, Lu Qin, Ying Zhang, Lijun Chang, and Xuemin Lin. “Exacting Eccentricity for Small-World Networks.” In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pp. 785-796. IEEE, 2018. (**Chapter 5**)
- **Wentao Li**, Miao Qiao, Lu Qin, Ying Zhang, Lijun Chang, and Xuemin Lin. “Eccentricities on Small-World Networks.” *The VLDB Journal* 28, no. 5 (2019): 765-792. (**Chapter 5**)
- **Wentao Li**, Min Gao, Fan Wu, Wenge Rong, Junhao Wen, Lu Qin. “Manipulating Black-Box Networks for Centrality Promotion.” In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021.

TABLE OF CONTENT

CERTIFICATE OF AUTHORSHIP/ORGINALITY	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	v
PUBLICATIONS	vii
TABLE OF CONTENT	viii
LIST OF FIGURES	xi
LIST OF TABLES	xiii
Chapter 1 INTRODUCTION	1
1.1 Parallelized Distance Labeling	1
1.2 Core-Tree Distance Labeling	5
1.3 Eccentricity Computation and Maintenance	7
1.4 Graph Model	10
1.5 Roadmap	11
Chapter 2 LITERATURE REVIEW	12
2.1 Distance Labeling	12
2.1.1 Exact Distance Labeling	12
2.1.2 Approximate Distance Labeling	13
2.1.3 Distance Labeling Using Tree Decomposition	14
2.2 Centrality Measures	15
2.2.1 Eccentricity Centrality	16
2.2.2 Betweenness Centrality	18
2.2.3 Other Centralities	19
Chapter 3 PARALLELIZED DISTANCE LABELING	20
3.1 Chapter Overview	20

3.2	Preliminary	21
3.2.1	Shortest Distance Problem	21
3.2.2	2-Hop Labeling for Distance Queries	21
3.2.3	Prune Landmark Labeling Approach	22
3.2.4	Node Order: Betweenness Centrality	25
3.3	Parallelized Distance Labeling	28
3.3.1	Label Property	28
3.3.2	Order Dependency	30
3.3.3	Distance Dependency	31
3.3.4	The Parallelized Labeling Method	35
3.4	Index Size Reduction	39
3.4.1	Equivalence Relation Reduction	39
3.4.2	Local Minimum Set Elimination	43
3.5	Index Optimization with Novel Node Order	48
3.5.1	Basic Sampling	48
3.5.2	Pool-Based Sampling	50
3.5.3	Improved Betweenness-based Node Order	55
3.6	Experimental Results	58
3.6.1	Test of Parallelism and Compression	58
3.6.2	Test on Node Ordering	65
3.7	Chapter Summary	71
Chapter 4 CORE-TREE DISTANCE LABELING		73
4.1	Chapter Overview	73
4.2	Problem Statement	73
4.3	Existing Solutions	74
4.3.1	2-Hop Labeling	75
4.3.2	Tree Decomposition and Treewidth	76
4.3.3	Hierarchical 2-Hop Labeling	79
4.3.4	Pruned Landmark Labeling	81
4.4	Core Tree Index	81
4.4.1	Limitation of 2-Hop Labeling	82
4.4.2	Core-Periphery Property	83
4.4.3	Core-Tree Decomposition	83
4.4.4	CT-Index Structure	85
4.4.5	CT-Index Query Processing	88
4.5	CT-Index Construction	93
4.6	Experimental Results	98
4.7	Chapter Summary	106

Chapter 5 ECCENTRICITY COMPUTATION AND MAINTENANCE	107
5.1 Chapter Overview	107
5.2 Preliminaries	107
5.2.1 Problem Definition	108
5.2.2 Indexing for Pair-Wise Distance Queries	110
5.3 Eccentricity Computation	111
5.3.1 The State of the Art	112
5.3.2 Problem Analysis	114
5.3.3 Exact Eccentricity Computation for a Node	116
5.3.4 Bound Update Optimization	126
5.4 Eccentricity Maintenance	134
5.4.1 Pairwise Distances Affected by Edge Updates	135
5.4.2 Scope the Maintenance Dirty Nodes	138
5.4.3 Refine the Eccentricity Update	139
5.5 Experiments	144
5.5.1 Eccentricity Computation	148
5.5.2 Eccentricity Maintenance	157
5.6 Chapter Summary	160
Chapter 6 EPILOGUE	161
BIBLIOGRAPHY	163

LIST OF FIGURES

3.1	Graph G	21
3.2	The Execution of PSL from $d = 0, d = 1$ to $d = 2$	38
3.3	Equivalence Relation Reduction	40
3.4	PLL: Degree and Label Size	43
3.5	Local Minimum Set	44
3.6	The Comparison of the Index Time on One Core	59
3.7	The Comparison of the Index Time on 45 Cores	60
3.8	The Comparison of the Index Size	61
3.9	The Comparison of the Query Time	61
3.10	The Effect of Core Number on the Index Time	62
3.11	The Test of Scalability for the Index Time	63
3.12	The Test of Scalability for the Index Size	64
3.13	The Test of Scalability for the Query Time	65
3.14	The Comparison of Node Order Degree and Betweenness	66
3.15	The Effect of the Node Order on the Index size	67
3.16	The Effect of the Node Order on the Query Time	68
3.17	The Effect of the Node Order on the Index Time	68
3.18	The Effect of the Hop Number k on the Index Size (PSL* ₁)	69
3.19	The Effect of the Sampling Time T on the Index Size (PSL* ₁)	70
4.1	Running Example	74
4.2	MDE-based Tree Decomposition of G	76
4.3	The Proof of Lemma 25	82
4.4	Core Tree Decomposition	84
4.5	The CT-Index	88
4.6	The Extended Label Set	90
4.7	The Comparison of the Index Size	99
4.8	The Comparison of the Index Time	101
4.9	The Comparison of the Query Time	101
4.10	The Effect of Bandwidth d	102
4.11	The Test of Scalability for the Index Size	103
4.12	The Test of Scalability for the Index Time	103

LIST OF FIGURES

4.13	The Test of Scalability for the Query Time	104
4.14	The Determination of d	105
5.1	Example Graph G	108
5.2	Eccentricity of Nodes in G	108
5.3	Pruned Landmark Labeling for All Nodes in G	111
5.4	Illustration of Bounded Set and Partial Set ($z = v_2$)	118
5.5	Illustration of $V_{\leq \lambda}^z$ and V' for $\lambda = 1$ ($z = v_2$)	119
5.6	The Process to Compute $ecc(v_9)$ ($z = v_2$)	121
5.7	Computing Eccentricity for All Nodes ($z = v_2$)	122
5.8	Eccentricity Computation ($pool = \{v_1, v_2\}$)	125
5.9	Local Spread: Before and After Updating the Bounds of v_{10}	127
5.10	Original Graph G	136
5.11	Updated Graph G'	136
5.12	$pecc'(v C^a)$	140
5.13	D-Rule: Lemma 53	140
5.14	D-Rule: Lemma 54	141
5.15	D-Rule: Lemma 55	141
5.16	I-Rules: Lemma 56 and 58	142
5.17	I-Rule: Lemma 57	142
5.18	Testing ECC (Varying # Reference Nodes)	149
5.19	Testing ECC-LS (Processing Time for Eccentricity)	150
5.20	Testing ECC-LS (# PLL Index Queries)	151
5.21	Scalability Testing	152
5.22	Testing Distribution of Distance to Reference Nodes	153
5.23	Testing Average Distance to Reference Nodes	153
5.24	Testing the Accuracy of kBFSEcc	155
5.25	Testing the Running Time of kBFSEcc and ECC	155
5.26	The Speedup of ECC-DY over ECC-LS	157
5.27	Scalability Testing for ECC-DY on Superuser	158
5.28	Testing the Average Number of Nodes in Each Step of ECC-DY for Edge Insertion	159

LIST OF TABLES

1.1	Commonly Used Notations	11
2.1	Labeling with Tree Decomposition	15
3.1	The Index of PLL and PSL	32
3.2	Reduce Index Size with Equivalence Relations	42
3.3	Reduced Index Size with Local Minimum Set	46
3.4	Local Minimum Set: Index and Query Time	47
3.5	The Description of the Datasets	60
3.6	The Description of Added Datasets	66
3.7	Overall Index Size Reduction Ratio	71
4.1	The Description of the Datasets	100
4.2	The Comparison Between CT-Index and CD	104
5.1	The Execution of BoundEcc	114
5.2	$ C^a \ll C^b $ on Small-World Networks	137
5.3	Dataset Description and Comparison with the State-of-the-art Methods	146
5.4	Dataset Description and the Results of HybridEcc	154
5.5	Testing ECC on Road Networks (Sec)	156

LIST OF TABLES
