SPRINGER BRIEFS IN COMPUTER SCIENCE

Yixiang Fang
Kai Wang
Xuemin Lin
Wenjie Zhang

Cohesive Subgraph Search Over Large Heterogeneous Information Networks

Springer

SpringerBriefs in Computer Science

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic.

Typical topics might include:

- A timely report of state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

Briefs allow authors to present their ideas and readers to absorb them with minimal time investment. Briefs will be published as part of Springer's eBook collection, with millions of users worldwide. In addition, Briefs will be available for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, easy-to-use manuscript preparation and formatting guidelines, and expedited production schedules. We aim for publication 8–12 weeks after acceptance. Both solicited and unsolicited manuscripts are considered for publication in this series.

**Indexing: This series is indexed in Scopus, Ei-Compendex, and zbMATH **

More information about this series at https://link.springer.com/bookseries/10028

Yixiang Fang • Kai Wang • Xuemin Lin •
Wenjie Zhang

Cohesive Subgraph Search Over Large Heterogeneous Information Networks

Springer

Yixiang Fang
School of Data Science
The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong, China

Kai Wang
Antai College of Economics & Management
Shanghai Jiao Tong University
Shanghai, China

Xuemin Lin
Antai College of Economics & Management
Shanghai Jiao Tong University
Shanghai, China

Wenjie Zhang
Computer Science and Engineering
The University of New South Wales
Sydney, NSW, Australia

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-3-030-97567-8 ISBN 978-3-030-97568-5 (eBook)
https://doi.org/10.1007/978-3-030-97568-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To the two groups of people who greatly support our research: ***our family members*** *and* ***our collaborators***

Preface

With the advent of a wide spectrum of recent applications, querying large heterogeneous information networks (HINs) has received a great deal of attention from both academic and industrial societies. HINs involve objects (vertices) and links (edges) that are classified into multiple types; examples include bibliography networks, social networks, knowledge networks, and user-item networks in E-business. An important component of these HINs is the cohesive subgraph, or subgraph containing vertices that are densely connected internally. Searching cohesive subgraphs over HINs has been found useful in many real-world applications, such as community search, product recommendation, and fraud detection. Consequently, how to design effective cohesive subgraph models (CSMs) and how to efficiently perform cohesive subgraph search (CSS) on large HINs has become important research topics in the era of big data.

The main purpose of this book is to thoroughly survey the recent technical developments on efficiently performing CSS, in view of the fact that many real graphs are usually large HINs. To have a whole picture of reviewing these works, we classify them according to the classic cohesiveness metrics such as core, truss, clique, connectivity, and density. Meanwhile, since the bipartite network is a special representative type of HINs that can often be processed in a special manner, we also classify HINs into two categories, namely bipartite networks and other general HINs (that are not only customized for bipartite networks). To review these works, we extensively discuss the specific models and their corresponding search solutions.

Moreover, we analyze and compare these CSMs and solutions systematically. Specifically, we first compare different groups of CSMs and analyze their common features and different features from multiple perspectives such as cohesiveness constraints, shared properties, and computational efficiency. Then, for the CSMs in each group, we analyze and compare their model properties (e.g., parameters and constraints) and high-level algorithm ideas. Note that since the bipartite network is a special case of HINs, all the models developed for general HINs can be directly applied to bipartite networks, but the models customized for bipartite networks may not be easily extended for other general HINs due to their restricted settings. Besides, we point out a list of promising research directions in this field.

We believe that this book does not only help researchers to have a better understanding of existing CSMs and solutions but also provides them interesting insights for future research. Consequently, the book can be used either as an extended survey for researchers who are interested in conducting research on cohesive subgraph computation over large HINs, or as a reference book for postgraduate students who are learning courses on the related topics, or as a guideline book for industry engineers to solve real problems using CSS solutions.

Organization The book is organized as follows:

In Chap. 1, we focus on providing the necessary background of CSS over large HINs and giving an introduction to the research field to highlight the popularity and applications in the topic of CSS. More specifically, we first show a list of example HINs (e.g., DBLP, Facebook, and Yago) and typical applications of CSS over HINs (e.g., fraud detection, community search, product recommendation, and biological data analysis). Then, we discuss the challenges of conducting CSS over large HINs. Finally, we make a classification of existing works according to the classic cohesiveness metrics (i.e., core, truss, clique, connectivity, and density).

In Chap. 2, we aim to present the preliminaries of performing cohesive subgraph search. We first formally introduce the data models of HINs and bipartite networks, and then we review typical classic cohesive subgraph models on homogeneous networks, including k-core, k-truss, k-clique, k-edge-connectivity component (k-ECC), and the densest subgraphs.

In Chap. 3, we extensively introduce the five groups of CSMs and solutions for the bipartite networks, which are core-, truss-, clique-, connectivity-, and density-based models and solutions.

In Chap. 4, we extensively introduce the four groups of CSMs and solutions for other general HINs, which are core-, truss-, clique-, and density-based models and solutions. We also review one model that is not covered by the groups above.

In Chap. 5, we perform a thorough analysis and comparison of different CSMs and their corresponding solutions on bipartite networks and other general HINs, respectively, by highlighting their advantages and disadvantages, such as analyzing their computational complexities and application scenarios.

In Chap. 6, we review the two groups of works that are highly related to the topic of our book, which are CSS on homogeneous networks and HIN clustering. In particular, for the first group, we mainly discuss the representative works of five CSMs on conventional homogeneous networks.

In Chap. 7, we discuss several promising future research directions about CSS over HINs, including novel application-driven CSMs, efficient search algorithms, parameter optimization, and an online repository for collecting HIN datasets, tools, and algorithm codes, which can provide researchers with some good starting points to work in this area. In addition, we draw a brief conclusion for the book.

Acknowledgments This book was partially supported by NSFC under grant 62102341, CUHK-SZ grant UDF01002139, National Key R&D Program of China under grant 2018AAA0102502, GuangDong Basic and Applied Basic Research Foundation 2019B1515120048, and Australian Research Council Discovery Projects (DP200101338, DP210101393, DP200101116).

Shenzhen, Guangdong, China — Yixiang Fang
Shanghai, China — Kai Wang
Shanghai, China — Xuemin Lin
Sydney, NSW, Australia — Wenjie Zhang
December 2021