Individual and Collective Graph Mining

Principles, Algorithms, and Applications

Synthesis Lectures on Data Mining and Knowledge Discovery

Editor

Jiawei Han, ***University of Illinois at Urbana-Champaign***
Lise Getoor, ***University of California, Santa Cruz***
Wei Wang, ***University of California, Los Angeles***
Johannes Gehrke, ***Cornell University***
Robert Grossman, ***University of Chicago***

Synthesis Lectures on Data Mining and Knowledge Discovery is edited by Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, and Robert Grossman. The series publishes 50- to 150-page publications on topics pertaining to data mining, web mining, text mining, and knowledge discovery, including tutorials and case studies. Potential topics include: data mining algorithms, innovative data mining applications, data mining systems, mining text, web and semi-structured data, high performance and parallel/distributed data mining, data mining standards, data mining and knowledge discovery framework and process, data mining foundations, mining data streams and sensor data, mining multi-media data, mining social networks and graph data, mining spatial and temporal data, pre-processing and post-processing in data mining, robust and scalable statistical methods, security, privacy, and adversarial data mining, visual data mining, visual analytics, and data visualization.

Individual and Collective Graph Mining: Principles, Algorithms, and Applications
Danai Koutra and Christos Faloutsos
2017

Phrase Mining from Massive Text and Its Applications
Jialu Liu, Jingbo Shang, and Jiawei Han
2017

Exploratory Causal Analysis with Time Series Data
James M. McCracken
2016

Mining Human Mobility in Location-Based Social Networks
Huiji Gao and Huan Liu
2015

Mining Latent Entity Structures
Chi Wang and Jiawei Han
2015

Probabilistic Approaches to Recommendations
Nicola Barbieri, Giuseppe Manco, and Ettore Ritacco
2014

Outlier Detection for Temporal Data
Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han
2014

Provenance Data in Social Media
Geoffrey Barbier, Zhuo Feng, Pritam Gundecha, and Huan Liu
2013

Graph Mining: Laws, Tools, and Case Studies
D. Chakrabarti and C. Faloutsos
2012

Mining Heterogeneous Information Networks: Principles and Methodologies
Yizhou Sun and Jiawei Han
2012

Privacy in Social Networks
Elena Zheleva, Evimaria Terzi, and Lise Getoor
2012

Community Detection and Mining in Social Media
Lei Tang and Huan Liu
2010

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions
Giovanni Seni and John F. Elder
2010

Modeling and Data Mining in Blogosphere
Nitin Agarwal and Huan Liu
2009

Individual and Collective Graph Mining: Principles, Algorithms, and Applications

Danai Koutra and Christos Faloutsos

ISBN: 978-03-031-00783-5 paperback
ISBN: 978-3-031-01911-1 ebook

DOI 10.1007/978-3-031-01911-1

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Lecture #14
Series Editors: Jiawei Han, *University of Illinois at Urbana-Champaign*
Lise Getoor, *University of California, Santa Cruz*
Wei Wang, *University of California, Los Angeles*
Johannes Gehrke, *Cornell University*
Robert Grossman, *University of Chicago*
Series ISSN
Print 2151-0067 Electronic 2151-0075

Individual and Collective Graph Mining

Principles, Algorithms, and Applications

Danai Koutra
University of Michigan, Ann Arbor

Christos Faloutsos
Carnegie Mellon University

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY #14

ABSTRACT

Graphs naturally represent information ranging from links between web pages, to communication in email networks, to connections between neurons in our brains. These graphs often span *billions* of nodes and interactions between them. Within this deluge of interconnected data, how can we find the most important structures and summarize them? How can we efficiently visualize them? How can we detect anomalies that indicate critical events, such as an attack on a computer system, disease formation in the human brain, or the fall of a company?

This book presents scalable, principled discovery algorithms that combine globality with locality to make sense of one or more graphs. In addition to fast *algorithmic methodologies*, we also contribute *graph-theoretical ideas and models*, and real-world *applications* in two main areas.

- **Individual Graph Mining**: We show how to interpretably *summarize* a single graph by identifying its important graph structures. We complement summarization with *inference*, which leverages information about few entities (obtained via summarization or other methods) and the network structure to efficiently and effectively learn information about the unknown entities.
- **Collective Graph Mining**: We extend the idea of individual-graph *summarization* to time-evolving graphs, and show how to scalably discover temporal patterns. Apart from summarization, we claim that *graph similarity* is often the underlying problem in a host of applications where multiple graphs occur (e.g., temporal anomaly detection, discovery of behavioral patterns), and we present principled, scalable algorithms for aligning networks and measuring their similarity.

The methods that we present in this book leverage techniques from diverse areas, such as matrix algebra, graph theory, optimization, information theory, machine learning, finance, and social science, to solve real-world problems. We present applications of our exploration algorithms to massive datasets, including a Web graph of 6.6 billion edges, a Twitter graph of 1.8 billion edges, brain graphs with up to 90 million edges, collaboration, peer-to-peer networks, browser logs, all spanning millions of users and interactions.

KEYWORDS

data mining, graph mining and exploration, graph similarity, graph matching, network alignment, graph summarization, pattern mining, outlier detection, anomaly detection, scalability, fast algorithms, models, visualization, social networks, brain graphs, connectomes

Contents

Acknowledgments

This research was sponsored by the National Science Foundation under grant numbers IIS-1151017415, IIS-1217559, and IIS-1408924, the Department of Energy/National Nuclear Security Administration under grant number DE-AC52-07NA27344, the Defense Advanced Research Projects Agency under grant number W911NF-11-C-0088, the Air Force Research Laboratory under grant number F8750-11-C-0115, and the US Army Research Lab under grant number W911NF-09-2-0053. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

The authors would also like to thank all their collaborators for contributing to the works presented in this book: Polo Chau, Wolfgang Gatterbauer, Brian Gallagher, Stephan Guennemann, U Kang, Tai-You Ke, David Lubensky, Hsing-Kuo (Kenneth) Pao, Neil Shah, Hanghang Tong, Jilles Vreeken, and Joshua Vogelstein. We especially thank Neil Shah for his major contributions to Chapter 4: Summarization of Dynamic Graphs.

We also thank the reviewers of this book—Austin Benson, George Karypis, and Philip Yu—for their thoughtful feedback and suggestions which helped improve this book. Last but not least, we thank the informal reviewer, Haoming Shen, for carefully reading the manuscript, pointing out inconsistencies, and helping edit it.

Danai Koutra and Christos Faloutsos
October 2017