Graph Mining

Laws, Tools, and Case Studies

Synthesis Lectures on Data Mining and Knowledge Discovery

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. The scope will largely follow the purview of premier computer science conferences, such as KDD. Potential topics include, but not limited to, 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.

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

Graph Mining: Laws, Tools, and Case Studies

D. Chakrabarti and C. Faloutsos

ISBN: 978-3-031-00775-0 paperback
ISBN: 978-3-031-01903-6 ebook

DOI 10.1007/978-3-031-01903-6

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

Lecture #6
Series Editors: Jiawei Han, *UIUC*
Lise Getoor, *University of Maryland*
Wei Wang, *University of North Carolina, Chapel Hill*
Johannes Gehrke, *Cornell University*
Robert Grossman, *University of Chicago*
Series ISSN
Synthesis Lectures on Data Mining and Knowledge Discovery
Print 2151-0067 Electronic 2151-0075

Graph Mining

Laws, Tools, and Case Studies

D. Chakrabarti
Facebook

C. Faloutsos
CMU

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY
#6

ABSTRACT

What does the Web look like? How can we find patterns, communities, outliers, in a social network? Which are the most central nodes in a network? These are the questions that motivate this work. Networks and graphs appear in many diverse settings, for example in social networks, computer-communication networks (intrusion detection, traffic management), protein-protein interaction networks in biology, document-text bipartite graphs in text retrieval, person-account graphs in financial fraud detection, and others.

In this work, first we list several surprising patterns that real graphs tend to follow. Then we give a detailed list of generators that try to mirror these patterns. Generators are important, because they can help with "what if" scenarios, extrapolations, and anonymization. Then we provide a list of powerful tools for graph analysis, and specifically spectral methods (Singular Value Decomposition (SVD)), tensors, and case studies like the famous "pageRank" algorithm and the "HITS" algorithm for ranking web search results. Finally, we conclude with a survey of tools and observations from related fields like sociology, which provide complementary viewpoints.

KEYWORDS

data mining, social networks, power laws, graph generators, pagerank, singular value decomposition.

Christos Faloutsos*: To Christina, for her patience, support, and down-to-earth questions; to Michalis and Petros, for the '99 paper that started it all.*

Deepayan Chakrabarti*: To Purna and my parents, for their support and help, and for always being there when I needed them.*

Contents

Acknowledgments

Several of the results that we present in this book were possible thanks to research funding from the National Science Foundation (grants IIS-0534205, IIS-0705359, IIS0808661, IIS-1017415), Defense Advanced Research Program Agency (contracts W911NF-09-2-0053, HDTRA1-10-1-0120, W911NF-11-C-0088), Lawrence Livermore National Laboratories (LLNL), IBM, Yahoo, and Google. Special thanks to Yahoo, for allowing access to the *M45* hadoop cluster. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors, and do not necessarily reflect the views of the National Science Foundation, DARPA, LLNL, or other funding parties.

D. Chakrabarti and C. Faloutsos
September 2012