

Graph Mining

Laws, Tools, and Case Studies

Synthesis Lectures on Data Mining and Knowledge Discovery

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*

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.

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

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2012

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

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	xv
1	Introduction	1
	 PART I Patterns and Laws	 7
2	Patterns in Static Graphs	9
2.1	S-1: Heavy-tailed Degree Distribution	9
2.2	S-2: Eigenvalue Power Law (EPL)	11
2.3	S-3 Small Diameter	12
2.4	S-4, S-5: Triangle Power Laws (TPL, DTPL)	15
3	Patterns in Evolving Graphs	19
3.1	D-1: Shrinking Diameters	19
3.2	D-2: Densification Power Law (DPL)	19
3.3	D-3: Diameter-plot and Gelling Point	21
3.4	D-4: Oscillating NLCCs Sizes	21
3.5	D-5: LPL: Principal Eigenvalue Over Time	24
4	Patterns in Weighted Graphs	27
4.1	W-1: Snapshot Power Laws (SPL)—“Fortification”	27
4.2	DW-1: Weight Power Law (WPL)	28
4.3	DW-2: LWPL: Weighted Principal Eigenvalue Over Time	30
5	Discussion—The Structure of Specific Graphs	31
5.1	The Internet	31
5.2	The World Wide Web (WWW)	31
6	Discussion—Power Laws and Deviations	35
6.1	Power Laws—Slope Estimation	35

6.2	Deviations from Power Laws	36
6.2.1	Exponential Cutoffs	37
6.2.2	Lognormals or the “DGX” Distribution	37
6.2.3	Doubly-Pareto Lognormal (<i>dPln</i>)	38
7	Summary of Patterns	41
 PART II Graph Generators		43
8	Graph Generators	45
8.1	Random Graph Models	46
8.1.1	The Erdős-Rényi Random Graph Model	46
8.1.2	Generalized Random Graph Models	51
9	Preferential Attachment and Variants	53
9.1	Main Ideas	53
9.1.1	Basic Preferential Attachment	53
9.1.2	Initial Attractiveness	56
9.1.3	Internal Edges and Rewiring	57
9.2	Related Methods	58
9.2.1	Edge Copying Models	58
9.2.2	Modifying the Preferential Attachment Equation	60
9.2.3	Modeling Increasing Average Degree	61
9.2.4	Node Fitness Measures	62
9.2.5	Generalizing Preferential Attachment	62
9.2.6	PageRank-based Preferential Attachment	63
9.2.7	The Forest Fire Model	64
9.3	Summary of Preferential Attachment Models	65
10	Incorporating Geographical Information	67
10.1	Early Models	67
10.1.1	The Small-World Model	67
10.1.2	The Waxman Model	69
10.1.3	The BRITE Generator	70
10.1.4	Other Geographical Constraints	71
10.2	Topology from Resource Optimizations	72

10.2.1	The Highly Optimized Tolerance Model	72
10.2.2	The Heuristically Optimized Tradeoffs Model	73
10.3	Generators for the Internet Topology	75
10.3.1	Structural Generators	75
10.3.2	The Inet Topology Generator	76
10.4	Comparison Studies	77
11	The <i>RMat</i> (Recursive MATrix) Graph Generator	81
12	Graph Generation by Kronecker Multiplication	87
13	Summary and Practitioner's Guide	91
 PART III Tools and Case Studies		93
14	SVD, Random Walks, and Tensors	95
14.1	Eigenvalues—Definition and Intuition	95
14.2	Singular Value Decomposition (SVD)	97
14.3	<i>HITS</i> : Hubs and Authorities	101
14.4	PageRank	103
15	Tensors	107
15.1	Introduction	107
15.2	Main Ideas	107
15.3	An Example: Tensors at Work	108
15.4	Conclusions—Practitioner's Guide	110
16	Community Detection	113
16.1	Clustering Coefficient	113
16.2	Methods for Extracting Graph Communities	115
16.3	A Word of Caution—"No Good Cuts"	119
17	Influence/Virus Propagation and Immunization	123
17.1	Introduction—Terminology	123
17.2	Main Result and its Generality	125
17.3	Applications	129

17.4	Discussion	131
17.4.1	Simulation Examples	131
17.4.2	λ_1 : Measure of Connectivity	132
17.5	Conclusion	133
18	Case Studies	135
18.1	Proximity and Random Walks	135
18.2	Automatic Captioning—Multi-modal Querying	137
18.2.1	The <i>GCap</i> Method	137
18.2.2	Performance and Variations	139
18.2.3	Discussion	140
18.2.4	Conclusions	140
18.3	Center-Piece Subgraphs—Who is the Mastermind?	140
PART IV Outreach—Related Work		145
19	Social Networks	147
19.1	Mapping Social Networks	147
19.2	Dataset Characteristics	148
19.3	Structure from Data	148
19.4	Social “Roles”	149
19.5	Social Capital	152
19.6	Recent Research Directions	152
19.7	Differences from Graph Mining	153
20	Other Related Work	155
20.1	Relational Learning	155
20.2	Finding Frequent Subgraphs	156
20.3	Navigation in Graphs	157
20.3.1	Methods of Navigation	157
20.3.2	Relationship Between Graph Topology and Ease of Navigation	158
20.4	Using Social Networks in Other Fields	160
21	Conclusions	161
21.1	Future Research Directions	162
21.2	Parting Thoughts	163

Resources	165
Bibliography	167
Authors' Biographies	191

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