

Springer Series in Statistics

Advisors:

P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger

Springer Series in Statistics

Athreya/Lahiri: Measure Theory and Probability Theory
Bilodeau/Brenner: Theory of Multivariate Statistics
Brockwell/Davis: An Introduction to Time Series and Forecasting
Carmona: Statistical Analysis of Financial Data in S-PLUS
Chow/Teicher: Probability Theory: Independence, Interchangeability, Martingales, 3rd ed.
Christensen: Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data;
Nonparametric Regression and Response Surface Maximization, 2nd ed.
Christensen: Log-Linear Models and Logistic Regression, 2nd ed.
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, 2nd ed.
Cryer/Chan: Time Series Analysis, Second Edition
DasGupta: Asymptotic Theory of Statistics and Probability
Davis: Statistical Methods for the Analysis of Repeated Measurements
Dean/Voss: Design and Analysis of Experiments
Dekking/Kraaikamp/Lopuhaä/Meester: A Modern Introduction to Probability and Statistics
Durrett: Essential of Stochastic Processes
Edwards: Introduction to Graphical Modeling, 2nd ed.
Everitt: An R and S-PLUS Companion to Multivariate Analysis
Gentle: Matrix Algebra: Theory, Computations, and Applications in Statistics
Ghosh/Delampady/Samanta: An Introduction to Bayesian Analysis
Gut: Probability: A Graduate Course
Heiberger/Holland: Statistical Analysis and Data Display; An Intermediate Course with
Examples in S-PLUS, R, and SAS
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II Categorical and Multivariate
Methods
Karr: Probability
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Kolaczyk: Statistical Analysis of Network Data
Lange: Applied Probability
Lange: Optimization
Lehmann: Elements of Large Sample Theory
Lehmann/Romano: Testing Statistical Hypotheses, 3rd ed.
Lehmann/Casella: Theory of Point Estimation, 2nd ed.
Longford: Studying Human Populations: An Advanced Course in Statistics
Marin/Robert: Bayesian Core: A Practical Approach to Computational Bayesian Statistics
Nolan/Speed: Stat Labs: Mathematical Statistics Through Applications
Pitman: Probability
Rawlings/Pantula/Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: From Decision-Theoretic Foundations to Computational
Implementation, 2nd ed.
Robert/Casella: Monte Carlo Statistical Methods, 2nd ed.
Rose/Smith: Mathematical Statistics with *Mathematica*
Ruppert: Statistics and Finance: An Introduction

(continued after index)

Eric D. Kolaczyk

Statistical Analysis of Network Data

Methods and Models

 Springer

Eric D. Kolaczyk
Department of Mathematics & Statistics
Boston University
111 Cummington St.
Boston MA 02215
USA

ISBN 978-0-387-88145-4 e-ISBN 978-0-387-88146-1
DOI 10.1007/978-0-387-88146-1

Library of Congress Control Number: 2009921812

© Springer Science+Business Media, LLC 2009

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

springer.com

To the ‘network’ that is my family ...
... those both near and far.

In Memorium

Br. Robert C. Ruhl, CSV

March 18, 1947 – Feb 2, 2008

“A teacher affects eternity; he can never tell where his influence stops.”

(Henry Brooks Adams)

Preface

In recent years there has been an explosion of *network data* – that is, measurements that are either of or from a system conceptualized as a network – from seemingly all corners of science. The combination of an increasingly pervasive interest in scientific analysis at a systems level and the ever-growing capabilities for high-throughput data collection in various fields has fueled this trend. Researchers from biology and bioinformatics to physics, from computer science to the information sciences, and from economics to sociology are more and more engaged in the collection and statistical analysis of data from a network-centric perspective.

Accordingly, the contributions to statistical methods and modeling in this area have come from a similarly broad spectrum of areas, often independently of each other. Many books already have been written addressing network data and network problems in specific individual disciplines. However, there is at present no single book that provides a modern treatment of a core body of knowledge for statistical analysis of network data that cuts across the various disciplines and is organized rather according to a statistical taxonomy of tasks and techniques. This book seeks to fill that gap and, as such, it aims to contribute to a growing trend in recent years to facilitate the exchange of knowledge across the pre-existing boundaries between those disciplines that play a role in what is coming to be called ‘network science.’

The book is written for students and researchers with a ‘mature’ knowledge of statistics and hence is intended not only for statisticians but also for people involved with network data in various other areas, like those mentioned above. Background in calculus and linear algebra and some reasonable foundation in statistics and probability are expected. Beyond that, I have attempted to build all necessary material as needed.

In an effort to reach this admittedly diverse audience successfully, I have aimed in each chapter to communicate the material in a manner that strikes an appropriate balance between concepts, on the one hand, and technical depth and rigor, on the other. It is expected that the interested reader will want – and, indeed, is encouraged – to pursue the relevant primary sources for details I may have chosen to omit. The book is in this sense intended to serve as an *entrée* to the larger literature. Copious use of references has been made throughout the book for this very purpose. In addi-

tion, the exercises at the end of each chapter provide further opportunities to explore some of the topics in greater depth. There are both analytical and computational exercises to be found, with the latter frequently designed to be fairly open-ended in nature, so as to encourage exploration. Finally, the methods and models presented herein are illustrated throughout the book with examples from a wide range of disciplines. I have found this overall approach to the pedagogy of the material to work well when I taught classes of precisely the diversity that I envision for the readership of this book.

The book itself would not have been possible without the help, feedback, and support of many. I thank my editor, John Kimmel, and the other folks at Springer, for their help and guidance throughout the publication process. The idea for the writing of this book arose while I was on sabbatical during the 2004-2005 academic year, in the Laboratoire d'Informatique Algorithmique, Fondements, et Applications (LIAFA), at l'Université Paris 7, courtesy of CNRS, and in the Department of Statistics, at Harvard University. My own work in this area has been generously supported by grants from the United States National Institutes of Health (NIH), National Science Foundation (NSF), and Office of Naval Research (ONR). Students and colleagues attending the courses I taught on this material, at Harvard University in Spring 2005 and at Boston University in Fall 2005, are gratefully acknowledged for their interest, questions, and hard work. Many of the datasets used in this book were generously shared by colleagues from around the world, including Ignacio Alvarez-Hamelin, Mark Coates, Mark Crovella, Tim Gardner, Sucharita Gopal, Boris Hayete, Mark Kramer, Emmanuel Lazergas, Naoki Nariai, Robert Nowak, Xiaoyu Jiang, and Fabien Viger. In addition, Ignacio Alvarez-Hamelin, Kevin Boyack, Ulrik Brandes, Rui Castro, Sucharita Gopal, and Mark Kramer were kind enough to produce and share some of the more stunning figures found herein. For graciously responding to my unabashed solicitations for feedback on material in various chapters, I am indebted to Ignacio Alvarez-Hamelin, Alain Barrat, Ulrik Brandes, Tom Britton, Hugh Chipman, Mark Crovella, Tim Gardner, Boris Hayete, Peter Hoff, David Hunter, Xiaoyu Jiang, Simon Kasif, Naoki Nariai, Robert Nowak, Pip Pattison, Mike Rabbat, Garry Robbins, Martin Steffen, Shu Yang, and Ji Zhu. Special thanks in this regard are due to Joe Whittaker, whose comments were extensive and invariably helpful. Notwithstanding all of this feedback, however, any typos, mistakes, and other similar errors are of course my own. Andrej Mrvar, of the Pajek software team, is also to be thanked for his help in responding to my numerous questions. Finally, although surely inadequate in measure, I offer my heartfelt thanks to my wife, Josée Dupuis, not only for her love, patience, and support over the many months that I was submersed in the writing of this book, and for her reading of the final manuscript, but also for serving as my **R** guru, fielding even my many late-night queries when she would surely have rather preferred to continue uninterrupted with her own work!

Eric D. Kolaczyk
Boston, Massachusetts
March, 2009

Contents

1	Introduction and Overview	1
1.1	Why Networks?	1
1.2	Examples of Networks	3
1.2.1	Technological Networks	3
1.2.2	Social Networks	5
1.2.3	Biological Networks	7
1.2.4	Information Networks	9
1.3	About this Book	11
2	Preliminaries	15
2.1	Background on Graphs	15
2.1.1	Basic Definitions and Concepts	16
2.1.2	Families of Graphs	18
2.1.3	Graphs and Matrix Algebra	20
2.1.4	Graph Data Structures and Algorithms	21
2.2	Background in Probability and Statistics	24
2.2.1	Probability	25
2.2.2	Principles of Statistical Inference	31
2.2.3	Methods of Statistical Inference: Tutorials	32
2.3	Statistical Analysis of Network Data: <i>Prelude</i>	42
2.4	Additional Related Topics and Reading	45
	Exercises	45
3	Mapping Networks	49
3.1	Introduction	49
3.2	Collecting Relational Network Data	50
3.2.1	Measurement of System Elements and Interactions	51
3.2.2	Enumerated, Partial, and Sampled Data	54
3.3	Constructing Network Graph Representations	56
3.4	Visualizing Network Graphs	58
3.4.1	Elements of Graph Visualization	58

3.4.2	Methods of Graph Visualization	60
3.5	Case Studies	63
3.5.1	Mapping 'Science'	65
3.5.2	Mapping the Internet	68
3.6	Mapping Dynamic Networks	74
3.7	Additional Related Topics and Reading	76
	Exercises	77
4	Descriptive Analysis of Network Graph Characteristics	79
4.1	Introduction	79
4.2	Vertex and Edge Characteristics	80
4.2.1	Degree	80
4.2.2	Centrality	88
4.3	Characterizing Network Cohesion	94
4.3.1	Local Density	94
4.3.2	Connectivity	97
4.3.3	Graph Partitioning	102
4.3.4	Assortativity and Mixing	111
4.4	Case Study: Analysis of an Epileptic Seizure	114
4.5	Characterizing Dynamic Network Graphs	116
4.6	Additional Related Topics and Reading	119
	Exercises	120
5	Sampling and Estimation in Network Graphs	123
5.1	Introduction	123
5.2	Background on Statistical Sampling Theory	126
5.2.1	Horvitz-Thompson Estimation for Totals	126
5.2.2	Estimation of Group Size	129
5.3	Common Network Graph Sampling Designs	131
5.3.1	Induced and Incident Subgraph Sampling	131
5.3.2	Star and Snowball Sampling	133
5.3.3	Link Tracing	136
5.4	Estimation of Totals in Network Graphs	137
5.4.1	Vertex Totals	137
5.4.2	Totals on Vertex Pairs	138
5.4.3	Totals of Higher Order	141
5.4.4	Effects of Design, Measurement, and Total	143
5.5	Estimation of Network Group Size	145
5.6	Other Network Graph Estimation Problems	149
5.7	Additional Related Topics and Reading	151
	Exercises	151

6	Models for Network Graphs	153
6.1	Introduction	153
6.2	Random Graph Models	154
6.2.1	Classical Random Graph Models	156
6.2.2	Generalized Random Graph Models	158
6.2.3	Simulating Random Graph Models	159
6.2.4	Statistical Application of Random Graph Models	162
6.3	Small-World Models	169
6.3.1	The Watts-Strogatz Model	169
6.3.2	Other Small-World Network Models	171
6.4	Network Growth Models	172
6.4.1	Preferential Attachment Models	173
6.4.2	Copying Models	176
6.4.3	Fitting Network Growth Models	178
6.5	Exponential Random Graph Models	180
6.5.1	Model Specification	180
6.5.2	Fitting Exponential Random Graph Models	185
6.5.3	Goodness-of-Fit and Model Degeneracy	187
6.5.4	Case Study: Modeling Collaboration Among Lawyers	188
6.6	Challenges in Modeling Network Graphs	191
6.7	Additional Related Topics and Reading	193
	Exercises	195
7	Network Topology Inference	197
7.1	Introduction	197
7.2	Link Prediction	199
7.2.1	Informal Scoring Methods	201
7.2.2	Probabilistic Classification Methods	202
7.2.3	Case Study: Predicting Lawyer Collaboration	205
7.3	Inference of Association Networks	207
7.3.1	Correlation Networks	209
7.3.2	Partial Correlation Networks	212
7.3.3	Gaussian Graphical Model Networks	216
7.3.4	Case Study: Inferring Genetic Regulatory Interactions	220
7.4	Tomographic Network Topology Inference	223
7.4.1	Tomographic Inference of Tree Topologies	225
7.4.2	Methods Based on Hierarchical Clustering	228
7.4.3	Likelihood-based Methods	231
7.4.4	Summarizing Collections of Trees	234
7.4.5	Case Study: Computer Network Topology Identification	236
7.5	Additional Related Topics and Reading	241
	Exercises	242

8	Modeling and Prediction for Processes on Network Graphs	245
8.1	Introduction	245
8.2	Nearest Neighbor Prediction	246
8.3	Markov Random Fields	249
8.3.1	Markov Random Field Models	249
8.3.2	Inference and Prediction for Markov Random Fields	252
8.3.3	Related Probabilistic Models	256
8.4	Kernel-based Regression	257
8.4.1	Kernel Regression on Graphs	258
8.4.2	Designing Kernels on Graphs	262
8.5	Case Study: Predicting Protein Function	266
8.6	Modeling and Prediction for Dynamic Processes	271
8.6.1	Epidemic Processes: An Illustration	272
8.6.2	Other Dynamic Processes	280
8.7	Additional Related Topics and Reading	281
	Exercises	282
9	Analysis of Network Flow Data	285
9.1	Introduction	285
9.2	Gravity Models	287
9.2.1	Model Specification	288
9.2.2	Inference for Gravity Models	292
9.3	Traffic Matrix Estimation	297
9.3.1	Static Methods	298
9.3.2	Dynamic Methods	306
9.3.3	Case Study: Internet Traffic Matrix Estimation	310
9.4	Estimation of Network Flow Costs	316
9.4.1	Link Costs from End-to-end Measurements	317
9.4.2	Path Costs from End-to-end Measurements	321
9.5	Additional Related Topics and Reading	328
	Exercises	330
10	Graphical Models	333
10.1	Introduction	333
10.2	Defining Graphical Models	334
10.2.1	Directed Graphical Models	335
10.2.2	Undirected Graphical Models	339
10.3	Inference for Graphical Models	342
10.4	Additional Related Topics and Reading	344
	Glossary of Notation	345
	References	347
	Author Index	373
	Subject Index	381