Network Models for Data Science

This text on the theory and applications of network science is aimed at beginning graduate students in statistics, data science, computer science, machine learning, and mathematics, as well as advanced students in business, computational biology, physics, social science, and engineering working with large, complex relational datasets. It provides an exciting array of analysis tools, including probability models, graph theory, and computational algorithms, exposing students to ways of thinking about types of data that are different from typical statistical data. Concepts are demonstrated in the context of real applications, such as relationships between financial institutions, between genes or proteins, between neurons in the brain, and between terrorist groups. Methods and models described in detail include random graph models, percolation processes, methods for sampling from huge networks, network partitioning, and community detection. In addition to static networks, the book introduces dynamic networks such as epidemics, where time is an important component.

Alan Julian Izenman is Professor of Statistics, Operations, and Data Science at Temple University. He received his Ph.D. from the University of California, Berkeley. He was a faculty member at Tel Aviv University and Colorado State University, and was a visiting faculty member at the University of Chicago, the University of Minnesota, Stanford University, and the University of Edinburgh. He was Program Director of Statistics and Probability at NSF (1992–94). A Fellow of the ASA, IMS, RSS, and ISI, he has served on the Editorial Boards of *JASA, Law, Probability, and Risk*, and *Statistical Analysis and Data Mining*. He is the author of *Modern Multivariate Statistical Techniques* (2013).

"Izenman offers readers an extensive set of descriptions of models for networks. Emphasis is on random networks, with a chapter devoted to parametric statistical models for dependent relational ties. The author is to be praised for pulling together different material into a single, very useful book."

– Stanley Wasserman, Indiana University

Network Models for Data Science

Theory, Algorithms, and Applications

Alan Julian Izenman

Temple University, Philadelphia

CAMBRIDGE UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108835763

DOI: 10.1017/9781108886666

First published 2023

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-83576-3 Hardback

Additional resources for this publication at www.cambridge.org/izenman

This book is dedicated
to the loves of my life,
Betty-Ann and Kayla

Contents

Preface

In recent years, we have witnessed many incredible technological breakthroughs, including huge reductions in the sizes of computational devices, major innovations in storage facilities that have allowed researchers to create enormous data repositories and databases at little cost, an explosion in the variety of types and sizes of data being collected, and vast improvements in computational speed to process such data. These amazing achievements have led, in turn, to the "big data revolution" and to the introduction of the discipline of *data science*.

As a rapidly growing new field, data science is generally viewed as a unification of the disciplines of probability, statistics, machine learning, data mining, database management systems, artificial intelligence, and algorithm development. Like statistics, data science deals with the acquisition, processing, analysis, and interpretation of a wide variety of types of digital data. Such data are obtained from multiple sources, where different data types require different techniques and computational tools. Today, the amount of digital information has grown exponentially, leading to the era of "big data," especially in the areas of healthcare, business, science, and government programs.

This book concentrates on an important data type encountered in machine learning, data mining, and data science, namely, *relational data*, which record connections, for example, between people who make up friendship networks. Relational data are visualized through a network graph and studied using probabilistic models and statistical analysis. The study of networks can help in understanding and visualizing complex information often buried in large relational datasets collected on financial, biological, physics, social, business, and technological applications. Understanding relational data is of great importance in characterizing the geometry and structure of complex data networks. The combination of probability models, statistical techniques, graph theory, and computational algorithms provides us with an exciting array of tools for the analysis of network data.

In a random network, the focus is on the interpretation and analysis of a graph consisting of a set of nodes and a set of edges. The *nodes* (or *vertices*) of a graph correspond to entities such as people, animals, websites, power stations, banks and other financial institutions, genes or proteins, and neurons in the brain. An *edge* (or a *link*) in a graph is a line joining a pair of nodes that indicates a connection (if any) between those entities. If no edge exists between a pair of nodes, then there is no connection between those entities. Because network models typically assume that the edges joining pairs of nodes are random variables (hence, the name "random

networks"), the observed network can be viewed as a realization from a probability distribution. With such a probability model, researchers have tried to explain how these edge-generating mechanisms work, how networks grow, and what are the conditions that can be damaging to networks. Some of these networks, such as those encountered in social network analysis, tend to be relatively small and fairly easy to manipulate, whilst other networks, such as technological and information networks – think of the World Wide Web and the Internet – are huge and computationally challenging.

This book was written to bring together many of the wide-ranging contributions of the interdisciplinary study of networks, whether theoretical, computational, or applied. Although there have already been a large number of books published on random networks (e.g., Wasserman and Faust, 1994; Easley and Kleinberg, 2010; Newman, 2010; Barrat, Barthélemy, and Vespignani, 2013; Barabási, 2016; Kolaczyk, 2017; Crane, 2018), there does not appear to be any other book that treats all the topics covered in this book. Some of the novel items in this book include the development of percolation models on various types of graphs, extensive discussions on network partitioning, methods for sampling "hard-to-access" networks, descriptions on how to deal with very large networks, and examples of many important and unusual applications.

Although there are technical theorems and mathematical derivations in this book, I have tried to cross over the invisible boundary between the theoretical and applied parts of the same discipline. I have tried to make this book as readable as possible, with historical and other informational remarks and footnotes in the text. In discussing specific applications of network theory, I have tried to give as complete a description of each application as possible in the hope that this book will also be viewed as educational and informative to a general audience.

Overview of Chapters

This book is divided into 17 chapters, which can be viewed as being arranged into eight sections:

I. Chapters 1–3 can be considered as introductory material. Chapter 1 introduces the terminology, notation, and basic research issues of graphs and networks. Chapter 2 describes seven different types of networks: technological, information, financial, social, biological, ecological, and terrorist networks. The last is a novel feature of this book and shows how the myriad of worldwide terrorist organizations are related to each other. Chapter 3 sets up definitions of the basic tools used in networks, namely, the adjacency matrix, paths, connectivity, clusters, and hubs, and also describes the real-world networks of telephone call graphs and Facebook social graphs, and provides various algorithms for graph searching and minimum spanning tree.

II. Chapter 4 begins the study of random graph models, starting with the Erdős–Rényi random graphs. Some technical tools, such as first- and second-moment methods, Chernoff bounds, Cayley's formula, and a brief account of branching processes, are given, followed by graph properties and a discussion of the emergence of a giant component.

III. Chapters 5 and 6 introduce the topic of percolation, which is closely related to the random graph models of Chapter 4. Chapter 5 restricts the percolation

process to a d-dimensional lattice and Chapter 6 allows percolation to live on more general state spaces. Several applications of percolation are then described: impurity doping of semiconductors, infectious diseases and epidemics, galactic structure and star formation, polymer gelation, and the modeling of microfabrication and cellular engineering in amorphous computing.

IV. Chapters 7 and 8 describe a variety of models for unstructured networks. Chapter 7 discusses the topology of networks, which includes small-world networks, the Watts–Strogatz model, degree distributions, the power law and scale-free networks. Chapter 8 describes models of network evolution and growth, which includes the configuration model, the expected-degree random graph model, the preferential attachment model, and the random copying model.

V. Chapter 9 describes how to sample from a huge network (by sampling nodes or sampling edges) and the various types of network sampling techniques (link-trace sampling, snowball sampling, respondent-driven sampling, random-walk sampling, and forest-fire sampling), illustrated by the problem of monitoring student health in an influenza study. Also described is what to do if sampling is to be carried out when the network is "difficult to access."

VI. Chapter 10 describes a variety of parametric statistical models and techniques of parameter estimation based upon the exponential family for unstructured networks. These models are the p_1, p_2, and p^* models, the last of which is also known as the exponential random graph model. There is also a description of latent space models. Examples described in this chapter involve a friendship network of lawyers and the important issue of bullying in schools.

VII. Chapters 11–14 deal with various methods of graph partitioning for networks of unknown structure. Chapter 11 describes the different types of graph-cutting algorithms, such as minimum cuts, ratio cuts, and normalized cuts for both binary and multiway cuts. An example of graph cuts discussed in this chapter is the difficult problem of legislative redistricting of a state to avoid gerrymandering. Chapter 12 discusses the notion of community detection, a type of cluster analysis for networks. Included are the concepts of stochastic equivalence and stochastic blockmodels, modularity, regularized stochastic blockmodels, and latent cluster random-effects models. Chapter 13 discusses spectral clustering by introducing unnormalized and normalized graph Laplacians, and the spectral clustering of graphs and its regularized version. These methods are illustrated by the problem of identifying different party affiliations in a political blogs network. Chapter 14 describes the problem of modeling the presence of overlapping communities, where nodes can be members of more than a single community.

VIII. Chapters 15–17 are more technical and discuss the difficult problem of how to work with large complex networks. Chapter 15 describes the nature of very large networks, similarity measures for comparing networks, exchangeable random structures (sequences and arrays), homomorphism densities, isomorphisms, the graph coloring problem, and property testing of networks. Chapter 16 describes discrete and continuous graphons, what networks look like in the limiting case, the process of generating networks by sampling graphons, how to estimate graphons, and how to compare graphon estimates. Chapter 17 deals with dynamic networks, networks that incorporate either continuous or

discrete time as an important component. Examples illustrate longitudinal social networks, with a focus on monitoring social networks for change, and dynamic biological networks, which discuss finding and counting motifs, comparing networks using graph distances, and building network models for tracking and analyzing epidemics and the spread of infectious diseases.

Audience

This book is meant for anyone interested in the theory and application of network science. It is primarily directed towards graduate students in statistics, data science, computer science, machine learning, mathematics, business, computational biology, physics, social science, and engineering, although advanced undergraduate students may find much that is interesting in the material. This book should also be useful to researchers in many diverse fields.

Readers should have some background in statistical theory and methods, probability, a good understanding of matrix/linear algebra, and multivariable calculus. Much of the necessary background material is detailed throughout the book. The applications used in this book are taken from a wide range of disciplines and a great deal of effort has been expended on describing those applications so that the reader will find the network methodology interesting and important.

Software Packages

It is very difficult to study networks and their structures without access to specialized computer software, especially graphics software. Fortunately, great graphics software is publicly available in the R computer package.

We also highly recommend the 3D graphics package `Persistence of Vision Raytracer` that should be more well known; its website is `povray.org`.

A file entitled Software Packages, which describes the major computer packages and software routines (if they are publicly available and can be readily downloaded) for carrying out the network analyses of each chapter, will be provided on the book's website. Updates will also be listed as they become available.

The datasets used to describe and analyze networks for this book are taken from numerous sources and disciplines. If data are acquired through the Internet, the data repository is listed in a footnote at the place where the data are used as illustration. There are classic datasets and some new datasets used in this book. As is the case with network data, there are some small datasets and some huge datasets, and it is suggested that the reader become familiar with both types.

Exercises

Exercises are listed at the end of each chapter. Some challenge the reader to solve theoretical problems, some illustrate methods on specific real data, and some ask the reader to write software to implement an algorithm described in the text. There is no uniformity to the level of difficulty of the exercises, and the reader is urged to try as many of them as possible.

Book Website

The book's website is located at `http://sites.temple.edu/alan`, where additional materials and the latest information will be available.

Acknowledgments

I owe a special acknowledgment to my association with John M. Hammersley, who played a major role in developing the material on percolation (Chapters 5 and 6). Hammersley visited the Statistical Laboratory at the University of California, Berkeley, as a guest of Jerzy Neyman, during the summers of 1968, 1969, and 1970. During that time, I was fortunate to collaborate with Hammersley on two papers: Hammersley et al. (1969) and Hammersley (1972).

I have enjoyed every moment of this book's gestation. Some parts of this book were written in draft form whilst I was on sabbatical during the Fall semester 2012 at the Department of Applied Mathematics and Statistics, Johns Hopkins University. I would especially like to thank Carey Priebe and his graduate students, especially Daniel Sussman, for many helpful discussions and meetings on the general topic of network modeling and for listening to some of my wild ideas. I would also like to acknowledge Daniel Naiman and John Weirman, both of whom contributed much to a most pleasant visit.

I would like to thank all those who helped, in person or by e-mail, to clear up points of my confusion, including Edo Airoldi, Harry Crane, Keith Crank, Marijtje van Duijn, Rick Durrett, Paul Krapivsky, Subhadeep Mukhopadhyay, Mark E.J. Newman, Ben Recht, C. Seshadhri, Vincent A. Traag, René R. Veenstra, and John Wierman.

I would especially like to thank my publisher, Lauren Cowles, at Cambridge University Press in New York, and her editorial assistant, Johnathan Fuentes, both of whom provided excellent and professional support during the final stages of preparing this book. I would also like to thank those individuals (Karen De Angeles, Elizabeth Sandler, Autumn Moss, Paloma Hammond, Lauren Aileen Briskman, and Melissa LeBoeuf) who helped with permissions to reproduce certain figures in the book.

The material in this book formed the basis for a graduate statistics course on Random Networks at Temple University given during the Fall semesters of 2015 and 2017, in which students in various Master's and Ph.D. programs participated. I thank the following students, who were instrumental in providing helpful comments, suggestions, and graphics for the book: Chen Chen, Patrick Coyle, Nooreen Dabbish, Nairong Fan, Lu Fang, Doug Fletcher, Lucas Glass, David Jungreis, Emily Lynch, Rich Nair, Shinjini Nandi, Abdul-Naseh Soale, Kaijun Wang, Zhentian Wei, Xu Zhang, and Lili Zhu. Others who helped me with the graphics and figures in this book include Richard M. Heiberger and student assistants Benjamin Evans and Safaniya Paul.

Last, but certainly not least, thanks go to my family, my wife Betty-Ann and my daughter Kayla, for their understanding, patience, and support during the time I spent researching and writing this book.