The Book Review Column[1]
by Frederic Green

Department of Mathematics and Computer Science
Clark University
Worcester, MA 02465
email: fgreen@clarku.edu

In this column we look at four books, ranging from the theoretical to the applied, in these three reviews:

1. **The Power of Networks: Six Principles that Connect our Lives**, by Christopher G. Brinton and Mung Chiang; and **Algorithms and Models for Network Data and Link Analysis**, by François Fouss, Marco Saerens, and Masashi Shimbo. Two books on networks, the first more high-level, the other a detailed mathematical introduction. Joint review by Panos Louridas.

2. **Game Theory, Alive,** by Anna R. Karlin and Yuval Peres. A lively textbook on the mathematics of game theory. Review by Amir Babak Aazami.

3. **Modern Data Science with R**, by Benjamin Baumer, Daniel T. Kaplan, and Nicholas J. Horton. A new textbook on the exciting emerging field of Data Science. Review by Allan M. Miller.

As always, warmest thanks our reviewers! Please let me know if you'd like to review something; take a look over the books on the next couple of pages, but if you have other ideas, don't hesitate to send me a note.

BOOKS THAT NEED REVIEWERS FOR THE SIGACT NEWS COLUMN

Algorithms

1. *Tractability: Practical approach to Hard Problems*, Edited by Bordeaux, Hamadi, Kohli
2. *Recent progress in the Boolean Domain*, Edited by Bernd Steinbach
3. *Finite Elements: Theory and Algorithms*, by Sahikumaar Ganesan and Lutz Tobiska
4. *Introduction to Property Testing*, by Oded Goldreich.
5. *Algorithmic Aspects of Machine Learning*, by Ankur Moitra.

Programming Languages

1. *Practical Foundations for Programming Languages,* by Robert Harper

Miscellaneous Computer Science

1. *Actual Causality*, by Joseph Y. Halpern
2. *Elements of Causal Inference: Foundations and Learning Algorithms*, by Jonas Peters, Dominik Janzing, and Bernhard Schölkopf.
3. *Elements of Parallel Computing*, by Eric Aubanel
4. *CoCo: The colorful history of Tandy's Underdog Computer* by Boisy Pitre and Bill Loguidice
5. *Introduction to Reversible Computing*, by Kalyan S. Perumalla
6. *A Short Course in Computational Geometry and Topology*, by Herbert Edelsbrunner
7. *Partially Observed Markov Decision Processes,* by Vikram Krishnamurthy
8. *Statistical Modeling and Machine Learning for Molecular Biology*, by Alan Moses
9. *Market Design: A Linear Programming Approach to Auctions and Matching,* by Martin Bichler.
10. *The Problem With Software: Why Smart Engineers Write Bad Code,* by Adam Barr.
11. *Language, Cognition, and Computational Models,* Theirry Poibeau and Aline Villavicencio, eds.

Computability, Complexity, Logic

1. *The Foundations of Computability Theory,* by Borut Robič
2. *Models of Computation*, by Roberto Bruni and Ugo Montanari
3. *Proof Analysis: A Contribution to Hilbert's Last Problem* by Negri and Von Plato.
4. *Applied Logic for Computer Scientists: Computational Deduction and Formal Proofs*, by Mauricio Ayala-Rincón and Flávio L.C. de Moura.
5. *Descriptive Complexity, Canonisation, and Definable Graph Structure Theory*, by Martin Grohe.

Cryptography and Security

1. *Cryptography in Constant Parallel Time,* by Benny Appelbaum
2. *Secure Multiparty Computation and Secret Sharing*, Ronald Cramer, Ivan Bjerre Damgård, and Jesper Buus Nielsen
3. *A Cryptography Primer: Secrets and Promises*, by Philip N. Klein

Combinatorics and Graph Theory

1. *Finite Geometry and Combinatorial Applications*, by Simeon Ball
2. *Introduction to Random Graphs*, by Alan Frieze and Michał Karoński
3. *Erdős –Ko–Rado Theorems: Algebraic Approaches*, by Christopher Godsil and Karen Meagher
4. *Combinatorics, Words and Symbolic Dynamics,* Edited by Valérie Berthé and Michel Rigo

Miscellaneous Mathematics and History

1. *Introduction to Probability*, by David F. Anderson, Timo Seppäläinen, and Benedek Valkó.
2. *The Banach-Tarski Paradox (2nd Ed.)*, by Grzegorz Tomkowicz and Stan Wagon.

Joint Review of[2]
The Power of Networks: Six Principles that Connect our Lives
by Christopher G. Brinton and Mung Chiang
Princeton University Press, 2016
List price: $35.00, Hardcover, 310 pages
and
Algorithms and Models for Network Data and Link Analysis
by François Fouss, Marco Saerens, and Masashi Shimbo
Cambridge University Press, 2016
List price: $83.99, Hardcover, 521 pages

Review by

Panos Louridas (`louridas@aueb.gr`)
Department of Management Science and Technology
Athens University of Economics and Business

1 Overview

These are two books that complement each other; they target different audiences, adopt different pedagogy, they even *look* very different, yet they treat largely overlapping material. In short, they both present material related to networks; not just computer networks, but any kind of networks that may be subsumed under the newly formed discipline of network science. One of the two books intends to provide a broad sweep and provide a high-level understanding of key concepts, presenting the topics in way accessible to an educated layperson. The other book goes into the nuts and bolts of network science, taking us to the realm of rigorous equations, proofs and algorithms.

That is not to say that the two books cannot be read by the same people; after all, this reviewer was happy to have both of them on his desk, and they could be read at different times and in different moods. *The Power of Networks* is written in an engaging style; somebody who already knows networks and graph theory will be able to go through the whole book in one or two sittings. Perhaps the specialist will not find something they don't already know; but perhaps they will, as the material the book covers spans traditional disciplines, going from graph theory to Internet routing. The non-specialist will find a first-class account of key network topics, and the variety of the subjects can only pique the interest of newcomers.

Algorithms and Models for Network Data and Link Analysis is written for computer scientists and engineers; most of the audience will be students in advanced undergraduate or postgraduate courses. It is not intended as an introduction to network science: the readers should already know what the field is about. In the book they will find detailed descriptions of concepts and algorithms. While *The Power of Networks* contains only some basic mathematics, this one kicks off with five pages of symbols and notation—nothing prohibitive, they are all within the standard material of theoretical network science—but they provide a hint as to who stands to benefit from the book.

2 Summary of Contents

The Power of Networks adopts the premise that a large part of the functionality of networks is based on the following six principles:

1. Sharing is Hard.
2. Ranking is Hard.
3. Crowds are Wise.
4. Crowds are not so Wise.
5. Divide and Conquer.
6. End to End.

Each principle corresponds to a separate part of the book, which therefore is split into six parts.

The first principle is about how to use a shared resource. This can be a spectrum, which we need to share in mobile telephony (Chapter 1). It can be a shared communication channel, as in WiFi (Chapter 2). Or it can be mobile data, as in different pricing plans (Chapter 3). The different approaches are explained lucidly, so that the reader will gain an understanding of CDMA and other mobile telephony standards, the basic ideas of collision detection protocols in WiFi, and flat vs. smart pricing for mobile data plans. There are three underlying themes running through this part: the importance of *negative feedback* as a mechanism for regulating access in a shared resource, *distributed coordination* that can enable coordination even when each entity has only a local view of the network, and *negative network effects* that occur when the overall well-being declines as more entities join a network.

The second principle moves to ranking a set of items found in a network. Two very different approaches (though both of them pioneered by the same company, Google) are presented. One (Chapter 4) is the bidding mechanism behind ad spaces in Google search. Bidders auction for ad space and the authors explain the difference between open auctions and closed envelope auctions before going on to generalized second-price auctions, used by Google AdWords. The second approach is PageRank (Chapter 5), the algorithm made famous as the backbone of the original Google Search engine. The idea of the webgraph is introduced, upon which the random surfer model is explained as the basis of Google Search.

The third principle deals with applications of the wisdom of crowds. These cover three areas: product ratings, as for example in Amazon (Chapter 6), product recommendations, like for example Netflix (Chapter 7), and social learning, in the form of Social Learning Networks and their role in Massive Open Online Courses (MOOCs, in Chapter 8). The wisdom of crowds is to a large effect the result of *positive network effects*, where additional entities joining the network bring along benefit to everybody there and *opinion aggregation*, where the unbiased and independent opinions of a diverse population are amalgamated to produce a reliable result.

In the fourth principle we make the transition from the wisdom to the folly of crowds: that's when inputs from many different agents contribute to a deleterious result. This is exemplified in information cascades (Chapter 9), influence, and network contagion (Chapter 10). In contrast to negative feedback, which regulates a system towards a stable state, we find that *positive feedback* amplifies network contributions, pushing the system away from the equilibrium.

The fifth principle brings us to what most people think of when they talk about networks: the Internet itself. Chapter 11 outlines its early history and then describes the fundamental idea of using packets instead

of circuits for the transfer of data. It then gives an overview of the layered architecture of the Internet, which segues in Chapter 12 to how network traffic is routed in the network (IP) layer using the Bellman-Ford shortest path algorithm; here we meet again a form of *distributed coordination*, which we first encountered in Chapter 2.

The sixth and final principle picks up from IP and goes up the protocol stack by describing the basic network congestion prevention mechanism of TCP, in Chapter 13. The last chapter of the book, Chapter 14, sheds light on the small world structure of many networks. Starting with random graphs and contrasting them with regular ring graphs, it shows that the Watts-Strogatz model can create graphs that resemble those we observe in many real-world settings.

A distinctive feature of the book is the interviews with luminaries in different fields of networks. We find discussions with Dennis Strigl, of Verizon, Eric Schmidt, of Google and Alphabet Inc., and Internet pioneers Robert Kahn and Vint Cert. The interviews appear at the end of some of the book parts and add historical background, as well as a personal dimension.

The book succeeds in remaining at an intuitive, non-technical level. There is some more advanced supplemental material in the book's website (`http://powerofnetworks.org/`); this again is written in an engaging style, so it should not be beyond the level assumed by the intended audience.

Algorithms and Models for Network Data and Link Analysis steps up the ante and provides a rigorous foundation for Network Science. The tone is given in the first chapter, where after the preliminary definitions of graphs we proceed with a brisk pace to material such as Laplacians, basic Markov Chain concepts, graph kernels, expectation-maximization, and the Floyd-Warshall algorithm.

The following chapters contain essential material for any serious student or researcher working on graphs. Chapter 2 covers similarity and proximity measures between nodes; Chapter 3 is a follow up on Chapter 2, with more advanced material on the distances and similarities between nodes in a graph. Centrality measures on nodes and edges are the topic of Chapter 4. The identification of prestigious nodes is the subject of Chapter 5. Chapter 6 introduces techniques for assigning labels to nodes based on the labels of other nodes and the graph structure. Nodes clustering, the partitioning of the nodes of the graph in different sets, is treated in Chapter 7. Then, in Chapter 8, the book covers methods for the identification of dense regions in a graph. Bipartite graphs and their analysis take up Chapter 9. The final chapter, Chapter 10, rounds up the book with a discussion on graph embedding, the association of each node in a graph with a position or a vector, thus placing it into an embedding space.

The authors cover a selection of approaches and algorithms in each chapter. It is unavoidable that a field so vast cannot fit in a single volume, so the authors had to make choices in what to include. While acknowledging that the choices they made reflect their own preferences, they have laid down their set of criteria for inclusion. The algorithms should be well-established, even though not necessarily widely known in computer science itself; calculations should be based on linear algebraic models and random walks; the mathematical techniques involved (e.g., least squares, maximum likelihood, expectation-maximisation) should be interesting in themselves; and the algorithms should scale to at least thousands of nodes.

There are many algorithms spread around the text, given as pseudocode. Keeping true to the mathematical slant of the book, the pseudocode assumes that we have a language with matrix computation facilities. Matlab/Octave code for many of the algorithms is gradually made available on the book's web page at `https://www.cambridge.org/9781107125773`.

The matrix-based presentation of algorithms make for succinct pseudocode. Readers with a more mathematical background would find the algorithms straightforward; those with a more practical programming grounding may need to get used to the way they are presented. For example, the Floyd-Warshall algorithm, which, as we mentioned, appears in the preliminary chapter, is presented in a matrix form, like this:

Algorithm: Computing the directed shortest-path distance matrix: Matrix form

Input:

- A weighted, possibly directed graph G containing n nodes.
- The $n \times n$ cost matrix $\mathbf{C}$ associated with G, containing nonnegative costs.

Output:

- The $n \times n$ shortest-path distance matrix $\mathbf{\Delta}$.

$\mathbf{\Delta} \leftarrow \mathbf{C}$
for $t = 0$ **to** n **do** ▷ enumerate all possible intermediate nodes
    $\mathbf{\Delta} \leftarrow \min(\mathbf{\Delta}, \mathbf{col}_t(\mathbf{\Delta})\mathbf{e}^{\mathrm{T}} + \mathbf{e}(\mathbf{row}_t(\mathbf{\Delta})\mathbf{e}))^{\mathrm{T}})$ ▷ recompute the distances when considering a new potential intermediate node t
end for
Set diagonal elements of $\mathbf{\Delta}$ to 0
return $\mathbf{\Delta}$

which probably looks familiar to those with a theoretical slant, less so to programmers-practitioners.

The book is akin to handbooks of algorithms, which have long been valuable references to computer scientists. The chapters have been written to be read independently, and are to a large extent self-contained. At the same time, this reviewer found that the material inside each chapter is better read in sequence; this is not a dictionary of algorithms and concepts, but rather an organized compendium of them.

3 Opinion

I was happy to go through both books, for different reasons. The writing style in *The Power of Networks* made it a pleasure to read; but what really puts it apart from other books on networks is the choice of topics. You can easily find books on networking protocols and *other* books on graph algorithms or auction theory; but this book really spans the network realm, moving smoothly from, say, PageRank to TCP; the authors have set out to show a set of basic principles underlying all kinds of networks, and they successfully integrate theoretical and engineering material. This is an excellent initiation to networks that will satisfy readers of different backgrounds. The authors point out that much of the material in the book has already been used successfully to teach over 100,000 students successfully from 2013 onwards, and indeed the book would be an ideal fit in an introductory course on networks.

Algorithms and Models for Network Data and Link Analysis has a more circumscribed scope, which keeps it faithful to its title. It covers a lot of material in its purview with a careful selection of algorithms. A casual browse through the pages of the book will convince the reader that this is serious stuff, in harmony with the advanced undergraduate or postgraduate level that was the authors' target. However, it would be a pity if the readership would stay limited to that. Network Science applications pop up frequently in big data processing, and programmers working on them would only benefit by taking the time to study a bit the different algorithms and choices involved, instead of quickly applying the first recommendation that pops up in a web search.

Review of[3]
Game Theory, Alive
by Anna R. Karlin and Yuval Peres
AMS, 2017
372 pages, hardcover, numerous figures, $74.98

Review by
Amir Babak Aazami (aaazami@clarku.edu)
Department of Mathematics & Computer Science
Clark University

1 Overview

Game Theory, Alive is a wonderful book and is to be highly recommended, either for teaching or self-study. By way of comparison, it covers fewer topics and is less advanced than the well known book *Game Theory,* by M. Maschler, E. Solan, and S. Zamir. For this reason, it seems ideal for a first course on Game Theory at the undergraduate level. Since *Game Theory, Alive* assumes some basic knowledge of probability theory (in addition to discrete math), students should have some probability theory as a background, though much of the book is still very accessible without it. This reviewer would not be surprised if *Game Theory, Alive* becomes the standard text for an introductory course on Game Theory. It is very well written and fun to read. The numerous figures, cartoons (see, e.g., Figure 6 on page xxi), photos, anecdotes, and especially, the historical summaries at the end of each chapter as well as the backgrounds of the mathematicians, statisticians, and economists whose results now go into Game Theory, is one of the loveliest features of this book. Wikipedia notwithstanding, we often don't learn enough about the 'players' themselves, their history and/or the evolution of how a result came into being (at least in this reviewer's primary field of study, which is not game theory or even probability theory), and this book very nicely bucks that trend. The most important caveat I should state here is that I am not myself in game theory or combinatorics, so my review of this book cannot be as deep as that of a practitioner in either of these fields. That is unfortunate, but the plus side of this is that I can make the following comment: *to me, the most beautiful feature of this book is how nicely — and yet how compactly — the authors convey intuition and motivation*, which is the most important thing a textbook can do for someone approaching the field for the first time. And to do so while at the same time presenting everything mathematically — definitions, theorems, proofs — is not easy.

2 Summary of Contents

Game Theory, Alive is divided into two parts (of which this reviewer looked over the first part more carefully than the second, with one exception, see below). Part I proceeds roughly as follows: it begins with games in which two players alternate turns, to zero-sum games in which two players take turns simultaneously. Random-turn games are also discussed towards the end of Part I. One nice feature here is that zero-sum games on graphs are also discussed, including a proof of Hall's Marriage Theorem. Finally, Part I discusses games in which there is no optimal strategy, in which the Nash equilibrium comes center stage. The authors prove Nash's Theorem via Brouwer's Fixed-Point Theorem, but they also prove Brouwer's Fixed-Point Theorem very nicely through the game Hex, which had been extensively discussed earlier. This chapter

[3]

is particularly well written, and its extensive discussion of Brouwer's Fixed-Point Theorem is very nicely done. Also, the way the authors differentiate between general-sum games with imperfect or incomplete information is nicely balanced. I also enjoyed the way the authors motivate the property of evolutionary stability and its connection to Nash equilibria (something I knew nothing about). Finally, the authors discuss the case when there is no good Nash equilibrium, and the effects of having a correlating device or a central planner. This leads nicely into their discussion of the price of anarchy. One final comment here: adaptive decision-making during repeated zero-sum games is also discussed, though in Part II of the book.

Part II discusses games that are not predefined, and overall I found their discussion here of voting systems, auctions, etc., to be very well written. *This is also the only area in which I can compare their work with that of other authors (not being a game theorist myself); see the next section below.* The outline is as follows. First, the authors discuss the need for stability in preference matching games, culminating in the Gale-Shapley algorithm. Two natural topics emerge from this context, namely, forming coalitions and bargaining, which lead naturally to social choice and voting theory[4]. Here Arrow's Impossibility Theorem is proved, due to J. Geanakoplos. A hallmark here is the elegant presentation and proof of something more advanced, the Gibbard-Satterthwaite Theorem. One apology from the reviewer is in order here: I'm afraid I cannot adequately comment on the remaining topics in Part II, namely the VCG mechanism and matching markets (knowing they might be useful to a computer scientist!), since I did not look through these chapters carefully.

3 Chapter Highlights and Historical Notes

Chapter 13 ("Social choice and voting") is the only chapter in Part II that this reviewer read all the way through. I am familiar with two other books on voting theory, namely, *The Mathematics of Elections and Voting,* by W.D. Wallis, and *The Mathematics of Voting and Elections: A Hands-On Approach,* by Jonathan K. Hodge and Richard Kilma. Both are excellent books, designed to be read with the very minimum of mathematical backgrounds (especially the latter). *Game Theory, Alive* is more mathematically sophisticated, but crucially, it conveys the same *intuition*, in less space, something that is not easy to do. A simple example: this reviewer remembers the 1998 gubernatorial election of Minnesota, in which the wrestler Jesse ("the body") Ventura won. But *Game Theory, Alive* points out that in fact he was the least favored candidate by a majority of Minnesotans. This is a very simple fact, but it is perfect motivation; it is all one needs to know in order appreciate the necessity of voting theory. This is one of the gems of the book: in very pithy statements just like this, scattered everywhere in the book, a great deal of motivation and/or intuition is very succinctly conveyed (given the authors of *Game Theory, Alive,* this is not so surprising!) There is, of course, a great deal of motivation presented in both *The Mathematics of Elections and Voting* and *The Mathematics of Voting and Elections: A Hands-On Approach*, especially the latter. In the latter, however, the discussion of this motivation is much more extensive. In *Game Theory, Alive*, the motivation is compact, but somehow conveys the same intuition. This is one of the gems of the book to me (and the lovely cartoons!), especially given that I do not work in game theory or combinatorics.

[4]Some readers may be interested in related literature covered in this column. See reviews of *Handbook of Computational Social Choice*, SIGACT News **48**(4), pp. 13-17; and *Trends in Computational Social Choice*, SIGACT News **49**(2), pp. 14-17. – Ed.

Review of[5]
Modern Data Science with R
by Benjamin Baumer, Daniel T. Kaplan, and Nicholas J. Horton
Chapman and Hall/CRC, 2017
ISBN 9781498724487, 556 pages, $99.95

Review by
Allan M. Miller[6] (allan.m.miller@berkeley.edu)

1 Introduction

Modern Data Science with R (MDSR) is one of the first textbooks to provide a comprehensive introduction to data science for students at the undergraduate level (it is also suitable for graduate students and professionals in other fields). The authors follow the approach taken by Garrett Grolemund and Hadley Wickham in their book, *R for Data Science*[7], and David Robinson in *Teach the Tidyverse to Beginners,* which emphasizes the teaching of data visualization and the tidyverse (using `dplyr` and chained pipes) before covering base R, along with using real-world data and modern data science methods.

The textbook includes end of chapter exercises (an instructor's solution manual is available), and a series of lab activities is also under development. The result is an excellent textbook that provides a solid foundation in data science for students and professionals alike.

2 Summary of the Book

Data Visualization:

After a brief introductory overview of data science in Chapter 1, Part I begins with coverage of data visualization. Chapter 2, *Data Visualization,* provides an in-depth discussion of data graphics **without using R**, covering the *general principles of data visualization*, such as visually distinguishing subgroups, graphing variation, networks, and examining the relationship between variables.

Chapter 3, *A Grammar of Graphics,* introduces Leland Wilkinson's grammar of graphics, a method for describing the components of a statistical graphic introduced in the first part of Chapter 2 in terms of geometries (types of graphs such as point, line, and bar), statistics (the mapping of data to visual elements), and aesthetics (such as color, shape, and texture) comprising different *layers* of a graph. Thus early in the text, the authors provide students with a framework and a systematic approach for designing statistical graphics.

The second half of the chapter provides in-depth coverage of `ggplot2`, which *implements* the grammar of graphics in an R package capable of generating statistical graphics. Several examples are presented and discussed in detail. After completing the exercises provided in chapters 2 and 3, students should be able to create complex plots based on the approach outlined in the chapters.

---

[6] Allan Miller, Ph.D., is a Data Scientist, educator, and active member of the San Francisco Bay Area data science community. He teaches popular courses in R, machine learning, statistics, and data science for the U.C. Berkeley Extension program, where he serves on their Data Science Advisory Committee. Dr. Miller is also the organizer of the Berkeley R users group.

[7] Reviewed in this column, SIGACT News **48**(3), pages 14-19. – Ed.

Data Wrangling:

The second half of Part I covers data wrangling, tidying data, and using R to implement vectorized operations.

Chapter 4, *Data Wrangling,* covers data manipulation using the R package `dplyr`. `dplyr` defines a set of verbs commonly used in data wrangling, such as `select` (for selecting columns), `filter` (for selecting rows based on logical criteria), `mutate` (for adding columns based on transformations and combinations of existing columns), `group_by` (for grouping observations into subgroups), and `summarize` (for computing statistics such as mean or ranking for subgroups). These verbs are combined with the "pipe operator" `%>%` in an intuitive style that facilitates complex manipulation or wrangling of data.

Chapter 5, *Tidy Data and Iteration,* introduces the tidy data concept, where datasets are transformed into "tidy" data format: every observation consists of a single row, and every column is a single attribute. The idea is to define a common (grid) format for datasets on which a set of packages (the "tidyverse"), encapsulating the most common data wrangling operations (for example reading and writing data), will work. The chapter also emphasizes using vectorized operations such as `dplyr::do` to iterate over tidy datasets, versus looping used in non-vectorized languages.

Part I of the text concludes with a chapter on Professional Ethics, covering some basic topics in data governance, privacy, and reproducibility. Upon completion of Part I, students will be able to load, wrangle, and visualize datasets using core tidyverse packages, reflecting the authors' goal of not initially focusing on R language syntax, and setting the stage by providing skills for the data science topics that follow.

Statistics and Modeling:

Part II, *Statistics and Modeling,* consists of four chapters.

Chapter 7, *Statistical Foundations,* covers some basic statistical concepts underlying data science. This chapter, despite its brevity (its coverage of statistics is high-level), is strong conceptually. The topics that are covered include sampling, the use of statistical models to explain variation, confounding factors, and the dangers of using p-values in hypothesis testing.

Chapters 8 - 10 provide a broad survey of statistical learning methods, essentially a very short course in statistical learning. Chapter 8, *Statistical Learning and Predictive Analytics,* briefly covers supervised learning, including classifiers (e.g., decision trees, random forests, and artificial neural networks), model parameter tuning, and ensemble methods. The second half of the chapter covers data preparation (cross validation) and model evaluation (RMSE, confusion matrix, and ROC curves), using an example taken from Census Bureau income data.

Chapter 9, *Unsupervised Learning,* provides short but effective coverage of clustering, highlighting the most commonly used unsupervised learning techniques (hierarchical and k-means clustering), and dimensionality reduction. Chapter 10, *Simulation,* provides valuable coverage of simulation techniques for complexity reduction not found in most data science textbooks.

Topics in Data Science

Part III, *Topics in Data Science,* consists of six chapters covering a variety of types of data and problems often encountered in modern data science projects.

Chapter 11, *Interactive Data Graphics,* provides rich coverage of R web-based interactive graphics using Leaflet for interactive spatial plots, `plotly` and `ggvis` for interactive visualizations, DataTables, an R interface to the JavaScript library DataTables, to display R matrices or data frames as tables on HTML pages, `dygraphs`, an R interface to the JavaScript charting library of the same name, providing rich facilities for charting time-series data in R, and `streamgraphs`, an htmlwidget that is based on the D3.js JavaScript library. Streamgraphs are a generalization of stacked area graphs where the baseline is free. By shifting the baseline, it is possible to minimize the change in slope (or wiggle) in individual series, thereby making it easier to perceive the thickness of any given layer across the data. Chapter 11 also introduces Shiny, the R package for developing rich web-based interactive data visualization applications.

Chapters 12 and 13, *Database Querying Using SQL and Database Administration,* cover the basics of designing, creating, implementing, and querying SQL databases, using both SQL and R. Several examples are presented using MySQL, `dbplyr` and DBI. The more recently developed `dbplyr` package is not covered.

Chapter 14, *Working with Spatial Data,* provides in-depth coverage of spatial data wrangling and mapping including projections, geocoding, routes, projections, and distances, using `ggmap` for static mapping and the R Leaflet package for interactive mapping.

Chapter 15, *Text as Data,* introduces natural language processing and computational linguistics applications in R. The chapter starts with using the text of Shakespeare's Macbeth to demonstrate basic R string processing operations, such as regular expressions, string parsing and searching. It goes on to cover foundational topics in NLP, including building corpora, word clouds, and document term matrices. Finally, the chapter ends with coverage of methods for obtaining text data, using R for web and twitter scraping.

Part III concludes with coverage of Network analysis (Chapter 16, Network Science) and big data (Chapter 17, Epilogue: Towards "big data").

Also included in the text are six appendices, covering the packages (msdr) and tools (RStudio) used in the text, an introduction to algorithms, reproducible research, regression modeling, and setting up a database server.

3 Conclusion

Modern Data Science with R is a breakthrough textbook for the following reasons:

- It adopts the "teach the tidyverse approach first," providing early coverage of data visualization using the grammar of graphics and `ggplot2`.
- It emphasizes coverage of the tidyverse versus starting with, and covering base R in detail.
- Provides comprehensive coverage of data science topics and subject material.
- Includes valuable coverage of special topics (skills and data).

Early coverage of data visualization allows allows students to generate plots with minimum coverage of detailed base R plotting syntax. The idea is to "get students to do powerful things quickly." MDSR implements this approach "using interesting case studies, generating informative and appealing visualizations, and drawing useful conclusions[8]." The idea is to engage students on real world analytical problems, demonstrate the power of graphics, and give them an approach (grammar of graphics) and tools (`ggplot2`) to implement appealing and informative plots.

Building on this approach, the authors then introduce data wrangling using `dplyr` with chained pipes. Much of objection to R as a teaching, learning, and statistical applications language is the supposedly arcane nature of base R syntax and the idiosyncrasies of the language. Using `dplyr` with chained pipes provides a highly intuitive approach to data wrangling, using data processing verbs (such as `filter`, `mutate`, `summarize`, and `group_by`) chained together in pipes to describe (and implement) data wrangling processes, designed from the start to work directly and easily with data visualization using `ggplot2`. By the third chapter in the text students are able to attack substantial analytics problems involving complex data wrangling and powerful visualizations. In and of itself, this makes MDSR a worthwhile text to adopt.

Part II, *Statistics and Modeling,* provides comprehensive coverage of these topics, the core of data science, but especially Chapter 8, Statistical Learning and Predictive Analytics, tries to cover too many topics too quickly, using what feels like a shotgun approach to covering the topics. In the end, it provides comprehensive, but rather superficial, coverage (some of the sections in that chapter are just a paragraph or two long). One has to wonder why there are two chapters and an appendix on databases, and one chapter on supervised statistical learning. The chapter on statistical learning could be expanded into two or three chapters, and include more detailed and coherent focus on methodology combined with coverage of algorithms. The shortcomings in this section can be easily compensated for by using a supplemental text and/or articles.

Part III, *Topics in Data Science,* is a rich treasure trove of topics, problems, and tools for data science students. It prepares students to tackle real world projects as data science professionals.

Modern Data Science with R, the first comprehensive data science textbook based on the highly effective "tidyverse first" approach to teaching and learning R, provides in-depth coverage of virtually all essential data science topics. While some adjustments in topics (expanded coverage of statistical learning, less database and other special topics), would improve the text, MDSR is a breakthrough text, definitely worth adopting.

[8] `http://varianceexplained.org/r/teach-tidyverse/`