Journal of Statistical Software
August 2015, Volume 66, Book Review 1. http://www.jstatsoft.org/

Reviewer: Matthew Nunes
Lancaster University

Statistical Analysis of Network Data with R

Eric D. Kolaczyk and Gábor Csárdi
Springer-Verlag, New York, 2014.
ISBN 978-1-4939-0983-4. 207 pp. GBP 39.99 (P).
http://www.springer.com/book/9781493909827

Statistical Analysis of Network Data with R is a recent addition to the growing *UseR!* series of computational statistics monographs using the R programming language (R Core Team 2015). It gives a practical introduction to the visualization, modeling and analysis of network data, a topic which has enjoyed a recent surge in popularity. The book brings together a partnership of two established researchers in the field: Eric Kolaczyk, author of a number of papers and a recent texts on statistical network analysis, and Gábor Csárdi, researcher of network data arising in biological applications and lead developer of a popular network analysis software suite. I was thus curious to see what this book has to offer, especially since such data is becoming more available and of interest in a wide range of scientific fields. In the preface, the authors state the aims of the book as "to provide an easily accessible introduction to the statistical analysis of network data", but flag to the reader that the book is "not a detailed manual for using the various R packages encountered [...] nor [...] provide exhaustive coverage of the conceptual and technical foundations of the topic area", but instead aims to "strike a balance between the two". I will discuss both the theoretical and computing aspects of the book below.

The book is divided into ten chapters on fundamentals of network analysis. The text starts with a short mathematical history of the field, and motivates its content by describing the main fields in which network data arise and accounts for the explosion of interest in analysis of such data. The next few pages in Chapter 1 then give a detailed overview of the material which is covered in the book, before giving another summary of where specific content in the book can be found. This organization of material is done according to what Kolaczyk and Csárdi call the *statistical taxonomy* of network analysis tasks: data manipulation, visualization, descriptive analysis, through to modeling and inference. Perhaps unsurprisingly, this structure mimicks that of Kolaczyk's recent book on statistical network analysis (Kolaczyk 2009), and in some sense the mathematical content of *Statistical Analysis of Network Data with R* can be seen as a condensed version of this more in-depth treatment of the area.

Chapter 2 describes network data manipulation, beginning with the construction of basic graphs using the `igraph` data structure from the **igraph** R package (Csárdi 2015), before describing set-theoretic operations on such structures (e.g., network union or intersection). It then moves onto graph *decoration*, that is, the process of assigning attributes to graphs (including, e.g., edge directionality or variables to aid visualization). With this structure, one can easily build increasingly complicated network objects by creating *layers* of vertex and edge attributes through use of the standard `$` assignment. The chapter also outlines various types of graphs and basic graph-theoretic properties in order to describe graphs (such as *connectedness*, *vertex degree* and movement around a graph). When reading this chapter, at first I found myself saying: "What about… X?", since – perhaps due to lack of a reassuring forward reference here of what is to come – at times the material seems a touch brief in explanation, as though the authors (admittedly like myself) were keen to get onto the more exciting chapters that follow. (I realized a few chapters later that my initial questioning was of course mistaken, since the material was covered there). The chapter nevertheless succeeds in alluding to the potential sophistication of network data descriptions, serving to whet the reader's appetite, particularly for the next chapter.

Whilst only totaling around 15 pages, Chapter 3 was arguably my favourite in the book. It deals with the visualization of networks and network data. The authors rightly take a somewhat philosophical stance in their treatment of visualization, emphasizing that the various ways of displaying network data can elucidate different information contained in the graph structure. It is here where the flexibility of **igraph** `layout` decoration comes to the fore: circular, spring-embedder or multidimensional scaling visualizations can all be generated quickly by changing the `layout` argument in the **igraph** `plot` command. The chapter really encourages the reader to put the book down momentarily to have a play with different layouts for the datasets accompanying the book or indeed their own network. The chapter left me eager to read on; the only thing I felt was missing was an example of interactive network graphics, perhaps using **rgl** graphics devices (Adler, Murdoch *et al.* 2015, e.g., called by the `gplot3d` function in the **sna** R package, Butts 2008). I also would have liked a bit more advice here on how to condense information and produce graphics for particularly large networks, where standard visualizations are prohibitive.

Chapter 4 elaborates on different vertex and edge characteristics, in particular describing graph-based quantification of vertex and edge 'importance' and their interpretation (so-called centrality measures). Following this, the chapter deals with the topic of network *cohesion* – the characterization of clustered neighborhood structures, and their measurement in a global graph structure. This leads naturally onto discussion of the important application of *graph partitioning* in Chapter 4.4, in which the authors introduce the various algorithms to identify such cluster formations in graphs.

In the second half of the book, the text moves onto network modeling. The authors follow more of a case study approach in this part of the book to demonstrate the different models. Chapter 5 is dedicated to mathematical network models, including the classical Erdös-Rényi random graphs and their generalizations, as well as so-called 'small-world' and 'preferential attachment' models. The section ends with a brief discussion of assessing characteristics of such mathematical network models. Whilst this chapter marks a change in the book's mood to one that's a bit more serious, the reader is helped with this transition through referrals back to previous chapters. The subsequent chapter is devoted to various statistical network models, focusing on model fitting and assessing goodness-of-fit for each class. More specifically, the chapter first introduces exponential random graph models, and the text leads the reader through their specification (helpfully managed through the familiar **base** `formula` syntax) and the model-fitting/goodness-of-fit capabilities of the **ergm** package (Hunter, Handcock, Butts, Goodreau, and Morris 2008). Following this, the chapter covers fitting network block models with the **mixer** package (Ambrose, Grasseau, Hoebeke, Latouche, Miele, Picard *et al.* 2015); the last class is the latent graph model, for which different model variants (and model fitting with MCMC) are demonstrated via the **eigenmodel** package (Hoff 2012). In Chapter 7 we are shown how to perform network topology inference, primarily focusing on link prediction, inference in association (e.g., correlation) networks, and *tomographic* network inference. Chapter 8 provides an overview of modeling and prediction of processes on graphs rather than modeling the networks themselves. After a brief look at nearest neighbor methods, the chapter has a nice treatment of Markov random fields using **ngspatial** (Hughes and Cui 2015), continuing with a similarly accessible exposition on graph kernel methods for network processes with **kernlab** (Karatzoglou, Smola, Hornik, and Zeileis 2004), both with interesting biological examples. Dynamic processes on graphs are explored in the last section, with a particular emphasis on epidemic models. This application-centred discussion is also followed in Chapter 9, where a telecommunications example is used to demonstrate modeling data flow through a (directed) network, concentrating on the particular statistical problem of origin-destination matrix estimation in different observed and unobserved data settings. The last chapter in the book is devoted to dynamic network analysis, mainly explored by using what is learnt in the previous chapters in the book applied to (time-indexed) multi-edge `igraph` objects. The chapter is unfortunately quite short (due to the lack of established methodology for dynamic networks currently in the literature), but hints at other possible R capabilities in this setting using the **networkDynamic** (Butts, Leslie-Cook, Krivitsky, and Bender-deMoll 2014) or **tergm** (Krivitsky, Handcock *et al.* 2015) packages.

The intended audiences of the book are listed as (i) statisticians looking to engage with statistical network analysis (ii) researchers from other quantitative fields working with network data seeking to familiarize themselves with statistical analyses and (iii) practitioners in applied areas wishing to get a foothold on a particular analysis. In terms of accessibility, the book is aimed at graduate students or advanced undergraduate students in quantitative fields, and with this in mind, the book assumes no specific prior knowledge of network data or graph theory. Outside of network analysis, the book touches on statistical methods such as ROC curves, cross-validation, FDR, and Monte Carlo methods etc., which crop up in natural places in the text when performing some data analyses. I broadly agree that the content is indeed accessible to these audiences, although some familiarity with mathematical and statistical concepts is helpful, due to the notation and terminology used in some of the later chapters. However, the authors provide additional reading on theoretical topics where appropriate and include bibliographic notes for further reading at the end of each chapter. The text is written in a precise but engaging conversational style, with side-comments helping the reader through the technicalities.

I now focus on the book's R content more generally. All code examples and datasets used in the book have been helpfully packaged into the **sand** package (Kolaczyk and Csárdi 2014), available on the CRAN R package repository. The coding elements of the book are pitched at and are accessible by readers with a basic familiarity with R object manipulation and data structures. Indeed, the authors do not provide any section or appendix devoted to introductory R commands – which, given the stated intended readers and also the wealth of introductory R programming resources available elsewhere – I think is a sensible omission. Most of the code snippets are based around the now well-developed **igraph** suite of network analysis tools (Csárdi and Nepusz 2006, more specifically the R port of the software, Csárdi 2015), designed and maintained by Gábor Csárdi. In particular, the book makes heavy use of `igraph` data representation and network layering. Using this strategy could have been overly-ambitious in view of the broad scope of models and application areas covered by the text – especially given the authors own admission on the somewhat steep learning curve of more advanced aspects of **igraph**. However, I found that having this underlying backbone to the code snippets throughout the book is for the most part useful in unifying the examples. Moreover, the authors manage to fairly successfully interweave code from other R packages where needed (including other network analysis software such as the **statnet** R suite, Goodreau, Handcock, Hunter, Butts, and Morris 2008), so that the snippets are not overly-dependent on **igraph**. The only exceptions where the code sometimes seems slightly inconveniently involved are the occasions when converting graph structure representations between packages, although to take advantage of different package functionality, I can see why this is sometimes necessary. As noted above, the theoretical content of the book increases in the second half; I noticed that the code complexity also grows towards the end of the book.

The datasets used in the book (many available in the **sand** package) are taken from a range of application areas to illustrate the described models and techniques appropriately. These include the now canonical *Karate Club* example dataset, protein interaction networks, telecommunications, and disease dynamics. I also note here that admirably all R featured in the book appears online at Eric Kolaczyk's GitHub repository `http://github.com/kolaczyk/sand`, annotated and reproduced so that the code chunks can be copy-and-paste'd directly into R from a browser. The website also has a space to track any code/text errata in the book and can ensure that any updates to the **sand** package are reflected in the example code.

There exist other books like De Nooy, Mrvar, and Batagelj (2011) or Cherven (2013) on software for network analysis, so why is *Statistical Analysis of Network Data with R* worth reading? Firstly, this is the first on the topic for the R programming language and thus arguably reaches a different audience. Secondly, whilst it may not feature in my pile of bedtime page-turners, I nevertheless enjoyed reading this book as it develops a solid instructive story. It doesn't tell you *what* to do to analyze network data, nor should it. Rather, by the end of the book the reader will have gained an introduction to the area through a well-constructed exposition, without being too bogged down by technicalities. Through the practical demonstrations he/she will also be equipped *how* to run possible analyses, fit common network models, and talk about network data using appropriate terminology and with a degree of confidence. In this sense Kolaczyk and Csárdi succeed in striking their intended balance between network theory and its implementation in R.

In summary, being just under two hundred pages without references, *Statistical Analysis of Network Data with R* allows a range of scientists to dive straight into network analysis – provided that other texts such as Wasserman and Faust (1994), Kolaczyk (2009), or Newman (2010) are within reach for additional reading. As the world of network data becomes more accessible and analyzing such complex datasets becomes more feasible, I expect that some of the material in this book (particularly on dynamic network analysis) may become out-of-date, but if this happens, I look forward to reading a sequel.

References

Adler D, Murdoch D, *et al.* (2015). ***rgl**: 3D Visualization Using OpenGL.* R package version 0.95.1247, URL http://CRAN.R-project.org/package=rgl.

Ambrose C, Grasseau G, Hoebeke M, Latouche P, Miele V, Picard F, *et al.* (2015). ***mixer**: Random Graph Clustering.* R package version 1.8, URL http://CRAN.R-project.org/package=mixer.

Butts CT (2008). "Social Network Analysis with **sna**." *Journal of Statistical Software*, **24**(6), 1–51. URL http://www.jstatsoft.org/v24/i06/.

Butts CT, Leslie-Cook A, Krivitsky PN, Bender-deMoll S (2014). ***networkDynamic**: Dynamic Extensions for Network Objects.* R package version 0.7.1, URL http://CRAN.R-project.org/package=networkDynamic.

Cherven K (2013). *Network Graph Analysis and Visualization with **Gephi**.* Packt Publishing Ltd.

Csárdi G (2015). ***igraph**: Network Analysis and Visualization.* R package version 1.0.1, URL http://CRAN.R-project.org/package=igraph.

Csárdi G, Nepusz T (2006). "The **igraph** Software Package for Complex Network Research." *InterJournal*, **Complex Systems**, 1695. URL http://igraph.org/.

De Nooy W, Mrvar A, Batagelj V (2011). *Exploratory Social Network Analysis with **Pajek***, volume 27. Cambridge University Press.

Goodreau SM, Handcock MS, Hunter DR, Butts CT, Morris M (2008). "A **statnet** Tutorial." *Journal of Statistical Software*, **24**(9), 1–26. URL http://www.jstatsoft.org/v24/i09/.

Hoff P (2012). ***eigenmodel**: Semiparametric Factor and Regression Models for Symmetric Relational Data.* R package version 1.01, URL http://CRAN.R-project.org/package=eigenmodel.

Hughes J, Cui X (2015). ***ngspatial**: Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data.* R package version 1.0-5, URL http://CRAN.R-project.org/package=ngspatial.

Hunter DR, Handcock MS, Butts CT, Goodreau SM, Morris M (2008). "**ergm**: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software*, **24**(3), 1–29. URL http://www.jstatsoft.org/v24/i03/.

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004). "**kernlab** – An S4 Package for Kernel Methods in R." *Journal of Statistical Software*, **11**(9), 1–20. URL http://www.jstatsoft.org/v11/i09/.

Kolaczyk E, Csárdi G (2014). ***sand**: Statistical Analysis of Network Data with R.* R package version 1.0.2, URL http://CRAN.R-project.org/package=sand.

Kolaczyk ED (2009). *Statistical Analysis of Network Data: Methods and Models.* Springer-Verlag.

Krivitsky PN, Handcock MS, *et al.* (2015). ***tergm**: Fit, Simulate and Diagnose Models for Network Evolution based on Exponential-Family Random Graph Models.* R package version 3.2.4, URL http://CRAN.R-project.org/package=tergm.

Newman M (2010). *Networks: An Introduction.* Oxford University Press.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Wasserman S, Faust K (1994). *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press.

Reviewer:

Matthew Nunes
Department of Mathematics and Statistics
Fylde College
Lancaster University
Lancaster, LA1 4YF, United Kingdom
E-mail: m.nunes@lancaster.ac.uk
URL: http://www.maths.lancs.ac.uk/~nunes/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 66, Book Review 1 *Published:* 2015-08-06
August 2015