Use R!

Douglas A. Luke

A User's Guide to Network Analysis in R

Springer

Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

More information about this series at http://www.springer.com/series/6991

Use R!

Albert: Bayesian Computation with R (2nd ed. 2009)
Bivand/Pebesma/Gómez-Rubio: Applied Spatial Data Analysis with R (2nd ed. 2013)
Cook/Swayne: Interactive and Dynamic Graphics for Data Analysis: With R and GGobi
Hahne/Huber/Gentleman/Falcon: Bioconductor Case Studies
Paradis: Analysis of Phylogenetics and Evolution with R (2nd ed. 2012)
Pfaff: Analysis of Integrated and Cointegrated Time Series with R (2nd ed. 2008)
Sarkar: Lattice: Multivariate Data Visualization with R
Spector: Data Manipulation with R

Douglas A. Luke

A User's Guide to Network Analysis in R

Springer

Douglas A. Luke
Center for Public Health Systems Science
George Warren Brown School
of Social Work
Washington University
St. Louis, MO, USA

ISSN 2197-5736 ISSN 2197-5744 (electronic)
Use R!
ISBN 978-3-319-23882-1 ISBN 978-3-319-23883-8 (eBook)
DOI 10.1007/978-3-319-23883-8

Library of Congress Control Number: 2015955739

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

To my most important social network—Sue, Alina, and Andrew

Preface

In early 2000, Stephen Hawking said that "...the next century will be the century of complexity." If his prediction is true, the implication is that we will need new scientific theories, data collection methods, and analytic techniques that are appropriate for the study of complex systems and behavior. Network science is one such approach that views the world through a network lens, where physical and social systems are made up of heterogeneous actors who are connected to one another through different types of relational ties. Network analysis is the set of analytic tools used to study these types of systems. Over the past several decades network analysis has become an increasingly important part of the analytic toolbox for social, health, and physical scientists.

Until recently, network analysis required specialized software, both for network data management and analyses. However, starting around 2000, network analytic tools became available in the `R` statistical programming environment. This not only made network analytic techniques more visible to the broader statistical community but also provided the breadth and power of `R`'s data management, graphic visualization, and general statistical modeling capabilities to the network analyst community.

As the title suggests, this book is a user's guide to network analysis in `R`. It provides a practical hands-on tour of the major network analytic tasks that can currently be done in `R`. The book concentrates on four primary tasks that a network analyst typically concerns herself with: network data management, network visualization, network description, and network modeling. The book includes all the R code that is used in the network analysis examples. It also comes with a set of network datasets that are used throughout the book. (See Chap. 1 for more details on the structure of the book, as well as instructions on how to obtain the network data.) The book is written for anybody who has an interest in doing network analysis in `R`. It can be used as a secondary text in a network science or analysis class or can simply serve as a reference for network techniques in `R`.

This book would not exist without the help, support, guidance, and mentoring I have received over the last 30 years from my own personal and professional social networks. In the mid-1980s I took a graduate network analysis class from Stan Wasserman at the University of Illinois in Champaign. I remember being excited

about this new way to analyze data, but thought that I was not likely to ever use it in my career. However, my colleagues in psychology and public health encouraged me in my early work exploring how network analysis could answer important research and evaluation questions. These include Julian Rappaport, Ed Seidman, Bruce Rapkin, Kurt Ribisl, Sharon Homan, Ross Brownson, and Matt Kreuter. Whether they know it or not, I have been inspired and encouraged by an amazing group of network and systems scientists, including Tom Valente, Steve Borgatti, Martina Morris, Tom Snijders, Scott Leischow, Patty Mabry, Stephen Marcus, and Ross Hammond. My best network ideas have come from my friends and colleagues at the Center for Public Health Systems Science, particularly Bobbi Carothers, Amar Dhand, Chris Robichaux, and Nancy Mueller. I am especially grateful to the students in my network analysis classes and workshops over the years; they have not only improved this book, but they have improved my thinking about network analysis. A very special thank you to Jenine Harris. Jenine was my first doctoral student, now I am inspired by the rigor and elegance of her own work in network science. I would also like to thank the Centers for Disease Control and Prevention, the National Institutes of Health, and the Missouri Foundation for Health for providing research and evaluation support that allowed me to develop and refine my approach to network analysis. Finally, my deepest thanks go to my family. They gave me specific suggestions about the content, provided me space and time to work hard on this book (including a crucial Father's Day gift), and cheered me on when I most needed it. Thank you, Sue, Ali, and Andrew.

St. Louis, MO, USA Douglas A. Luke
July, 2015

Contents

Chapter 1
Introducing Network Analysis in R

> *Begin at the beginning,* the King said, very gravely, *and go on till you come to the end: then stop. (*Lewis Carroll, *Alice in Wonderland)*

1.1 What Are Networks?

This book is a user's guide for conducting network analysis in the `R` statistical programming language. Networks are all around us. Humans naturally organize themselves in networked systems. Our families and friends form personal social networks around each of us. Neighborhoods and communities organize themselves in networked coalitions to advocate for change. Businesses work with (and against) each other in complex, interlocking networks of trade and financial partnerships. Public health is advanced through partnerships and coalitions of governmental and NGO organizations (Luke and Harris 2007). Nations are connected to one another through systems of migration, trade, and treaty obligations.

Moreover, non-human networks exist almost anywhere you look. Our genes and proteins interact with one another through complex biological networks. The human brain is now viewed as a complex network, or 'connectome' (Sporns 2012). Similarly, human diseases and their underlying genetic roots are connected as a 'diseasome' (Barabási 2007). Animal species interact in many complex ways, one of which is a networked food-web that describes interactions in 'who-eats-whom' relationships. Information itself is networked. Our legal system is built on an interconnecting network of prior legal decisions and precedents. Social and scientific progress is driven by a diffusion of innovation process by which information is disseminated across connected social systems, whether they are Iowa corn farmers (Rogers 2003) or public health scientists (Harris and Luke 2009). It appears that one of the ways the universe is organized is with networks.

So what is a network? Figures 1.1 and 1.2 present two examples of important and interesting social networks. Figure 1.1 presents the contact network of the 19 9–11 hijackers, based on the work of Valdis Krebs (2002). Every social network is made up of a set of actors (also called nodes) that are connected to one another via some type of social relationship (also called a tie). In the figure, nodes are the circles and the ties are the lines connecting some of the nodes. The network shows

D.A. Luke, *A User's Guide to Network Analysis in R*, Use R!,
DOI 10.1007/978-3-319-23883-8_1

us that the hijackers had some contact with one another before September 11th, but the network is not very densely connected and there appears to be no prominent network member who is connected to all or even most of the other hijackers.

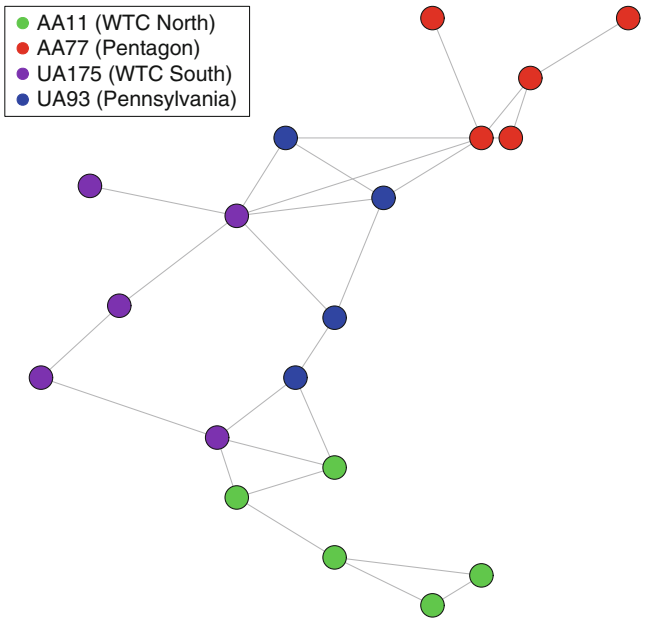

Fig. 1.1 Network of 9–11 hijackers

The second example in Fig. 1.2 is from a very different sort of social network. Here the nodes are members of the 2010 Netherlands FIFA World Cup team, who went on to lose in the final to Spain. The ties represent passes between the different players during the World Cup matches. The arrows show the directional pattern of the passes. We can see that the goalkeeper passed primarily to the defenders, and the forwards received passes primarily from the midfielders (except for #6, who appears to have a different passing pattern than the other two forwards).

These two examples may appear to have little in common. However, they both share a fundamental characteristic common to all social networks. The social patterns that are displayed in the network figures are not random. They reflect underlying social processes that can be explored using network science theories and methods. The terrorist network has no prominent leader and is not tightly interconnected because it makes the network harder to detect or disrupt. The pattern of passing ties in the soccer network reflects the assigned positions of the players, the rules of the game, and the strategies of the coach. The network analysis does not 'know' about any of those rules or strategies. Yet, network analysis can be used to reveal these patterns that reflect the underlying rules and regularities.