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.

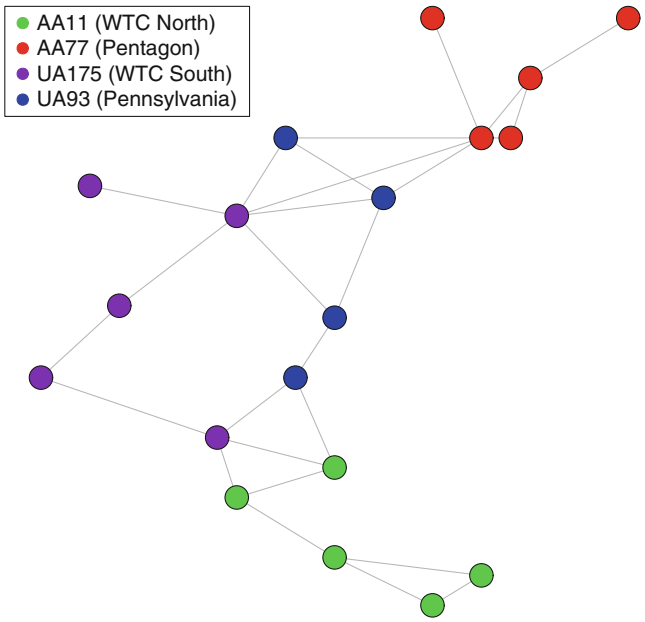

Fig. 1.1 Network of 9–11 hijackers

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.

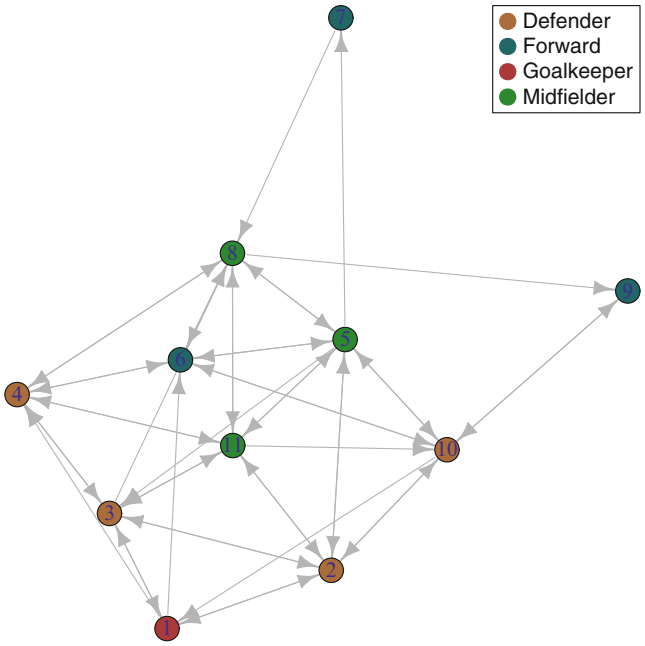

Fig. 1.2 Network of Netherlands 2010 World Cup soccer team

1.2 What Is Network Analysis?

Network science is a broad approach to research and scholarship that uses a relational lens to study and understand biological, physical, social, and informational systems. The primary tool for network scientists is network analysis, which is a set of methods that are used to (1) visualize networks, (2) describe specific characteristics of overall network structure as well as details about the individual nodes, ties, and subgroups within the networks, and (3) build mathematical and statistical models of network structures and dynamics. Because the core question of network science is about relationships, most of the methods used in network analysis are quite distinct from the more traditional statistical tools used by social and health scientists.

Network analysis as a distinct scientific enterprise with its own theories and methods grew out of developments in many other disciplines, particularly graph theory and topology in mathematics, the study of kinship systems in anthropology, and social groups and process from sociology and psychology. Although network analysis was not invented by one person at a specific place and time, the initial development of what we now recognize as modern network analysis can be traced back to the work of Jacob Moreno in the 1930s. He defined the study of social relations as sociometry, and founded the journal *Sociometry* that would publish the early studies in this area. He also invented the sociogram, which was a visual way to display

network structures. The first published sociogram appeared in the New York Times in 1933, and it was a network diagram of the friendship ties among a 4th grade class. (These data are available as part of the network dataset package that accompanies this book, see Sect. 1.4.3 below.)

The theories and methods of network analysis were developed throughout the rest of the twentieth century, with important contributions from sociology, psychology, political science, business, public health, and computer science. Network science as an empirical practice was propelled by the development of a number of network specific software tools and packages, including UCINet, STRUCTURE, Negopy, and Pajek. The interest in network science has exploded in the last 20–30 years, driven by at least three different factors. First, mathematicians, physicists, and other researchers developed a number of influential theories of network structure and formation that brought attention and energy to network science (see Chap. 10 for some discussion of these theories). Second, advances in computational power and speed allowed network methods to be applied to large and very large networks, such as the internet, the population of the planet, or the human brain. Finally, advances in statistical network theory allowed analysts for the first time to move beyond simple network description to be able to build and test statistical models of network structures and processes (see Chaps. 11 and 12).

1.3 Five Good Reasons to Do Network Analysis in R

As the title suggests, this book is designed as a general guide for how to do network analysis in the R statistical language and environment. Why is R an ideal platform for developing and conducting network analyses? There are at least five good reasons.

1.3.1 Scope of R

The R statistical programming language and environment comprise a vast integrated system of thousands of packages and functions that allow it to handle innumerable data management, analysis, or visualization tasks. The R system includes a number of packages that are designed to accomplish specific network analytic tasks. However, by performing these network tasks within the R environment, the analyst can take advantage of any of the other capabilities of R. Most other network analysis programs (e.g., Pajek, UCINet, Gephi) are stand-alone packages, and thus do not have the advantages of working within an integrated statistical programming environment.

1.3.2 Free and Open Nature of R

One of the important reasons for R's popularity and success is its free and open nature. This is formally ensured via the GNU General Public License (GPL) that R-code is released under. More informally, there is a vast R user and developer community which is continually working to enhance and improve R base code and the thousands of R packages that can be freely accessed. The social network capabilities of R described in this book have, in fact, been developed by the R user community. This open nature of R facilitates faster (and arguably, cleaner and more powerful) development and dissemination of new statistical and data analytic techniques, such as these network analytic tools.

1.3.3 Data and Project Management Capabilities of R

Although there are many good network analysis programs available which can handle a wide variety of network descriptive statistics and visualization tasks, no other network package has the same power to handle often complex data and project management tasks for larger-scale network analyses compared to R. First, as suggested above, network analysis in R can take advantage of the powerful data management, cleaning, import and export capabilities of base R. As described in Chap. 3, network analysis often starts by importing and transforming data from other sources into a form that can be analyzed by network tools. All network packages have some data management capabilities, but no other program can match R's breadth and depth.

Second, when conducting sophisticated scientific or commercial network analyses, it is important to have the right project management tools to facilitate code storage and retrieval, managing analysis outputs such as statistical results and information graphics, and producing reports for internal and external audiences. Traditional statistical analysis platforms such as SAS and SPSS have these sorts of tools, but most network programs do not. By pairing R up with an integrated development environment (IDE) such as RStudio (`http://rstudio.org/`) and taking advantage of packages such as `knitr` and `shiny`, the user has the ability to manage any type of complex network project. In fact, the development and availability of these tools has been one of the driving forces of the *reproducible research* movement (Gentleman and Lang 2007), which emphasizes the importance of combining data, code, results, and documentation in permanent and shareable forms. As one example of the power of the reproducible research tools accessible in R is this book, which was created entirely in RStudio.

1.3.4 Breadth of Network Packages in R

The primary reason R is ideal for network analysis is the breadth of packages that are currently available to manage network data and conduct network visualization, network description, and network modeling. There are dozens of network-related packages, and more are being created all the time. R network data can be managed and stored in R native objects by the `network` and `igraph` packages, and the data can be exchanged between formats with the `intergraph` package. Basic network analysis and visualization can be handled with the `sna` package contained within the much broader `statnet` suite of network packages, as well as within `igraph`. More sophisticated network modeling can be handled by `ergm` and its associated libraries, and dynamic actor-based network models are produced by `RSiena`. Free-standing network analysis programs have many strengths (e.g., the visualization capabilities of Gephi), but no single program matches the combined power of the social network analysis packages contained in R.

1.3.5 Strength of Network Modeling in R

Finally, the particular network modeling strengths of R should be mentioned. R is the only generally available software package that includes comprehensive facilities to do stochastic network modeling (e.g., exponential random graph models), dynamic actor-based network models that allow study of how networks change over time, and other network simulation procedures.

1.4 Scope of Book and Resources

1.4.1 Scope

As the title suggests, the goal of this book is to provide a hands-on, practical guide to doing network analysis in the `R` statistical programming environment. It is hands-on in the sense that the book provides guidance primarily in the form of short network analysis code snippets applied to realistic network data. The results of the analyses follow immediately. All the code and data are available to the reader, so that it is easy to replicate what is shown in the book, experiment with your own data or code extensions, and thus facilitate learning.

The practical goal of the book is to demonstrate network analytic techniques in `R` that will be useful for a wide variety of data analysis and research goals. This includes data management, network visualization, computation of relevant network descriptive statistics, and performing mathematical, statistical, and dynamic

modeling of networks. The intended audiences include students, analysts and researchers across a wide variety of disciplines, particularly the social, health, business, and engineering domains.

It is also useful to state what this book is not designed to do. First, it does not provide an in-depth treatment of network science theories or history. There are many good books, papers, training courses, and online resources available that cover this material. For good general overviews, the classic text by Wasserman and Faust (1994) is still relevant, and John Scott provides a good, more current treatment (2012). For more in-depth treatment of network science and statistical theory, see Newman (2010) or Kolaczyk (2009). Finally, two edited volumes that have good coverage of the recent history of network science as well as well-executed examples of empirical network research are Newman et al. (2006) and Scott and Carrington (2011).

Second, this book is not in any way an adequate introduction to `R` programming and statistical analysis. Although every attempt is made to make each code example clear and succinct, a novice `R` user will find some of the techniques and code syntax hard to follow. In particular, understanding `R`'s capabilities for data management, graphics, and the object-oriented approach to statistical modeling will be very helpful for getting the most out of this user-guide.

Thus, the book is designed for the interested student, analyst, or researcher who is familiar with `R` and has some understanding of network science theories and methods. It could serve as a secondary text for a graduate level class in network analysis. It also could be useful as a primer for an experienced `R` analyst who wants to incorporate network analysis into her programming and analytic toolbox.

1.4.2 Book Roadmap

The book is organized into four main sections, which correspond to the four fundamental tasks that network analysts will spend most of their time on: data management, network visualization, network description, and network modeling. The first section has two chapters that cover both a simple introduction to basic network techniques, then a more in-depth presentation of data management issues in network analysis. The three chapters in the Visualization section cover basic network graphics layout, network graphic design suggestions, and some discussion of advanced graphics topics and techniques. The Description and Analysis section has three chapters that cover the most widely used techniques for describing important network characteristics, including actor prominence, network subgroups and communities, and handling affiliation networks. The final section, Modeling, includes four chapters that present advanced techniques for mathematical modeling, statistical modeling, modeling of dynamic networks, and network simulations. Table 1.1 presents this roadmap.

Chapter	Packages	Datasets
Introduction		FIFA_Nether, Krebs
5 number summary	statnet, sna	Moreno
Network data	statnet, network, igraph	DHHS, ICTS
Basic visualization	statnet, sna	Moreno, Bali
Graphic design	statnet, sna, igraph	Bali
Advanced graphics	arcdiagram, circlize, visNetwork, networkD3	Simpsons, Bali
Prominence	statnet, sna	DHHS, Bali
Subgroups	igraph	DHHS, Moreno, Bali
Affiliation networks	igraph	hwd
Mathematical models	igraph	lhds
Stochastic models	ergm	TCnetworks
Dynamic models	RSiena	Coevolve
Simulations	igraph	

Table 1.1 User's Guide roadmap

1.4.3 Resources

The most important resource for this user guide is a collection of network datasets that have been curated and made available to the readers of this book. Over a dozen network datasets are included in the form of an R package called `UserNetR`. These datasets are used throughout the book to support the coding and analysis examples. The network data included in the `UserNetR` package mostly come from published network studies, while a few are created to help illustrate particular analytic options. Table 1.1 lists the names of the datasets that are featured in each chapter.

The `UserNetR` package is maintained on GitHub, and must be downloaded and installed to make the network data available. This can be done using the following code. (The `devtools` package must also be installed if it is not on your system.)

```
library(devtools)
install_github("DougLuke/UserNetR")
```

Once this is done, the package must be loaded to make the various datafiles available. This can be done with the `library()` function, just like for any `R` package. This command will not always be explicitly shown throughout the book, so make sure to load the package prior to executing any of the included `R` code.

```
library(UserNetR)
```

Finally, the documentation for the `UserNetR` package can be viewed through the `R` help system.

```
help(package='UserNetR')
```