Chapter 2
The Network Analysis 'Five-Number Summary'

There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after. (J.R.R. Tolkien – *The Hobbit)*

2.1 Network Analysis in R: Where to Start

How should you start when you want to do a network analysis in R? The answer to this question rests of course on the analytic questions you hope to answer, the state of the network data that you have available, and the intended audience(s) for the results of this work. The good news about performing network analysis in R is that, as will be seen in subsequent chapters, R provides a multitude of available network analysis options. However, it can be daunting to know exactly where to start.

In 1977, John Tukey introduced the five-number summary as a simple and quick way to summarize the most important characteristics of a univariate distribution. Networks are more complicated than single variables, but it is also possible to explore a set of important characteristics of a social network using a small number of procedures in R.

In this chapter, we will focus on two initial steps that are almost always useful for beginning a network analysis: simple visualization, and basic description using a 'five-number summary.' This chapter also serves as a gentle introduction to basic network analysis in R, and demonstrates how quickly this can be done.

2.2 Preparation

Similar to most types of statistical analysis using R, the first steps are to load appropriate packages (installing them first if necessary), and then making data available for the analyses. The `statnet` suite of network analysis packages will be used here for the analyses. The data used in this chapter (and throughout the rest of the book) are from the `UserNetR` package that accompanies the book. The specific dataset used here is called `Moreno`, and contains a friendship network of fourth grade students first collected by Jacob Moreno in the 1930s.

D.A. Luke, *A User's Guide to Network Analysis in R*, Use R!,
DOI 10.1007/978-3-319-23883-8_2

```
library(statnet)
library(UserNetR)
data(Moreno)
```

2.3 Simple Visualization

The first step in network analysis is often to just take a look at the network. Network visualization is critical, but as Chaps. 4, 5 and 6 indicate, effective network graphics take careful planning and execution to produce. That being said, an informative network plot can be produced with one simple function call. The only added complexity here is that we are using information about the network members' gender to color code the nodes. The syntax details underlying this example will be covered in greater depth in Chaps. 3, 4 and 5.

```
gender <- Moreno %v% "gender"
plot(Moreno, vertex.col = gender + 2, vertex.cex = 1.2)
```

The resulting plot makes it immediately clear how the friendship network is made up of two fairly distinct subgroups, based on gender. A quickly produced network graphic like this can often reveal the most important structural patterns contained in the social network.

2.4 Basic Description

Tukey's original five-number summary was intended to describe the most important distributional characteristics of a variable, including its central tendency and variability, using easy to produce statistical summaries. Similarly, using only a few functions and lines of R code, we can produce a network five-number summary that tells us how *large* the network is, how *densely* connected it is, whether the network is made up of one or more distinct *groups*, how *compact* it is, and how *clustered* are the network members.

2.4.1 Size

The most basic characteristic of a network is its size. The size is simply the number of members, usually called nodes, vertices or actors. The `network.size()` function is the easiest way to get this. The basic summary of a `statnet` network object also provides this information, among other things. The Moreno network has 33

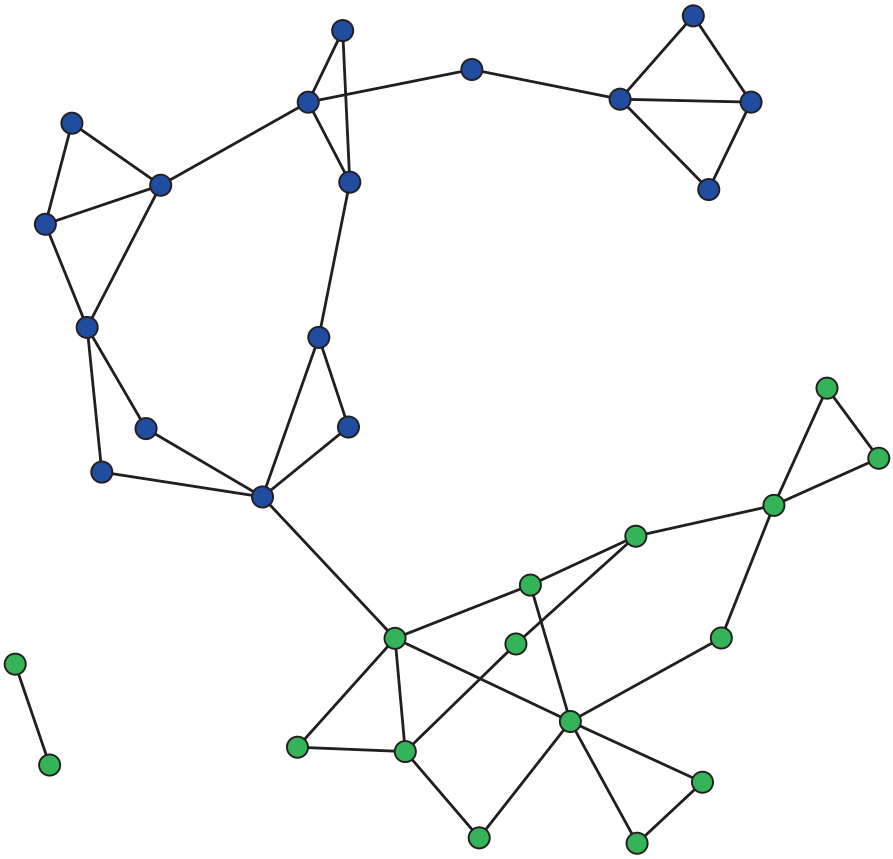

Fig. 2.1 Moreno sociogram

members, based on the network.size and summary calls. (Setting the `print.adj` to false suppresses some detailed adjacency information that can take up a lot of room.)

```
network.size(Moreno)

  ## [1] 33

summary(Moreno,print.adj=FALSE)

  ## Network attributes:
  ##   vertices = 33
  ##   directed = FALSE
  ##   hyper = FALSE
  ##   loops = FALSE
  ##   multiple = FALSE
```

```
##   bipartite = FALSE
##  total edges = 46
##    missing edges = 0
##    non-missing edges = 46
##  density = 0.0871
##
## Vertex attributes:
##
##  gender:
##    numeric valued attribute
##    attribute summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    2.00    1.52    2.00    2.00
##   vertex.names:
##    character valued attribute
##    33 valid vertex names
##
## No edge attributes
```

2.4.2 Density

Of all the basic characteristics of a social network, density is among the most important as well as being one of the easiest to understand. Density is the proportion of observed ties (also called edges, arcs, or relations) in a network to the maximum number of possible ties. Thus, density is a ratio that can range from 0 to 1. The closer to 1 the density is, the more interconnected is the network.

Density is relatively easy to calculate, although the underlying equation differs based on whether the network ties are directed or undirected. An undirected tie is one with no direction. Collaboration would be a good example of an undirected tie; if A collaborates with B, then by necessity B is also collaborating with A. Directed ties, on the other hand, have direction. Money flow is a good example of a directed tie. Just because A gives money to B, does not necessarily mean that B reciprocates. For a directed network, the maximum number of possible ties among k actors is $k*(k-1)$, so the formula for density is:

$$\frac{L}{k \times (k-1)},$$

where L is the number of observed ties in the network. Density, as defined here, does not allow for ties between a particular node and itself (called a loop).

For an undirected network the maximum number of ties is $k*(k-1)/2$ because non-directed ties should only be counted once for every dyad (i.e., pair of nodes). So, density for an undirected network becomes:

$$\frac{2L}{k \times (k-1)}.$$

The information obtained in the previous section told us that the Moreno network has 33 nodes and 46 non-directed edges. We could then use R to calculate that by hand, but it is easier to simply use the `gden()` function.

```
den_hand <- 2*46/(33*32)
den_hand

  ## [1] 0.0871

gden(Moreno)

  ## [1] 0.0871
```

2.4.3 Components

A social network is sometimes split into various subgroups. Chapter 8 will describe how to use R to identify a wide variety of network groups and communities. However, a very basic type of subgroup in a network is a component. An informal definition of a component is a subgroup in which all actors are connected, directly or indirectly. The number of components in a network can be obtained with the `components` function. (Note that the meaning of components is more complicated for directed networks. See `help(components)` for more information.)

```
components(Moreno)

  ## [1] 2
```

2.4.4 Diameter

Although the overall size of a network may be interesting, a more useful characteristic of the network is how compact it is, given its size and degree of interconnectedness. The diameter of a network is a useful measure of this compactness. A path is the series of steps required to go from node A to node B in a network. The shortest path is the shortest number of steps required. The diameter then for an entire network is the longest of the shortest paths across all pairs of nodes. This is a measure of compactness or network efficiency in that the diameter reflects the 'worst

case scenario' for sending information (or any other resource) across a network. Although social networks can be very large, they can still have small diameters because of their density and clustering (see below).

The only complicating factor for examining the diameter of a network is that it is undefined for networks that contain more than one component. A typical approach when there are multiple components is to examine the diameter of the largest component in the network. For the Moreno network there are two components (see Fig. 2.1). The smaller component only has two nodes. Therefore, we will use the larger component that contains the other 31 connected students.

In the following code the largest component is extracted into a new matrix. The geodesics (shortest paths) are then calculated for each pair of nodes using the `geodist()` function. The maximum geodesic is then extracted, which is the diameter for this component. A diameter of 11 suggests that this network is not very compact. It takes 11 steps to connect the two nodes that are situated the furthest apart in this friendship network.

```
lgc <- component.largest(Moreno,result="graph")
gd <- geodist(lgc)
max(gd$gdist)

  ## [1] 11
```

2.5 Clustering Coefficient

One of the fundamental characteristics of social networks (compared to random networks) is the presence of clustering, or the tendency to formed closed triangles. The process of closure occurs in a social network when two people who share a common friend also become friends themselves. This can be measured in a social network by examining its transitivity. Transitivity is defined as the proportion of closed triangles (triads where all three ties are observed) to the total number of open and closed triangles (triads where either two or all three ties are observed). Thus, like density, transitivity is a ratio that can range from 0 to 1. Transitivity of a network can be calculated using the `gtrans()` function. The transitivity for the 4th graders is 0.29, suggesting a moderate level of clustering in the classroom network.

```
gtrans(Moreno,mode="graph")

  ## [1] 0.286
```

In the rest of this book, we will examine in more detail how the power of R can be harnessed to explore and study the characteristics of social networks. The preceding examples show that basic plots and statistics can be easily obtained. The meaning of these statistics will always rest on the theories and hypotheses that the analyst brings to the task, as well as history and experience doing network analysis with other similar types of social networks.

http://www.springer.com/978-3-319-23882-1

A User's Guide to Network Analysis in R

Luke, D.A.

2015, XII, 238 p. 92 illus., 81 illus. in color., Softcover
ISBN: 978-3-319-23882-1