The purpose of social network analysis is to investigate complex relationships and interactions amongst entities within a network. Advanced social network analysis techniques can be used to answer question such as:

* With whom will an individual develop relationships?
* How does an individual interact with others via such relationships?
* Who is central or an opinion leader in the social network to which an individual belongs?

Social network analysis techniques can be categorized as either descriptive techniques or statistical inference techniques. One of the main limitations of descriptive techniques is that it is difficult to analyze certain characteristics of social networks in a relative sense. For example, if a particular social network has a certain number of ties among the members, with descriptive methods we can hardly determine the degree to which the members of the network are densely connected. On the other hand, statistical inference techniques enable us to produce appropriate conclusions about network data that goes beyond the immediate scope of the data.

The aim of this tutorial is to give you a basic introduction to statistical inference of network data with RStudio. The tutorial will be structured as follows:

* Understanding network data in relation to statistics.
* Analysing the dataset.
* Describing one network.
* Exercise one: calculating the density of the network
* Understanding the questions – how to compare datasets to get characteristics such as density, how to do regression with datasets.
* Answering the question 1.
* Answering question 2.

UNDERSTANDING NETWORK DATA IN RELATION TO STATISTICS

Before we start, there are two essential concepts that you must understand about applying statistics to network data.

Firstly, social network analysis is about relations among actors, not about relations between variables. This is extremely important to understand because it means that rather than describing distributions of attributes of actors (or “variables), we are describing the distributions of relations among actors. Therefore, in applying statistics to network data, we are concerned with issues like the average strength of relations between actors, for example:

*“Is the strength of ties between actors in a network correlated with the centrality of the actors in the network?”*

Secondly, many of the standard inferential statistic tools that are concerned with the distribution of attributes cannot be applied directly to network data. This is because, unlike attribute analysis, the observations in network data are not independent samplings from populations. For example, in network data, we might observe Chad’s tie with Darren, while another observation might be Darren’s tie with Ryan. Network analysis says that we cannot suppose that these relations are independent because Darren is involved in two of the observations and there is a strong likelihood that we will infer another observation which depicts Chad’s tie with Ryan. Standard inferential tests assume independent observations and therefore, if we apply standard inferential statistics to observations that are not independent, it will produce seriously misleading results.

In order to overcome these challenges, when applying statistics to network data, we need to use alternative numerical approaches. These approaches are known as “bootstrapping” and permutations but for now, let’s start with the tutorial.

UNDERSTANDING THE DATA

This tutorial is purely for introductory purposes and so, the following sections’ objective is to walk you through two exercises which will give you a basic understanding of how to apply inferential statistics to network data.

For this tutorial, we will be using the Knoke and Knoki datasets which describes two relations; the exchange of information and the exchange of money among 10 organizations. The following code exhibits the money exchange matrix and the information exchange matrix:

As you can see, network data matrices consist of a square array of measurements. The rows of the array are the same set of observations, while each cell of the array describes a relationship between actors. In our dataset, each cell represents either the exchange of money or information between two organisations as a 1 or no exchange as a 0. By comparing rows of the array we can see which actors are similar to which other actors in whom they choose to exchange information or money. By looking at the columns, we can see who is similar to whom in terms of being chosen by others. This helps us to see which actors have similar positions in the network which brings us to the first major emphasis of network analysis: how are actors located or embedded within the overall network?

However, a network analyst will also look at the data in a more holistic way. For example, they may notice that there are about equal numbers of ones and zeros in the matrix which suggests that there is a moderate density of liking overall. However, the analyst may want to test prior made hypotheses about the density or mean tie strength of a network. For example, we might want to test:

* Whether the hypothesis that, the proportion of binary ties present, differs from .50?
* Whether the hypothesis that, the average strength of a valued tie, differs from 3?

This will bring us to the first exercise which involves finding the degree to which the members of a network are densely connected.

DESCRIBING ONE NETWORK

Before we start with the first exercise, let’s do an in-depth analysis of one of the matrices, we will use the Knoke matrix, in order to solidify our understanding of the various statistical values which we may observe.

*Knoke dataset*

Firstly, it is important to note that this particular dataset is asymmetric and binary. The scale of measurement, binary or valued, matters in making proper choices about interpretation and application of many statistical tools. The data which we are observing is the relations between the actors. So, in each matrix we have 10 x 10 = 100 observations or cases. For many analyses, the ties of actors with themselves (the main diagonal) are not meaningful and so there is actually N \* N-1 = 90 observations. If the data were symmetric (Xij = Xji), half of the observations would be redundant and so there would be N \* N-1 / 2 = 45 observations, however, this is not the case for our dataset.

The first thing we are going to do is summarise some of the most basic characteristics of the distribution of these scores. To do this, we are going to generate the most commonly used statistical measures for the Knoke matrix.

*Show the mean, SD, sum , variance, Minimum, Maximum and Num of observations for the Knoke matrix.*

By looking at the data we can see that the number of observations is 90 which range from a minimum score of 0 to a maximum score of 1. The sum of the ties is “x” and the average value (mean) of the ties is x/90 = i. Due to the fact that our relations have been coded as a “dummy” variables (0 for no relation, 1 for a relation), the mean is also the proportion of possible ties that are present (or the density) within the network. What we can also calculate is the coefficient of variation, SD / mean x 100 = “t”. This suggests quite a lot of variation as the percentage of the average score.

Adding onto this, we might also want to examine the distribution of ties for each actor, opposed to the distribution of ties for the entire network.

*Show the mean for each actor in Knoke matrix*

By analysing the network of rows for the Knoke matrix we can see that actor 1 has a mean (or density) of tie sending of “x”. This means that this actor sent “x x 100”

In order to discover the degree to which the members of a network are densely connected, we need to be able to compare the network to a randomly generated network of the same size. This is because we need to be able to compare the value of the density or average tie strength of the network against a test value.

This brings us back to the numerical approaches of “bootstrapping” and permutations.

The reason for this is that if we are able to compare particular characteristics of a social network with another network that possesses similar characteristics, it makes it possible for us to say whether the observed network is more or less likely to have the same characteristics of the compared network. Therefore, by comparing our network to a randomly generated network, we can determine the degree to which our network has the characteristic of interest compared to a network generated by chance.

If our network has a higher degree than another network then or network degree is stronger but you repeat these randomly generated networks so that you can see.

The analyst might also compare the cells above and below the diagonal to see if there is reciprocity in choices (e.g. Bob chose Ted, did Ted choose Bob?). This is the second major emphasis of network analysis: seeing how the whole pattern of individual choices gives rise to more holistic patterns.