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 1:
  + Univariate analysis: calculating the density of 1 network against a hypothesis.
  + Bivariate analysis: calculating the density of 2 networks against a hypothesis.
* Exercise 2:correlation between networks.
* Exercise 3: network regression.
* Conclusion

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.

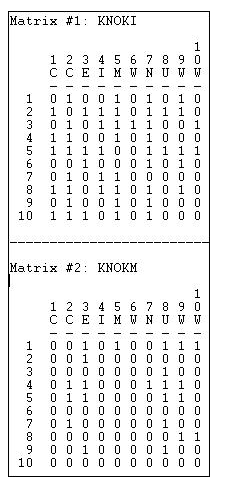
In order to overcome these challenges, when applying statistics to network data, we need to use alternative numerical approaches. These approaches are known as “bootstrapping” and permutations but for now, let’s start with the tutorial.

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 how to create a money exchange matrix and how to create an information exchange matrix:

*Put Matrices here so that it the end output look something like this:*

**

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:

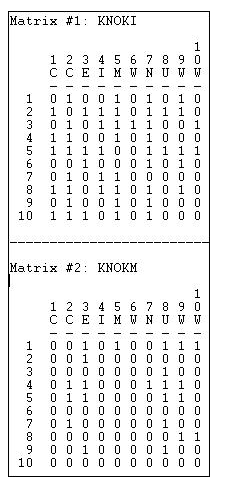
* Whether the hypothesis that, the proportion of binary ties present, differs from .50?
* Whether the hypothesis that, the average strength of a valued tie, differs from 3?

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 and 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.

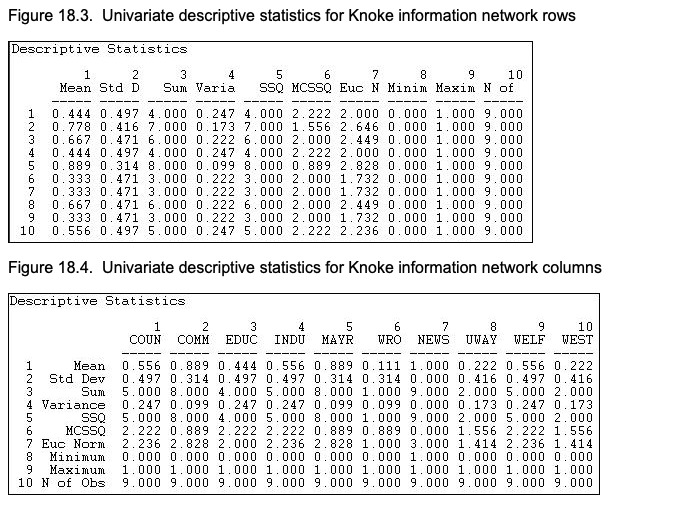
*Show the mean, SD, sum , variance, Minimum, Maximum and Num of observations for the Knoke matrix so that we have something like this:*

**

By looking at the data we already know that the number of observations is 90 which range from a minimum score of 0 to a maximum score of 1. By analysing the summary of the matrix we can observe that the sum of the ties (total edges) is 22 and the average value (density) of the ties is 22/90 = 0.244. 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.

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 (specify row network and column network) so that the end output looks something like this (THE VALUES WILL BE DIFFERENT BECAUSE THIS IS FOR THE* ***KNOKI*** *AND WE ARE DOING* ***KNOKE****):*

**

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(0,444)”. This means that the actor sent “t(4)” ties to the available 9 actors- this can be inferred by looking at the ‘sum’ column for actor 1 in the row network. If we look at the column network we can see that actor 1 received somewhat more money than they sent because their column mean is “y(0,556)”. Adding onto this, if we scan down the column of means we can see that there is a considerable amount of variability across actors for sending and receiving money- some send more money while others get more money.

It is helpful to note that if we were working with valued data, rather than binary data, the measure of variability would be more informative across actors. This is because the variability of binary data is strictly a function of its mean. The main objective of this section was to emphasise the first essential concept that when we use statistics to describe network data, we are describing properties of the distribution of relations (or ties among actors) rather than properties of the distribution of attributes across actors.

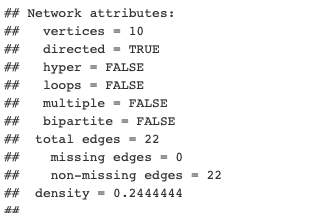
EXERCISE ONE: UNIVARIATE ANALYSIS

Now that we have a better understanding of a single network, and the general statistical measures of that network, we can begin with the first exercise. As previously mentioned, network analysts are usually interested in a more holistic view of the network which usually involves the density, or average tie strength, of the network. It is typical for an analyst to want to prove or disprove a theory about the density of a network and for this we need to create a hypothesis (for the purpose of this exercise we will be working with the Knoki matrix):

Hypothesis*: “Let’s suppose that we think that all organisations have the tendency to want to directly distribute information to all other organisations in the network, as a way of legitimizing themselves.”*

This hypothesis suggests that all the organisations will exchange information with all other organisations and therefore each cell in the matrix will = 1. Therefore, if this hypothesis/theory is correct, then the density of the Knoki network should be 1.0. We are now able to compare the true value of the network density, or average tie strength, against a test value (density = 1.0).

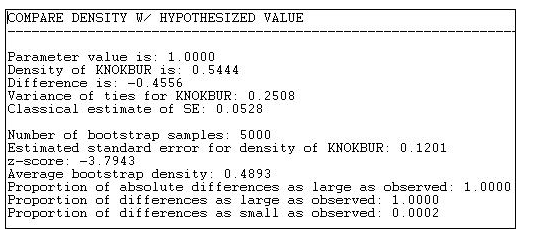
*Calculate mean of Knoki: give the summary of the Knoki dataset (summary(Knonki)) so that they can see what the value for density is, it should be 0,544. So the below picture, but for the KNOKI dataset:*

**

We can see that this theory is not true because the true network density of Knoki is not equal to 1.0, it is equal to 0.544. However, perhaps the difference between what we see (density = 0.544) and what the theory predicts (density = 1.0) is due to random variation (which occurs when the information is collected). Therefore, we need to find a way to ask the data to convince us that we can be confident in rejecting the hypothesis. To do this, we need to be able to compare the network to a randomly generated network of the same size. By comparing our network to a randomly generated network, we can determine the degree to which our network has the same density compared to a network generated by chance, allowing us to reduce the effect of random variation. This brings us back to the numerical approaches of “bootstrapping” and permutations.

Bootstrapping calculates random sub-samples from a network, for a specified number of times, to produce a sample distribution of density measures. We know that the difference between the test value (1,00) and the observed value (0,544) is equal to (0,456). Now, using the sample distribution of density measures, we can observe how often a difference this large (0,456), happens by random sample variation, if the null hypothesis (density = 1.0) was really true in the population. If the difference happens often then the null hypothesis is true, if it does not happen often then it is false.

*Run the Knoki network against the randomly generated networks and explain the code. You want output that looks something like this:*

**

We can now observe that the mean of our sampling distribution is 0.4893, the standard deviation (standard error) is 0,1201 and the test statistic (z-score) is -3,7943. ‘p’ values are significant if they fall between -0.05 and 0.05, they are insignificant if they fall outside of these values. As we can see, the first p value for the “proportion of absolute differences as large as observed” is equal to 1.0, therefore, it is insignificant and this means that the null hypothesis is not true. The last p value, which infers that “the proportion of difference is smaller than the one observed” is valued at p = 0.0002, therefore it is significant. This means that, throughout the tests, the proportion of difference is smaller than the one we observed (0,456) and this also proves that the null hypothesis is false because the proportion of difference would have to be greater than or equal to the one we observed to reach the hypothesized density of 1.0.

Remember, the reason why we use bootstrapping rather than the classical formula for the standard error of a mean (s / sqr(N)) is because the classical formula is based on the notion that all observations are independent, which is not the case for network data. If we used the standard inferential formulas we would observe unrealistically small values for our network data which would result in two kinds of errors, the false positive or rejecting the null hypothesis when we shouldn’t.

EXERCISE ONE: BIVARIATE ANALYSIS

In the above example we were working with univariate analysis (analysis on one network), we will now move on to bivariate analysis (analysis of two networks). Bivariate analysis is important because it enables network analysts to continue to infer holistic information about the network. In fact, bivariate analysis brings up the second major emphasis of network analysis: seeing how the whole pattern of individual choices gives rise to more holistic patterns. The basic question of bivariate analysis on network data is:

*“Do the pattern of ties for one relation among a set of actors align with the pattern of ties for another relation among the same set of actors?”*

In summary, this question is asking whether the relations between two networks, with the same set of actors, correlate or not? We can answer this by analysing different kinds of relations within the network, such as:

* Does the central tendency (mean) of one relation differ significantly from the central tendency (mean) of another?
* If we know that the relation of one type exists between two actors, how much does this increase (or decrease) the likelihood that the relation of another type exists between them?

We will work through the first question now but the following question will be covered in exercise 2 of this tutorial. Let’s relate the first question to our networks. Our networks describe the money and information exchange among organisations (the 2 separate matrices: Knoke and Knoki). Therefore, in relation to the first question, we are asking:

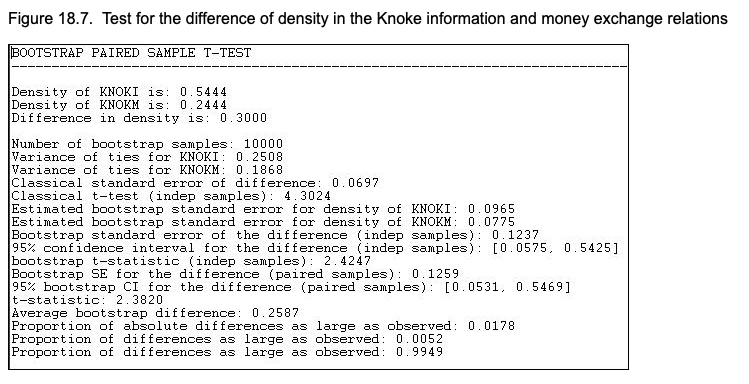
*“Which ties are more prevalent: money exchange ties or information exchange ties?”*

In other words, “which network has the higher density?”. This question is answered by analysing the densities of both networks and comparing them. However, similarly to the univariate analysis, we must make use of “bootstrapping” to ensure that our observation is accurate and not a result of random variation. We can already observe that the density of the information exchange matrix (0,544) is higher than the density of the monetary exchange matrix (0,244). Therefore, our hypothesis should be:

*Information exchange ties are more prevalent than monetary exchange ties”.*

The density of Knoki (0,544) minus the density of Knoke (0,244) is equal to 0,3. Use the following code to observe how often a difference this large (0,3) happens by random sample variation:

*Code comparing both knoki and knoke together to randomly generated networks. It must look something like this:*

**

By analysing the output we can see that we increased the number of permutations, for this test, to 10 000 (if we increase the number of random samples, I don’t know?) and that the standard deviation (classical standard error of difference) is 0,0697. Adding onto this, we can see that the first p value is equal to 0.0178, making it significant. The first p value states that “the proportion of absolute differences is as large as the one observed (0,3)” and because it is significant, it results in our hypothesis being true. This is because it maintains the observation that the Knoki matrix density is larger than the Knoke matrix density.

EXERCISE 2: CORRELATION BETWEEN NETWORKS

The objective of this exercise is to come to a conclusion about whether the relations of the two networks correlate or not. We will do this by answering the previously stated question: “If we know that the relation of one type exists between two actors, how much does this increase (or decrease) the likelihood that the relation of another type exists between them?”. In other words, if two actors have a strong tie of one type, are they also likely to have a strong tie of another? Let’s start by relating this question to our networks:

*“If two organizations exchange information, how likely is it that they will also exchange money?”*

If organizations exchange information, this may create a sense of trust, making monetary exchange relations more likely. This could also be inferred the other way, if two organisations exchange money, this may facilitate more open communications and, as a result, the exchange of information. Therefore, we might hypothesise that:

*“The matrix of information relations would be positively correlated with the matrix of monetary relations”.*

Otherwise stated, pairs that engage in one type of exchange are more likely to engage in the other. A positive correlation might be true, but it is possible that we could also observe:

* A negative correlation- where money flows in one direction and information in the other direction.
* No correlation- the two relations have nothing to do with one another.

To find out if the positive correlation is true we need to calculate the measures of nominal, ordinal and interval association between the relations in two matrices and use quadratic assignment procedure (QAP) to develop standard errors to test for the significance of an association. QAP, found in Statnet library, is a permutation test that identifies and measures the association between two relations. Essentially, QAP shuffles the dependent variable /actor data, in our case the KNOK matrices, through several permutations and then compares these coefficients with the sampling distribution of coefficients from the permuted matrices. QAP controls for network structure through these permutation. The QAP estimates standard errors for R-squared, regression coefficients and p-values, which are the proportion of time the null coefficient is smaller or equal to the observed estimate. A small drawback to performing QAP is that it is not scalable, as the iteration increases so does the memory and network size, this inevitably delays overall performance.

To perform QAP the matrices must be square with the same number of rows as columns (e.g. 10x10), the matrices must be the same size and number of nodes and only one relationship between nodes can be measured.

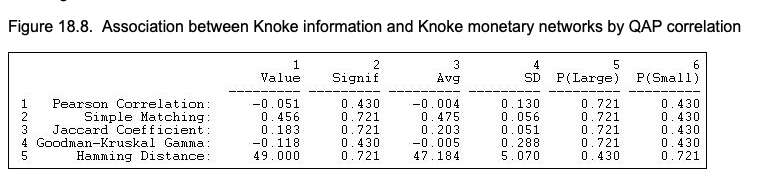
There are two QAP procedures that we will be performing in our tutorial. Namely QAP Correlation and Multiple Regression.

QAP Correlation

The QAP Correlation analysis test, tests if a statistically significant association between given actors exists, in our case two Knok network matrices representing money and information. The Pearson correlation is a standard measure for matrices which have valued relations measured at interval levels. Our QAP correlation analysis is performed using the GCOR function measuring levels of linear relationships ranging from -1 to 1. To interpret this output, a second step has to be performed where the rows and columns of the dependent matrices has to be randomly and repeatedly permuted. For our tutorial our matrices were permuted 2000 times. This creates distribution of correlations namely, p>= (0.7285), which is the probability that the correlation between the two networks are smaller or equal to the random graph/ matrix, and for p<= (0.4295), which is the probability that the correlation between the two networks are bigger than or equal to the random graph/ matrix. The distribution of correlation values also range from -1 to 1, the closer to 0 the value is the more significant the value is. These p-values are then compared to the initial Pearson correlation value of -0.0508 to determine whether there is a strong correlation or not. The QAP correlation is deemed significant if the Pearson correlation is greater than at least 95% of the distribution correlations (p-values). Stated differently, if less than 5% of the random correlations are larger than the Pearson correlation value then the QAP correlation suggests a significant correlation.

We can run the following code to correlate the Knoki matrix with the Knoke matrix:

*Code that compares Knoki to Knoke, the output must look something like this:*

**

The first column of the output (value) shows the values of 5 alternative measures of association:

* The Pearson coefficient is a standard measure when both matrices have valued relations (our dataset does not have any valued relations).
* Simple matching and the Jaccard coefficient are reasonable measures when both relations are binary (this relates to our dataset).
* Gamma would be a reasonable choice if one or both relations were measured on an ordinal scale.
* The Hamming distance is a measure of dissimilarity or distance between the scores in one matrix and the scores in the other

The third column (Avg) shows the average value of the measure of association across a large number of trials in which the rows and columns of the two matrices have been randomly permuted. In other words, what would the correlation be, on average, if we matched random actors? If we look at this value for the simple matching measure of association we can see that it is equal to 0,475. Therefore, this means that if there is a 1 in a cell in matrix one, there is a 47,5% chance there will be a 1 in the corresponding cell of matrix 2. However, due to the fact that we are analysing binary data and not valued relations, the average density of the matrices is strictly a function of its mean and therefore, the observed measures will hardly differ from a random result.

Rather, to test the hypothesis that there is an association, we will look at the proportion of random trials (p values) that would generate a coefficient as large as (or as small as, depending on the measure) the statistic actually observed. These measures are reported in the columns labelled as “P(large)” and “P(small)”, while the most appropriate value to use to test the null hypothesis (which will always state that there is no association) is shown in the “Signif” column.

Therefore, if we look at the simple matching measure we can see that the signif value is equal to 0,721. This refers to the P(large) value and because the value is > 0,05, the test is insignificant. Therefore our hypothesis is not true, meaning that the matrix of information relations is not positively correlated to the matrix of money relations and, as a result, pairs that engage in information exchange are not more likely to engage in money exchange. I know that the hypothesis is false but I don’t know if how I explained why it is false, is correct. Also, I don’t know if this means that there is a negative correlation or that there is no correlation?

EXERCISE 3: NETWORK REGRESSION

The last exercise of this tutorial will deal with network regression. Rather than correlating one relation with another, we may wish to predict one relation knowing the other. The standard tool for this type of asymmetric association is linear regression which is commonly used to predict a quantitative outcome variable/ actor or a relation based variable/ actor, which is called predictor variables. A linear regression assumes that there exists a linear relationship between the quantitative variable and the predictor variables.

The NETLM function outputs various results, a crucial output is F-statistic which should be less than 0.05 to be deemed a useful model. Additionally the coefficient of determination or R-squared is defined by the proportion of total variability explained by the regression model. The closer the R-squared value is to 1 the stronger the model fits the data.

Suppose we want to make the following hypothesis:

*“Institutionally similar organizations are more likely to exchange information”.*

Therefore we want to be able to predict which of the organisations sent information to which others. To do this, we are going to treat the information exchange network (Knoki) as our dependant network with N = 90 (there are 90 observations) and we are going to create a new matrix which we will treat as one of our independent networks (linear regression may be extended to using more than one independent variable). Use the following code to make the new matrix:

*Put in the code to make the private/non-private organisations matrix. Knokp*

As you can see, we have created another 10 x 10 matrix, coding the cell to be 1 if the interacting organisations are both private or both non-private and 0 if the interacting organisations are a mix (a private organisation interacting with a non-private organisation). We can now perform a standard multiple regression analysis by regressing each element in the information network (Knoki) on its corresponding elements in the monetary network (Knoke) and the institution network (Knokp). Knoki will be our dependant variable and Knoke and Knokp will be our two independent variables. We will again be using QAP.

*Put in the code where Knoki is the DEPENDANT variable and Knoke AND Knokp are the INDEPENDENT variables. The output should look something like this:*

**

Through analysis of the output we can observe that the R-square value is “x/0,018”. This indicates that knowing whether one organisation sends money to another, and whether the two organisations are institutionally similar (both private/non-private), reduces uncertainty in predicting an information tie by is about “x/2%” (0.018 x 100 = 1.8% so ~ 2%). From the regression coefficients we see that the intercept value is “y/0,613719”. This indicates that if two organizations are not of the same institutional type, and one does not send money to the other, the probability that one sends information to the other is “y/61%” (0,613719 x 100 = 61%).

By looking at the Knoke value within the regression coefficients we can see that it is equal to “i/-0,46”, while the Knokp value is equal to “t/-0,12”. This means that if one organization does exchange money with the other, this reduces (because of the minus sign) the probability of an information link by “i/46%”, while if two organisations are of the same institutional type, the probability of information sending is reduced (because of the minus sign) by “t/12%”. This gives us some interesting results, it suggests that monetary and informational linkages are alternative rather than re-enforcing and that institutionally similar organisations are less likely to communicate.

CONCLUSION

This brings us to the end of the tutorial. Let’s summarise what we have learnt:

* Two essential concepts about network data:
  + Network analysis is about relations among actors, NOT about relations between variables.
  + Observations in network data are NOT independent samplings from populations.
* Understanding matrices and binary data compared to valued data (the variability of binary data is strictly a function of its mean).
* Univariate analysis:
  + Comparing 1 network to randomly generated networks of the same size to prove/disprove a hypothesis.
* Bivariate analysis:
  + Comparing 2 networks to randomly generated networks of the same size to answer the question: *“Does mean/density of one relation differ significantly from the mean/density of another?”.*
  + Correlating 2 networks with the same set of actors to answer the question: *“If we know that the relation of one type exists between two actors, how much does this increase (or decrease) the likelihood that the relation of another type exists between them?”.*
* Network regression:
  + Being able to predict one relation knowing the other.