than a picture itself. As a result, it can be redrawn at any resolution/size. Not all graphics programs can read metafiles, but all Microsoft Office applications can.

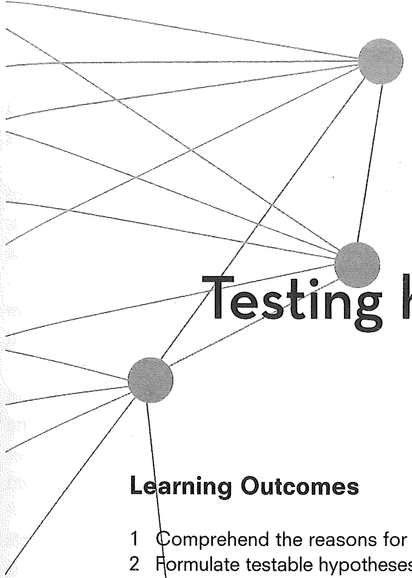
In addition to saving a graphics file, NetDraw allows the user to save the data to a text file using a format called 'VNA' (short for 'Visual Network Analysis'). The VNA format allows you to save both network and attribute information together, and also visual information such as the color and size of nodes and lines. This is handy because it means the next time you start up NetDraw, you can open a previously saved VNA file and have everything look exactly as you left it in your last session. This can save a lot of time.

7.9 Closing comments

One of the most important constraints on the valid graphical representation of social networks concerns human limitations of perception. There has been a great deal of work on this topic, and an in-depth discussion was beyond the scope of this chapter. For a good review of issues concerning human color perception and the communication of network graphical information see Krempel (2002), or for a more general review see Munzner (2000). For further reading on the history of network visualization a good source is Freeman (2000). In this chapter we discussed a number of ways in which network structural and compositional information can be communicated in network graphs through the visual manipulation of the various properties of nodes and edges and their spatial relations. In addition, we examined various means for reducing the complexity of network structures through methods for reducing the graphical complexity of networks using such mechanisms as turning on and off nodes, ties, clusters, subgroups or components in order to reveal potentially hidden structural properties. More on this can be found in Chapter 14, which discusses strategies for dealing with large networks.

7.10 Summary

The ability to visualize a social network is one of the attractive features of social network analysis. When done correctly, visualization allows the researcher to obtain a simple qualitative understanding of the network structure. The use of good layouts to emphasize properties of the network is key, and graph layout algorithms are highly effective and widely used. In addition, the use of shape, size and color to capture nodal attribute properties can further enhance the effectiveness of any visualization. In a similar way, color, thickness and line style can be used to emphasize properties of edges. We do not always want to view the whole network (particularly if it is large or complex), so we often filter nodes or edges to reveal portions of the network which are of particular interest. Changes over time provide additional challenges, but these can be addressed by using techniques such as stacked correspondence analysis.



8 Testing hypotheses

Learning Outcomes

1 Comprehend the reasons for using, and principles of, permutation tests
2 Formulate testable hypotheses at the dyadic, monadic and whole-network level
3 Understand when SIENA and exponential random graph models may be appropriate

8.1 Introduction

Padgett and Ansell (1993) collected data on the relations between Florentine families during the Renaissance. One social relation they recorded was marriage ties between families. Another one was business ties among the same set of families. An obvious hypothesis for an economic sociologist might be that economic transactions are embedded in social relations, so that those families doing business with each other will also tend to have marriage ties with one another. One might even speculate that families of this time strategically intermarried in order to facilitate future business ties (not to mention political coordination).

How would we test this? Essentially, we have two adjacency matrices, one for marriage ties and one for business ties, and we would like to correlate them. We cannot do this in a standard statistical package for two reasons. First, programs like SPSS and Stata are set up to correlate vectors, not matrices. This is not too serious a problem, however, since we could just reshape the matrices so that all the values in each matrix were lined up in a single column with $N \times N$ values.[1]

[1] Of course, we might ignore the diagonal values, yielding vectors of length $N2 - N$, and for undirected data we might ignore the redundant top half of each adjacency matrix, yielding $N(N - 1)/2$ values per variable.

We could then correlate the columns corresponding to each matrix. Second – and this is a serious problem – the significance tests used in standard statistical packages make a number of assumptions about the data which are violated by network data. For example, standard inferential tests assume that the observations are statistically independent, which, in the case of adjacency matrices, they are not. To see this, consider that all the values along one row of an adjacency matrix pertain to a single node. If that node has a special quality, such as being very antisocial, it will affect all of their relations with others, introducing a lack of independence among all those cells in the row. Another typical assumption of classical tests is that the variables are drawn from a population with a particular distribution, such as a normal distribution. Often times in network data, the distribution of the population variables is not normal or is simply unknown. Moreover, the data is probably not a random sample, and may not be a sample at all, but rather a population (e.g., you are studying the pattern of collaboration among all film studios in the world).

So we need special methods. One approach is to develop statistical models specifically designed for studying the distribution of ties in a network. This is the approach taken by those working on exponential random graph models (ERGMs) and actor-oriented longitudinal models, as exemplified by the SIENA model. Both of these are complex subjects in their own right and a detailed discussion is beyond the scope of this book. However, we will provide a highly simplified introduction to give a flavor of what is involved.

An alternative approach is to use the generic methodology of randomization tests (also called permutation tests) to modify standard methods like regression. These methods are easy to use and interpret, and can be customized for different research questions. UCINET provides a number of techniques of this type, and we begin our discussion with them.

8.2 Permutation tests

Classical significance tests are based on sampling theory and have the following logic. You measure a set of variables (say, two variables) on a sample of cases drawn via a probability sample from a population. You are interested in the relationship between the variables, as measured, say, by a correlation coefficient. So you correlate the variables using your sample data, and get a value like 0.384. The classical significance test then tells you the probability of obtaining a correlation that large given that in the population the variables are actually independent (correlation zero). When this probability is really low (less than 0.05), we call it significant and are willing to claim that the variables are actually related in the population, and not just in your sample. When the probability is higher, we feel we cannot reject the null hypothesis that the variables are independent in the population and just happen to be correlated in the sample. Note that if you have a biased sample, or you do not have a sample at all, it does not make sense to use the classical test.

The logic of randomization tests is different and does not involve samples, at least not in the ordinary sense. Suppose you believe that tall kids are favored by your particular math teacher and as a result they learn more math than short kids. So you think height and math scores in this teacher's class will be correlated. You have the teacher give all the kids a math test, measure their height, and then correlate the two variables. You get a correlation of 0.384. Hypothesis confirmed? In the world of classical statistics we would say yes, because you have a population, and the correlation is not zero, which is what you wanted to know. But let us think about this a little more. Just for fun, instead of actually giving the math test, suppose you write down a set of math scores on slips of paper, and then have each kid select his or her math score by drawing blindly from a hat. Now, you know that in this experiment a kid's math score and height are totally independent because it was completely arbitrary who got what score. And yet, could it not happen, by chance alone, all the high scores happened to go to the tall people? It may be unlikely, but it could happen. In fact, there are lots of ways (permutations) that scores could be matched to kids such that the correlation between height and score was positive (and just as many such that the correlation was negative). The question is, what proportion of all the ways the scores could have come out would result in a correlation as large as the one we actually observed (the 0.384)? In short, what are the chances of observing such a large correlation even when the values of the variables are assigned independently of each other? If the probability is high, say, 20%, we probably do not want to conclude that the teacher is biased toward tall kids. In other words, even in a population, we still want a statistical test in order to guard against spurious correlations.

The permutation test essentially calculates all the ways that the experiment could have come out given that scores were actually independent of height, and counts the proportion of random assignments yielding a correlation as large as the one actually observed. This is the '*p*-value' or significance of the test. The general logic is that one wants to compare the observed correlation against the distribution of correlations that one could obtain if the two variables were in fact independent of each other.

In the following sections we consider how randomization tests can be used to test a variety of network hypotheses. Before we start, however, it is important to remember that we may be interested in testing hypotheses at various levels of analysis. For example, one kind of hypothesis is the node-level or monadic hypothesis, such as the hypothesis that more central people tend to be happier. This kind of hypothesis closely resembles the hypotheses you encounter in non-network data analysis. The cases are single nodes (e.g., persons), and basically you have one characteristic of each node (e.g., centrality) and another characteristic of each node (e.g., test score), and you want to correlate them. That is just a matter of correlating two vectors – two columns of data – which seems simple enough, but as we will explain, there are a few subtleties involved.

Another kind of hypothesis is the dyadic one that we opened the chapter with. Here, you are hypothesizing that if a pair of persons (or, in the example, families) has a certain kind of relationship, it is more likely they will also have another kind of relationship. For instance, you might expect that the shorter the distance between people's offices in a building, the more they communicate over time. So the cases are pairs of persons (hence the label 'dyadic'), normally organized as $N \times N$ matrices, and you want to correlate the two matrices. Clearly, this is not something you would ordinarily do in a traditional statistics package.

We may also want to test a hypothesis in which one variable is dyadic, such as friendship, and the other is monadic, such as gender. The research question being asked might be something like 'does the gender of each person affect who is friends with whom?'. In this question, the monadic variable is on the independent side and the dyadic variable is on the dependent side. Another research question might be 'are people's attitudes affected by who they interact with?'. Here it is the independent variable that is dyadic and it is the dependent variable that is monadic. As we shall see, we typically test these kinds of hypotheses by rephrasing them as purely dyadic hypotheses.

Finally, another kind of hypothesis is a group- or network-level hypothesis. For instance, suppose you have asked 100 different teams to solve a problem and you have measured how long it takes them to solve it. Time-to-solution is the dependent variable. The independent variable is a measure of some aspect of the social structure of each team, such as the density of trust ties among team members. The data file looks just like the data file for node-level hypotheses, except the cases here are entire networks rather than individual nodes.

We now consider how to test each of the four kinds of hypotheses, starting with the one involving the most numerous and least aggregate cases (dyadic) and ending with the one involving the least numerous and most aggregate cases (whole networks).

8.3 Dyadic hypotheses

Network analysis packages such as UCINET provide a technique called QAP correlation that is designed to correlate whole matrices. The QAP technique correlates the two matrices by effectively reshaping them into two long columns as described above and calculating an ordinary measure of statistical association such as Pearson's r. We call this the 'observed' correlation. To calculate the significance of the observed correlation, the method compares the observed correlation to the correlations between thousands of pairs of matrices that are just like the data matrices, but are known to be independent of each other. To construct a p-value, it simply counts the proportion of these correlations among independent matrices that were as large as the observed correlation. As elsewhere, we typically consider a p-value of less than 5% to be significant (i.e., supporting the hypothesis that the two matrices are related).

To generate pairs of matrices that are just like our data matrices and yet known to be independent of each other, we a simple trick. We take one of the data matrices and randomly rearrange its rows (and matching columns). Because this is done randomly, we know that the resulting matrix is independent of the data matrix it came from. And because the new matrix is just a rearrangement of the old, it has all the same properties of the original: the same mean, the same standard deviation, the same number of 2s, the same number of cliques, etc. In addition, because we are rearranging whole rows and columns rather than individual cells, more subtle properties of the matrices are also preserved. For example, suppose one of the matrices shows the physical distance between people's homes. A property of physical distance is that if the distance from i to j is 7, and the distance from j to k is 10, then the distance from i to k is constrained to lie between 3 and 17. That means that in the matrix, the (i, j), (j, k) and (i, k) cells are not independent of each other. Given the values of any two of them, the value of the third cell cannot be just anything. When we permute the rows and columns of such a matrix, these kinds of autocorrelational properties are preserved, so when we compare the observed correlation against our distribution of correlations we can be sure we are comparing apples with apples.

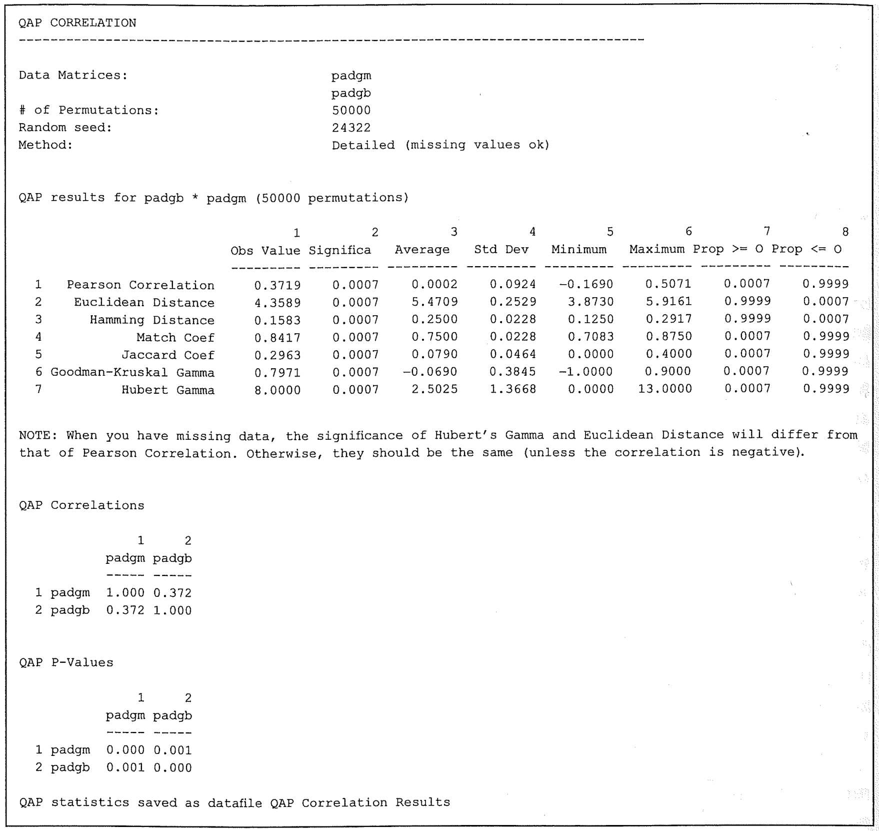
To illustrate QAP correlation, we run it on the Padgett and Ansell data described in the introduction. As shown in Figure 8.1, the correlation between the network of marriage ties and the network of business ties is 0.372, and it is highly significant ($p = 0.0007$). The results support the hypothesis that the two kinds of ties are related.

One thing to note in the output is that 50,000 permutations were used in this run. It is important to run a large number like this in order to stabilize the p-value. Since the permutations are random, if we only used a handful of them, each time we ran the program we would get a slightly different p-value (but the correlation would always be the same). The larger the sample of permutations, the less the variability in p-values.

8.3.1 QAP regression

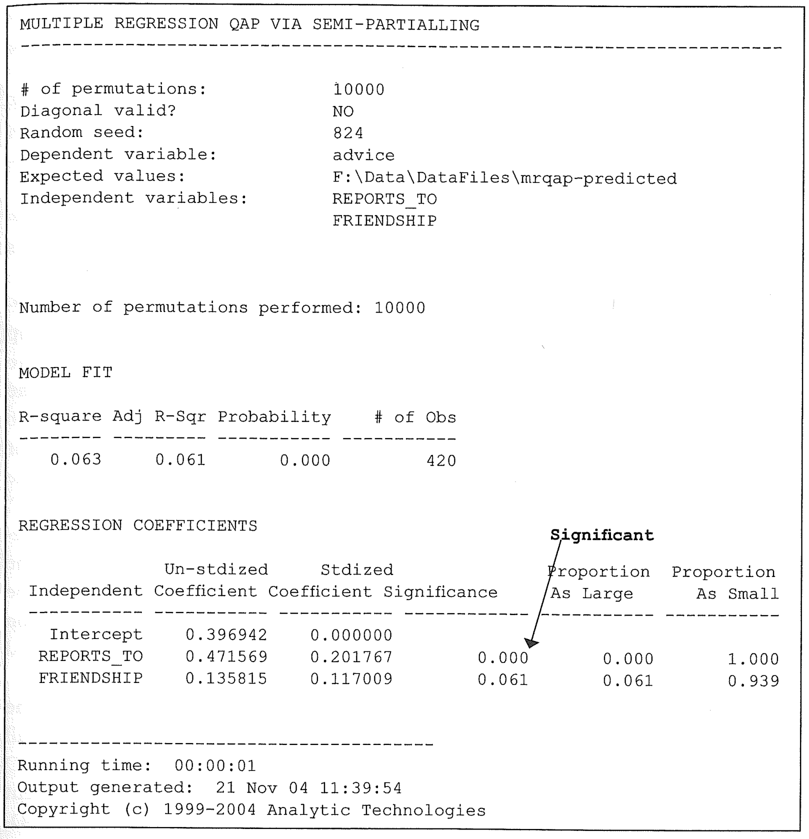
The relationship between QAP regression (also known as MR-QAP) and QAP correlation is the same as between their analogues in ordinary statistics. QAP regression allows you to model the values of a dyadic dependent variable (such as business ties) using multiple independent variables (such as marriage ties and some other dyadic variable, such as friendship ties or physical proximity of homes).

For example, suppose we are interested in advice-seeking within organizations. We can imagine that a person does not seek advice randomly from others. One factor that may influence who one seeks advice from is the existence of prior friendly



```
QAP CORRELATION
--------------------------------------------------------------------------------

Data Matrices:                          padgm
                                        padgb
# of Permutations:                      50000
Random seed:                            24322
Method:                                 Detailed (missing values ok)


QAP results for padgb * padgm (50000 permutations)

                                   1         2         3         4         5         6         7         8
                           Obs Value Significa   Average   Std Dev   Minimum   Maximum Prop >= O Prop <= O
                           --------- --------- --------- --------- --------- --------- --------- ---------
 1     Pearson Correlation    0.3719    0.0007    0.0002    0.0924   -0.1690    0.5071    0.0007    0.9999
 2      Euclidean Distance    4.3589    0.0007    5.4709    0.2529    3.8730    5.9161    0.9999    0.0007
 3        Hamming Distance    0.1583    0.0007    0.2500    0.0228    0.1250    0.2917    0.9999    0.0007
 4              Match Coef    0.8417    0.0007    0.7500    0.0228    0.7083    0.8750    0.0007    0.9999
 5            Jaccard Coef    0.2963    0.0007    0.0790    0.0464    0.0000    0.4000    0.0007    0.9999
 6 Goodman-Kruskal Gamma      0.7971    0.0007   -0.0690    0.3845   -1.0000    0.9000    0.0007    0.9999
 7            Hubert Gamma    8.0000    0.0007    2.5025    1.3668    0.0000   13.0000    0.0007    0.9999

NOTE: When you have missing data, the significance of Hubert's Gamma and Euclidean Distance will differ from
that of Pearson Correlation. Otherwise, they should be the same (unless the correlation is negative).


QAP Correlations

              1     2
          padgm padgb
          ----- -----
 1 padgm  1.000 0.372
 2 padgb  0.372 1.000


QAP P-Values

              1     2
          padgm padgb
          ----- -----
 1 padgm  0.000 0.001
 2 padgb  0.001 0.000

QAP statistics saved as datafile QAP Correlation Results
```

Figure 8.1 Results of QAP correlation.



```
MULTIPLE REGRESSION QAP VIA SEMI-PARTIALLING
----------------------------------------------------------------------

# of permutations:            10000
Diagonal valid?               NO
Random seed:                  824
Dependent variable:           advice
Expected values:              F:\Data\DataFiles\mrqap-predicted
Independent variables:        REPORTS_TO
                              FRIENDSHIP


Number of permutations performed: 10000


MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.063     0.061       0.000         420


REGRESSION COEFFICIENTS
                                                            Significant
                  Un-stdized      Stdized                    Proportion  Proportion
 Independent Coefficient Coefficient Significance            As Large     As Small
 ----------- ----------- ----------- ------------ ----------- -----------
   Intercept    0.396942    0.000000
  REPORTS_TO    0.471569    0.201767        0.000       0.000       1.000
  FRIENDSHIP    0.135815    0.117009        0.061       0.061       0.939


----------------------------------------
Running time:  00:00:01
Output generated:  21 Nov 04 11:39:54
Copyright (c) 1999-2004 Analytic Technologies
```

Figure 8.2 Results of MR-QAP regression.

relations – one is less likely to ask advice from those one does not know or does not like. Another factor might be structural position – whether they are in a position to know the answer. For example, we might predict that employees will often seek advice from those to whom they report. Krackhardt (1987) collected advice, friendship and reporting relationships among a set of managers in a high-tech organization, and this data is available in UCINET, allowing us to test our hypotheses.

To do this, we run one of the QAP multiple regression routines in UCINET. The result is shown in Figure 8.2. The *R*-square value of 6.3% suggests that neither who one reports to nor friendship is a major factor in determining who a person decides to seek advice from. In other words, there are other more important variables that we have not measured, perhaps including the amount of expertise that the other person has relative to the person looking for advice. Still, the 'reports to' relation is significant ($p < 0.001$), so it seems that it is at least a piece of the puzzle. Friendship is interestingly not significant: this is not quite in line with Casciaro and Lobo's (2005) finding that people prefer to seek advice from people they like even when there are more qualified – but less nice – people available.

It should be noted that, in our example, the dependent variable is binary. Using ordinary regression to regress a binary variable would be unthinkable if we were not using permutation methods to calculate significance. Since we are, though, the *p*-values on each coefficient are valid and interpretable. But it is important to keep in mind that the regression coefficients mean what they mean in ordinary least squares regression: they have not been magically transformed into, say, odds, such that you could say that an increase in one unit of the *X* variable is associated with a certain increase in the odds of that case being a 1 on the dependent variable. To have this interpretation, we would need to have run a logistic regression QAP (LR-QAP).[2] This can also be done in UCINET, although it is more time-consuming than MR-QAP.

As another example of how QAP regression can be used, we examine the Newcomb (1961) fraternity data in UCINET. This dataset consists of 15 matrices recording weekly sociometric preference rankings among 17 men attending the University of Michigan in the fall of 1956. The men were recruited to live in off-campus (fraternity) housing, rented for them as part of the Michigan Group Study Project supervised by Theodore Newcomb from 1953 to 1956. All were incoming transfer students with no prior acquaintance. We shall examine the first two time periods to study reciprocity and transitivity. We are interested to know if new friendship ties formed in Week 1 are a result of reciprocity and/or transitivity of ties formed in week 0. One way to do this is to construct the dependent variable as the cell-by-cell difference between the matrix for Week 1 (called NEWC1) and week 0 (called NEWC0). Alternatively, we can simply predict NEWC1 and include NEWC0 as a control variable. In order to illustrate the LR-QAP procedure, we choose the second approach and also dichotomize the matrices so that the (*i*, *j*) entry for each matrix equals 1 if person *i* ranked person *j* among their top three choices and 0 otherwise. We refer to the dichotomized matrices as NEWC0D and NEWC1D.

We now form two further matrices from NEWC0D. The first is simply the transpose of NEWC0D which, for ease of interpretation later, we shall call NEWC0D-Reciprocity. A value of 1 for cell (*i*, *j*) of the transpose of NEWC0D indicates that in Week 0, *i* received a nomination from person *j*. To the extent that people tend to reciprocate incoming ties, we should see that a 1 in NEWC0D-Reciprocity is matched by a 1 in the corresponding cell of NEWC1D.

Our second matrix will be the friends of friends matrix that has a 1 in the (*i*, *j*) entry if actor *j* is 2 steps or less away from actor *i* by the shortest path. We name

this matrix NEWC0D-Transitivity. To the extent that one tends to become friends with one's friends' friends, we should see that a 1 in the (*i*, *j*) cell of NEWC0D-Transitivity should be matched to a 1 in the (*i*, *j*) cell of NEWC1D. The transitivity martrix also has direct ties, but these are controlled for by including NEWC0D in the regression.

We then run a QAP-based logistic regression using NEWC1D as the dependent variable, and NEWC0D, NEWC0D-Reciprocity, and NEWC0D-Transitivity as the independent variables. The results are shown in Figure 8.3. We can see from the *p*-values in the output (in the column labeled 'Sig') that NEWC0D is significant, which is what we would expect since it would be surprising if the social structure at time *T* was wholly unrelated to the social structure a week earlier. The reciprocity parameter is positive and significant ($p = 0.008$), indicating a greater-than-chance tendency to reciprocate ties, but the transitivity parameter is not significant ($p = 0.071$), indicating no particular tendency to become friends with friends of friends.

```
Dependent variable: newc1D

   Overall fit of the logistic regression model

                             1           2           3           4           5
                      Log Lik Pseudo Rsq         Sig         Obs       Perms
                   ---------- ---------- ---------- ---------- ----------
      1 Statistics:   -100.906       0.259       0.000         272       10000

   1 rows, 5 columns, 1 levels.

   LR Coefficients

                                            1        2        3        4        5
                                          Coef OddsRat      Sig   StdErr      Avg
                                       -------  -------  -------  -------  -------
      1                     Intercept  -2.614   -8.654    0.000    0.210   -1.462
      2                        newc0D   2.290    9.880    0.000    0.426   -0.011
      3            newc0D-Reciprocity   0.818    2.267    0.008    0.362   -0.009
      4 newc0D-Transitivity (Closure)   0.598    1.818    0.071    0.398    0.010
```

Figure 8.3 Logistic regression results.

8.4 Mixed dyadic–monadic hypotheses

In this section we consider ways of relating node attributes to relational data. For example, when we look at the diagram in Figure 8.4, in which gender is indicated by the shape of the node, it is hard to avoid the conclusion that the pattern of ties

[2] As an aside, we can interpret the coefficients from MR-QAP on binary data as follows. In our output, the 0.472 value for the 'reports to' coefficient means that when the *X* variable is one unit higher, the dependent variable will, on average, be 0.472 units higher. This does not mean each case is 0.472 units higher, but that in any batch of 1000 dyads where *i* reports to *j*, we expect to see about 472 more cases of advice-seeking than when *i* does not report to *j*. This is not too difficult to understand. The trouble comes when we consider dyads in which *i* does not report to *j* ($X = 0$) but does seek advice from *j* ($Y = 1$), and compare these with dyads in which *i* does report to *j* (*X* is a unit higher). *Y* is already at its maximum value, so for this batch of dyads, the expectation that *Y* will be an additional 0.472 units higher does not make sense.

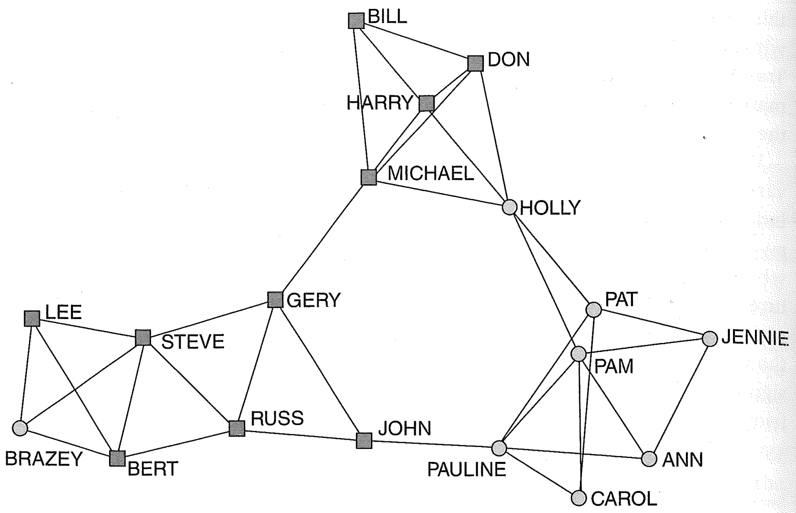


Figure 8.4 Campnet dataset showing top three choices among a set of men and women.

is related to gender. Specifically, there are more ties between members of the same gender than you would expect by chance. It would appear that actors have a tendency to interact with people of the same gender as themselves, a phenomenon known as 'homophily'. Homophily is an instance of a larger class of frequently hypothesized social processes known as 'selection', in which actors choose other actors based on attributes of those actors.

Another common type of hypothesis that links dyadic data with monadic attributes is the diffusion or influence hypothesis. Diffusion is the idea that people's beliefs, attitudes and practices come about in part because of interaction with others who already have those beliefs. So the fact that I own an iPhone may be in part due to the fact that my friend has one. I am more likely to have conservative political beliefs if everyone around me has conservative beliefs.

Both diffusion and selection hypotheses relate a dyadic variable (the network) with a monadic variable (the node attribute). The difference between diffusion and selection hypotheses is just the direction of causality. In diffusion, the dyadic variable causes the monadic variable, and in the selection the monadic variable causes the dyadic variable. We should note that, if the data is cross-sectional rather than longitudinal, we will not normally be able to distinguish empirically between diffusion and selection, although in the case of Figure 8.4 we tend to be confident that it is not a case of gender diffusion but rather people selecting friends based on gender.

The standard approach to testing the association between a node attribute and a dyadic relation is to convert the problem into a purely dyadic hypothesis by constructing a dyadic variable from the node attribute. Different techniques are needed depending on whether the attribute is categorical, such as gender or department, or continuous, such as age or wealth.

8.4.1 Continuous attributes

In traditional bureaucracies, we expect that employees have predictable career trajectories in which they move to higher and higher levels over time. As such, we expect managers to be older (in terms of years of service to the organization) than the people who report to them. In modern high-tech organizations, however, we expect more fluid career trajectories based more on competence than on years of service. Hence, in this kind of organization we do not necessarily expect employees to be younger (in years of service) than their bosses.

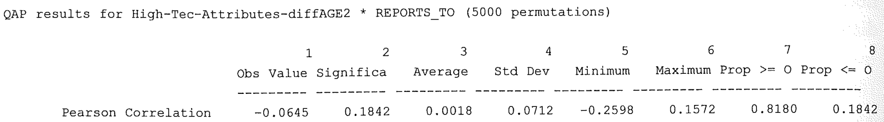
One way to test this idea in the organization studied by Krackhardt (1987) would be to construct a node-by-node matrix of differences in years of service, and then use QAP correlation to correlate this matrix with the 'reports to' matrix. As discussed in Chapter 5, in UCINET we can construct a node-by-node matrix of differences in years of service using the Data|Attribute-to-Matrix procedure. This creates a matrix in which the (*i*, *j*) cell gives the tenure of node *j* subtracted from the tenure of node *i* – that is, it is the row node's value minus the column node's value. Matrix 8.1 shows the node-level age variable, along with the dyadic difference in age matrix computed by UCINET.

```
Age        1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21
         --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
33   1     0  -9  -7   0   1 -26 -22  -1 -29  -4 -13  -1 -15 -10  -7   6   3   0   1  -5  -3
42   2     9   0   2   9  10 -17 -13   8 -20   5  -4   8  -6  -1   2  15  12   9  10   4   6
40   3     7  -2   0   7   8 -19 -15   6 -22   3  -6   6  -8  -3   0  13  10   7   8   2   4
33   4     0  -9  -7   0   1 -26 -22  -1 -29  -4 -13  -1 -15 -10  -7   6   3   0   1  -5  -3
32   5    -1 -10  -8  -1   0 -27 -23  -2 -30  -5 -14  -2 -16 -11  -8   5   2  -1   0  -6  -4
59   6    26  17  19  26  27   0   4  25  -3  22  13  25  11  16  19  32  29  26  27  21  23
55   7    22  13  15  22  23  -4   0  21  -7  18   9  21   7  12  15  28  25  22  23  17  19
34   8     1  -8  -6   1   2 -25 -21   0 -28  -3 -12   0 -14  -9  -6   7   4   1   2  -4  -2
62   9    29  20  22  29  30   3   7  28   0  25  16  28  14  19  22  35  32  29  30  24  26
37  10     4  -5  -3   4   5 -22 -18   3 -25   0  -9   3 -11  -6  -3  10   7   4   5  -1   1
46  11    13   4   6  13  14 -13  -9  12 -16   9   0  12  -2   3   6  19  16  13  14   8  10
34  12     1  -8  -6   1   2 -25 -21   0 -28  -3 -12   0 -14  -9  -6   7   4   1   2  -4  -2
48  13    15   6   8  15  16 -11  -7  14 -14  11   2  14   0   5   8  21  18  15  16  10  12
43  14    10   1   3  10  11 -16 -12   9 -19   6  -3   9  -5   0   3  16  13  10  11   5   7
40  15     7  -2   0   7   8 -19 -15   6 -22   3  -6   6  -8  -3   0  13  10   7   8   2   4
27  16    -6 -15 -13  -6  -5 -32 -28  -7 -35 -10 -19  -7 -21 -16 -13   0  -3  -6  -5 -11  -9
30  17    -3 -12 -10  -3  -2 -29 -25  -4 -32  -7 -16  -4 -18 -13 -10   3   0  -3  -2  -8  -6
33  18     0  -9  -7   0   1 -26 -22  -1 -29  -4 -13  -1 -15 -10  -7   6   3   0   1  -5  -3
32  19    -1 -10  -8  -1   0 -27 -23  -2 -30  -5 -14  -2 -16 -11  -8   5   2  -1   0  -6  -4
38  20     5  -4  -2   5   6 -21 -17   4 -24   1  -8   4 -10  -5  -2  11   8   5   6   0   2
36  21     3  -6  -4   3   4 -23 -19   2 -26  -1 -10   2 -12  -7  -4   9   6   3   4  -2   0
```

Matrix 8.1 Age of each node (left) and differences in ages between all pairs of nodes (right).

The 'reports to' matrix is arranged such that a 1 in the (i, j) cell indicates that the row person reports to the column person. Hence, if the organization were a traditional bureaucracy, we would expect a negative correlation between this matrix and the age-difference matrix, since the row person should have a smaller number of years of service than the column person. But since the organization is a modern high-tech company, we are actually expecting no correlation. The result is shown in Figure 8.5 The correlation is negative, but it is not significant (r = 0.0645), just as we expected.

However, there are a couple of problems with our analysis. First of all, it is always difficult to test a hypothesis of no relationship, because if you do observe no relationship it could be simply because your statistical test lacks power (e.g., your sample size is too small). Second, our test implicitly assumes that every person could potentially report to anyone older than themselves. But our common-sense knowledge of the 'reports to' relation tells us that each person only reports to one manager. This creates a lot of cases where A is younger than B, but A fails to report to them. A better test would examine just pairs of nodes in which one reports to the other, and then test whether age difference is correlated with who reports to whom. We can do this by placing missing values for all cells in which neither party reports to the other. When we do this and rerun the analysis, we get a stronger correlation of −0.320, but the p-value is 0.147, which is non-significant. In this company, who you report to is simply not a function of relative age.



```
QAP results for High-Tec-Attributes-diffAGE2 * REPORTS_TO (5000 permutations)

                                1         2         3         4         5         6         7         8
                        Obs Value Significa   Average   Std Dev   Minimum   Maximum Prop >= 0 Prop <= 0
                        --------- --------- --------- --------- --------- --------- --------- ---------
   Pearson Correlation    -0.0645    0.1842    0.0018    0.0712   -0.2598    0.1572    0.8180    0.1842
```

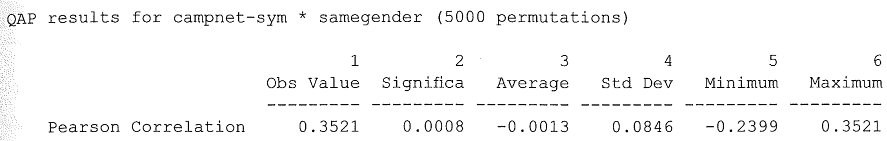
Figure 8.5 QAP correlation between age difference and who reports to whom.

8.4.2 Categorical attributes

Borgatti et al. (2012) collected ties among participants in a 3-week workshop. As noted earlier, a visual display of the Campnet dataset seems to suggest that gender affects who interacts with whom (see Figure 8.4). However, the human brain is notorious for seeing patterns and focusing on confirmatory evidence while ignoring contradictory data. Therefore, we would like to statistically test this homophily hypothesis.

An approach that is closely parallel to the way we handled age earlier is to construct a node-by-node matrix in which the (i, j) cell is 1 if nodes i and j belong to the same gender, and 0 if they belong to different genders. In UCINET this is done using the same Data|Attribute-to-Matrix procedure we used for continuous attributes, but selecting the 'Exact matches' option instead of 'Difference'. We can then use QAP correlation to correlate the matrix of actual network ties with the 'is the same gender' matrix. The result (not shown) is a strong correlation of 0.33 with a p-value of 0.0006, indicating support for the homophily hypothesis.

We should note, though, that we got a little lucky in this example. The independent variable, 'same gender', is a symmetric matrix – if I am the same gender as you, you must be the same gender as me. Yet the dependent variable is not symmetric. This data is of the forced choice type in which each person lists the top three people they interact with. This tends to force asymmetry because a popular person will be listed by many more than three others, yet the respondent is only allowed to reciprocate three of these. Further, there is no way for a symmetric independent variable to perfectly predict a non-symmetric dependent variable (this is handled by the QAP significance test, but the R-square value may be misleadingly low). In this case, it might make more sense to symmetrize the Campnet matrix via the maximum method, which means that a tie is said to exist between two nodes if either lists the other as one of their top three interactors. If we take this approach and rerun the correlation, we obtain a correlation that is a little bit higher at 0.352, and of course still significant.



```
QAP results for campnet-sym * samegender (5000 permutations)

                                1         2         3         4         5         6
                        Obs Value Significa   Average   Std Dev   Minimum   Maximum
                        --------- --------- --------- --------- --------- ---------
   Pearson Correlation     0.3521    0.0008   -0.0013    0.0846   -0.2399    0.3521
```

Figure 8.6 QAP correlation with symmetrized Campnet data.

A node-level hypothesis is one in which the variables are characteristics of individual nodes, such as persons. For example, we might investigate whether the number of top management friends a person has predicts the size of her bonus at the end of the year. In some ways, this is an easy one: just run an ordinary regression. Indeed, this is the way most hypotheses of this type are tested in the literature. But suppose our research site is a small company of, say, 20 individuals and we have surveyed all of them. If we are being careful, we might note that the sample size is small and, while small sizes can be conservative (in the sense that if the results are significant on a small sample size it must be a pretty strong effect), if they get too small the assumptions of the classical significance test will no longer hold. We might also realize that we do not have an actual sample. We have the entire population of organization members, and the organization itself is a sample of one chosen non-randomly from the population of firms. This also is not a situation that the classical significance test for regression coefficients is meant to handle.

The safer thing to do is run a randomization test. For example, we could run ordinary least squares as usual to obtain the regression coefficients, but then use the permutation technique to construct the p-values. Figure 8.6 shows the results of testing a simple hypothesis that men will have more friends who are not friends of other friends. In other words, the hypothesis is that there will be fewer connections among men's friends then among women's friends. To test this, we constructed the dependent variable by running UCINET's Egonet density procedure, which gave us the proportion of each node's friends that were friends with each other. The independent variable was simply gender (coded 1 = women, 2 = men) and we also controlled for the individual's role (1 = participant, 2 = instructor). We then ran UCINET's node-level regression to produce Figure 8.7. As you can see, the hypothesis was not supported. Whatever determines the degree of connection among one's friends, it is not one's gender, nor one's role in the organization.

```
NOTE: All probabilities based on randomization tests.


MODEL FIT

            Adjusted                  One-Tailed
R-square  R-square       F Value     Probability
-------------------      -------     -----------
   0.090     -0.080        0.744           0.485


REGRESSION COEFFICIENTS

                    Un-stdized      St'dized     Proportion  Proportion  Proportion
  Independent      Coefficient     Coefficient     As Large    As Small  As Extreme
  -----------  --------------- ----------- ----------- ----------- -----------
    Intercept         0.462500     0.000000       1.000       0.000       1.000
       Gender         0.138889     0.292305       0.166       0.834       0.334
         Role        -0.168056    -0.295918       0.831       0.169       0.327
```

Figure 8.7 Ordinary least squares regression with *p*-values calculated via a randomization test.

8.5 Whole-network hypotheses

A whole-network hypothesis is one in which the cases are collectivities such as teams, firms or countries, and the variables are characteristics of the network of ties within the units. For instance, Athanassiou and Nigh (2000) studied a sample of 37 firms, and looked at how a firm's degree of internationalization affected the density of advice ties among members of its top management team.

Assuming the firms are obtained via a random sample, to test a hypothesis like this we can just run a normal correlation in a standard statistical package such as SPSS. The classical significance test would be perfectly valid. Of course, if we did use a randomization test, the results would also be perfectly valid, but would take more time to compute and require the use of a network analysis software package such as UCINET, or a specialized statistical package such as StatXact. On the other hand, if the data was not collected via a random sample, it would be wise to use a randomization test.

Randomization tests provide an elegant and powerful way to deal with some of the special issues posed by social network data. A key advantage is that, if one has programming capability, one can construct a suitable significance test for any test statistic, including new ones developed specifically for the research at hand. One thing to remember, however, is that while a randomization test will allow you to test for significance even when you have a non-random sample or population, it does not magically create generalizability. A significant result relating *X* to *Y* in Mrs Smith's third-grade classroom tells you that, in that classroom, *X* and *Y* are probably not independent, but it does not make up for the fact that you did not randomly sample children from all over the world, nor did you sample from the set of all classrooms. Generalizability comes from your research design, not from significance statistics.

8.6 Exponential random graph models

QAP regressions are about comparing two (or more) networks. We may only have one network that is actual data but, as we have seen, it is sometimes possible to construct a second hypothetical structure matrix based on some underlying concept of a social process, such as homophily or transitive closure. The QAP regression then assesses the fit between the actual data and an ideal matrix consistent with a hypothesized social process. However, another way to conceptualize the problem is in terms of identifying micro-configurations (such as transitive triples, 4-cycles, etc.) that represent the theoretical social process, and then counting them in the data to see if there are more of them than one would expect if the process were not happening. The baseline model can also take into account constraints such as limits on the number of ties that each node could have. This is the approach taken by exponential random graph models (ERGMs, also known as 'p* models').

As these models are not in UCINET, our discussion is more about getting a general idea of what they are and what they can be used for. More complete descriptions can be found in Robins et al. (2007) Robins (2011) and Lusher et al. (2013). The models are related to the general linear models of standard statistics but have important modifications to deal with the fact that we cannot assume independence of observations – in our case the edges. A key concept is the notion of conditional dependence. If two edges share a common vertex then they are dependent, conditional on the rest of the graph. Models based on conditional