Testing Network Hypotheses

The first thing to understand about testing hypotheses using network data is that, most of the time, classical statistical methods don't work. This is for a number of related reasons, including: the observations are not independent; you don't usually have a random sample; you may not have a sample at all (you have a population); and the variables not known to be normally distributed. As a result, special methods have to be used. The class of methods we will use here is called "randomization tests" or "permutation tests".

Classical significance tests are based on sampling theory and have the following logic. You measure a set of variables (let's say two variables) on a sample of cases drawn from a population. You are interested in the relationship between the variables, as measured, let's 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 tells you the probability of obtaining a correlation that large given that in the population the variables are actually independent (correlation zero). When the 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. When the probability is higher, we feel we can't reject the null hypothesis that the variables actually independent. Note that if you have a weird sample, or you don't have a sample at all, it doesn't 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 math teacher and as a result they learn more math than short kids. So you think height and math scores will be correlated. So you give all the kids math test, you measure their height, and you correlate the two and get a correlation of 0.384. Hypothesis confirmed? Not so fast. Suppose height and math ability are in actuality unrelated. In fact, just for fun, instead of actually giving the math test 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 bowl containing all the slips. So you know that math score and height are totally unrelated. But couldn't it happen that 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 in which 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 been matched to kids 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?

The permutation test essentially calculates all the ways that the experiment could have come out by chance (i.e., every possible assignment of scores to students), and counts the proportion of random assignments that yield 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 independent of each other.

The second thing to understand about testing hypotheses involving network variables is that there are more possibilities for different kinds of hypotheses with network data than with simple attribute data. For example, one kind of hypothesis is the node-level or monadic hypothesis, such as the hypothesis that more central people are happier. This kind of hypothesis closely resembles 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. So that's just correlating two vectors – two columns of data – which is simple enough.

Another kind of hypothesis is the dyadic one. Here, you are hypothesizing that the more a pair of persons has a certain kind of relationship, the more they will also have another kind 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"). So each variable is an entire person-by-person matrix, and you want to correlate the two matrices. Clearly, this is not something you would ordinarily do in SPSS.

Still another kind of hypothesis is at the group or network level. 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-until-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 ties. The data file looks just like the data file for node-level hypotheses, except the cases here are entire networks rather than individual nodes.

Finally, there is the kind of hypothesis in which one variable is dyadic, such as friendship, and the other is monadic (node-level), such as gender. The question being asked is something like whether gender affects who is friends with whom. For example, the gender homophily hypothesis says that people are more likely to be friends with people of the same gender as themselves.

In this chapter, we consider how to test each of the four kinds of hypotheses, starting with the one involving the least aggregate cases (dyadic) and ending with the one involving the most aggregate cases (whole networks).

10.1 Dyadic Hypotheses

Padgett and Ansell (19xx) 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 statistical package for two reasons. One, statistical packages are set up to correlate vectors, not matrices. This one is not too serious a problem, however, because we can just reshape the matrices so that all the values in each matrix are lined up in a single column that is NxN values long, as shown in Figure xx. Two, the significance test in a statistical package assumes that the values within a variable are statistically independent, which, in the case of 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 anti-social, it will affect all of their relations with others, introducing a lack of independence of all those cells in the matrix.

10.1.1 QAP Correlation

Network analysis packages like UCINET provide a technique called QAP Correlation that is designed expressly to correlate two adjacency matrices. The QAP technique correlates the two matrices by effectively reshaping them into two long columns as described above and calculating ordinary measures of statistical association such as Pearson's r. To calculate the significance of the correlation, the method randomly permutes the ordering of the rows (and corresponding columns) of one matrix relative to the other, and then recomputes the correlation. This effectively creates a new matrix that whose values are assigned independently of the other matrix. Then the rows and columns are again randomly permuted and the correlation is recomputed. This is done thousands of times to generate a distribution of correlation coefficients for pairs of matrices very similar to the data matrices but which are known to be independent of each other. The proportion of these correlations that are as large as the correlation between the two data matrices is the p-value of the test. By convention, the p-value should be less than 5% to be considered significant (i.e., supporting the hypothesis that the two matrices are related).

When we run QAP correlation on the Padgett data, we obtain the results shown in Output 10.1.

{annotate the output}

<note about how the p-value can vary>

10.1.2 QAP Regression

The relationship between QAP Correlation and QAP Regression (also known as MRQAP) is the same as between their analogues in ordinary statistics. QAP Regression allows you to model the values of a dependent variable (such as business ties) using multiple independent variables (such as marriage ties and some other social relation such as friendship ties).

For example, suppose we are interested advice seeking within organizations. We can imagine that a person doesn't seek advice randomly from others. One factor that may influence who one seeks advice from is the existence of a prior friendly relations – one doesn't normally ask advice from those one doesn't know or has unfriendly relations with. Another factor might be structural position – whether they are in a position to know the answer. This suggests that employees will often seek advice from those they report to. Krackhardt () collected advice, friendship and reporting relationships among a set of managers in a high-tech organization, and these data are 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 Output 10.2.

```
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

                                                                  Significant
REGRESSION COEFFICIENTS

              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
```

The R-squared value of 6.3% suggests that neither who one reports to nor friendship is the determining factor in how a person decides who to ask advice of. In other words, there are other variables that we have not measured, such as having relevant expertise. Still, the "reports to" relation is significant ($p < 0.001$), so it is a piece of the puzzle. Friendship is not significant, indicating that it may be unrelated to choice of advice source.

10.2 Mixed Dyadic-Monadic Hypotheses

In this section we consider quantitative methods of relating node attributes to relational data. One standard hypothesis of this type is the diffusion hypothesis. Diffusion (or social homogeneity) is the idea that, as people interact, they influence or transmit things to each other, and as a result come to have similar ideas, practices or qualities. Thus, interaction (a relational or dyadic variable) explains a node's value on a particular attribute (a monadic variable). Another standard hypothesis is the selection hypothesis. Selection is the idea that nodes choose whom to interact with based on the compatibility of their respective attributes. A classic example is the phenomenon of homophily, which is the often-seen tendency for actors to interact more with members of their own kind than with others – e.g., in a playground, boys are more likely to play with other boys, while girls are more likely to play with other girls.

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 are cross-sectional rather than longitudinal, we will not normally be able to distinguish between diffusion and selection.

Different techniques are needed depending on the measurement level of the monadic variable (the node attribute). In particular, we must distinguish between nominal-scale (categorical) variables, such as gender or department, which essentially classify nodes into discrete groups, and interval-scale (continuous) variables, such as age or wealth, which locate nodes along a continuum of values.

10.2.1 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 years of service. Hence, in this kind of organization we don't 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 () 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 3, in UCINET we can construct a node-by-node matrix of differences in years of service using the Data>Attribute procedure. This program creates a matrix in which the I,jth cell gives the tenure of node j subtracted from the tenure of node I – i.e., it is the row-node's value minus the column-node's value. The reports-to matrix is arranged such that a 1 in the I,jth cell indicates that the row person reports to the column person. Hence, according to our hypothesis, we expect a negative correlation between the two matrices, since the row person should have a smaller number of years of service than the column person.

The result is shown in Output 10.3. The correlation is indeed negative, as expected, but it is not significant ($r = 0.092$).

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

Observed matrix:              tendiff
Structure matrix:             reports_to
# of Permutations:           10000
Random seed:                  370


Univariate statistics

                     1         2
                tendiff reports_t
              --------- ---------
  1     Mean     -0.000     0.048
  2  Std Dev     11.369     0.213
  3      Sum      0.000    20.000
  4 Variance    129.263     0.045
  5      SSQ  54290.324    20.000
  6    MCSSQ  54290.324    19.048
  7 Euc Norm    233.003     4.472
  8  Minimum    -29.750     0.000
  9  Maximum     29.750     1.000
 10 N of Obs    420.000   420.000


Hubert's gamma: -103.167


Bivariate Statistics
                                                    Not significant!
                                1         2         3         4         5         6         7
                            Value    Signif       Avg        SD  P(Large)  P(Small)     NPerm
                          --------- --------- --------- --------- --------- --------- ---------
  1   Pearson Correlation:   -0.101     0.092    -0.000     0.071     0.909     0.092 10000.000
  2      Simple Matching:    0.000     1.000     0.000     0.000     1.000     0.955 10000.000
  3   Jaccard Coefficient:    0.048     1.000     0.048     0.000     1.000     1.000 10000.000
  4 Goodman-Kruskal Gamma:              0.000     0.000     0.000     0.000     0.000     0.000
  5      Hamming Distance:  420.000     1.000   419.913     4.204     0.955     1.000 10000.000


---------------------------------------
Running time:  00:00:01
Output generated:  21 Nov 04 12:23:25
Copyright (c) 1999-2004 Analytic Technologies
```

<what about the fact that one variable is zeros and ones>

10.2.1 Categorical Attributes

Borgatti et al (19xx) collected ties among participants in a 3-week workshop. A visual display of the CAMPNET dataset, using node shape to indicate gender, seems to suggest that gender affects who interacts with whom (see Figure 10.1). In particular, there appear to be more ties within genders than between – i.e., homophily. However, the human brain is notorious for focusing on confirmatory evidence and ignoring contradictory data. Therefore, we would like to statistically test this homophily hypothesis.

An approach that is closely parallel to the way we handle continuous variables is to construct a node-by-node matrix in which the i,jth 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 procedure we used for continuous attributes, but selecting a different option. We can then use QAP correlation to correlate the matrix of network ties with the "is the same gender" matrix. The result, as shown in Output 10.xx, is a strangely small, non-significant correlation (r = -0.065, p = 0.239).

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

Observed matrix:              samegender
Structure matrix:             campnet
# of Permutations:           10000
Random seed:                  783


Univariate statistics

                     1       2
              samegen campnet
              ------- -------
  1     Mean    0.477   0.176
  2  Std Dev    0.499   0.381
  3      Sum  146.000  54.000
  4 Variance    0.249   0.145
  5      SSQ  146.000  54.000
  6    MCSSQ   76.340  44.471
  7 Euc Norm   12.083   7.348
  8  Minimum    0.000   0.000
  9  Maximum    1.000   1.000
 10 N of Obs  306.000 306.000


Hubert's gamma: 22.000


Bivariate Statistics

                                1         2         3         4         5         6         7
                            Value    Signif       Avg        SD  P(Large)  P(Small)     NPerm
                         --------- --------- --------- --------- --------- --------- ---------
  1   Pearson Correlation:    -0.065     0.239    -0.001     0.076     0.837     0.239 10000.000
  2       Simple Matching:     0.490     0.837     0.514     0.029     0.837     0.239 10000.000
```

Why is this correlation so small? One reason is that the “same gender” matrix is symmetric – if I am the same gender as you, you must be the same gender as me. Yet the CAMPNET matrix is not symmetric. These data are of the forced-choice type in which each person lists the top 3 people they interact with. This tends to force asymmetry because a popular person will be listed by many more than 3 others, yet the respondent is only allowed to reciprocate 3 of these. In this case, it might make more sense to symmetrize the CAMPNET matrix via the maximum method so that a tie is said to exist between two nodes if either lists the other as one of their top 3 interactors. If take this approach and rerun the correlation, we obtain the result given in Output 10.3. Now the correlation is 0.351 and it is significant.

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

Observed matrix:              samegender
Structure matrix:             symcampnet
# of Permutations:            10000
Random seed:                  54


Univariate statistics

                    1       2
              samegen symcamp
              ------- -------
  1     Mean    0.477   0.227
  2  Std Dev    0.499   0.419
  3      Sum  146.000  69.000
  4 Variance    0.249   0.175
  5      SSQ  146.000  69.000
  6    MCSSQ   76.340  53.339
  7 Euc Norm   12.083   8.307
  8  Minimum    0.000   0.000
  9  Maximum    1.000   1.000
 10 N of Obs  306.000 304.000


Hubert's gamma: 55.000


Bivariate Statistics

                                1         2         3         4         5         6         7
                            Value    Signif       Avg        SD  P(Large)  P(Small)     NPerm
                          --------- --------- --------- --------- --------- --------- ---------
  1   Pearson Correlation:     0.351     0.001     0.001     0.084     0.001     0.999 10000.000
  2       Simple Matching:     0.661     0.001     0.513     0.035     0.001     1.000 10000.000
  3   Jaccard Coefficient:     0.348     0.001     0.183     0.035     0.001     0.999 10000.000
  4 Goodman-Kruskal Gamma:     0.731     0.001     0.001     0.194     0.001     0.999 10000.000
  5      Hamming Distance:   103.000     0.001   148.031    10.753     1.000     0.001 10000.000


----------------------------------------
Running time:  00:00:01
Output generated:  21 Nov 04 15:05:20
Copyright (c) 1999-2004 Analytic Technologies
```

Another way to look at how gender patterns interactions is through a density matrix. In UCINET we can obtain a density matrix by running the “Anova/Density Models” procedure located in the Tools>Statistics menu. We will need to supply two inputs: the network dataset, and a node attribute such as gender. In addition, we shall need to specify “constant homophily” as the model to be run. (The significance of this is explained further along.) The result is given in Output 10.45.

```
NETWORK AUTOCORRELATION WITH CATEGORICAL ATTRIBUTES
------------------------------------------------------------------------------

Network/Proximities:          F:\Data\DataFiles\symcampnet
Attribute(s):                 campattr2 col 1
Method:                       Constant Homophily
# of Permutations:           10000
Random seed:                  674

Density Table

           1     2
           1     2
       ----- -----
  1 1  0.429 0.087
  2 2  0.087 0.356


MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.124     0.124       0.000         306


REGRESSION COEFFICIENTS

              Un-stdized     Stdized               Proportion  Proportion
 Independent Coefficient Coefficient Significance    As Large    As Small
 ----------- ----------- ----------- ------------ ----------- -----------
   Intercept    0.087500    0.000000        1.000       1.000       0.000
    In-group    0.296062    0.352057        0.000       0.000       1.000


----------------------------------------
Running time:  00:00:01
Output generated:  21 Nov 04 15:32:15
Copyright (c) 1999-2004 Analytic Technologies
```

The first thing to look at is the table labeled “Density Table”, which gives the density of ties within and between each gender. For example, the 43% in the top left cell indicates that all nearly half of all pairs of women in the network chose were chosen by another woman. Similarly, 36% of all pairs of men have a (symmetrized) tie. In contrast, only about 9% of all possible cross-gender pairings were actually realized. This pattern of large numbers along the main diagonal and small numbers off-diagonal is indicative of homophily.

The next bit of output, labeled “MODEL FIT” tests whether the on-diagonal values (within group densities) are significantly greater than the off-diagonal values (between group densities). In this case, the R-squared statistic is modest ($r2 = 0.124$), but significant ($p < 0.001$). The r-squared value is (within rounding error) the square of the correlation obtained in the QAP regression earlier, showing that the two approaches are equivalent.

The advantage of the Anova approach over the simple QAP approach presented earlier is that we can use the Anova approach to fit more interesting models than the “constant homophily” model we have just fit. This model is so named because it assumes that each group (each gender in our case) has the same tendency to prefer its own kind. However, it is possible that some groups have only a small preference for their own kind, while others are wholly xenophobic. We call this model “variable homophily”.

```
NETWORK AUTOCORRELATION WITH CATEGORICAL ATTRIBUTES
------------------------------------------------------------------------------

Network/Proximities:          F:\Data\DataFiles\symcampnet
Attribute(s):                 campattr2 col 1
Method:                       Variable Homophily
# of Permutations:           10000
Random seed:                  567

Density Table

           1     2
           1     2
       ----- -----
  1 1  0.429 0.087
  2 2  0.087 0.356


MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.127     0.124       0.000         306


REGRESSION COEFFICIENTS

              Un-stdized     Stdized               Proportion  Proportion
 Independent Coefficient Coefficient Significance    As Large    As Small
 ----------- ----------- ----------- ------------ ----------- -----------
   Intercept    0.087500    0.000000        1.000       1.000       0.000
     Group 1    0.341071    0.313982        0.001       0.001       1.000
     Group 2    0.268056    0.290782        0.000       0.000       1.000


----------------------------------------
Running time:  00:00:01
Output generated:  21 Nov 04 15:59:21
Copyright (c) 1999-2004 Analytic Technologies
```

Running the variable homophily model in UCINET gives the result shown in Output 10.46. The density table is the same since the data haven't changed. The r-squared is slightly larger indicating that this model, which utilizes more parameters, fits slightly (but negligibly) better. The table labeled "Regression Coefficients" gives information about the relative levels of homophily in each group. In particular, the un-standardized coefficient gives the increase in density seen in each group relative to ties between groups. For example, for group 1, the unstandardized coefficient is 0.341071, which indicates that the density of ties among the women (which happen to be group 1) is .341071 greater than then the density of ties between men and women (0.0875, labeled "intercept"). As a check, we can see that adding .0875 to 0.341071 gives us 0.429, as reported in the density table.

The regression coefficients table also gives us the significance for each group, which indicates whether the group's density is significantly larger than the density between groups. In this case, both group's densities are significant, indicating that both are homophilous.

Sometimes the relationship between a categorical node attribute and a relational variable is more complicated than the patterns implied by homophily and variable homophily. For example, consider the case of a communication network in an organization in which the nodes belong to different organizational departments (e.g., marketing, accounting, human resources, etc.). While we probably do expect more communication within departments, we also expect significant communication between certain departments. For example, we might expect the bridge construction unit in an engineering firm to communicate closely with the quality control department. Other departments may have little or nothing to do

with each other. The research question is simply whether the distribution of ties between departments is uniform, which would indicate that department membership had no effect on communication, or whether there was significant variance in interdepartmental densities.

	BHS	CCG	DCL	ES	HEW	IS	MS	SRG	STAT	TAS
BHS	0.10	0.13	0.01	0.06	0.05	0.01	0.04	0.06	0.17	0.01
CCG	0.13	0.40	0.10	0.15	0.15	0.08	0.11	0.11	0.20	0.12
DCL	0.01	0.10	0.14	0.02	0.04	0.09	0.04	0.02	0.02	0.07
ES	0.06	0.15	0.02	0.09	0.03	0.02	0.04	0.04	0.12	0.02
HEW	0.05	0.15	0.04	0.03	0.10	0.01	0.03	0.04	0.10	0.02
IS	0.01	0.08	0.09	0.02	0.01	0.14	0.02	0.02	0.02	0.06
MS	0.04	0.11	0.04	0.04	0.03	0.02	0.07	0.03	0.10	0.05
SRG	0.06	0.11	0.02	0.04	0.04	0.02	0.03	0.08	0.13	0.01
STAT	0.17	0.20	0.02	0.12	0.10	0.02	0.10	0.13	0.36	0.04
TAS	0.01	0.12	0.07	0.02	0.02	0.06	0.05	0.01	0.04	0.17

10.2 Node-level (Monadic) Hypotheses

A node-level hypothesis is one in which the variables are characteristics of individual nodes such as persons. For example, you might ask whether a person’s degree centrality at the beginning of the year predicts the size of their raise at the end of the year.