PREDICTING WITH NETWORKS: NONPARAMETRIC MULTIPLE REGRESSION ANALYSIS OF DYADIC DATA *

David KRACKHARDT **

Johnson Graduate School of Management, Cornell University

This paper argues that the quadratic assignment procedure (QAP) is superior to OLS for testing hypotheses in both simple and multiple regression models based on dyadic data, such as found in network analysis. A model of autocorrelation is proposed that is consistent with the assumptions of dyadic data. Results of Monte Carlo simulations indicate that OLS analysis is statistically biased, with the degree of bias varying as a function of the amount of structural autocorrelation. On the other hand, the simulations demonstrate that QAP is relatively unbiased. The Sampson data are used to illustrate the QAP multiple regression procedure and a general method of testing whether the results are statistically biased.

1. Introduction

We have witnessed tremendous advances over the past 15 years in the area of formal network analysis (e.g. Breiger and Pattison 1986; Burt and Minor 1983). Most of these advances have focused on sophisticated techniques for better understanding and describing network structures, often by reducing them to more parsimonious forms. As Coleman (1983) has lamented, there has been a relative dirth of work formally relating these (now elegantly described) structures to their antecedents and consequences. He called for more complete models that explored the relationship among network variables and other dependent/independent variables of social interest.

Unfortunately, there are barriers to taking up this challenge. One of the most serious problems is that the unit of analysis is the dyad—and

dyads, it is reasonably argued, cannot be assumed to be independent of one another. [1] How can we perform inferential tests on data that are (potentially) highly interdependent?

We wish to propose a general procedure for answering this question. We suggest that Hubert's quadratic assignment procedure (QAP) can be extended to the multiple regression model. We first frame the problem as an autocorrelation problem and then demonstrate how QAP provides unbiased tests of significance of both simple and multiple regression coefficients. To illustrate how the procedure may be used in practice, we provide a simple example from Sampson's (1968) monastery data. Finally, we suggest that, to be safe, the degree of bias in any given application of this procedure should be tested by simulating the particular kind of data being analyzed. We illustrate this by simulating the Sampson data under the null hypothesis.

2. Structural autocorrelation

One way to approach this problem of non-independence of observations is to frame it as econometricians have—as an autocorrelation problem. [2] If one had data that were temporally interdependent, or autocorrelated, one could use any number of time series analytic tools available (Judge *et al.* 1985) to estimate the autocorrelation and to assess the significance of the various parameter estimates. But the nature of the autocorrelation here is far more complex and intractable than found in most econometric models.

To formalize this problem, consider the following simple regression model:

$$y = \beta X + \epsilon, \quad E(\epsilon\epsilon') = \sigma^2 \Omega$$

where y and X represent vectors of observations on some variables of

---

[1] We are excluding from our discussions here analysis of attributes of actors that are derived from networks, such as centrality (Freeman 1978) or position (Burt 1982). Inferential statements about autocorrelation problems in such attribute-level data have been explored elsewhere (Dow *et al.* 1982; Ord 1975).

[2] Lincoln (1984) has proposed a similar framing of this problem, although his solution, following from Doreian, Teuter and Wang (1984) and Ord (1975), is very different from the one proposed here. Kraemer and Jacklin's (1979) solution poses even more restrictive assumptions about the a priori knowledge of the structure of interdependence.

interest. If we assume that $\Omega = I$, that is, that the error terms in the model are independently and identically distributed, then the customary approach to testing data consistent with this model is to use ordinary least squares (OLS) analysis. [3]

But we are interested in an autocorrelated models. In the general case, autocorrelation can be represented by the scaled correlation matrix in the error terms:

$$\Omega = \sigma^2 \begin{array}{c} \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{array} \begin{array}{cccc} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \\ \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{21} & 1 & \cdots & \rho_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \cdots & 1 \end{pmatrix} \end{array}$$

For network data, variables y and x, represent relations between two actors, i (the "sender" of a relation) and j (the "receiver" of the relation). With this in mind, we can write a simple network model as follows:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij} \quad \text{for all } i \neq j.$$

And the general autocorrelation structure for this model is given as:

$$\Omega_{ij,kl} = \sigma^2 \begin{array}{c} \\ \epsilon_{12} \\ \epsilon_{13} \\ \vdots \\ \epsilon_{n(n-1)} \end{array} \begin{array}{cccc} \epsilon_{12} & \epsilon_{13} & \cdots & \epsilon_{n(n-1)} \\ \begin{pmatrix} 1 & \rho_{12,13} & \cdots & \rho_{12,n(n-1)} \\ \rho_{13,12} & 1 & \cdots & \rho_{13,n(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n(n-1),12} & \rho_{n(n-1),13} & \cdots & 1 \end{pmatrix} \end{array}$$

What makes network data particularly troublesome is represented in this autocorrelation matrix. Network data are assumed not to consist of independent observations, but rather have varying amounts of dependence on one another according to which row or column they "belong" to. That is, the error terms can be assumed to be autocorrelated, to at least some (unknown) degree, within rows and columns.

[3] The remaining assumptions of OLS are of less interest to us at this point in the paper. At the moment, all data are assumed to be continuously and normally distributed. We will drop these restrictions later when we specifically model the Sampson data.

In particular, one way of representing network data is to assume that they have an autocorrelation structure as follows:

$$\rho_{ij,kl} = \begin{cases} 1 & \text{if } i = k \text{ and } j = l; \quad \text{(diagonals of } \Omega\text{)} \\ \rho_{i,jl} & \text{if } i = k \text{ and } j \neq l; \quad \text{(row autocorrelation parameters)} \\ \rho_{j,ik} & \text{if } i \neq k \text{ and } j = l; \quad \text{(column autocorrelation} \\ & \qquad \qquad \qquad \qquad \qquad \text{parameters)} \\ 0 & \text{otherwise.} \end{cases}$$

That is, there is a set of row parameters $(\rho_{i,jl})$ that describe the non-independence of observations that occur in the same row of the original data matrix, and a set of column parameters $(\rho_{j,ik})$ that describe the non-independence of observations that occur in the same column of the original data matrix.

The reader familiar with econometrics might be tempted to apply general least squares (GLS) to this problem (a solution discussed by Proctor 1969). There are certain advantages to such an approach. If one knew what the set of row and column autocorrelation parameters were, one could use GLS to transform the original observations and derive appropriate estimates of the means and variances of α and β. However, I know of no theory to guide us in choosing such values. And, as Engle (1974) has demonstrated, it can be far worse to assume an incorrect autocorrelation structure than to incorrectly use OLS in the first place. To those who wish to attack this problem by estimating the parameters in a two-step estimation procedure (Judge *et al.* 1985), it is worth noting that there are $N\binom{N-1}{2}$ row parameters and an equal number of column parameters (total $= N(N-1)(N-2)$) to estimate. Given that there are only $N(N-1)$ observations, the autocorrelation parameters "... cannot be satisfactorily estimated" (Judge *et al.* 1985: 174).

3. The quadratic assignment procedure

A nonparametric answer to this problem of testing the null hypothesis that two network variables are uncorrelated has been proposed (Mantel 1967) and developed at length (Hubert and Schultz 1976; Hubert 1983; Hubert 1985; Hubert and Golledge 1981). [4] By generating all correla-

tions that result from permuting the rows and columns of one of the structural matrices, one can determine the distribution of all possible correlations given the structures of the two matrices. Thus, it builds into the test statistic the kind of row/column interdependence that is assumed in network data. This permutation procedure, referred to as the quadratic assignment procedure (QAP), is one answer to the aforementioned autocorrelation question. [5]

Our intent here is to compare OLS and QAP under various network assumptions. It has been shown elsewhere that, even under extreme autocorrrelation conditions in some models, OLS does not do badly in estimating the first two moments of regression coefficients (e.g. Engle 1974; Kramer 1980). The question we explore here is how well or poorly each does in providing a test of null hypotheses in structurally autocorrelated models.

4. Monte Carlo simulations of structurally autocorrelated data: The case of normally distributed data

Monte Carlo simulations were used to assess the effects of networks autocorrelation on Type I errors. Data were generated that were consistent with the null hypothesis (i.e. there is no correlation between y and x) and tested using both OLS and QAP to determine whether each test "reports" that the observed sample correlation was significantly different from zero. For these tests, α was arbitrarily chosen to be 0.10 (two-tailed).

---

[4] Accessible and more detailed descriptions of this technique appear elsewhere (Baker and Hubert, 1981; Krackhardt, 1987). For a clear description of the calculations, see Mantel (1967); for a thorough, mathematical treatment and review, see Hubert and Schultz (1976) and Hubert (1985).

[5] In his original formulation, Mantel (1967) provided an analytical solution to the problem of calculating the mean and variance of all the correlations under all row-column permutations. This direct, relatively simple method results in a Z-statistic which can be compared to the usual normal distribution to ascertain the probability of observing such a correlation under the null hypothesis that each permutation was an equally likely event. Of course, the interpretation of the Z-statistic depends on the assumption that the correlations will be normally distributed under all permutations. As has been shown elsewhere (Faust and Romney, 1985; Mielke, 1979), this assumption is questionable under certain conditions Costanzo *et al.* (1983) provide an alternative method to cover cases where the Z-statistic is not normally distributed. For our purposes, however, we will restrict ourselves to normally distributed cases that do not require such special treatment. All QAP tests performed in this paper are based on Mantel's Z statistic.

The key issue is whether these results might be affected by the size of the autocorrelation parameters. We could vary each one of the $N(N - 1)(N - 2)$ parameters separately and record the effects of each on Type I errors, but such effects would be difficult to report in any comprehensible way. For purposes of demonstration, then, we will vary all the row parameters and column parameters together. This will provide a simplified, tractable picture of the effects of degrees of structural autocorrelation.

The data for the model were generated simply as follows:

$$x_{ij} = \epsilon_{x_{ij}}$$

$$y_{ij} = K_{yx} x_{ij} + \epsilon_{y_{ij}}$$

or, since we are generating data from the null hypothesis, and thus $K_{yx} = 0$, simply

$$y_{ij} = \epsilon_{y_{ij}}.$$

Autocorrelation was created with the following generators:

$$\epsilon_{y_{ij}} = K_R u_{y_i} + K_C u_{y_j} + K_{ij} u_{y_{ij}}$$

$$\epsilon_{x_{ij}} = K_R u_{x_i} + K_C u_{x_j} + K_{ij} u_{x_{ij}}$$

with the following constraints:

$$0 \leq K_{ij} \leq 1$$

$$K_R = 1 - K_{ij}$$

$$K_C = K_R.$$

The random variables u_{y_i}, u_{y_j}, $u_{y_{ij}}$, u_{x_i}, u_{x_j}, and $u_{x_{ij}}$, are all independently drawn from a $N(0, 1)$ distribution. The constants K_R and K_C represent the degree of weight given to the row and column autocorrelations, respectively. The constant K_{ij} represents the degree of weight given to the "purely" random term for each observation. It is the relative size of K_R (and K_C) to K_{ij} that determines the autocorrelation in the rows (and columns). By constraining K_R to equal $K_C = 1 - K_{ij}$,

we can use the K_R constant to represent the strength of autocorrelation in the population as a whole. When K_R equals 0, there will be no autocorrelation in the data. When K_R equals 1, row and column autocorrelations will be the maximum possible.

To summarize, we generate sample data sets from a well-defined population. Each data set consists of two network variables, x and y, each of size $N \times (N-1)$ dyadic observations. To construct y, N random row components, N random column components, and $N \times (N-1)$ random error components are generated. Each y_{ij} observation, then, is the weighted sum of the ith row component, the jth column component, and the ijth error component. This procedure is then repeated with different random generators to create x. By this construction, we assure that the null hypothesis is always true in these populations (i.e. x and y are uncorrelated) and that the degree of structural autocorrelation is a direct, monotonic function of the constants K_R and K_C.

Results for the simple regression model

Forty-one populations were explored, with each succeeding population differing from the preceeding one by a small degree of row/column autocorrelation (increment in $K_R = 0.025$). From each population, 1,000 samples of y and x of size $N = 18$ were generated (the sample size of the Sampson data to be analyzed shortly). For each x–y pair of samples, test statistics based on OLS and on QAP were calculated. If the test indicated that the observed correlation between x and y was significant at the 0.10 level (two-tailed), then a Type I error was recorded for that test statistic for that particular simulation.

The results for all 41 populations are plotted in Figure 1 and summarized in Table 1. The solid line in the plot represents the proportion of samples where the OLS F-test was significant; the dotted line represents the proportion of samples where the QAP test was significant ($\alpha \leq 0.10$, two-tailed).

Ideally, one would hope that a test statistic would indicate a significant correlation about α fraction of the time (in this case, 0.1 of the time). As can be seen, both QAP and OLS hover around this "ideal" line when the autocorrelation is weak ($K_R < 0.25$). As the autocorrelation increases, however, the F-test starts to increase its Type I error rate considerably. When the autocorrelation is only moderate ($K_R = $
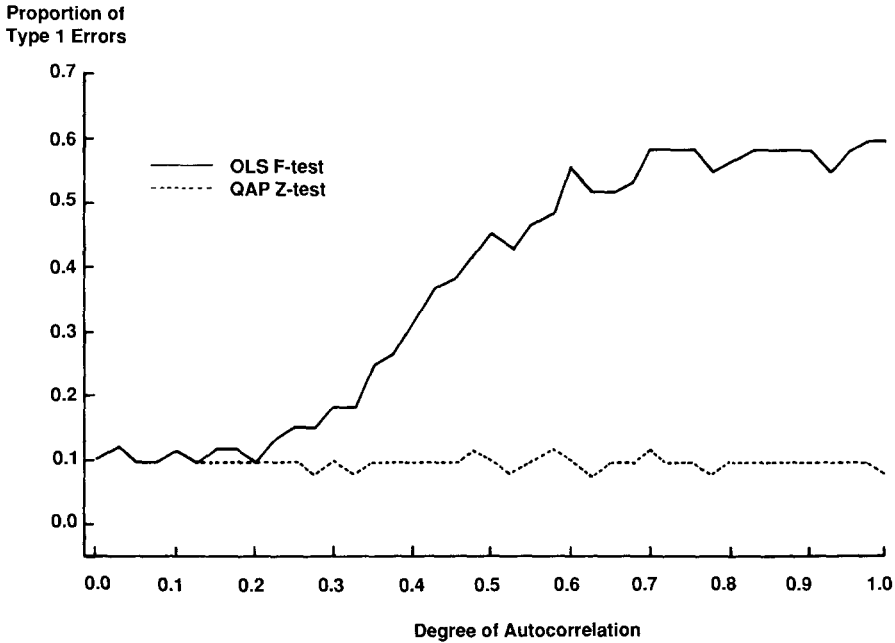
**Proportion of
Type 1 Errors**

Fig. 1. Plot of Type 1 error rates of significance tests for simple regression coefficients.

0.5), samples have better than a 40 percent chance of appearing significant to OLS. As the row autocorrelation becomes strong ($K_R >$ 0.8), the OLS tests show significant correlations in excess of 60 percent of the time.

Table 1
Summary of Type 1 error rates in tests of simple regression coefficients in simulated normal samples

	Type 1 errors for OLS test (%)	Type 1 errors for QAP test (%)	Total number of samples generated
Autocorrelated weight			
0 to 0.2	10.71	10.34	9000.00
0.225 to 0.4	20.01	9.76	8000.00
0.425 to 0.6	44.39	10.00	8000.00
0.625 to 0.8	55.63	9.94	8000.00
0.825 to 1.0	58.63	9.98	8000.00

In stark contrast, the QAP test of significance consistently finds approximately 10 percent of the samples to be significantly correlated at the 0.10 level. This is true no matter how large the autocorrelation.

5. Multiple regression analysis of network data

QAP is inherently a two-variable test of significance. Of more interest to most social scientists is the multiple variable case. How can we test whether a particular beta coefficient in a multiple regression equation is significantly different from zero when the data are structurally autocorrelated? To answer this question, we digress to discuss alternative forms for calculating multiple regressions.

5.1. The unstandardized multiple regression coefficient

Assume we are concerned with a model of the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \epsilon.$$

The appropriate OLS solution to solving for the vector of beta weights, β, is given by the following equation:

$$\beta = (X'X)^{-1} X'y.$$

The usual interpretation given to each beta is that, controlling for all the other independent variables in the equation, the beta represents the unique contribution that this particular variable makes toward one unit of the dependent variable, y. Another way to "control for" these other independent variables is to statistically extract the information predicted by the other independent variables, and then conduct a simple regression on the residuals (cf. Krackhardt 1987). Let $y^*_{234\ldots n}$ represent the residuals $y - \hat{y}_{234\ldots n}$ calculated by solving the regression equation predicting y from the set of x variables $x_2, x_3, x_4, \ldots, x_n$. Let $x^*_{1.234\ldots n}$ represent the residuals $x_1 - \hat{x}_{1_{284\ldots n}}$ calculated by solving the regression equation predicting x_1 from the set of x variables $x_2, x_3, x_4, \ldots, x_n$.

Then, if we calculate a new, simple regression coefficient between the two sets of residuals:

$$y^*_{234\ldots n} = \beta^*_0 + \beta^*_1 x^*_{1.234\ldots n} + \epsilon$$

the resulting β_1^* will be precisely equal to the β_1 in multiple regression model above. Similarly, each β can calculated as the simple regression coefficient between the residuals on y and the residuals on the appropriate x variable. The advantage of this form of calculation for our purposes is that the problem of calculating a multiple partial regression coefficient has been reduced to a simple regression. In this bivariate form, QAP can assess whether the correlation is significantly different from zero, and, by implication, whether the corresponding multiple regression β is significantly different from zero.

While QAP can *mechanically* assess the significance of the β coefficient, the question still remains as to whether this procedure is biased in the multiple regression case. To determine its bias, we conducted further simulations, this time on a three variable model. Observations were generated as follows:

$$x_{1_{ij}} = \epsilon_{x_{1_{ij}}}$$

$$x_{2_{ij}} = \epsilon_{x_{2_{ij}}}$$

$$y_{ij} = K_{yx_1} x_{1_{ij}} + \epsilon_{y_{ij}}$$

or, again since we are generating data from the null hypothesis, and thus $K_{yx_1} = 0$,

$$y_{ij} = \epsilon_{y_{ij}}.$$

In a similar manner to the first set of simulations, autocorrelation in the error terms was created using the following generators:

$$\epsilon_{y_{ij}} = K_R u_{y_i} + K_C u_{y_j} + K_{ij} u_{y_{ij}}$$

$$\epsilon_{x_{1_{ij}}} = K_R u_{x_{1_i}} + K_C u_{x_{1_j}} + K_{ij} u_{x_{1_{ij}}}$$

$$\epsilon_{x_{2_{ij}}} = K_R u_{x_{2_i}} + K_C u_{x_{2_j}} + K_{ij} u_{x_{2_{ij}}}$$

with the following constraints:

$$0 \leqslant K_{ij} \leqslant 1$$

$$K_R = 1 - K_{ij}$$

$$K_C = K_R$$

And, as before, u_{y_i}, u_{y_i}, $u_{y_{ij}}$, $u_{x_{1_i}}$, $u_{x_{1_i}}$, $u_{x_{1_{ij}}}$, $u_{x_{2_i}}$, $u_{x_{2_j}}$, and $u_{x_{2_{ij}}}$ are independently drawn from a Normal(0, 1) distribution.

Again, two tests were computed: (1) an OLS F-test on the multiple regression coefficient β_1 to determine whether the observed coefficient is significantly different from zero; and (2) a comparable QAP test on the β_1^* coefficient as calculated using the aforementioned residuals method.

5.2. Results of multiple regression simulations

Figure 2 plots the Type I error rates for each of the tests as a function of increasing structural autocorrelation; Table 2 summarizes these results. As can be seen by comparing the first two figures and tables, the pattern of biases for both the OLS and QAP tests in the multiple regression simulations are very close to those in the simple regression simulations above. The OLS tests are relatively unbiased (that is, their Type I error rate is approximately equal to $\alpha = 0.10$) when the autocor-
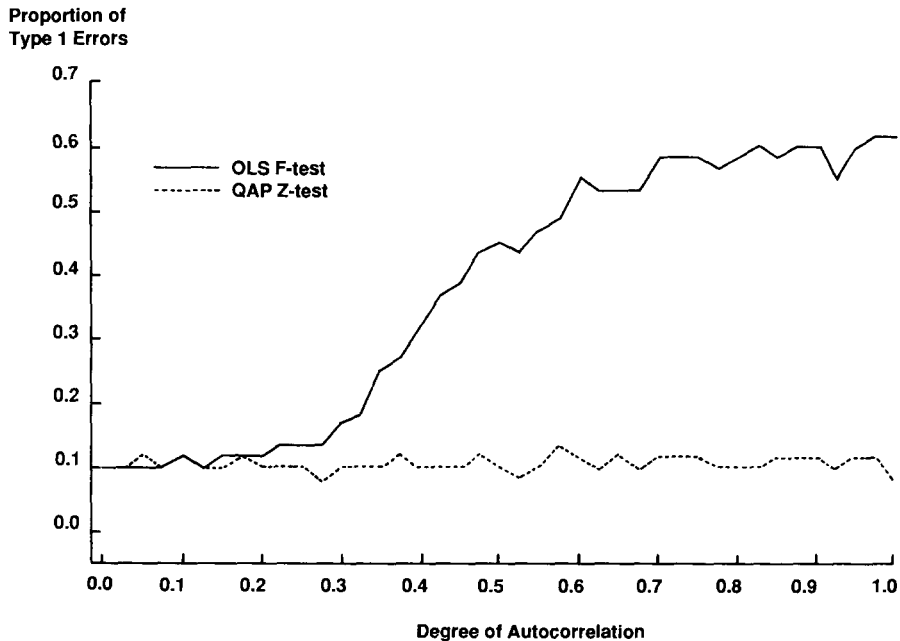

Fig. 2. Plot of Type 1 error rates of significance tests for multiple regression coefficients.

Table 2
Summary of Type 1 error rates in tests of multiple regression coefficients in simulated normal samples

	Type 1 errors for OLS test (%)	Type 1 errors for QAP test (%)	Total number of samples generated
Autocorrelation weight			
0 to 0.2	10.70	10.43	9000.00
0.225 to 0.4	19.84	9.90	8000.00
0.425 to 0.6	44.34	10.44	8000.00
0.625 to 0.8	56.19	10.49	8000.00
0.825 to 1.0	59.30	10.51	8000.00

relation is weak ($K_R < 0.25$), but then the error rate accelerates rapidly after that. On the other hand, the QAP test of the multiple regression coefficient is virtually unbiased.

6. Example: Sampson's monastery study

To illustrate how this procedure might work in practice, we provide an analysis of the well-known Sampson data (1968). It was Sampson's intent to explore the effects of social relationships on perceptions and judgment (using a classic autokinetic effect set of experiments) in a total institution, a Catholic monastery. As he was collecting his data, he discovered that 18 newcomers to the monastery were embroiled in an ideological and political struggle that mirrored tensions found in the Catholic Church following Vatican II. Sampson's careful data collection was followed, serendipitously, by a small revolution. Four of the eighteen were expelled from the monastery. Over the ensuing months, ten others quit in a well-documented sequence (Sampson 1968: 373–382).

Many researchers have taken advantage of Sampson's carefully documented data collection during this time (e.g. White et al. 1976). Much of this sophisticated analysis focused on the set of network questions Sampson asked, accompanied by his rich, ethnographic description of the events that followed, including the order in which people left (see Rice and Richards, 1985 for a review). Typically, this rich description was used to make sense of the structures found by the techniques being demonstrated.

Our tack here will depart somewhat from this tradition. We will explore this relationship, between social structure and order of leaving, more formally and statistically. In particular, we will use order of leaving as a dependent variable and the set of social relationships as independent variables. We will ask the following dyadic question: Were people who nominated others more likely to leave around the same time as those others? And, conversely, did people who were not tied to others tend to leave at very different times (much sooner or much later) than those others?

6.1. Dependent variable: Similarity in leaving order

The order in which the 18 novices left the monastery is well-documented in Sampson's original dissertation (pp. 373–382). The order is reproduced in Table 3. The four who were expelled were the first to leave (Gregory, Basil, Elias, and Simplicus), 10 others quit over the next seven months during which Sampson collected his data, and the remaining four novices (Bonaventure, Berthold, Ambrose, and Louis) presumably did not leave until they "graduated" from the order in good standing.

Since the question addressed is a dyadic one (similarity in exit order), a square matrix of similarities was created from the vector of ranks. Each cell in this matrix was simply the absolute difference between the rank of i and the rank of j (see Table 3 for the completed matrix).

6.2. Independent variable: Social ties

Eight different social relations were collected. Sampson asked each participant to nominate and rank three people in both the positive and negative categories for each relation. A portion of the questionnaire used to collect the sociometric information follows:

1. List those three brothers [in order] whom you personally liked the most.
2. List those three brothers [in order] whom you personally liked the least.
3. List those three brothers [in order] whom you most esteemed.
4. List those three brothers [in order] whom you esteemed least.

Table 3
Dependent variable: Order of exit from monastery

ID	Name and ID number provided by Sampson	Order of Exit
1	18 John Bosco	5.0
2	19 Gregory	2.5
3	20 Basil	2.5
4	24 Peter	13.0
5	25 Bonaventure	16.5
6	26 Berthold	16.5
7	30 Mark	9.0
8	32 Victor	10.5
9	33 Ambrose	16.5
10	34 Romuald	12.0
11	35 Louis	16.5
12	36 Winfrid	14.0
13	37 Amand	10.5
14	38 Hugh	8.0
15	39 Boniface	7.0
16	40 Albert	6.0
17	41 Elias	2.5
18	42 Simplicus	2.5

Matrix of Absolute Differences in Order of Exit

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0.0	2.5	2.5	8.0	11.5	11.5	4.0	5.5	11.5	7.0	11.5	9.0	5.5	3.0	2.0	1.0	2.5	2.5
2	2.5	0.0	0.0	10.5	14.0	14.0	6.5	8.0	14.0	9.5	14.0	11.5	8.0	5.5	4.5	3.5	0.0	0.0
3	2.5	0.0	0.0	10.5	14.0	14.0	6.5	8.0	14.0	9.5	14.0	11.6	8.0	5.5	4.5	3.5	0.0	0.0
4	8.0	10.5	10.5	0.0	3.5	3.5	4.0	2.5	3.5	1.0	3.5	1.0	2.5	5.0	6.0	7.0	10.5	10.5
5	11.5	14.0	14.0	3.5	0.0	0.0	7.5	6.0	0.0	4.5	0.0	2.5	6.0	8.5	9.5	10.5	14.0	14.0
6	11.5	14.0	14.0	3.5	0.0	0.0	7.5	6.0	0.0	4.5	0.0	2.5	6.0	8.5	9.5	10.5	14.0	14.0
7	4.0	6.5	6.5	4.0	7.5	7.5	0.0	1.5	7.5	3.0	7.5	5.0	1.5	1.0	2.0	3.0	6.5	6.5
8	5.5	8.0	8.0	2.5	6.0	6.0	1.5	0.0	6.0	1.5	6.0	3.5	0.0	2.5	3.5	4.5	8.0	8.0
9	11.5	14.0	14.0	3.5	0.0	0.0	7.5	6.0	0.0	4.5	0.0	2.5	6.0	8.5	9.5	10.5	14.0	14.0
10	7.0	9.5	9.5	1.0	4.5	4.5	3.0	1.5	4.5	0.0	4.5	2.0	1.5	4.0	5.0	6.0	9.5	9.5
11	11.5	14.0	14.0	3.5	0.0	0.0	7.5	6.0	0.0	4.5	0.0	2.5	6.0	8.5	9.5	10.5	14.0	14.0
12	9.0	11.5	11.6	1.0	2.5	2.5	5.0	3.5	2.5	2.0	2.5	0.0	3.5	6.0	7.0	8.0	11.5	11.5
13	5.5	8.0	8.0	2.5	6.0	6.0	1.5	0.0	6.0	1.5	6.0	3.5	0.0	2.5	3.5	4.5	8.0	8.0
14	3.0	5.5	5.5	5.0	8.5	8.5	1.0	2.5	8.5	4.0	8.5	6.0	2.5	0.0	1.0	2.0	5.5	5.5
15	2.0	4.5	4.5	6.0	9.5	9.5	2.0	3.5	9.5	5.0	9.5	7.0	3.5	1.0	0.0	1.0	4.5	4.5
16	1.0	3.5	3.5	7.0	10.5	10.5	3.0	4.5	10.5	6.0	10.5	8.0	4.5	2.0	1.0	0.0	3.5	3.5
17	2.5	0.0	0.0	10.5	14.0	14.0	6.5	8.0	14.0	9.5	14.0	11.5	8.0	5.5	4.5	3.5	0.0	0.0
18	2.5	0.0	0.0	10.5	14.0	14.0	6.5	8.0	14.0	9.5	14.0	11.5	8.0	5.5	4.5	3.5	0.0	0.0

Table 4

Correlations among independent variables and principal components analysis

	LIKE	ESTEEM	INFLUENC	PRAISE	DISLIKE	DISESTEE	NOTINFLU	BLAME
LIKE	1.00000	0.65306	0.65054	0.55413	-0.16842	-0.17595	-0.16490	-0.05481
ESTEEM	0.65306	1.00000	0.79708	0.62915	-0.14500	-0.18623	-0.17108	-0.14131
INFLUENC	0.65054	0.79708	1.00000	0.64192	-0.11341	-0.14180	-0.16961	-0.15074
PRAISE	0.55413	0.62915	0.64192	1.00000	-0.13614	-0.14812	-0.14107	-0.12638
DISLIKE	-0.16842	-0.14500	-0.11341	-0.13614	1.0000	0.68878	0.52165	0.33974
DISESTEE	-0.17595	-0.18623	-0.14180	-0.14812	0.68878	1.00000	0.66022	0.42456
NOTINFLU	-0.16490	-0.17108	-0.16961	-0.14107	0.52165	0.66022	1.00000	0.35523
BLAME	-0.05481	-0.14131	-0.15074	-0.12638	0.33974	0.42456	0.35523	1.00000

Rotated factor pattern

	FACTOR1	FACTOR2
LIKE	0.81931	-0.09804
ESTEEM	0.89293	-0.10866
INFLUENC	0.89971	-0.08116
PRAISE	0.80967	-0.08623
DISLIKE	-0.07089	0.81963
DISESTEE	-0.09215	0.89168
NOTINFLU	-0.10366	0.80610
BLAME	-0.07389	0.61606

Eigenvalues for each factor:

	FACTOR1	FACTOR2
	2.963366	2.531664

5. List those three brothers [in order] who had the most influence upon you.
6. List those three brothers [in order] who had the least influence upon you.
7. List those three brothers [in order] whom you went out of your way to support, praise and/or help because their behavior was consistent with your view of the Spirit of the Order.
8. List those three brothers [in order] whom you went out of your way to correct, encourage and/or help because their behavior was not consistent with your view of the Spirit of the Order.

The nominations were converted to scores as follows: first choice was recorded as '3', second choice as a '2', and third choice as a '1'.

As might be expected, there was considerable overlap in nominations. That is, a person who chose another as someone they liked also tended to choose that same person as someone who had influence over them, and whom they esteemed and praised. In fact, the correlations among these eight independent variables reveal a strong clustering pattern (see Table 4). The four positively phrased questions are strongly correlated with each other; the four negatively phrased questions are also strongly correlated with each other; and, as would also be expected, there is a small negative correlation between the positively phrased items and the negatively phrased items.

A factor analysis of the variables confirms these two clusters (see Table 4). The first two factors extracted from a principle components analysis account for 69 percent of the total variance in the data (the eigenvalues for the remaining six factors were substantially less than 1.0). A varimax rotation of the two factors shows clearly that the positive variables define the first factor and the negative variables define the second. Therefore, we summed the four positive choice questions into a combined choice matrix; and, similarly, we summed the four negative choice questions into a combined negative choice matrix. The individual scores in each cell of each summary matrix could range from 0 (i did not choose j in any of the four questions) to 12 (i chose j as a first choice in all four questions). [6] Chronbach's

---

[6] In fact, the range for the sum of scores in the positive matrix was 0 to 12, and the sum of scores in the negative matrix was 0 to 6.

Table 5
Results of QAP multiple regression analysis of Sampson data

Dependent variable: Difference in exit order	Unstandardized regression coefficients and associated QAP test		
Independent variable: Positive ties	Model 1 -0.4025 $Z = -3.981$ $p < 0.0001$	Model 2	Model 3 -0.3482 $Z = -3.487$ $p < 0.0005$
Negative ties		0.3628 $Z = 3.236$ $p < 0.005$	0.2800 $Z = 2.466$ $p < 0.05$

alphas for the positive and negative combined scores were 0.88 and 0.80, respectively.

6.3. Results of QAP multiple regression analysis of Sampson data

Table 5 contains the results of this analysis. In the first model, positive links contribute significantly (negatively) to dissimilarity. The -0.40 coefficient is significant beyond the 0.0001 level. Recall that a high difference score in the dependent variable means that the pair of people are dissimilar in their exit position. The negative coefficient, therefore, suggests that when person i reports a positive connection to person j, he is more likely to exit around the same time as j than if i did not report such a positive connection. Similarly, in Model 2, negative nominations (such as blame, dislike, etc.) contribute to the two people leaving at different points in time (*beta* = 0.36, $p \leq 0.005$). In the multiple regression model (Model 3), the coefficients and significance levels lose a little strength, but they tell much the same story. One may interpret these results hierarchically by simply stating that positive and negative ties each add significantly to the variance explained by the other.

7. Modeling the Sampson data

As was shown in the simulations above, if dyadic data are continuous and normally distributed, QAP does a reasonable job of testing the significance of both simple and multiple regression coefficients no

matter what degree of structural autocorrelation there might be. However, in the current example, the Sampson data are neither continuous nor normally distributed. In fact, the dependent variable, because of its construction as a set of differences, has a unique pattern of autocorrelations — a pattern that was not modeled in the above simulation. The independent variables were also not modeled accurately in the previous simulations. The underlying scores were skewed (only three nominations were permissible in each underlying relation). And, as Faust and Romney (1985) point out, relying on QAP to test relationships between data that may be skewed can be misleading.

As a general procedure, it may be wise to heed Faust and Romney's warning by modeling the particular characteristics inherent in the data being analyzed. To account for these possible peculiarities, another simulation was conducted to test the bias of the QAP multiple regression in the Sampson data. This time, the dependent variable was created by generating a vector of length 18 of random numbers. These numbers were converted to ranks, y. A dependent matrix, Y, was calculated as follows:

$$Y_{ij} = |y_i - y_j|$$

To be conservative, the independent variables, X_1, X_2, were generated with maximum skew. This was done by reducing the summary matrix to one relation, with each row containing one '1', one '2', one '3', and fifteen '0's. The location of the '1', '2' and '3' in each row was determined by generating a vector (length 18) of random numbers, ranking them, converting the highest number to a '3', the next highest to a '2', the next to a '1', and the remaining fifteen to '0'.

The opportunity to introduce autocorrelation presented itself in these independent variables. It is quite possible, under the null hypothesis that the time of exits is unrelated to social ties, that some j's would be popular choices (e.g., "everyone likes Sam" or "Steve could never influence anyone"). In other words, there is room for a considerable amount of column autocorrelation. As before, we varied the degree of autocorrelation in the independent variables as follows:

$$\epsilon_{X_{1_{ij}}} = K_C u_{X_{1_j}} + K_{ij} u_{X_{1_{ij}}}$$

$$\epsilon_{X_{2_{ij}}} = K_c u_{X_{2_j}} + K_{ij} u_{X_{2_{ij}}}$$

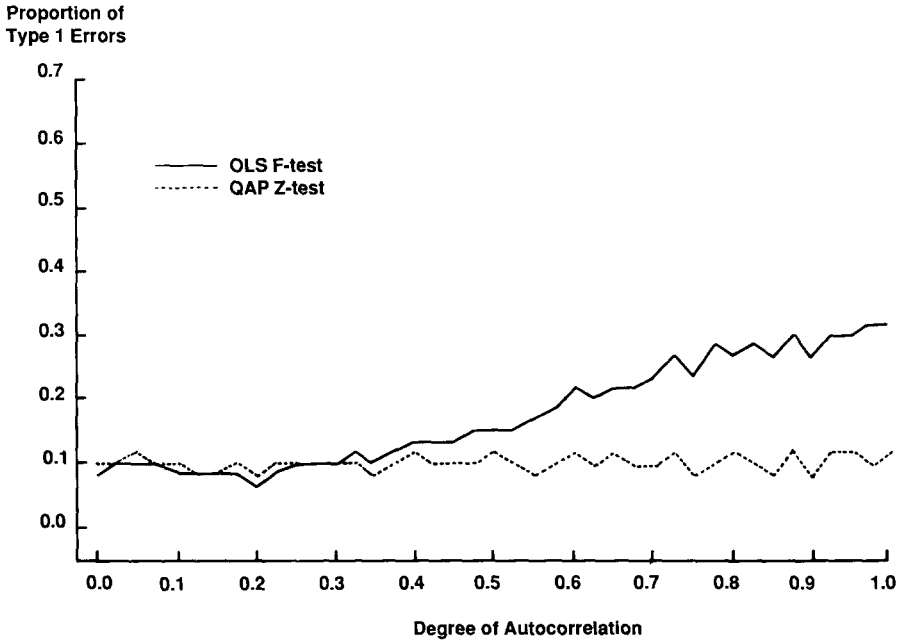**Proportion of**
Type 1 Errors

Fig. 3. Plot of Type 1 error rates of significance tests for multiple regression coefficients in simulations of Sampson-type data.

with the following constraints:

$$0 \le K_{ij} \le 1$$

$$K_C = 1 - K_{ij}.$$

Again, all u's are independently and normally distributed. These autocorrelated scores were converted to ranks (as previously described) and the ranks converted to Sampson-type scores. As before, 41,000 samples were generated from 41 populations, each with a slightly different degree of autocorrelation as determined by the size of the constant K_C.

The results of the simulation are depicted in Figure 3 and summarized in Table 6. Once again, the F-test in the OLS analysis becomes substantially biased with an increase in autocorrelation. The QAP test does not.

Table 6
Summary of Type 1 error rates in tests of multiple regression coefficients in simulations of Sampson-type data

	OLS Type-1 errors (%)	QAP Type-1 errors (%)	Total number of samples generated
Autocorrelation weight			
0 to 0.2	8.86	9.72	9000.00
0.225 to 0.4	10.75	9.80	8000.00
0.425 to 0.6	16.08	10.16	8000.00
0.625 to 0.8	23.98	10.28	8000.00
0.825 to 1.0	29.49	10.24	8000.00

8. Conclusion

In summary, the results of these simulations suggest that, when the assumptions behind OLS analysis are met (specifically, there is no autocorrelation in the error terms), either OLS or QAP may be used to test the significance of simple or multiple regression coefficients. When structural autocorrelation exists, however, QAP provides a relatively unbiased test of the coefficients, whereas OLS can become severely positively biased. Given these results, and given that the degree of autocorrelation in any set of network data cannot be reliably estimated, then we conclude that the QAP test of regression coefficients are interpretable and caution against using OLS procedures.

We wish to make clear, however, that QAP is not a universal answer to all dyadic analysis problems. For example, the QAP test does not adjust for the number of independent variables in the multiple regression equation, as the OLS analysis does. If the ratio of independent variables to number of actors gets too large, and the data are strongly autocorrelated, then the estimates of β's and significance levels might be severely biased. And, as mentioned before, other cases of potential bias could arise if the data are highly skewed, although no apparent bias was uncovered in the current Sampson example. In general, then, it is advised that the researcher cautiously test for bias in questionable cases by running simulations to model the specific data structures being analyzed. Over time, with sufficient examination of the boundary conditions for QAP multiple regression analysis, firmer guidelines for its recommended use are likely to emerge.

References

Baker, F.B. and L.J. Hubert
 (1981) "The analysis of social interaction data: A nonparametric technique". *Sociological Methods and Research 9* (3), 339–361.
Breiger, R.L. and P.E. Pattison
 (1986) "Cumulated social roles: The duality of persons and their algebras". *Social Networks 8*, 215–256.
Burt, R.S. and M.J. Minor (Eds.)
 (1983) *Applied Network Analysis: A Methodological Introduction*. Beverly Hills: Sage Publications.
Burt, R.S.
 (1982) *Towards a Structural Theory of Action*. New York: Academic Press.
Coleman, J.A.
 (1983) Purposive action embedded in social networks. Keynote address at 3rd Annual Social Network Conference, San Diego, California.
Costanzo, C.M., L.J. Hubert and R.G. Golledge
 (1983) "A higher moment for spatial statistics". *Geographical Analysis 15*(4), 347–351.
Doreian, P., K. Teuter and C. Wang
 (1984) "Network autocorrelation models, some Monte Carlo results". *Sociological Methods and Research 13*(2), 155–200.
Dow, M., M. Burton, D. White and K. Reitz
 (1984) "Galton's problem as network autocorrelation". *American Ethnologist 11*, 754–770.
Engle, R.F.
 (1974) "Specifications of the disturbance term for efficient estimation". *Econometrica 42*, 135–146.
Faust, K. and A.K. Romney
 (1985) "The effect of skewed distributions on matrix permutation tests". *The British Journal of Mathematical and Statistical Psychology 38*, 152–160.
Freeman, L.C.
 (1978) "Centrality in social networks: Conceptual clarification". *Social Networks 1*, 215–239.
Hubert, L.J.
 (1983) Inference procedures for the evaluation and comparison of proximity matrices. In J. Felsenstein (ed.), *Numerical Taxonomy*. New York: Springer-Verlag.
Hubert, L.J.
 (1985) "Combinatorial data analysis: Association and partial association". *Psychometrika 50*(4), 449–467.
Hubert, L.J. and R.G. Golledge
 (1981) "A heuristic method for the comparison of related structures". *Journal of Mathematical Psychology 23*, 214–226.
Hubert, L.J. and J. Schultz
 (1976) "Quadratic assignment as a general data analysis strategy". *British Journal of Mathematical and Statistical Psychology 29*, 190–241.
Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lutkeponl and T. Lee (Eds.)
 (1985) *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.
Krackhardt, D.
 (1987) "QAP partialling as a test of spuriousness". *Social Networks 9*, 171–186.
Kraemer, H. and C. Jacklin
 (1979) "Statistical analysis of dyadic social behavior". *Psychological Bulletin 86*, 217–224.

Kramer, W.
    (1980) "Finite sample efficiency of ordinary least squares in the linear regression model and autocorrelated errors". *Journal of the American Statistical Association 75*, 1005–1009.
Lincoln, J.R.
    (1984) "Analyzing relations in dyads". *Sociological Methods and Research 13*(1), 45–76.
Mantel, N.
    (1967) "The detection of disease clustering and a general regression approach". *Cancer Research 27*(2), 209–220.
Mielke, P.W.
    (1979) "On asymptotic non-normality of null distributions of MRPP statistics". *Communications in Statistics-Theory and Methods A8*(15), 1541–1550.
Ord, K.
    (1975) "Estimation methods for models of spatial interaction". *Journal of the American Statistical Association 70* (March), 120–126.
Proctor, C.H.
    (1969) "Analyzing pair data and point data on social relationships, attitudes and background characteristics of Costa Rican Census Bureau employees". *Proceedings of the Social Statistics Section, American Statistical Association*, 457–465.
Rice, R.E. and W.D. Richards
    (1985) An overview of network analysis methods and programs. In B. Dervin and M. Voigt (eds.), *Progress in Communication Sciences, VI*. Norwood, NJ: Ablex Publishing Corp.
Sampson, S.F.
    (1968) A Novitiate in a period of change: An experimental and case study of social relationships. Unpublished Doctoral Dissertation, Cornell University.
White, H.C., S.A. Boorman and R.L. Breiger
    (1976) "Social structure from multiple networks. I. Blockmodels of roles and position". *American Journal of Sociology 81*, 730–780.