6

CORRELATION AND ASSOCIATION MODELS FOR STUDYING MEASUREMENTS ON ORDINAL RELATIONS

*Katherine Faust**
Stanley Wasserman†

This paper describes and illustrates correlation models (correspondence analysis and canonical correlation analysis) and association models for studying the order and spacing of categories of ordinal relational variables. Both correlation models and association models study departures from independence in two-way contingency tables. One result of fitting these models is the possibility of assignment of scores to the categories of the row and/or the column variables to reflect the relative spacing of these categories. If the model fitting is done using statistical procedures, then restricted versions of these models allow one to test hypotheses about the spacing, linearity, or equality of the categories. Correlation and association models are especially useful for studying discrete ordinal variables, which arise quite frequently in the social and behavioral sciences.

We illustrate correlation and association models using two empirical examples in which respondents used ordered

*University of South Carolina
†University of Illinois

This paper was presented at the Conference on Measurement Theory and Social Networks, University of California, Irvine, August, 1991. We would like to thank Carolyn Anderson, Clifford Clogg, Linton Freeman, Peter Marsden, James A. McRae, Alaina Michaelson, and several anonymous reviewers for helpful comments. Special thanks go to Carolyn Anderson for computational assistance.

categories to rate the strength of their liking for, or acquaintance with, others in a social network. In this paper we describe how to use both correlation models and association models to test specific hypotheses about the spacing of these response categories.

1. INTRODUCTION

This paper describes and illustrates correlation and association models for studying the order and spacing of categories on ordinal relational variables. Both correlation and association models study the nature and strength of the relationship between rows and columns in a contingency table. Correlation models (including correspondence analysis and canonical correlation models) and association models have been the focus of considerable research in the last decade or so (Anderson 1992; Becker and Clogg 1989; Becker 1990; Böckenholt and Böckenholt 1990; Clogg 1982*a*, 1982*b*, 1986; Gilula 1986; Gilula and Haberman 1986, 1988; Goodman 1979, 1981*a*, 1981*b*, 1985, 1986, 1991; Greenacre 1984; Haberman 1981; Nishisato 1980; van der Heijden and de Leeuw 1985, 1989; van der Heijden and Meijerink 1989).

Correlation and association models both study departures from independence in contingency tables; however, the models differ in how they measure the strength and nature of the relationship between the rows and columns. Correlation models focus on departures from independence using the correlations between row categories and column categories. Both correspondence analysis and canonical correlation analysis are often referred to as *correlation* models. Alternatively, one could study departure from independence using other measures of the relationship between rows and columns, such as odd ratios for two-by-two subtables in a two-way cross-classification. Such models are referred to as *association* models. We describe both correlation and association models in detail below.

Correlation models and association models involve the assignment of scores to the categories of the row and column variables in order to maximize the relevant measure of relationship (the correlation coefficient in the correlation models or the measure of intrinsic association in association models). One can then use the scores pertaining to the row or column categories to study the order and spacing of these categories. Both models are especially interesting when the row and/or column variables are ordinal. A few examples of substantive problems for which these models have been used to study ordinal variables include: measures of well-being or happiness (Clogg 1982*a*, 1982*b*; Goodman 1985, 1986) attitudes toward treatment of criminals (Clogg 1982*b*), socioeconomic status (Goodman 1985; 1991), and levels of donations from corporations to not-for-profit agencies (Wasserman, Faust, and Galaskiewicz 1990).

Restricted versions of correlation and association models place constraints on the values of the scores assigned to the row and/or the column categories. If the model fitting is done using statistical procedures, one can then use the restricted models to test specific hypotheses about the dimensionality of the solution (the number of sets of scores needed), and about the spacing of row and/or column categories (such as their equality, uniform spacing, or other a priori spacing).

In this paper we describe and illustrate both correlation models and association models, including versions of these models that place restrictions on the scores for the row and/or column categories. The specific problem that we focus on is the assignment of scores to peoples' ratings of the strength of their acquaintance or friendship with others in a social network. The goal is to use models of correlation and association to study the order, spacing and equality of response categories that respondents use to indicate their degree of acquaintance or friendship. We use two social network data sets: one on observed and reported interactions among members of a fraternity (Bernard, Killworth, and Sailer 1979–80) and the second on friendship and message sending among members of a computer network (S. Freeman and L. Freeman 1979; L. Freeman and S. Freeman 1980; L. Freeman 1986). In both cases we use the measures of interactions among people as predictors of the strength of their friendship or acquaintance. We conclude the paper with a general comparison of correlation and association models.

Our illustrations use the specific example of ordered relational variables measuring the strength of ties among actors in a social network. However, it is important to note that the correlation and association models described here are applicable to any two-way contingency table of counts or frequencies, not just to social network data. We begin by describing our application before moving on to a discussion of the models.

2. ORDERED RELATIONAL VARIABLES

In recent decades social network analysis has become widely accepted as an approach for modeling social systems as collections of relational ties linking actors. The actors in the network are social units (such as people, nations, corporations, and so on), and the relational ties are substantive connections among the actors (such as friendships among people, imports and exports among nations, or interlocking boards of directors among corporations). The ties among actors may have values or strengths indicating the intensity, frequency, closeness, or amount of the relational tie between a pair of actors. Valued relational variables are almost always discrete and are often measured on an ordinal scale.

When actors are people in a group, relational ties can be measured by having people evaluate the strength of their ties to others within the group. An important question is how the various responses people give in evaluating the strength of their relational ties to others indicate the relative intensity of the relational ties. In this paper we describe models for studying these responses directly.

Researchers have considered the strength of network ties from several perspectives. Authors such as Granovetter (1973), Winship (1977), and more recently Freeman (1992), among others, consider the implications of the distribution of strong and weak ties for social structural patterns and processes. Other authors, notably Marsden and Campbell (1984) and Friedkin (1990) have considered factors that influence whether strong versus weak ties will occur between people. Relatively less attention has been paid to studying the strength of ties directly (however, see Burt and Guilarte 1986). In this paper we take this third perspective, by proposing and illustrating models to assign scores to categories of tie strength so that we can study directly the strength of relational ties.

Granovetter (1973) was among the first to discuss the theoretical importance of tie strength, distinguishing between strong, weak, and absent ties. He argued that

> The strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie. (1973, p. 348)

Granovetter also discussed the implications of tie strength for social structural processes such as the diffusion of novel information and community integration. Following the arguments of Granovetter, certain patterns of strong and weak ties are permitted, or forbidden within a network. For example, if actor i has a strong tie to actor j, and actor j in turn has a strong tie to actor k, then the tie from actor i to actor k should not be absent. Freeman (1992) argues that one way to determine which level of tie is strong versus weak is to describe properties of networks that hold at each level of a valued relation. Theoretically important properties (for example, transitivity) should hold for strong ties but not necessarily for weak ties.

In their discussion of how to measure the strength of relational ties, Marsden and Campbell (1984) distinguish between *indicators* and *predictors* of tie strength. Indicators are "actual components of tie strength" (p. 485) as specified by Granovetter, whereas predictors are variables such as context and attribute similarity that are related to the strength of ties. Friedkin (1990) argues that components of tie strength (discussion, seeking help, and friendship) form a Guttman scale, rather than an additive function (as argued by Granovetter) in that "the claim of friendship implies the claims of help seeking and frequent discussion; the claim of help seeking implies the claim of frequent discussion" (p. 250).

Few studies have focused directly on the strength of relational ties linking pairs of individuals. A notable exception is Burt and Guilarte (1986), who studied tie strength in the General Social Survey ego-centered network data by looking at reported properties of ties from respondent to the alters named, and among pairs of alters named by the respondent. In these data, respondents evaluated the relational tie between each pair of alters named as "especially close," "acquainted," or "strangers."

Burt and Guilarte (1986) propose that response categories indicating the strength of relational ties can be scaled by considering how the probability of a given level of a second variable changes across categories of the relational variable that is being scaled. For example, if one is scaling categories of friendship strength, then one could compare the probability of a specific amount of behavioral interaction across the several categories of friendship. In Burt and Guilarte's model, the spacing between two friendship categories would be proportional to the difference in the probabilities of a specific level of interaction between the friendship categories. Their paper includes details on how to estimate these values.

Using this method to scale response categories for acquaintanceship in the General Social Survey network data, Burt and Guilarte concluded that "the middle category of interalter relations lies about 0.2 of the distance from total strangers to people being especially close" (p. 391). On the other hand, they found that respondents make no distinction between alters with whom they are "especially close" and alters to whom they are "equally close," whereas alters who are "less close" are about 0.7 the strength of "especially close" or "equally close" (p. 395). Thus their method results in a set of scale values describing the spacing or intervals between response categories on an ordinal relational variable. The models we describe in this paper provide an alternative method for assigning scores to the categories of an ordinal relational variable.

An important property of a relational variable is whether it is dichotomous (taking on only two values) or whether it is valued. Valued relational variables usually have values indicating the strength, intensity, or frequency of the relational tie.

Social network data are often collected by asking respondents to *rate* the strength or intensity of their relational ties to others in the group (for example, their degree of friendship or respect for each person in the group). Responses may take the form of labeled categories, for example "close personal friend," "friend," "acquaintance," "someone I have met," "someone I have heard of but not met," "someone I have not heard of" (for example, see S. Freeman and L. Freeman 1979; L. Freeman and S. Freeman 1980). Or, the responses may be numerical values indicating the intensity of the relational tie. Whether verbal labels or numerical values are used, the resulting relation is measured by responses on a number of ordered response categories.

It is important to contrast the *rating* response format, where respondents use a limited number of response categories, with a *complete rank order* format. In a complete rank order format, respondents typically are asked to rank order the other people in the network from most to least in terms of the intensity of the respondent's relational tie to each other person. If there are g people in the group, then respondents are asked to use all integers from 1 to $g - 1$ to rank order the strength of their ties to others. By contrast, in a rating format the response categories may be reused by a respondent. In fact, if the number of response categories, C, is less than the number of other people in the group, $g - 1$, then a respondent must reuse some response categories.[1]

Consider the example of the five ordered response categories for measuring friendship that we described above. Although it seems likely that degree of friendship is ordered from "close personal friend" through "someone I have met" to "someone I have not heard of," it is important to study the relative spacing of these categories. It might be the case that respondents see very little difference between "a friend" and "a close personal friend" but both responses are quite different from "a person I have met." One of the results of the models described here is the assignment of scores to response categories of ordinal relational variables to reflect the order and spacing of the categories.

Let us consider a respondent choosing among the response categories on a given relation to indicate the strength of her relational tie to each other person in the group. In complete social network studies, each respondent judges her own relational ties to all other people in the group. For example, each respondent judges her degree of friendship with each member of the group. We assume that individuals are presented with stimuli (relational ties) that vary in terms of important determinants of degree of the response relation

[1]In order to aggregate responses across people, we are assuming that all respondents use the response categories in the same way. For example, if a rating of "1" means that a person is disliked by the respondent, then this response should tend to go with infrequent interactions for all respondents who use the response category "1." However, in a full rank order format all respondents are forced to use the category "1" for their least-liked person in the group, regardless of their absolute degree of liking for, or frequency of interaction with, that person. Thus responses from a full rank order format are not expected to be associated with predictor relational variables in the ways required by our approach. In a more general context, Nishisato (1980) discusses correspondence analysis models for rank order data coded as paired comparisons.

(say friendship). A given respondent, when presented the list of others in the group, is faced with people whom they have known a long time, people they have met recently, and others they have never met (ties vary in duration). In addition, there are some people whom the respondent sees quite often, and others whom the respondent sees only occasionally (ties vary in frequency of contact). Also, some of the people may be the respondent's family members, coworkers, or neighbors (ties vary in context). Thus duration, frequency, and context are also properties of the relational ties to which the respondent is assigning an evaluation of strength of friendship.

We assume that a respondent's assessment of strength of a relational tie depends primarily on properties of the relational tie from respondent to alter, and not on the attributes of the respondent or of the alter. Thus the response category that is used by a respondent to describe the strength of their relational tie should be associated with other aspects of the relationship from respondent to alter. For example, the degree of friendship expressed by a respondent for alter is likely to be associated with the length of time they have known each other, the frequency with which they interact, and so on (see Marsden and Campbell 1984).

It seems unreasonable to assume that all respondents have the same degree of friendship with a specific other person. Thus we do not assume that all respondents will use the same response category (for example, "friend") for a given person. Rather, we assume that in general people use the same response category on a relational variable to describe *relational ties* that are similar on other relational variables. These other relational variables are considered to be *predictors* of the strength or value of a relational tie (Marsden and Campbell 1984). Therefore, in order to study response categories on a given relation, we must have (at least) a second relational variable measured on the same pairs of actors. We will distinguish between the *relational response* variable (whose categories we are attempting to scale), and the *relational predictor* variable(s) that we use to understand the response categories on the response relation.

For example, we can study response categories for different degrees of friendship by examining how the categories are associated with categories for other predictor relations, such as the frequency, duration, intensity, and context of the relational tie.

3. NOTATION AND DATA ARRAYS

Both correlation and association models study contingency tables. We will denote the contingency table as **F**, where f_{kl} is the observed frequency in row k column l of **F**. Commonly, the frequencies in **F** record the responses from g respondents to two (or more) questionnaire items. All of the models described in this paper can be used to study such a contingency table. However, to study social network data, the particular example we will use, it is necessary to focus on *pairs* of people, rather than individual respondents. In this section we describe the particular data arrays that are required to study social network data.

We begin with a set of g actors, and two (or more) relations $\mathcal{X}_1$, $\mathcal{X}_2, \ldots, \mathcal{X}_R$, defined on these actors. We will designate $\mathcal{X}_1$ as the relational response variable, whose response categories we are studying. In addition we will have $\mathcal{X}_2$ (and possibly other $\mathcal{X}$'s) as relational predictor variable(s). There are R relations in total. Let x_{ijr} be the value of the relational tie from actor i to actor j on relation $\mathcal{X}_r$. We will assume that these values are ordered and discrete. In general, we let C_r be the number of levels of relational variable $\mathcal{X}_r$.

The most common data representation for social network data is a *sociomatrix*. A sociomatrix for a single relation, $\mathrm{X} = \{x_{ij}\}$, is a matrix with g rows and g columns, indexing actors and partners. The (i, j)th entry of **X** codes the value of the tie from row actor i to column actor j. However, when the focus of the analysis is the strengths or values of the relational ties, a sociomatrix is not the appropriate data array to analyze. In this section we describe a two-way array that codes the relational ties among a set of actors on two or more relations, which will allow us to fit proper models. (The appendix to this chapter describes in detail the relationship between this array and other common data arrays that are used for fitting the models that we describe in later sections.)

We are interested in studying the distribution of relational response categories for ties defined on ordered pairs of actors, across different levels of one (or possibly more) predictor relations. The idea is to code the state of each of the $g(g - 1)$ dyads (or ordered pairs of actors) defining the relational ties in a network data set. This state is defined by two quantities: the category (or strength) of the relational response variable and the combination of categories of the relational predictor variables.

First, consider the state of an ordered pair of actors on the relational response variable, X_1. Since this variable has C_1 categories, each ordered pair of actors can be in one of C_1 states on this variable. Now consider the number of states for the relational predictor variable(s). This number depends on the number of relational predictor variables that are included and the number of categories of each. We will let L be the total number of states on the relational predictor variables. In the simplest case there is a single relational predictor variable, X_2, with C_2 levels, and there are as many possible states as there are levels of X_2: namely, $L = C_2$. If there is more than one relational predictor variable, then we consider the combination of possible states on all predictor variables. This state can be coded by the cross-classification of these predictor variables. In general, there are $R - 1$ relational predictor variables, $X_2, X_3, \ldots, X_R$, with C_2, $C_3, \ldots, C_R$ categories, respectively. The state of an ordered pair of actors on these variables is given by the cross-classification of these variables, and the number of possible states is equal to the number of entries in the cross-classification. Since there are $C_2 \times C_3 \times \ldots \times C_R$ cells in this cross-classification, there are $L = C_2 \times C_3 \times \ldots \times C_R$ possible states for an ordered pair of actors on the relational predictor variables.

To study the state of an ordered pair of actors on both the relational response variable and the relational predictor variable(s), we focus on the cross-classification of the relational variable whose response categories we are studying, with one or more other relational predictor variables. For example, we can look at the cross-classification of level of friendship and frequency of contact for pairs of actors in a group. This cross-classification, which we denote by **F**, has C_1 rows coding the state of the ordered pair on the relational response variable and L columns coding the state of the ordered pair on the relational predictor variable(s).

The **F** array is a C_1 by L table, whose entries code the state of each ordered pair of actors. Since there are $g(g - 1)$ ordered pairs of actors, there are $g(g - 1)$ observations classified in **F**. For a single relational predictor variable, the entry in cell (k,l) of **F** counts the number of times $x_{ij1} = k$ and $x_{ij2} = l$, for $i,j = 1,2 \ldots ,g$, and $i \neq j$: the number of times response category k is used at level l of the relational predictor variables. For $R - 1$ relational predictor variables with $L = C_2 \times C_3 \times \ldots \times C_R$ states, each entry in **F** counts the number of times category k of the relational response variable is used for each state of the combined relational predictor variables.

We will now describe correlation models (correspondence analysis and canonical correlation analysis), including restricted versions of these models, and illustrate how these models can be used to study the spacing, linearity, and equality of relational response categories. We then describe and illustrate association models.

4. CORRELATION MODELS: CORRESPONDENCE AND CANONICAL CORRELATION ANALYSIS

In this section we present a general description of the mathematics of correlation models, including both canonical correlation analysis and correspondence analysis for a two-way contingency table. More extensive discussions of correspondence analysis and canonical correlation analysis can be found in Gilula and Haberman (1986, 1988), Goodman (1985, 1986), Greenacre (1984), van der Heijden and de Leeuw (1985, 1989), van der Heijden and Meijerink (1989), Wasserman and Faust (1989), and Wasserman, Faust, and Galaskiewicz (1990).

We define:

- $\mathbf{F} = \{f_{kl}\}$ a two-way cross-classification, with C_1 rows and L columns,
- $P_{kl} = \dfrac{f_{kl}}{f_{++}}$ the probability that an observation is in row k, column l,
- $P_{k\bullet} = \dfrac{f_{k+}}{f_{++}}$ the probability that an observation is row k,
- $P_{\bullet l} = \dfrac{f_{+l}}{f_{-+}}$ the probability that an observation is in column l,
- $t = \min(C_1 - 1, L - 1)$.

The $\{P_{kl}\}$ probabilities are observed, rather than theoretical probabilities. The *canonical decomposition* of **F** is defined as:

$$P_{kl} = P_{k\bullet}P_{\bullet l}\left[1 + \sum_{m=1}^{t} p_m u_{km} v_{lm}\right]. \tag{1}$$

We will refer to this as the $C(t)$ model. Goodman (1986) refers to the theoretical version of this as the *saturated RC canonical correlation* model.

Canonical correlation analysis of $\mathbf{F}$ results in three sets of information:

- A set of C_1 row scores, $\{u_{km}\}$ for $m = 1,2, \ldots t$,
- A set of L column scores, $\{v_{lm}\}$ for $m = 1,2, \ldots t$,
- A set of t principal inertias (squared canonical correlations), $\{\rho_m^2\}$ for $m = 1,2, \ldots t$; ρ_m measures the correlation between the row scores, u_{km}, and the column scores, v_{lm}.

The canonical variables $\mathbf{u}_m$ and $\mathbf{v}_m$ are constrained as follows:

$$\sum_{k=1}^{C_1} u_{km}P_{k\bullet} = \sum_{l=1}^{L} v_{lm}P_{\bullet l} = 0 \tag{2}$$

$$\sum_{k=1}^{C_1} u_{km}^2 P_{k\bullet} = \sum_{l=1}^{L} v_{lm}^2 P_{\bullet l} = 1. \tag{3}$$

When scaled in this way, the $\mathbf{u}$'s and $\mathbf{v}$'s are referred to as *standard coordinates* (Greenacre 1984). For distinct m and m', $\mathbf{u}_m$ and $\mathbf{u}_{m'}$ are uncorrelated, as are $\mathbf{v}_m$ and $\mathbf{v}_{m'}$:

$$\sum_{k=1}^{C_1} u_{km}u_{km'}P_{k\bullet} = \sum_{l=1}^{L} v_{lm}v_{lm'}P_{\bullet l} = 0, \qquad \text{for } m \neq m'. \tag{4}$$

For a given m, the correlation between canonical variables $\mathbf{u}_m$ and $\mathbf{v}_m$ is equal to the canonical correlation ρ_m. The canonical correlation, ρ_m, can be expressed as:

$$\rho_m = \sum_{k=1}^{C_1} \sum_{l=1}^{L} P_{kl} u_{km} v_{lm}. \tag{5}$$

The canonical variables $\mathbf{u}_m$ and $\mathbf{v}_m$ are the scores for the rows and columns, respectively, that maximize the correlation, ρ_m, in equation (5). It is important to note that the scores for the row categories are optimal with respect to maximizing the correlation with the specific column variable being studied, and vice versa.

A rescaling of the canonical scores in equation (1) is equivalent to Goodman's (1986) *saturated RC correspondence analysis model.* We will denote these rescaled canonical scores by $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$, where:

$$\sum_{k=1}^{C_1} \tilde{u}_{km} P_{k\bullet} = \sum_{l=1}^{L} \tilde{v}_{lm} P_{\bullet l} = 0 \tag{6}$$

$$\sum_{k=1}^{C_1} \tilde{u}_{km}^2 P_{k\bullet} = \sum_{l=1}^{L} \tilde{v}_{lm}^2 P_{\bullet l} = \rho_m^2 \tag{7}$$

for each set $m = 1,2, \ldots ,t$. When the row and column scores are scaled as in equations (6) and (7), they are referred to as *principal coordinates* (Greenacre 1986, 1984). The relationship between the principal coordinates and the standard coordinates is straightforward:

$$\begin{aligned} \tilde{u}_{km} &= u_{km}\rho_m \\ \tilde{v}_{jm} &= v_{jm}\rho_m. \end{aligned} \tag{8}$$

Therefore equation (1) can be rewritten in terms of principal coordinates as:

$$P_{kl} = P_{k\bullet}P_{\bullet l}\left[1 + \sum_{m=1}^{t} \tilde{u}_{km}\tilde{v}_{lm} \Big/ \rho_m\right]. \tag{9}$$

The advantage of the principal coordinates scaling is that the variance of each set of scores, within each of the t sets, is equal to the principal inertia, ρ_m^2, for that dimension. This scaling is standard output of Greenacre's correspondence analysis program, *SIMCA* (Greenacre 1986).

Correspondence analysis or canonical correlation analysis of the **F** array results in scores that pertain directly to the categories of the row and column variables. When the rows of **F** code the state of the relational response variable and the columns code the state of the relational predictor variable(s), then the row scores, u_{km} for $k = 1,2, \ldots ,C_1$ and $m = 1,2, \ldots ,t$ pertain to the C_1 categories of the relational response variable and the column scores, v_{lm} for $l = 1,2, \ldots ,L$ and $m = 1,2, \ldots ,t$ pertain to the L states of the relational predictor variable(s). The canonical correlations, ρ_m, describe the correlation between the scores for the relational response categories and the scores for the relational predictor categories.

5. AN EXAMPLE

We illustrate correspondence analysis for studying the response categories on an ordinal relational variable using a data set collected by Bernard, Killworth, and Sailer (1979–80). Bernard et al. measured liking, observed interactions, and reported interactions among 58 students in a fraternity. There are three relations in this data set:

- $\mathcal{X}_1$—*liking:* rating by a fraternity member of how well he likes each person in the group, on an 11-point scale, where 11 means most liked, and 1 means least liked; $C_1 = 11$.
- $\mathcal{X}_2$—*observed interactions:* recorded as the number of times each pair of actors was observed interacting over a several week period. The modal number of interactions is 0, the median is 1, and 75 percent of all pairs of actors were observed interacting 2 times or less. We have recoded this variable to $C_2 = 5$ levels:
  - 0
  - 1
  - 2
  - 3 or 4
  - 5 or more.
- $\mathcal{X}_3$—*reported interactions:* rating by each fraternity member of their recalled amount of interaction with each member of the group, measured on a five-point scale, where 5 means most and 1 means least amount of interaction; $C_3 = 5$.

We will focus on the 11-point rating of liking as the relational response variable. We would like to study how people use these 11 categories of liking, across the combined levels of observed and reported interactions. For this analysis, $L = 5 \times 5 = 25$—the cross-classification of observed and reported interactions. The **F** array for this example is the 11 by 25 cross-classification of liking by the combination of observed and reported interactions. This array is given in Table 1.

Results of correspondence analysis of the data in Table 1 (using *SIMCA,* Greenacre 1986) give $\rho_1^2 = 0.4917$, $\rho_2^2 = 0.1915$, and $\rho_3^2 = 0.0652$, and account for 59.90, 23.32, and 7.95 percent of the total inertia, respectively. The complete model has $t = 11 - 1 = 10$ sets of

TABLE 1
Fraternity Liking by Reported and Observed Interactions

	Observed Interaction																								
Reported	0					1					2					3 or 4					5 or more				
Interaction	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
LIKING																									
1	20	2	3	0	0	6	4	0	0	0	1	4	0	0	0	3	2	0	0	0	1	2	1	0	0
2	15	6	1	0	0	10	2	0	0	0	1	2	0	0	0	0	0	0	0	0	0	2	0	0	0
3	8	5	0	0	0	8	1	3	0	0	0	3	0	0	0	0	0	0	0	0	3	1	2	0	0
4	16	7	4	0	0	2	5	1	0	0	1	3	2	0	0	2	1	1	0	0	2	3	0	1	0
5	28	30	9	0	0	8	17	6	0	1	3	4	2	0	0	1	6	2	1	0	1	3	2	0	0
6	88	124	83	9	1	24	44	26	7	2	9	28	15	5	0	4	15	12	2	1	6	12	10	3	2
7	28	105	65	7	1	7	48	36	8	0	1	19	25	5	0	0	7	13	0	2	0	18	24	8	1
8	25	92	92	8	1	8	46	42	20	3	0	17	44	9	0	1	9	12	8	0	0	8	29	13	1
9	18	55	93	36	8	5	22	59	42	8	2	9	30	34	8	0	1	18	24	7	0	6	30	44	12
10	3	27	81	56	9	1	8	41	49	14	0	2	22	31	12	0	4	18	20	9	0	2	29	43	40
11	2	8	17	33	43	1	7	13	29	48	0	2	8	18	27	0	0	4	9	25	0	1	5	37	132

TABLE 2
First Set of Scores from Correspondence Analysis Correlation Model for Fraternity Data

Liking	$\mathbf{u}_1$
1	1.10085
2	1.11731
3	1.01451
4	0.91727
5	0.81895
6	0.64371
7	0.45940
8	0.32805
9	−0.15377
10	−0.50014
11	−1.34569

	Observed Interactions				
	0	1	2	3 to 4	5 or more
Reported					
1	0.88209	0.94399	0.91409	1.14852	1.16942
2	0.52737	0.52306	0.65881	0.68777	0.63951
3	0.16311	0.05563	0.12204	0.05411	0.06372
4	−0.63474	−0.50832	−0.51014	−0.46958	−0.64389
5	−1.40720	−1.30854	−1.32189	−1.22053	−1.49746

scores; however, the last seven sets of scores account for only 8.83 percent of the total inertia. Table 2 presents only the first set of scores (corresponding to ρ_1^2) for the 11 categories of liking, and the 25 categories of the cross-classification of observed and reported interactions. Although these results show negative scores for high levels of liking and high levels of reported interactions (and positive scores for low levels of these variables), equivalent solutions exist in which the signs of all scores are reversed. The first set of scores for the 11 categories of liking is displayed in Figure 1, where it is clear that the 11 levels of liking are, with one reversal, ordered from most to least. However, they are not equally spaced. There are relatively small distinctions among the lowest levels of liking (levels 1, 2, 3, 4 and 5) and there are larger distinctions among the higher levels (9, 10, and 11).

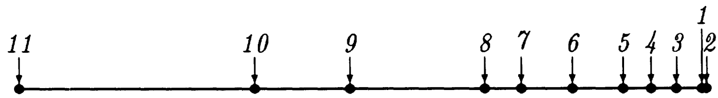

FIGURE 1. Scores for fraternity liking response categories, from correspondence analysis.

The canonical correlation, ρ_m, measures the correlation between the friendship response categories and the combination of observed and reported interactions. For this fraternity example, higher levels of liking are associated with higher levels of reported interaction and with higher levels of observed interaction, but liking appears to be more strongly related to reported interaction than to observed interaction.

In this example we used correspondence analysis to study a two-way table constructed by "stacking" levels of a three-way array. As van der Heijden and de Leeuw (1985) have noted, there are many ways to use correspondence analysis to study three-way problems by analyzing two-way tables derived from the three-way array by different aggregations. As they note, some of these approaches are equivalent, or are equivalent when the derived row/column scores are appropriately rescaled. Other methods are difficult to compare.

Consider the C_1 by $L = C_2 \times C_3$ array, **F**. This array is identical to "stacking" the C_1 by C_2 cross-classification of variables X_1 by X_2 next to each other for each of the C_3 levels of variable X_3. Correspondence analysis of this array gives scores for the C_1 categories of variable X_1, and the $L = C_2 \times C_3$ categories of the cross-classification of variables X_2 and X_3. As van der Heijden and de Leeuw (1985) and van der Heijden and Meijerink (1989) have observed, correspondence analysis of this array can be interpreted as a decomposition of the residuals from the log-linear model of the independence of variable X_1 from X_2 and X_3 jointly: [1][23].

Thus we can interpret correspondence analysis of the relationship between liking response categories and the cross-classification of observed and reported interactions (Table 2) as a decomposition of the residuals from the log-linear model of the independence of liking ratings and the joint effects of observed and reported interactions.

Now let us turn to more parsimonious versions of correlation models.

6. RESTRICTED CANONICAL CORRELATION ANALYSIS

The canonical decomposition described in equation (1) completely describes the data in **F**. It is a saturated model and uses all available degrees of freedom. As noted above, we refer to this as the $C(t)$ model. Goodman (1985) refers to theoretical versions of this model as the *saturated RC(t) canonical correlation model.*

In this section we will consider restricted versions of $C(t)$ that describe **F** in a considerably more parsimonious manner. First, we will describe models that use fewer than t dimensions (or sets of scores). We will then describe models that place restrictions on the scores for the categories of the row and/or column variables. These restricted models have natural and interesting interpretations, and they are most useful for studying the order, spacing, equality, and linearity of the response categories on ordinal relations. Restricted correspondence analysis and restricted correlation models are described in Gilula (1986), Gilula and Haberman (1986, 1988), Böckenholt and Böckenholt (1990), and Takane, Yanai, and Mayekawa (1991). Theoretical versions of these models can be fit using statistical methods, such as maximum likelihood estimation as in Gilula and Haberman's program *CANON,* (Gilula and Haberman, 1986; Gilula 1986). One can then compare the fit of these restricted models with the saturated model, to see whether the more restricted model provides an adequate description of the data.

Standard statistical theory, including the use of maximum likelihood estimation, assumes that observations are independent and identically distributed. Thus, as is standard in many statistical analyses of social network data (Holland and Leinhardt 1975; Fienberg and Wasserman 1981, and so on), we assume that dyads (ordered pairs of actors along with the relational ties between them) are independent. In the following sections we report parameter estimates and goodness-of-fit statistics (X^2 and G^2), but not the p-values for the hypotheses tested by these statistics. Researchers who are uncomfortable with the assumption of dyadic independence can, nevertheless, use correspondence analysis and related approaches to study network data in an exploratory vein (as in Gifi 1990).

First, let us consider models with fewer than the full t sets of scores in the saturated model $C(t)$ presented in equation (1).

6.1. *Fewer Than* t *Dimensions*

The simplest restricted models take equation (1), but they include fewer than the full set of t canonical correlations. So, for $w < t$ we have model C(w):

$$P_{kl} = P_{k\bullet}P_{\bullet l}\left[1 + \sum_{m=1}^{w} \rho_m u_{km} v_{lm}\right]. \tag{10}$$

C(w) has $(C_1 - 1 - w)(L - 1 - w)$ degrees of freedom. Canonical correlations $\rho_{w+1}, \ldots, \rho_t$ are equal to 0 in this model. The same conditions and interpretations of the canonical correlations and canonical scores hold as in the saturated model C(t) (see equations (2) and (3)). The model C(0) implies independence of the rows and columns.

When statistical procedures, such as the maximum likelihood estimation procedure described by Gilula and Haberman (1986, 1988) and Goodman (1987), are used to fit C(w), one gets the usual goodness-of-fit statistics, Pearson's X^2 and the likelihood-ratio test statistic G^2. One can then test whether the data may be modeled by the more parsimonious model, C(w), compared to C(t).

Of more interest for studying response categories on ordinal variables are models that place restrictions on the scores associated with the response categories. We will describe three such models: equality of response categories, uniform spacing (linearity) of categories, and a priori scores for categories.

6.2. *Equality of Response Categories*

The model for equality of response categories examines whether two (or more) categories of the row (or column) variables are equivalent in terms of the conditional distributions within the equivalent rows (or columns). This model stipulates that the canonical scores for equivalent row categories (k and k'), or column categories (l and l'), are equal. Specifically, for $m = 1,2, \ldots, w$, we have for row categories

$$u_{km} = u_{k'm} \qquad k \neq k', \tag{11}$$

and for column categories

$$v_{lm} = v_{l'm} \qquad l \neq l'. \tag{12}$$

This restriction (equating two categories on w sets of scores) has w degrees of freedom associated with it. In terms of the probabilities of observations in **F**, stating that two rows, k and k', have equivalent scores, $u_{km} = u_{k'm}$, stipulates that for a given column, say l, differences between cell probabilities, P_{kl} and $P_{k'l}$, are attributable to differences in marginal row probabilities, $P_{k\bullet}$ and $P_{k'\bullet}$. If row categories k and k' have equivalent scores, then $P_{kl}/P_{k\bullet} = P_{k'l}/P_{k'\bullet}$ for $l = 1,2, \ldots, L$. For example, a model equating two categories of friendship ratings would stipulate that a difference in probabilities of levels of interaction (a predictor relation) across equivalent friendship categories is due to different marginal probabilities of the friendship response categories.

Goodman (1981*b*) and Gilula (1986) have used this model to study the *homogeneity* of rows or columns in a table.

6.3. *Uniform Spacing (Linearity) of Response Categories*

This model states that the interval between adjacent row (or column) categories is constant. For rows

$$u_{mk} - u_{m,k+1} = \mathrm{d}_u \qquad (k = 1,2, \ldots, C_1 - 1), \tag{13}$$

or for columns,

$$v_{ml} - v_{m,l+1} = \mathrm{d}_v \qquad (l = 1,2, \ldots, L - 1), \tag{14}$$

where d_u and d_v are constants. For restrictions on the rows, this model has $(C_1 - 1 - w)(L - 2 - w) + (C_1 - 2)$ degrees of freedom; for restrictions on the columns, it has $(C_1 - 1 - w)(L - 2 - w) + (L - 2)$ degrees of freedom; and for restrictions on both rows and columns, it has $(C_1 - 1 - w)(L - 2 - w) + (C_1 - 2) + (L - 2)$ degrees of freedom.

Uniform spacing of both row and column canonical scores implies that C(w) (equation (10)) may be restated as:

$$P_{kl} = P_{k\bullet}P_{\bullet l}\left[1 + \sum_{m=1}^{w} \rho_m kl\mathrm{d}_u\mathrm{d}_v\right]. \tag{15}$$

Goodman (1987) refers to this as the *U* or *uniform RC correlation model.* One could use this model to test whether response categories on an ordinal variable are equally spaced.

6.4. A priori *Scores for Response Categories*

Instead of equal spacing between row (column) scores, one might have a prior hypothesis about the spacing. In this model a priori scores are proposed for row (and/or column) categories to reflect the relative spacing of these categories. For example, scores $\dot{u}_{km}$ and $\dot{u}_{k'm}$ can be assigned to row categories k and k' respectively so that

$$u_{km} = a + b\dot{u}_{km} \quad \text{and} \quad u_{k'm} = a + b\dot{u}_{k'm}\,(2 \leq k \leq C_1, k \neq k'). \tag{16}$$

There are $C_1 - 2$ degrees of freedom associated with the restrictions on the C_1 row scores. This model implies that

$$\frac{u_{km} - u_{1m}}{u_{k'm} - u_{1m}} = \frac{\dot{u}_{km} - \dot{u}_{1m}}{\dot{u}_{k'm} - \dot{u}_{1m}}. \tag{17}$$

The model of uniform spacing, described above, is a special case of the model of a priori scores in which all intervals between adjacent categories are specified to be equal.

One could use the model of a priori scores to study hypotheses about the intervals among relational response categories. For example, we illustrate this model by evaluating both suggestions by Burt and Guilarte (1986) about the spacing: first, that a relational tie described as "less close" is about 0.7 the strength of the relational tie described as "especially close"; and second, that the middle level of ties between alters is 0.2 of the distance from "strangers" to "especially close."

Restricted versions of correlation models place constraints on the values of the canonical scores in equations (1) and (10). These models are more parsimonious than unrestricted correlation models in that they use fewer degrees of freedom. If one model is a restricted version of another, then the fit of the more restricted model may be compared with the fit of the less restricted model in order to assess whether the restricted model provides an adequate description of the data. Using maximum likelihood estimation techniques, the associated goodness-of-fit statistics from the more and less restricted models can be compared (with the degrees of freedom equal to the difference between the degrees of freedom associated with the two models). We have found Gilula and Haberman's (1986) program *CANON* to be useful for fitting restricted correlation models.

7. AN EXAMPLE

Now let us consider a different example to illustrate restricted versions of correlation models. We will illustrate these models using data collected by Freeman from a computer conference among social science researchers, the Electronic Information Exchange System (EIES) (S. Freeman and L. Freeman 1979; L. Freeman and S. Freeman 1980; Freeman 1986). There are two relations measured on 32 people in this group:

- $\mathcal{X}_1$–*friendship:* a person's reported friendship with each member of the group on a five-point scale:
  - "Unknown"
  - "Person I've heard of"
  - "Person I've met"
  - "Friend"
  - "Close personal friend."
- $\mathcal{X}_2$–messages: the number of messages sent from a person to each other person. The median number of messages is 0, and 75 percent of all ordered pairs of people sent 12 or fewer messages. We have recoded this to three levels:
  - 0
  - 1 to 11
  - 12 or more.

Table 3 presents the cross-classification of Friendship and Message sending for Freeman's EIES data (the **F** array). Since there are $L = 3$ levels of message sending, and $C_1 = 5$ levels of friendship, this

TABLE 3
EIES Messages and Friendship

	Messages		
Friendship	0	1 to 11	12 or more
Unknown	228	68	46
Heard of	96	25	16
Met	168	70	122
Friend	51	24	36
Close personal friend	9	4	29

TABLE 4
Correlation Models for EIES Messages and Friendship

	MODEL	X^2	G^2	df	ρ_1	ρ_2
(1)	C(2) (saturated)	0	0	0	0.322	0.049
	Restrict number of dimensions					
(2a)	C(1)	2.647	2.775	3	0.313	–
(2b)	C(0) (independence)	105.153	102.337	8	–	–
	Restrict canonical scores C(1)					
(3)	Uniform spacing of response categories	20.603	22.040	6	0.274	–
	Equality of response categories C(1)					
(4a)	*Unknown* = *Heard of*	2.991	3.113	4	0.313	–
(4b)	*Met* = *Friend*	2.675	2.805	4	0.313	–
(4c)	*Unknown* = *Heard of* and *Met* = *Friend*	3.020	3.143	5	0.313	–
	A priori scores C(1)					
(5a)	A priori scores (0 0 .7 .7 1)	11.102	11.266	6	0.296	–
(5b)	A priori scores (0 0 .2 .2 1)	17.054	16.812	6	0.301	–

5×3 table is completely explained by $t = 3 - 1$ sets of scores. Thus C(2) is a saturated model for this table.

Table 4 presents the results of several restricted correlation models of the relationship between friendship and message sending from Table 3. We first consider models that include fewer than the full $t = 2$ sets of scores. The independence model, C(0), does not fit these data; therefore, there is some relationship between friendship and message sending in this group. The model with a single set of scores, C(1), does fit these data.

Scores for the friendship response categories, and message sending levels for C(1) (model 2a in Table 4) are presented in Table 5. The scores for the Friendship response categories from this model are displayed in Figure 2.

Now, consider placing restrictions on the set of scores for the relational response categories from C(1). We present the goodness-of-fit statistics for several models in Table 4 (even though a glance at Figure 2 leads us to expect that some of these are unlikely to fit well). First, consider uniform (equal) spacing of the response categories. This model stipulates that the interval between adjacent categories is a constant. This model (model 3) does not fit these data.

TABLE 5
Scores from Correlation Model C(1) of EIES Messages and Friendship

Friendship	$\mathbf{u}_1$
Unknown	−0.8761
Heard of	−1.0200
Met	0.6808
Friend	0.6149
Close personal friend	3.0006
Messages	$\mathbf{v}_1$
0	−0.6934
1 to 11	−0.2100
12 or more	1.6983

Some models for equality of response categories fit quite well. Model (4a) equating *Unknown* with *Heard of,* model (4b) equating *Met* with *Friend,* and model (4c) equating *Unknown* with *Heard of* and *Met* with *Friend* at the same time, all seem to fit these data quite well. Model (4c) suggests that the five response categories of friendship can be summarized in three levels: {*Unknown* or *Heard of*}, {*Met* or *Friend*}, and {*Close personal friend*}. In addition, collapsing the five levels of friendship to three levels suggests that we could combine the corresponding rows in **F**, and represent the frequencies in a smaller, 3 × 3 table (see Gilula 1986). However, model (4c) does not specify the relative spacing of these three levels of friendship.

The spacing of the response categories, shown in Figure 2, suggests that the middle level of friendship (*Met* and *Friend*) is about midway between the lower level (*Unknown* and *Heard of*) and the higher level (*Close personal friend*). We thus fit both of the models of a priori scores suggested by Burt and Guilarte:(0, 0, 0.7, 0.7, 1); and (0, 0, 0.2, 0.2, 1). These models both equate the two lowest categories of friendship (*Unknown* and *Heard of*), and the two middle

FIGURE 2. Scores for EIES friendship response categories, from correlation model C(1).

categories (*Met* and *Friend*). The first model proposes that the middle level of friendship is 0.7 of the distance from the lowest to the highest level; the second model proposes that the middle level of friendship is 0.2 of the distance from the lowest to the highest. Models (5a) and (5b) in Table 4 show that neither set of a priori scores provides a good fit for these data.

The next section describes association models and restricted association models. Following that, we return to the EIES data and illustrate these models.

8. ASSOCIATION MODELS

Models of association study the relationship between rows and columns in a cross-classification using odds-ratios. We let F_{kl} be the expected frequency under the model, and denote the odds-ratio by θ, where:

$$\theta_{kl,k'l'} = \frac{F_{kl}F_{k'l'}}{F_{k'l}F_{kl'}}. \tag{18}$$

This focus on odds-ratios (rather than the correlation ρ) gives rise to models of *association*. There has been considerable research on models of association since the late 1970s (Goodman 1979, 1985, 1986, 1991; Haberman 1981; Clogg 1982*a*, 1982*b*, 1986; Gilula and Haberman 1988; Becker and Clogg 1989; Becker 1990).

As above, we have the cross-classification of observed frequencies, $\mathbf{F} = \{f_{kl}\}$. In addition, for association models, we define:

- $\tau_k^{(R)}$, for $k = 1,2,\ldots,C_1$, a set of row effects,
- $\tau_l^{(C)}$, for $l = 1,2,\ldots,L$, a set of column effects,
- τ, an "overall" effect.

The τ's are main effects for the rows, the columns, and the sample size, respectively, and are of little substantive interest. The model of association for $\mathbf{F}$ is defined as:

$$F_{kl} = \tau\tau_k^{(R)}\tau_l^{(C)}\exp\Big(\sum_{m=1}^{t}\phi_m\mu_{km}\nu_{lm}\Big). \tag{19}$$

One can also express the model in equation (19) in terms of the natural logarithm of F_{kl}. Letting $\lambda_k^{(R)} = \log \tau_k^{(R)}$, $\lambda_l^{(C)} = \log \tau_l^{(C)}$, and $\lambda = \log \tau$, and taking natural logarithms of both sides of equation (19), gives:

$$\log F_{kl} = \lambda + \lambda_k^{(R)} + \lambda_l^{(C)} + \sum_{m=1}^{t} \phi_m \mu_{km} \nu_{lm}. \tag{20}$$

We will refer to either model (19) or model (20) as the A(t) association model.

Association analysis of **F** results in three important sets of information:

- A set of C_1 row scores, $\{\mu_{km}\}$ for $m = 1,2, \ldots, t$,
- A set of L column scores, $\{\nu_{lm}\}$ for $m = 1,2, \ldots, t$,
- A set of t measures of intrinsic association, ϕ_m for $m = 1,2, \ldots, t$, that measure the association between the row scores μ_{km}, and the column scores, ν_{lm}.

The row and column variables $\boldsymbol{\mu}_m$ and $\boldsymbol{\nu}_m$ are scaled as follows:

$$\sum_{k=1}^{C_1} \mu_{km} P_{k\bullet} = \sum_{l=1}^{L} \nu_{lm} P_{\bullet l} = 0 \tag{21}$$

$$\sum_{k=1}^{C_1} \mu_{km}^2 P_{k\bullet} = \sum_{l=1}^{L} \nu_{lm}^2 P_{\bullet l} = 1, \tag{22}$$

for all m. In addition, for distinct m and m', $\boldsymbol{\mu}_m$ and $\boldsymbol{\mu}_{m'}$ are uncorrelated, as are $\boldsymbol{\nu}_m$ and $\boldsymbol{\nu}_{m'}$:

$$\sum_{k=1}^{C_1} \mu_{km} \mu_{km'} P_{k\bullet} = \sum_{l=1}^{L} \nu_{lm} \nu_{lm'} P_{\bullet l} = 0, \qquad \text{for } m \neq m'. \tag{23}$$

Goodman (1991) refers to the theoretical version of the association model with row and column scores scaled as in equations (21) and (22) as the *weighted association model,* where the weights are the marginal row and column proportions ($P_{k\bullet}$ and $P_{\bullet l}$). Association models with other weights are also possible (see Becker and Clogg 1989 and Anderson 1992, for example). The scaling in equations (21) and

(22) is most useful for comparing results of correlation and association models.

The intrinsic association, ϕ_m, can be expressed as:

$$\phi_m = \sum_{k=1}^{C_1} \sum_{l=1}^{L} (\log P_{kl}) P_{k\bullet} P_{\bullet l} \mu_{km} \nu_{lm}. \tag{24}$$

Scores ν_m and μ_m, for rows and columns respectively, maximize the intrinsic association ϕ_m in equation (24).

One can also consider the natural logarithm of the odds-ratio as a function of the intrinsic association, ϕ_m, and the relevant row and column scores. For the two-by-two subtable of rows k and k' and columns l and l', $\log \theta_{kl,k'l'}$ can be expressed as a function of the intrinsic association, ϕ_m, and the differences between the row category scores for rows k and k' and between the column category scores for columns l and l':

$$\log \theta_{kl,k'l'} = \sum_{m=1}^{t} \phi_m (\mu_{km} - \mu_{k'm})(\nu_{lm} - \nu_{l'm}). \tag{25}$$

From equation (25) we see that the intrinsic association, ϕ_m, can be interpreted as the expected log-odds-ratio for the two-by-two subtable comparing rows and columns that are one unit apart (Goodman 1986).

The A(t) model, equations (19) or (20), is a saturated model. As with the correlation models, we can consider more parsimonious versions of A(t) that either include fewer than the full set of t sets of scores, or place restrictions on the row scores, the $\boldsymbol{\mu}_m$'s, and/or the column scores, the $\boldsymbol{\nu}_m$'s. These models are parallel to the restricted correlation models that we described in Section 6. More extensive discussion of restricted association models can be found in Clogg (1982*a*, 1982*b*), Goodman (1981*a*, 1985, 1986, 1991), Gilula (1986), and Gilula and Haberman (1988).

When statistical procedures, such as maximum likelihood estimation described by Gilula and Haberman (1986, 1988) and Goodman (1987), are used to fit restricted association models, one obtains the usual goodness-of-fit statistics, Pearson's X^2, and the likelihood-ratio test statistic G^2. One can then test whether the data may be fit by more parsimonious models.

8.1. *Fewer Than* t *Dimensions*

The simplest restricted association models take equation (19), but include fewer than the full set of t dimensions. For $w < t$, we have model A(w):

$$F_{kl} = \tau\tau_k^{(R)}\tau_l^{(C)}\exp\left(\sum_{m=1}^{w}\phi_m\mu_{km}\nu_{lm}\right). \tag{26}$$

Model A(w) has $(C_1 - 1 - w)(L - 1 - w)$ degrees of freedom. Intrinsic associations $\phi_{w+1}, \ldots, \phi_t$ are equal to 0 in this model. The model A(0) is the model of independence.

Now let us consider models that place restrictions on the row and/or the column scores. We will describe three models that are parallel to restricted correlation models described above: equality of response categories, uniform spacing (linearity) of categories, and a priori scores for categories. It is important to note that the interpretations of these restricted association models are similar to, but not identical to, the restricted correlation models. (We discuss these differences below.)

8.2. *Equality of Response Categories*

The model for equality of response categories examines whether two (or more) categories of the row (or column) variables are equivalent in terms of their odds-ratios. This model stipulates that the scores for equivalent categories are equal. Specifically, for row categories,

$$\mu_{km} = \mu_{k'm} \qquad k \neq k', \tag{27}$$

and for column categories,

$$\nu_{lm} = \nu_{l'm} \qquad l \neq l', \tag{28}$$

for all m. A restriction such as one of the two above, equating two categories across w sets of scores, has w degrees of freedom associated with it. Goodman (1981*b*) has used this model to study the homogeneity of row (or column) categories in a cross-classification.

Returning to equation (25), we can see that stating that two rows, k and k', have equivalent scores, $\mu_{km} = \mu_{k'm}$, stipulates that all odds-ratios involving rows k and k', $\theta_{kl,k'l'}$, are equal to one, for columns $l = 1,2, \ldots, L - 1$.

A model equating two categories, k and k', of a relational response variable stipulates that the odds of response k to response k' is the same for all values of the predictor relational variable(s), l and l'; $F_{kl}/F_{k'l} = F_{kl'}/F_{k'l'}$ for $l = 1,2, \ldots, L - 1$. For example, a model equating two categories of friendship ratings with respect to a relational predictor variable (for example, amount of interaction) would stipulate that the odds of using one friendship response category to the other equivalent friendship response category is the same across all levels of interaction.

8.3. *Uniform Spacing (Linearity) of Response Categories*

This model states that the interval between adjacent row (or column) categories is equal to a constant. For rows,

$$\mu_{mk} - \mu_{m,k+1} = \mathrm{d}_\mu \qquad (k = 1,2, \ldots C_1 - 1), \tag{29}$$

or for columns,

$$\nu_{ml} - \nu_{m,l+1} = \mathrm{d}_\nu \qquad (l = 1,2, \ldots, L - 1) \tag{30}$$

for all m. For restrictions on the rows, this model has $(C_1 - 1 - w)(L - 2 - w) + (C_1 - 2)$ degrees of freedom. For restrictions on the columns, this model has $(C_1 - 1 - w)(L - 2 - w) + (L - 2)$ degrees of freedom. And, for restrictions on both rows and columns, this model has $(C_1 - 1 - w)(L - 2 - w) + (C_1 - 2) + (L - 2)$ degrees of freedom. Uniform spacing of both row and column scores in the association model constrains odds ratios between adjacent rows/columns to be a constant: $\theta_{kl,k+1l+1} = \theta$ for $k = 1, \ldots, C_1 - 1$, and $l = 1, \ldots, L - 1$.

For uniform spacing of *both* row and column scores, A(w) (equation 19) may be restated as:

$$F_{kl} = \tau\tau_k^{(R)}\tau_l^{(C)}\exp\left(\sum_{m=1}^{w} \phi_m kl \mathrm{d}_\mu \mathrm{d}_\nu\right). \tag{31}$$

This model has been referred to as the U, or *uniform RC association model* (Goodman 1986, 1987; Clogg 1982*a*).

8.4. A Priori *Scores for Response Categories*

A priori scores can also be proposed for the row and/or column category scores to reflect the relative spacing of these categories (for example, see Clogg 1982*b*). We will let scores $\dot{\mu}_{km}$ and $\dot{\mu}_{k'm}$ be assigned to row categories k and k', respectively, so that

$$\mu_{km} = a + b\dot{\mu}_{km} \quad \text{and} \quad \mu_{k'm} = a + b\dot{\mu}_{k'm} (2 \leq k \leq C_1, k \neq k'). \tag{32}$$

There are $C_1 - 2$ degrees of freedom associated with the restrictions on the C_1 row scores.

The association model with these a priori scores assigned to row categories k and k' implies that ratios of differences of scores are fixed; that is,

$$\frac{\mu_{km} - \mu_{1m}}{\mu_{k'm} - \mu_{1m}} = \frac{\dot{\mu}_{km} - \dot{\mu}_{1m}}{\dot{\mu}_{k'm} - \dot{\mu}_{1m}} \tag{33}$$

for all k, l, and m. Such constraints affect the interpretation of fitted models.

As with the correlation models, restricted versions of association models place constraints on the values of the row and column category scores in equations (19), (20), or (26). These models are more parsimonious than unrestricted association models in that they use fewer degrees of freedom, so that conditional tests can be made to test for parsimonious, nested models. For fitting these models, we have found Gilula and Haberman's program ASSOC (Gilula and Haberman 1986) and Eliason's program CDAS (Eliason 1990) to be quite useful.

9. EXAMPLE

Let us return to Freeman's EIES example of friendship and message sending (the cross-classification in Table 3) to illustrate association models. Table 6 presents results of restricted versions of association models. From the analysis of these data using the correlation model, we already know that the model of independence does not fit these data (C(0) and A(0) give identical results). The association model with a single dimension, A(1), (model 2a in Table 6) does fit these data. The scores for row categories, the $\boldsymbol{\mu}_m$'s, and the scores for the

TABLE 6
Association Models for EIES Messages and Friendship

	MODEL	X^2	G^2	df	ϕ_1	ϕ_2
(1)	A(2) (saturated)	0	0	0	0.327	0.033
	Restrict number of dimensions					
(2a)	A(1)	0.791	0.804	3	0.331	–
(2b)	A(0) (independence)	105.153	102.337	8	–	–
	Restrict canonical scores A(1)					
(3)	Uniform spacing of response categories	17.865	18.731	6	0.301	–
	Equality of response categories A(1)					
(4a)	Unknown = Heard of	1.231	1.245	4	0.330	–
(4b)	Met = Friend	0.827	0.837	4	0.332	–
(4c)	Unknown = Heard of and Met = Friend	1.265	1.281	5	0.329	–
	A priori scores A(1)					
(5a)	A priori scores (0 0 0.7 0.7 1)	8.197	8.362	6	0.326	–
(5b)	A priori scores (0 0 0.2 0.2 1)	21.972	20.562	6	0.297	–

column categories, the ν_m's, are presented in Table 7. The scores for the friendship response categories from model A(1) are displayed in Figure 3.

Comparing the results of the association models (Table 6) with the results of correlation models (Table 4), we can see that for this

TABLE 7
Scores from Association Model A(1) of EIES Messages and Friendship

Friendship	μ_1
Unknown	−0.8736
Heard of	−1.1311
Met	0.7338
Friend	0.6785
Close personal friend	2.7203
Messages	ν_1
0	−0.7266
1 to 11	−0.0848
12 or more	1.6757

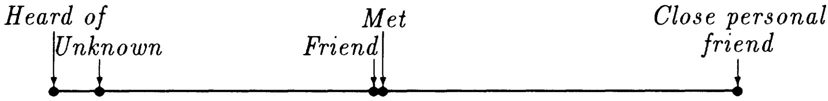

FIGURE 3. Scores for EIES friendship response categories, from association model A(1).

example the results are quite similar. The model with a single dimension fits these data using either the association model, A(1), or the correlation model, C(1). In addition, the model that equates the friendship responses *Unknown* with *Heard of*, and *Met* with *Friend*, fits these data, using either the asociation model or the correlation model. Neither set of a priori scores fits these data, using either the correlation or the association model.

We turn now to some general comparisons of correlation and association models.

10. COMPARISON OF CORRELATION AND ASSOCIATION MODELS

Several recent papers have compared models for association and models for correlation (including correspondence analysis and canonical correlation analysis), and they have commented on situations in which the two models would be expected to give similar results, and situations in which the two models would be expected to give different results (Goodman 1981*a*, 1985, 1986, 1991; Clogg 1986; Gilula and Haberman 1988; Gilula, Krieger, and Ritov 1988; Haberman 1981). We summarize some of these comparisons in this section.

First, consider the parameters in each model. The ρ_m in the correlation model, equation (1), measures the correlation between row variables and column variables in terms of the Pearson product moment correlation coefficient. As Goodman (1991) shows, for a two-by-two table, ρ_1 is equal to the correlation between the variables:

$$\rho_1 = \frac{F_{11}F_{22} - F_{12}F_{21}}{\sqrt{F_{1\bullet}F_{2\bullet}F_{\bullet 1}F_{\bullet 2}}}. \tag{34}$$

In the correlation model, the row scores $\{\mathbf{u}_m\}$ and the column scores $\{\mathbf{v}_m\}$ maximize the correlations ρ_m, defined in equation (5).

On the other hand, the measure of intrinsic association, ϕ, in the association model, equation (19) or (20), is based on the log-odds-ratio for the two-by-two subtables within the table. For a two-by-two table, Goodman (1991) shows that:

$$\phi_1 = \frac{|\log \theta_{11,22}|}{2}. \tag{35}$$

Goodman (1991) gives examples of tables that have the same value of ϕ_1 but different values of ρ_1, and of tables that have the same value of ρ_1 but different values of ϕ_1.

Both the correlation models and the weighted association model are sensitive to marginal distributions; ϕ and ρ are not invariant under multiplicative changes in row/column marginal totals. However, *unweighted* association models—in which $P_{k\bullet}$ and $P_{\bullet l}$ in equations 19 and 20 are replaced by unity—are "margin free" (Goodman 1991; Clogg 1986; Clogg and Rao 1991).

Consider the models of independence, C(0) and A(0). In the association model, $\phi_m = 0$ for $m = 1,2, \ldots ,t$, and similarly in the correlation model, $\rho_m = 0$ for $m = 1,2, \ldots ,t$. Models C(0) and A(0) are identical.

If we now consider models C(w) and A(w), with $1 \leq w < t$, both association models and correlation models study departures from independence (Clogg 1986; Clogg and Rao 1991; Goodman 1985, 1986, 1991). However, the two models represent this departure in different ways. The correlation model focuses on the residuals from the model of independence (van der Heijden and de Leeuw 1985; Goodman 1991 and commentary following). On the other hand, association models represent the "intrinsic" association present in the entire table using odds-ratios (Clogg 1986; Goodman 1991).

Goodman (1991) describes conditions under which the association model is especially appropriate. If the row and column variables are assumed to adhere to a bivariate normal distribution (or one that can be transformed to a joint normal distribution by separately transforming the marginal distributions), and if the "discretizations" of the row and column variables are not too coarse, then the association model gives a better approximation to this distribution than does the correlation model.

Several authors have noted that fitting the association model is somewhat easier than fitting the correlation models, in that the association model cannot give rise to negative fitted values. Certain constraints on the $\mathbf{u}_m$ and $\mathbf{v}_m$, equation (10), are necessary in order to avoid negative fitted values in the correlation models (Goodman 1985).

Finally, in practice, association models often seem to give better fits (lower G^2 and X^2) than do correlation models. For example, our results on the EIES data on friendship and message sending show slightly better fits for the association models for all but model (5b) specifying a priori scores (0, 0, 0.2, 0.2, 1) (compare Tables 4 and 6).

On the other hand, computer programs for correspondence analysis are more readily available than are computer programs for association models. In addition, correspondence analysis (a correlation model) is widely used in exploratory analysis. In this context, correspondence analysis can be used to study many different kinds of data arrays, including data arrays that are not contingency tables of counts or frequencies (such as incidence matrices, response pattern matrices, and so on; see Nishisato 1980, Greenacre 1984, and Weller and Romney 1990, for example). Correspondence analysis results are also more likely to be used for graphical displays of the row and column scores (see Carroll, Green, and Schaeffer 1986; and Greenacre and Hastie 1987, for example). Association models are less widely used for graphical display (but see Goodman 1991; Clogg 1986; and Clogg and Rao 1991).

11. DISCUSSION

In this paper we have described the use of correlation models, including both correspondence analysis and restricted versions of canonical correlation analysis and association and restricted association models for directly studying the order and spacing of response categories of ordinal relational variables. These models were illustrated on two social network data sets. Correspondence analysis of the 11-point rating scale of liking from Bernard et al.'s (1979–80, 1982) study of a fraternity showed that respondents make greater distinctions among the higher levels of liking than among lower levels of liking. In the second example, restricted versions of correlation models and restricted versions of association models using maximum likelihood estimation on ratings of friendship from Freeman and Freeman's

(Freeman 1986; S. Freeman and L. Freeman 1979; L. Freeman and S. Freeman 1980) study of an electronic information exchange network, showed that five levels of friendship could be well represented by just three levels.

APPENDIX

In this paper we have described correspondence and canonical correlation analysis of the matrix **F**, the cross-classification of the relational response variable, and the relational predictor variable(s). However, in previous papers we have argued that the most appropriate array for (multiple) correspondence analysis of social network data is a *response pattern matrix,* also called an *indicator matrix,* that codes the state of each of the dyads in a network data set (Wasserman and Faust 1989; Wasserman, Faust and Galaskiewicz 1990). In addition, many general discussions of multiple correspondence analysis use the indicator matrix (Greenacre 1984; van der Heijden and de Leeuw 1989; van der Heijden and Meijerink 1989). Thus it is important to note that correspondence analysis and canonical correlation analysis of the **F** array are equivalent to analysis of a specific indicator matrix coding the state of each case in a data set.

An indicator matrix, denoted **M**, has cases (here ordered pairs of actors) defining the rows and indicator variables defining the columns. The appropriate indicator matrix for a social network with g actors has $g(g - 1)$ rows, and two sets of columns. The first set of columns is a collection of C_1 indicator variables, coding the state of the relational tie from actor i to actor j on the relational response variable. A single entry of "1" in the appropriate column codes the level of the relational tie from actor i to actor j on the relational response variable (there are "0"'s in the remaining $C_1 - 1$ columns). The second set of columns codes the state of the ordered pair of actors on the relational predictor variable(s). As described above, the state of an ordered pair of actors on these variables is given by the cross-classification of the $R - 1$ relational predictor variables, with $L = C_2 \times C_3 \times \ldots \times C_R$ cells. Thus the second set of columns in the indicator matrix is a collection of L indicator variables. A single "1" in this set of columns codes the state of the ordered pair of actors on the relational predictor variables.

Thus there are $C_1 + L$ columns in **M**. The state of each ordered pair of actors is coded by two entries in its corresponding row of **M**. All row marginal totals of **M** are equal to 2. The column marginal totals indicate the total number of ordered pairs of actors in each state or level of the relational variables.

The matrix **M** consists of two submatrices: $\mathbf{M}_1$, a $g(g - 1) \times C_1$ matrix coding the state of the relational response variable, and $\mathbf{M}_2$, a $g(g - 1) \times L$ matrix coding the state of the relational predictor variables. Schematically, we can represent this matrix as:

$$\mathbf{M} = [\mathbf{M}_1|\mathbf{M}_2].$$

Occasionally correspondence analysis is described for a "Burt" matrix, which we denote by **B**. There are simple relationships among the indicator matrix, **M**, with submatrices $\mathbf{M}_1$ and $\mathbf{M}_2$, the "Burt" matrix, **B**, and the cross-classification, **F**:

- Indicator matrix, $\mathbf{M} = [\mathbf{M}_1|\mathbf{M}_2]$, of size $g(g - 1) \times (C_1 + L)$,
- Cross-classification, $\mathbf{F} = \mathbf{M}_1'\mathbf{M}_2$, of size $C_1 \times L$,
- "Burt matrix", $\mathbf{B} = \mathbf{M}'\mathbf{M}$, of size $(C_1 + L) \times (C_1 + L)$.

In addition, if we denote the row and column marginal totals of **F** as f_{i+}, and f_{+j}, respectively, we can then construct two diagonal matrices: a $C_1 \times C_1$ matrix $\mathbf{C} = \text{diag}(f_{i+})$, with row totals of **F** on the diagonal and zeroes elsewhere, and an $L \times L$ matrix $\mathbf{L} = \text{diag}(f_{+j})$, with column totals of **F** on the diagonal and zeroes elsewhere. The "Burt" matrix has the form:

$$\begin{bmatrix} \mathbf{C} & \mathbf{F} \\ \mathbf{F}' & \mathbf{L} \end{bmatrix}.$$

The indicator matrix, **M**, or some function of it, is one of the most common data arrays for multiple correspondence analysis (van der Heijden and de Leeuw 1989; van der Heijden and Meijerink 1989), and it is useful when one is interested in studying individual cases (here, ordered pairs of actors). However, when one is not interested in individual cases, equivalent scores are obtained from correspondence analysis or canonical correlation analysis of **F**, **M**, or **B** (once scores are appropriately scaled).

Correspondence analysis (or canonical correlation analysis), of **F**, **M**, or **B** results in equivalent sets of scores for the C_1 categories of the relational response variable, and the L states of the relational predictor variable(s), once scores are rescaled within sets. Thus, when one is not interested in scores for the individual cases (here the ordered pairs of actors), analyzing the $C_1 \times L$ array **F** is likely to be more efficient than analyzing the $g(g - 1) \times (C_1 + L)$ array, **M**.

REFERENCES

Anderson, C. J. 1992. "The Analysis of Multivariate Categorical Data by Canonical Correlation and Association Models." Ph.D. diss., University of Illinois.

Becker, M. P. 1990. "Maximum Likelihood Estimation of the RC(M) Association Model." *Applied Statistics* 39:152–67.

Becker, M. P., and C. C. Clogg. 1989. "Analysis of Sets of Two-Way Contingency Tables Using Association Models." *Journal of the American Statistical Association,* 84:142–51.

Bernard, H. R., P. D. Killworth, and L. Sailer. 1979–80. "Informant Accuracy in Social-Network Data IV: A Comparison of Clique-Level Structure in Behavioral and Cognitive Network Data." *Social Networks* 2:191–218.

———. 1982. "Informant Accuracy in Social-Network Data V: An Experimental Attempt to Predict Actual Communication from Recall Data." *Social Science Research* 11:30–66.

Böckenholt, U., and I. Böckenholt. 1990. "Canonical Analysis of Contingency Tables with Linear Constraints." *Psychometrika* 55:633–39.

Burt, R. S., and M. G. Guilarte. 1986. "A Note on Scaling the General Social Survey Network Item Response Categories." *Social Networks* 8:387–96.

Carroll, D. C., P. E. Green, and C. M. Schaffer. 1986. "Interpoint Distance Comparisons in Corresponding Analysis." *Journal of Marketing Research* 24:271–80.

Clogg, C. C. 1982*a*. "Some Models for the Analysis of Association in Multiway Cross-Classifications Having Ordered Categories." *Journal of the American Statistical Association* 77:803–15.

———. 1982*b*. "Using Association Models in Sociological Research: Some Examples." *American Journal of Sociology* 88:114–34.

———. 1986. "Statistical Modelling Versus Singular Value Decomposition." *International Statistical Review* 54:284–88.

Clogg, C. C., and C. R. Rao. 1991. "Comment." *Journal of the American Statistical Association* 86:1118–20.

Eliason, S. R. 1990. *Categorical Data Analysis System: User's Manual.* University of Iowa, Department of Sociology.

Fienberg, S. E., and S. Wasserman. 1981. "Categorical Data Analysis of Single Sociometric Relations." In *Sociological Methodology 1981,* edited by S. Leinhardt, 156–92. San Francisco: Jossey-Bass.

Freeman, L. C. 1986. "The Impact of Computer-Based Communication on the Social Structure of an Emerging Scientific Specialty." *Social Networks* 6:201–22.

———. 1992. "The Sociological Concept of "Group": An Empirical Test of Two Models." *American Journal of Sociology* 98:55–79.

Freeman, L. C., and S. C. Freeman. 1980. "A Semi-Visible College: Structural Effects of Seven Months of *EIES* Participation by a Social Networks Community. In *Electronic Communication: Technology and Impacts,* edited by M. M. Henderson and M. J. MacNaughton, 77–85, AAAS Symposium 53. Washington, D.C.: American Association for the Advancement of Science.

Freeman, S. C., and L. C. Freeman. 1979. "The Networkers Network: A Study of the Impact of a New Communications Medium on Sociometric Structure." Social Science Research Reports No. 46. Irvine, Calif.: University of California.

Friedkin, N. E. 1990. "A Guttman Scale for the Strength of an Interpersonal Tie." *Social Networks* 12:239–52.

Gifi, A. 1990. *Nonlinear Multivariate Analysis.* New York: Wiley.

Gilula, Z. 1986. "Grouping and Association in Contingency Tables: An Exploratory Canonical Correlation Approach. *Journal of the American Statistical Association* 81:773–79.

Gilula, Z., and S. J. Haberman. 1986. "Canonical Analysis of Contingency Tables by Maximum Likelihood." *Journal of the American Statistical Association* 81:780–88.

———. 1988. "Analysis of Multivariate Contingency Tables by Restricted Canonical and Restricted Association Models." *Journal of the American Statistical Association* 83:760–71.

Gilula, Z., A. M. Krieger, and Y. Ritov. 1988. "Ordinal Association in Contingency Tables: Some Interpretative Aspects." *Journal of the American Statistical Association* 83:540–45.

Goodman, L. A. 1979 "Simple Models for the Analysis of Association in Cross-classifications Having Ordered Categories." *Journal of the American Statistical Association* 74:537–52.

———. 1981*a*. "Association Models and Canonical Correlation in the Analysis of Cross-Classifications Having Ordered Categories." *Journal of the American Statistical Association* 76:320–34.

———. 1981*b*. "Criteria for Determining Whether Certain Categories in a Cross-Classification Table Should Be Combined, with Special Reference to Occupational Categories in an Occupational Mobility Table." *American Journal of Sociology* 87:612–50.

———. 1985. "The Analysis of Cross-Classified Data Having Ordered and/or Unordered Categories: Association Models, Correlation Models, and Asymmetry Models for Contingency Tables with or without Missing Entries." *Annals of Statistics* 13:10–69.

———. 1986. "Some Useful Extensions of the Usual Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables." *International Statistical Review* 54:243–309.

———. 1987. "New Methods for Analyzing the Intrinsic Character of Qualitative Variables Using Cross-Classified Data." *American Journal of Sociology* 529–83.

———. 1991. Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data." *Journal of the American Statistical Association* 86: 1085–138.

Granovetter, M. S. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78:1360–80.

Greenacre, M. J. 1984. *Theory and Application of Correspondence Analysis.* New York: Academic Press.

———. 1986. "*SIMCA:* A Program to Perform Simple Correspondence Analysis." *Psychometrika* 51:172–73.

Greenacre, M. J., and T. Hastie. 1987. "The Geometric Interpretation of Correspondence Analysis." *Journal of the American Statistical Association* 82:437–47.

Haberman, S. J. 1981. "Tests for Independence in Two-Way Contingency Tables Based on Canonical Correlation and on Linear-by-Linear Interaction." *The Annals of Statistics* 9:1178–86.

Holland, P. W., and S. Leinhardt. 1975. "Local Structure in Social Networks." In *Sociological Methodology 1976,* edited by D. R. Heise, 1–45. San Francisco: Jossey-Bass.

Marsden, P. V., and K. E. Campbell. 1984. "Measuring Tie Strength." *Social Forces* 63:482–501.

Nishisato, S. 1980. *Analysis of Categorical Data: Dual Scaling and Its Applications.* Toronto: University of Toronto Press.

Takane, Y., H. Yanai, and S. Mayekawa. 1991. "Relationships Among Several Methods of Linearly Constrained Correspondence Analysis." *Psychometrika* 56:667–84.

van der Heijden, P. G. M., and J. de Leeuw. 1985. "Correspondence Analysis Used Complementary to Loglinear Analysis." *Psychometrika* 50:429–47.

———. 1989. "Correspondence Analysis, with Special Attention to the Analysis of Panel Data and Event History Data." *Sociological Methodology, 1989,* 43–87. Cambridge, Mass.: Basil Blackwell.

van der Heijden, P. G. M., and F. Meijerink. 1989. "Generalized Correspondence Analysis of Multi-way Contingency Tables and Multi-way (Super-) Indicator Matrices." In *Multiway Data Analysis,* edited by R. Coppi and S. Bolasco, 185–202. Amsterdam: North-Holland.

Wasserman, S., and K. Faust. 1989. "Canonical Analysis of the Composition and Structure of Social Networks." In *Sociological Methodology, 1989,* edited by C. C. Clogg, 1–42. Cambridge, Mass.: Basil Blackwell.

Wasserman, S., K. Faust, and J. Galaskiewicz. 1990. "Correspondence and Canonical Analysis of Relational Data." *Journal of Mathematical Sociology* 15:11–64.

Weller, S. C., and A. K. Romney. 1990. *Metric Scaling: Correspondence Analysis.* Newbury Park, Calif.: Sage.

Winship, C. 1977. "A Distance Model for Sociometric Structure." *Journal of Mathematical Sociology* 5:21–39.