1990

Journal of Mathematical Sociology, 1989, Vol. 1(1), pp. 11–64
Reprints available directly from the publisher
Photocopying permitted by license only

CORRESPONDENCE AND CANONICAL ANALYSIS OF RELATIONAL DATA

STANLEY WASSERMAN

Department of Psychology and Department of Statistics
University of Illinois

KATHERINE FAUST

Department of Sociology
University of South Carolina

JOSEPH GALASKIEWICZ

Department of Sociology
University of Minnesota

May, 1988

Correspondence analysis, a data analytic technique used to study two-way cross-classifications, is applied to social relational data. Such data are frequently termed "sociometric" or "network" data. The method allows one to model forms of relational data and types of empirical relationships not easily analyzed using either standard social network methods or common scaling or clustering techniques. In particular, correspondence analysis allows one to model:

—two-mode networks (rows and columns of a sociomatrix refer to different objects)

—valued relations (e.g. counts, ratings, or frequencies).

In general, the technique provides scale values for row and column units, visual presentation of relationships among rows and columns, and criteria for assessing "dimensionality" or graphical complexity of the data and goodness-of-fit to particular models. Correspondence analysis has recently been the subject of research by Goodman, Haberman, and Gilula, who have termed their approach to the problem "canonical analysis" to reflect its similarity to canonical correlation analysis of continuous multivariate data. This generalization links the technique to more standard categorical data analysis models, and provides a much-needed statistical justification.

We review both correspondence and canonical analysis, and present these ideas by analyzing relational data on the 1980 monetary donations from corporations to nonprofit organizations in the Minneapolis-St. Paul metropolitan area. We also show how these techniques are related to dyadic independence models, first introduced by Holland, Leinhardt, Fienberg, and Wasserman in the early 1980's. The highlight of this paper is the relationship between correspondence and canonical analysis, and these dyadic independence models, which are designed specifically for relational data. The paper concludes with a discussion of this relationship, and some data analyses that illustrate the fact that correspondence analysis models can be used as approximate dyadic independence models.

The focus of this paper is on *correspondence analysis*, a data analytic technique used to study two-way cross-classifications. We are especially interested in the application of correspondence analysis to *relational data*. Correspondence analysis allows one to model forms of relational data and types of empirical relationships not easily analyzed using either standard social network methods, or common scaling or clustering techniques. In particular, correspondence analysis allows one to model both two-mode networks (rows and columns of a sociomatrix refer to different objects) and valued relations (e.g. counts, ratings, or frequencies).

In general, the technique provides scale values for row and column units, visual presentation of relationships among rows and columns, and criteria for assessing the complexity of the data and goodness-of-fit to particular models. We will describe these uses of correspondence analysis in later sections of this paper. Correspondence analysis has recently been the subject of research by Goodman, Haberman, Gilula and others, who have termed a generalization of the technique "canonical analysis" to reflect its similarity to canonical correlation analysis of continuous multivariate data. This generalization links the technique to more standard categorical data analysis models, and provides a much-needed statistical justification.

As mentioned above, we are primarily interested in the analysis of relational data. Relational data, often called *social network data*, *social interaction data*, or, in some special instances, *sociometric data*, consist of a set of social entities or actors and information about the linkages among them. The observations which give rise to the data are typically on the relationships among the social entities. The entities in the network may be individuals (for example, people in a social group) or collective bodies or aggregates (for example, corporations, departments within a corporation, or political units). Such data are commonly used as a flexible and powerful tool for the measurement and modelling of the structure of social relations.

The flexibility of a relational approach to the study of social structure is illustrated by the range of interactional contents which may be represented using relational data. The relations among units may be *affective* (liking, blaming, esteem, respect, admiration), *interactional* (communication, sharing social activities, exchange of information), *instrumental* (giving aid or advice), *economic* (donating or loaning money or other resources, transacting business), *political* (forming a coalition or an alliance among nations, speaking in favor of some person or policy), or of several other types.

We will review both correspondence and canonical analysis, showing how these methods can be applied to two-mode relational data, and present these ideas by analyzing data on the 1980 monetary donations from corporations to nonprofit organizations in the Minneapolis-St. Paul metropolitan area. We also show how these techniques are related to dyadic independence models, first introduced by Holland, Leinhardt, Fienberg, and Wasserman in the early 1980's.

1. INTRODUCTION TO THE PROBLEM, CONCEPTS, AND NOTATION

Among the many goals of social network research are the modelling of the structure of social interactions and the testing of hypotheses about the nature of this structure. Specifically, models for social structure include a) those that reveal subsets of actors in the network who are closely related to one another, b) those that locate sets of actors (and partners, in two-mode networks) who are similar in their interactions and relationships with others, and c) those that simplify the information in relational data by grouping equivalent actors. These goals are certainly not distinct, and many methods will accomplish two or even all three simultaneously. There is much interest in methods that describe the pattern or intensity of relations among groups and the overall complexity or "dimensionality" of group structure. If one has more than a single measured relational variable, then of prime importance are questions concerning how associated these variables are. Here, one would like to simplify the information in the data by grouping together equivalent relations. The researcher might also like to know whether one of the relations can be "predicted" from the others.

Statistical significance tests have been developed for a variety of substantive hypotheses, including how strong or weak are the interactions between groups, and how associated are various structural properties (such as prestige, prominance, and centrality) with actor characteristics. Many of these tests are new, having been developed during the past decade.

Techniques for approaching and answering these questions have been reviewed in many articles, chapters, and books. Notable references include Holland and Leinhardt (1977, 1979), Marsden and Lin (1982), Burt and Minor (1983), Freeman, White, and Romney (1989), and especially the comprehensive review papers and monographs by Burt (1980), Knoke and Kuklinski (1982), Frank (1981), and Berkowitz (1982).

Despite all of these research efforts, methods presently in use are typically limited since the majority are not statistically based (and hence, do not allow for significance tests). Others are not generalizable to network data sets containing actor attribute variables or multiple relations. Typical methods focus on particular structural properties, such as density, centrality, degree, structural equivalence, and so forth, and do not allow the investigator to get a good "picture" of global network structure. Non-standard data structures, such as two-mode networks, and egocentric networks (which are very popular in community psychology studies of social support), simply cannot be analyzed with standard methods. Recent statistical methods for the global analysis of networks introduced by Holland and Leinhardt (1981), Fienberg and Wasserman (1981), Fienberg, Meyer, and Wasserman (1985), Wasserman and Iacobucci (1986), Wang and Wong (1987) and others, are potentially useful but can be computationally difficult. There is still a need for simple, generalizable methods that are capable of simultaneously answering many substantative research questions.

The purpose of this paper is to present a technique for both the exploratory and statistical analysis of categorical data, and to show how it can be applied to relational data. The technique, correspondence or canonical analysis, is known by many names, including optimal scaling, dual scaling, and reciprocal averaging. Recent statistical research, which we review here, has linked this technique to more standard log-linear modelling of discrete categorical data. We will demonstrate how correspondence analysis (CA) is related to the statistical methods for relational data of Holland, Leinhardt, Fienberg, Wasserman, Meyer, Iacobucci, and others, and we will demonstrate the technique using an extensive network data set gathered by Galaskiewicz. First, we will review relational data and give some necessary notation and related concepts.

1.1. Relational Data

In a formal representation of relational data, social relations are presented as a set of ties linking pairs of social units. The variables linking the units are usually termed *relations*. The information coded for each pair may be simply the presence or absence of the type of tie (a *binary* relation), or it may have a value representing the strength or intensity of the tie. For example, a binary relation may be defined as whether a contribution was made from a corporation to a nonprofit agency, or simply whether there was an association between actors. More generally, the same relation may be measured as the dollar amount of the contribution. Relations may also be discrete-valued, and take a value from a set of ordered categories representing the degree or level of the relation. Relations that are not binary, will be termed *valued*.

The important features which distinguish a relational situation from the more common (in sociology) actor by attribute situation are two in number: the researcher has a set (or sets) of social entities, and information about the linkages among the entities. Typical sociological data sets lack information on the linkages. As mentioned above, linkages exist between pairs of entities; thus the pair of entities and the potential links between them define a dyad. The dyad is the focus of the analysis of relational data. An important assumption of the approaches we discuss here is that the observational (sampling) unit for these analyses is the *dyad*. The collection of dyads, typically consisting of all pairs of social units, taken with the relational information and possibly with attribute information on the actors and partners, make up a collection of data which we will refer to as a *social relational system*.

We define r_{ij} as the strength of the tie from actor i to partner j, which may take on C values ($r_{ij} = 0,1,2,\ldots,C-1$). Dyads defined in this way will be referred to as a *one-mode network* (see Tucker, 1964) since the same set of units serve as both actors and partners. We use the term *two-mode* (or *rectangular*) network when actors and partners are from different sets, or are different types of entities. In some cases, social network researchers may wish to view actors as initiators of choices or relations as distinctly different from actors as recipients of choices. In this case, initiators and recipients could be viewed as two distinct modes, in the Tucker tradition.

Note that to fully and unambiguously describe the information recorded for a dyad in this minimal social relational system we need: the identity (ID) of the actor (i), the identity (ID) of the partner (j), the strength of the tie from i to j (r_{ij}), and the strength of the tie from j to i (r_{ji}). Each dyad can then be represented as a quadruplet:

$$(i,j,r_{ij},r_{ji}) \quad i=1,2,\ldots,g-1; \quad j=i+1,i+2,\ldots,g; \quad r_{ij}=0,1,\ldots,C-1; \quad r_{ji}=0,1,\ldots,C-1. \tag{1}$$

For a one-mode relational system where we observe relations among all actors and partners, there are $g(g-1)/2$ dyads, since relational variables are undefined if the actor and the partner are the same ($i=j$). For two-mode networks, we define two sets of social entities: **G**, which contains actors (initiators of relations) ($\mathbf{G} = \{1,2,\ldots,g\}$), and **H**, which contains partners (recipients of relations) ($\mathbf{H} = \{1,2,\ldots,h\}$). We let $R = \{r_{ij};\ i \in \mathbf{G};\ j \in \mathbf{H}\}$. Since we have distinct actor and partner sets, only three (rather than four) pieces of information are necessary to unambiguously describe the dyad: (1) the actor ID (i) and (2) the partner ID (j), as before; and (3), the relation r_{ij}, since actors (i) may only be initiators of relations and partners (j) may only be recipients. Thus, the minimal dyadic information for a two-mode network now consists of the triple (i,j,r_{ij}). If we have linkages defined for all actors and partners, the social relational system will consist of a collection of gh dyads.

When a single relation is measured for the dyads in a one-mode network, the data may be presented in a $g \times g$ *sociomatrix*, **Z**. Such data representations are familiar to social network researchers. Rows and columns of the sociomatrix index individual actors, arranged in identical order. Self-self ties (the diagonal entries of **Z**) are usually undefined. The value (intensity or strength) of the tie from actor i to actor j is represented by the (i,j)th element of **Z**:

$$Z_{ij} = \text{the value of the tie from actor } i \text{ to actor } j. \tag{2}$$

In general, Z_{ij} has discrete values, so we let this random variable have C possible realizations:

$$z_{ij} = 0,1,2,\ldots,C-1 \quad \text{(the maximum value)}. \tag{3}$$

One can think of the elements of **Z** simply as the coded values of the relation **R**. An important generalization of our social relational system is the allowance for rectangular, or two-mode networks in which the set of sending actors (the first dimension or rows of **Z**) differs from the set of receiving partners (the second dimension or columns of **Z**). In this instance, **Z** is of size $g \times h$, where g = # of actors in the sending set and h = # of partners in the receiving set. For example, the actors may be corporations in a major metropolitan area and the partners, the nonprofit organizations that are supported through charitable contributions from the corporations. We note that while the sociomatrix is a useful format for presenting social relational data, it is not the best format for correspondence analysis.

Regardless of whether the network is one-mode or two-mode, the information in the minimal social relational system can be presented in other ways. One way is via a matrix usually termed a *response pattern matrix*. This array is an indicator matrix with as many rows as there are dyads (either $g(g-1)/2$ or gh), and several sets of columns. There is one set of columns for each of the variables, or pieces of information necessary to code the minimal dyadic relation (either four for a one-mode network or three for a two-mode network). This array will be discussed in detail later.

Focusing on the dyad as the observational unit and representing relational data in this way will be useful to us in our discussion of how correspondence analysis and multiple correspondence analysis (the generalization of correspondence to more than two categorical variables) may be applied to relational data (see section 4). Our presentation will employ response pattern matrices, which is now a standard approach.

In addition to information about relations among individuals, we may also have information about the actors and the partners themselves. We can accomodate such attribute variables into our relational system by defining Q attribute variables as follows:

$$a_{iq} = \text{the value for individual } i \text{ on characteristic } q. \tag{4}$$

If the network is two-mode, we can generalize the definition (4) to handle both actor attributes (Q_1 in number) and partner attributes (Q_2).

2. AN EXAMPLE: CORPORATE-NONPROFIT DONATIONS

The data we will be analyzing come from two studies conducted by Galaskiewicz of a corporate grants economy. The first, conducted in 1981, looked at corporate giving in 1980 and 1981 (Galaskiewicz, 1985) and the second, conducted in 1985, examined the response of nonprofit organizations and corporate funders to cutbacks in government funding during the early years of the Reagan administration (Galaskiewicz and Bielefeld, 1986). Our focus in this paper will be primarily on the relational data on donations made by corporations to nonprofit agencies in 1980 in the Minneapolis/St. Paul metropolitan area. In addition, we will consider characteristics of both the corporations and the nonprofit agencies. A detailed explanation of the sampling procedures used by Galaskiewicz is presented in Galaskiewicz (1985; see also Galaskiewicz, 1987). Galaskiewicz and Wasserman (1986, 1989) studied this network in its entirety. We will limit our attention to 67 nonprofit organizations which received at least one 1980 donation from one of the corporations in the sample. On the corporate side, we limit our attention to 75 firms with over 200 employees, which remained in business and were headquartered in Minneapolis/St. Paul area through 1984.

2.1. Corporate-Nonprofit Donations: Three-way Relational Structure

Our primary focus in this paper will be on the donative relationship between corporations and nonprofit agencies. The data on this relationship come from interviews conducted with the nonprofit administrator of each nonprofit agency. Each administrator was asked to look at a list of all publicly-held firms in the Twin Cities area and indicate the amount of corporate contributions received from each. Response categories, rather than exact dollar amounts, were gathered (i.e., the original data obtained were categorical, not continuous). The categories are: (1) No Donation; (2) More than zero, but less than $1,000; (3) $1,000–$2,999; (4) $3,000–$6,999; (5) $7,000–$14,999; (6) $15,000–$30,999; (7) $31,000–$62,999; (8) $63,000–$126,999; and (9) $127,000 and over.

Since transactions are from corporations to nonprofit agencies, there are two distinct kinds of social entities, and we have a two-mode (or rectangular) network. Thus, following equation (1), for each of the $75 \times 67 = 5{,}025$ dyads, we have a triple of information containing the identity of the corporation (1–75), the identity of the nonprofit (1–67), and the value of the donative relation (1–9). This network may be presented in a sociomatrix with 75 rows for the corporations, 67 columns for the nonprofit agencies, and a value in each cell indicating the amount of the contribution from the corporation (indexed by the row) to the nonprofit agency (indexed by the column). However, we should keep in mind that this is actually a three-way structure, in which the level of the contribution is the third variable. Table 1 shows how relational data on levels of contribution can be presented as a sociomatrix for a subset of the full network consisting of ten corporations and twenty nonprofits. We will use this subset to illustrate different uses of correspondence analysis throughout the paper. We will return to the full set of data in the final section when we demonstrate statistical applications of CA.

TABLE 1
Sociomatrix of donations from corporations to nonprofit agencies

	Nonprofit Agency																			
Corporation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	3	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	1	2
2	1	1	1	1	1	4	1	2	1	1	1	1	1	1	1	1	1	1	1	4
3	5	1	5	6	1	4	1	1	6	1	9	1	5	1	1	1	3	6	3	9
4	1	3	1	1	1	1	1	1	1	1	2	2	1	1	1	3	1	1	1	2
5	5	1	5	6	1	4	1	3	5	1	9	1	4	1	5	1	3	6	4	9
6	1	4	1	1	1	1	1	1	1	6	8	3	1	1	1	3	1	1	1	7
7	3	2	3	2	2	2	2	2	2	1	6	1	1	2	2	1	1	1	2	4
8	1	6	1	1	1	2	2	1	4	5	9	1	3	1	1	1	1	1	1	6
9	1	1	5	3	1	3	3	2	4	1	9	1	1	1	1	1	1	1	1	8
10	1	1	1	3	4	1	4	1	1	1	9	1	1	3	1	1	1	1	1	9

2.2. Attribute Variables

In order to understand the pattern of donations from corporations to nonprofit agencies we will want to consider characteristics of both. There are Q_2 = two characteristics of nonprofit agencies that we examined. First, the nonprofit agencies are involved in one of ten different types of activity: recreation, legal services, health and welfare, media, housing and urban development, education, environmental, cultural, civic and other. Second, nonprofit agencies differ in whether they are perceived as providing essential and outstanding services. During interviews with the 26 people who had an influence on corporate donations, each was given a list of the 326 nonprofit agencies in the (full) sample and asked to indicate which agencies they recognized, thought essential, or regarded as outstanding. Each agency was then given a score of 1, 2, or 3 (recognize only; recognize, and essential or outstanding; or recognize, and essential and outstanding) for each of the 26 individuals. These scores were summed to give an overall measure for each nonprofit.

For the corporations, we considered Q_1 = three attributes. First, corporations are involved in different industries which are coded into seven general categories based on standard industry codes (SICs): crude petroleum and natural gas producers (SIC 8), manufacturing (13–64), transportation and communication (65–67), utilities (68), wholesale and retail trade (69), financial or real estate (70–71), and other services

(72–77). Second, corporations differ in economic activity, which we measure as pretax income in 1980 (3 categories—low, medium, and high). Third, we can consider the extent to which corporations are linked to prestigious members of the local business elite. This is measured as a combination of the degree to which prestigious individuals know officers or board members in a corporation, and the degree to which chief executive officers belong to the same prestigious clubs and cultural boards as the business elite (also 3 categories). Details on the construction of this measure are presented in Galaskiewicz and Wasserman (1989).

2.3. Two-way Aggregations

Consider the social relational system on donations from corporations to nonprofit agencies (that is, a single relation without any information on corporation or nonprofit characteristics). As we have mentioned, this is a three-way structure consisting of the identity of the corporation, the identity of the nonprofit agency, and the level of the 1980 contribution from the specific corporation to the specific nonprofit. This three-way structure can be viewed in a number of different ways, as is common with higher-dimensional contingency tables. Here, we would like to focus on the three two-way relationships contained in this table found by looking at just two of the three variables. These three relationships are: a) corporation ID by level of contribution made; b) nonprofit agency ID by level of contribution received; and c) corporation ID by nonprofit agency ID. Each two-way aggregation is a two-way marginal table from the three-way structure. In general, one can form three two-way margins by summing over the third variable in the system. The tables will be of size $g \times C$, $h \times C$, and $g \times h$.

Due to the design of the relational system, table (c) is a $g \times h$ table and (for our example) gives the cross-classification of corporation ID by nonprofit ID. The table has a one in every cell of the table, indicating that we have coded some level of donation between each corporation and each nonprofit agency. The other two-way tables are much more interesting, and are given in Tables 2a and 2b for the data presented in Table 1. Remember that these data are a subset from the full corporate-nonprofit donation data set and contain $g = 10$ corporations and $h = 20$ nonprofits. These tables, giving information on how the corporations donate their money and on how the nonprofits receive their money, ignore either where the money goes (Table 2b) or from whom the money comes (Table 2a). Nevertheless, these two-way tables can be studied with correspondence analysis and much can be learned about how similar the actors are and how similar the partners are. We will look at these analyses after a discussion of correspondence analysis.

3. INTRODUCTION TO CORRESPONDENCE ANALYSIS–CLASSICAL MODEL

The problems that arise when analyzing relational data and the frequent structural complexity of networks make correspondence analysis an attractive tool. As we will show, it is easy to use and applicable in a variety of ways. While there are many standard social network methods, ranging from graph-theoretic approaches, to clustering algorithms, and finally to techniques designed to test various structural

TABLE 2a
Two-way contingency table of nonprofit agency by level of donations received

	Level of Donation								
Nonprofit	1	2	3	4	5	6	7	8	9
1	7	0	1	0	2	0	0	0	0
2	5	1	2	1	0	1	0	0	0
3	6	0	1	0	3	0	0	0	0
4	5	1	2	0	0	2	0	0	0
5	8	1	0	1	0	0	0	0	0
6	4	2	1	3	0	0	0	0	0
7	6	2	1	1	0	0	0	0	0
8	6	3	1	0	0	0	0	0	0
9	5	1	0	2	1	1	0	0	0
10	7	0	0	1	1	1	0	0	0
11	2	1	0	0	0	1	0	1	5
12	8	1	1	0	0	0	0	0	0
13	7	0	1	1	1	0	0	0	0
14	8	1	1	0	0	0	0	0	0
15	8	1	0	0	1	0	0	0	0
16	8	0	2	0	0	0	0	0	0
17	8	0	2	0	0	0	0	0	0
18	8	0	0	0	0	2	0	0	0
19	7	1	1	1	0	0	0	0	0
20	0	2	0	2	0	1	1	1	3

TABLE 2b
Two-way contingency table of corporations by level of donations made

	Level of Donation								
Corporation	1	2	3	4	5	6	7	8	9
1	17	1	1	1	0	0	0	0	0
2	17	1	0	2	0	0	0	0	0
3	9	0	2	1	3	3	0	0	2
4	15	3	2	0	0	0	0	0	0
5	7	0	2	3	4	2	0	0	2
6	14	0	2	1	0	1	1	1	0
7	6	10	2	1	0	1	0	0	0
8	12	2	1	1	1	2	0	0	1
9	12	1	3	1	1	0	0	1	1
10	14	0	2	2	0	0	0	0	2

hypotheses, most of these are applicable just to one-mode, binary relational data. Some are useful only when the relations are symmetric. There are very few methods which allow one to analyze valued relations or two-mode networks. In addition, standard scaling or clustering procedures usually require a one-mode, symmetric matrix, and therefore are useful for two-mode relational systems only after the data have been "pre-processed". Others have discovered the methodological features offered by correspondence analysis. We are not the first to apply it to relational data.

3.1. An Historical Perspective

The correspondence analysis method was first described by Hirschfeld (who later changed his name to the more well-known H. O. Hartley) (1935). Hirschfield was interested in the simultaneous linear regressions of the rows and columns in a contingency table on derived score values. The technique appears to have been "discovered" by several researchers, particularly Guttman (1941, 1946, 1959) and Fisher (1940), but was largely neglected until the mid-1970's. Fisher was not aware of Hirschfeld's research, and has been regarded as the method's first inventor. Fisher analyzed a (by now) classical example due to Maung (1941) consisting of a two-way table of hair color and eye color in Scottish schoolchildren. Fisher sought to replace the two discrete variables indexing the table with derived variables (each level of each discrete variable receives a "score") so that the correlation between the derived variables is maximized.

Some of the "discoverers" of the technique introduced CA as a method of scaling the variables of the two-way table, rather than simply a method for contingency table analysis. Guttman (1941) and Torgerson (1958) both discuss CA as a scaling technique (see also Nishisato, 1978). Kendall and Stuart, in the second volume of their classic work *The Advanced Theory of Statistics*, discuss the technique from the viewpoint of a *canonical analysis* of a two-way table (Kendall and Stuart, 1973, page 588). The most thorough history of CA can be found in the introductory chapter of Nishisato (1980). We describe the many approaches to the CA solution in the next section of this paper.

Most of the early applications of CA to social network data have been exploratory, primarily using the results of correspondence analysis to graphically display the patterns of social relationships in small groups. The networks have been one-mode and the relations not symmetric. The examples we review here are of two different types. One line of research has applied standard correspondence analysis to relational data in standard ways (see Romney and Boyd, 1986, and Wasserman and Anderson, 1987). Another line of work employs an independently developed method, centroid scaling, which is conceptually and formally similar to correspondence analysis, but which is applicable to a specific form of relational data (see Levine, 1979, and Noma, 1982a, 1982b). We review both of these here.

Noma and Smith (1985) were the first to use CA (or at least a variant of it) to study relational data. They showed how CA row and column scores could be used to permute rows and columns of a sociomatrix to group together structurally similar actors. Such groupings usually make the social structure of the group more apparent. Wasserman and Anderson (1987) suggested that the row and column scores be used to form groups of structurally similar actors, which could then be used in further statistical analysis. In particular, Wasserman and Anderson noted that actors with similar row (column) scores could be grouped to form a blockmodel (White, Boorman, and Breiger, 1976) of the data. Romney and Boyd (1986) used CA row and column scores from the analysis of a one-mode not symmetric sociomatrix to show how one could look at the dual roles of actors both as initiators and as recipients of choices within the same group. In addition they used correspondence analysis to explore the duality of social structure as the simultaneity of actors linked by attendance at social events and events linked by their participants (see Breiger, 1974).

These studies (Noma and Smith, Wasserman and Anderson, and Romney and Boyd) are similar with respect to their usage of correspondence analysis. These

authors applied CA directly to the sociomatrices studied. We have found that there are other ways to apply CA to relational data. One of the purposes of this paper is to clarify when and how CA should be applied, and to describe the variety of relational data that can be analyzed with it.

Levine independently devised a method termed *centroid scaling*, which is quite similar to correspondence analysis. Centroid scaling, discussed by Levine (1979), is a technique to scale "pick any" data. "Pick any" data arise when subjects are asked to select items from a range of possibilities. Subjects may choose different numbers of items, and the alternatives they consider in making their selections may be different. For example, subjects might be asked to name historical figures they admire. Similarly, corporations select people to sit on their boards of directors in a "pick any" manner. This latter example results in relational data. The goal of centroid scaling is to assign scores to subjects and items simultaneously so that a subject's score is the mean of the scores of the items they select, and an item's score is the mean of the scores of the subjects who select it. As we will see below, this goal of centroid scaling is very close to one of the goals of CA, or as it is sometimes referred to, *dual scaling*. The mathematical solution, for which we will refer the reader to Levine (1979), parallels the calculations necessary to find row and column scores in CA.

Centroid scaling has been applied to two quite different substantive problems: a) interlocks among corporate boards of directors (Levine 1979, 1984); and b) citation networks among important scientific articles (Noma 1982a, 1982b). In both cases the data are expressed as relational and are gathered in a "pick-any" manner. Both studies represented the data as a binary two-mode sociomatrix.

These applications demonstrate the desirability and advantages of using correspondence analysis on relational data. Before we discuss how to apply correspondence analysis to relational data, we present the purpose and associated mathematics of CA in more detail.

3.2. Mathematics

We will let **F** denote a two-dimensional, cross-classified contingency table, with dimensions of I rows and J columns. Table 2 gives two examples of such a table. The primary goal of correspondence analysis is to explore the relationship between the rows and columns of a contingency table by assigning scores to the rows and to the columns of **F**. One can then study the relationship between these row and column scores either by examining tables of the scores, or (as is usually done) by graphing the first two dimensions or sets of scores on a two-dimensional graph. These graphical representations can then be relied upon to uncover both the relationship between rows and columns and the "dimensionality" of the table. The dimensionality of the table is equal to the number of sets of scores that must be used to approximately reproduce the frequencies in the table. We note that the term *correspondence analysis* was first used by Benzecri (1969), who designated the method "analyse factorielle des correspondances". Benzecri examined primarily incidence data.

Since 1980, a number of books on CA have been published. We have found Greenacre (1984) to be quite useful because of his geometric perspective, but also recommend Nishisato (1980), Lebart, Morineau, and Warwick (1984), as well as reviews by Hill (1974, 1982), Greenacre (1981), de Leeuw (1973), Tenenhaus and Young (1985), van der Heijden and de Leeuw (1985), and Goodman (1986). CA can be viewed as a *discrete principal components analysis* and can be used for a number of different purposes, including assigning scores to ordinal categorical variables with unknown scale values, grouping together equivalent rows and equivalent columns (i.e., deciding which levels of the row (column) variables are similar with respect to the levels of the column (row) variable), and determining whether the scores assigned to the rows (columns) are linear. Interest in CA has increased lately because of recent discussions of its relation to log-linear models for ordinal categorical variables (see Agresti, 1984; Goodman, 1985).

While we want to emphasize the importance of recent statistical research by Goodman, Gilula, Haberman, and others, our interest in CA is not statistical. Rather, we see CA as an exploratory tool that should prove useful to researchers in a number of different ways. Thus, our focus is on the exploratory, rather than the statistical, side of correspondence analysis.

We will now describe how to correspondence analyze the data matrix $\mathbf{F}$, and show how these mathematics arise from the various criteria which have the CA algorithm as a solution. For historical (as well as pedagogical purposes) we will adopt Fisher's (1940) perspective and consider the problem of simultaneously assigning scores $\mathbf{x} = \{x_i;\ i = 1,2,\ldots,I\}$ and $\mathbf{y} = \{y_j;\ j = 1,2,\ldots,J\}$ to the rows and columns (respectively) of $\mathbf{F} = \{f_{ij}\}$ in such a way that

$$x_i \text{ is proportional to } \sum_j \frac{f_{ij}}{f_{i\cdot}} y_j \quad \text{and}$$
$$y_j \text{ is proportional to } \sum_i \frac{f_{ij}}{f_{\cdot j}} x_i; \tag{5}$$

i.e., the x's are weighted means of the columns scores and the y's are weighted means of the row scores. The $\{f_{i\cdot}\}$ and $\{f_{\cdot j}\}$ are the row sums and column sums of $\mathbf{F}$. The weights are simple functions of the elements of $\mathbf{F}$. One can see why the term *reciprocal averaging*, used by Hill and other ecologists, has been used synonomously with correspondence analysis.

Let us denote the solution of the problem (5) by the triple $(\eta,\mathbf{x},\mathbf{y})$ where

$$\eta x_i = \sum \frac{f_{ij}}{f_{i\cdot}} y_j$$
$$\eta y_j = \sum \frac{f_{ij}}{f_{\cdot j}} x_i \tag{6}$$

The constant η is the proportionality factor from (5). The solution to this problem is straightforward. We define the arrays $\mathbf{R}$ and $\mathbf{C}$ as diagonal arrays containing the row sums and column sums of $\mathbf{F}$: $\mathbf{R} = \text{diag}\{f_{i\cdot}\}$ and $\mathbf{C} = \text{diag}\{f_{\cdot j}\}$. The equations (6) then can be re-written

$$\eta\mathbf{x} = \mathbf{R}^{-1}\mathbf{F}\mathbf{y}$$
$$\eta\mathbf{y} = \mathbf{C}^{-1}\mathbf{F}'\mathbf{x}. \tag{7}$$

Equations (7) can be combined by substituting the first equation into the second (or the second into the first). The latter operation yields the single equation

$$\eta\mathbf{x} = \mathbf{R}^{-1}\mathbf{F}(\eta^{-1}\mathbf{C}^{-1}\mathbf{F}'\mathbf{x}) \tag{8}$$

or, after some rearranging and square rooting,

$$\eta^2(\mathbf{R}^{1/2}\mathbf{x}) = (\mathbf{R}^{-1/2}\mathbf{F}\mathbf{C}^{-1/2})(\mathbf{R}^{-1/2}\mathbf{F}\mathbf{C}^{-1/2})'(\mathbf{R}^{1/2}\mathbf{x}).$$

If we define $\mathbf{x}^*$ as $\mathbf{R}^{1/2}\mathbf{x}$, and let $\mathbf{F}^* = \mathbf{R}^{-1/2}\mathbf{F}\mathbf{C}^{-1/2}$ be a scaled version of $\mathbf{F}$, then we must solve

$$[\mathbf{F}^*\mathbf{F}^{*\prime} - \eta^2\mathbf{I}]\mathbf{x}^* = \mathbf{0}. \tag{9}$$

One can see that $\mathbf{x}^*$ is the eigenvector of the square, symmetric matrix $\mathbf{F}^*\mathbf{F}^{*\prime}$ (which is of size $I \times I$), corresponding to the largest eigenvalue η^2. Actually, the largest eigenvalue is unity, because the equations admit the trivial solutions $\mathbf{x} = \mathbf{1}$ and $\mathbf{y} = \mathbf{1}$. For this reason, the largest eigenvalue ($\eta_1^2 = 1$) and its associated eigenvector are referred to as the "trivial" solution and are ignored. One starts with the second eigenvalue, which will be less than (or equal to) one.

A few comments should help. First, note that the array $\mathbf{F}^*$ has elements $\{f_{ij}/(\sqrt{f_{i\cdot}}\sqrt{f_{\cdot j}})\}$. This implies that the array $\mathbf{F}^*\mathbf{F}^{*\prime}$ can be viewed as containing ratios of observed values to the square roots of the fitted values (multiplied by the total sample size) from a chi-squared test for independence. Secondly, since the eigenvector corresponding to the first eigenvalue is $\mathbf{1}$ (the trivial solution), both $\mathbf{x}^*$ and $\mathbf{y}^*$ must sum to zero to be orthogonal to $\mathbf{1}$. Therefore, the means of the CA scores are zero. Lastly, it is easy to find $\mathbf{x}$ from $\mathbf{x}^*$, since $\mathbf{x} = \mathbf{R}^{-1/2}\mathbf{x}^*$. Further, one can then calculate $\mathbf{y} = (1/\eta)\mathbf{C}^{-1}\mathbf{F}'\mathbf{x}$.

There is another way to view correspondence analysis. Suppose we consider the row variable of $\mathbf{F}$ as a collection of I dummy-coded variables, one for each category or level, and the column variable as a collection of J such variables. Every unit in the data will have a value on each of these dummy-coded variables. Thus, units falling into the (i,j)th cell of $\mathbf{F}$ would have a value of unity on the ith variable in the first set and the jth variable in the second set, and zeros on the other $I + J - 2$ variables. Consider the problem of comparing the first set of variables and the second set. One solution to this problem is the well-known *canonical correlation analysis* of the two sets of variables which seeks to find a linear combination of the first set and a linear combination of the second set with maximal correlation. Guttman (1959) (see also Lancaster, 1966) was the first to prove that the scores $\mathbf{x}$ and $\mathbf{y}$, when viewed as transformations, or simply linear combinations, of the row and column variables, maximize the first eigenvalue of their correlation matrix; i.e., η is the canonical correlation between the two sets of variables. Thus, many authors refer to CA as *canonical analysis*, or *optimal scaling*, a term first used by Guttman. Recent interest in this method for the analysis of two-way contingency tables utilizes this interpretation of the triple $(\eta,\mathbf{x},\mathbf{y})$ (see section 5). We will actually consider the $I + J$ variables defined here when we discuss multiple correspondence analysis, a generalization of CA to higher-dimensional contingency tables.

The mention of the term "scaling" above should give the reader the idea that there is an interesting geometrical interpretation of the CA solution. Guttman

(1941) and Torgerson (1958) introduced CA as a method of scaling rather than of contingency table analysis. Benzecri (1973) utilized such a geometric approach in his development, seeking directions of maximum "inertia" in a multidimensional space with a special metric. We will define *inertia* as simply the *dispersion* of points in a multidimensional space and will measure it by the total distances of the points to the derived space. One metric to choose, since we are working with contingency tables, is the chi-squared metric (see equation 11).

To define this metric, we must consider the notion of a row (column) profile. A profile of a given row is simply the entries for the row, divided by the row total: $\mathbf{a}_i = (f_{i1}/f_{i\cdot}, f_{i2}/f_{i\cdot}, \ldots, f_{iJ}/f_{i\cdot})$ for the ith row. The average row profile for the table is the profile of the column totals, which is simply the weighted average of the row profiles using the row sums as weights (or the column totals divided by the grand total). One can also define column profiles in an identical manner. Each row (column) profile is a point in $J(I)$ dimensional space.

Now consider two row profiles, and label the points associated with the rows i and i', $\mathbf{a}_i$ and $\mathbf{a}_{i'}$. Define $\mathbf{D}_c$ as a diagonal $J \times J$ matrix containing the average row profile: $\mathbf{D}_c = \mathrm{diag}\{f_{\cdot j}/f_{\cdot\cdot}\}$. Now define the distance between these two points as

$$d^2(\mathbf{a}_i, \mathbf{a}_{i'}) = (\mathbf{a}_i - \mathbf{a}_{i'})' \mathbf{D}_c^{-1} (\mathbf{a}_i - \mathbf{a}_{i'})$$
$$= \sum_j [(a_{ij} - a_{i'j})^2 / (f_{\cdot j}/f_{\cdot\cdot})]. \quad (10)$$

One can define a similar metric for the distance between two column profile points. Rewriting (10) in terms of the entries, row sums, and column sums of $\mathbf{F}$ will verify that this metric is indeed a "chi-squared" metric, since the distance weights inversely by "chi-squared expectations" (estimated expected values from a test of independence). Given $d(\cdot,\cdot)$, the goal is to find a p-dimensional subspace which is closest to all the points, where p is as small as possible. By "close", we mean minimizing the squared chi-squared distances from the points to the subspace (see Heiser and Muelman 1983).

If we focus on only one-dimensional subspaces ($p = 1$), then we are seeking the line of "closest fit" to the row profile points. With some algebra, and the definition (10), one can show that this line is a multiple of the eigenvector corresponding to the largest (non-unity) eigenvalue of $\mathbf{F}^*\mathbf{F}^{*\prime}$, which is $\mathbf{x}^*$. This result should not be surprising, since $\mathbf{F}^*\mathbf{F}^{*\prime}$ can be written $\mathbf{R}^{-1/2}\mathbf{F}\mathbf{C}^{-1}\mathbf{F}'\mathbf{R}^{-1/2}$ which is an $I \times I$ matrix containing the chi-squared distances between row profiles. Thus, one sees that this geometric approach to the problem is equivalent to reciprocal averaging and optimal scaling. With larger p's, the consequent eigenvectors give the directions of maximum dispersion of the profiles.

There is a duality to this geometric problem, which arises by considering the column profiles. To be brief, equation (10) can be restated in terms of column profiles weighted by the inverse of a diagonal matrix containing the average row profile. The subspace of closest fit can be found by studying the eigenvectors of the matrix $\mathbf{C}^{-1/2}\mathbf{F}'\mathbf{R}^{-1}\mathbf{F}\mathbf{C}^{-1/2}$ which contains the chi-squared distances between column profiles. It happens that the subspace for the row profiles is identical to the subspace for the column profiles (the two decomposed matrices have identical eigenvalues) and that one can find the fitted line for the row profiles from the fitted line for the column profiles (and vice versa). This duality highlights why CA is occasionally referred to (by Nishisato, 1980, and others) as *dual scaling*. So, a graphical representation of the rows and columns in the same space is justified, and gives CA a uniqueness among standard scaling procedures.

The fact that the geometric interpretation of the correspondence [analysis problem relies on a chi-squared distance metric can be used to view CA as a technique to decompose Pearson's chi-squared statistic for testing independence in the $I \times J$ contingency table. In fact, the total inertia is equal to Pearson's statistic. We consider the problem of minimizing

$$X^2 = \sum\sum \{(f_{ij} - u_i v_j)^2 / (f_{i\cdot} f_{\cdot j} / f_{\cdot\cdot})\} \quad (11)$$

If we write u_i as $f_{i\cdot} x_i$ and v_j as $f_{\cdot j} y_j$, then the first (trivial) solution is simply fitting $\mathbf{x} = \mathbf{1}$ and $\mathbf{y} = \mathbf{1}$. Subsequent solutions, orthogonal to the first, are simply the eigenvectors of $\mathbf{F}^*\mathbf{F}^{*\prime}$. If we denote the solutions as $\{x_{im},\ m = 1,2,\ldots,K\}$ and $\{y_{jm},\ m = 1,2,\ldots,K\}$ (where $K = \min(I-1, J-1)$), then one obtains a standard spectral decomposition of f_{ij} as follows:

$$f_{ij} = \sum_m \eta_m u_{im} v_{jm} = \sum_m \eta_m f_{i\cdot} x_{im} f_{\cdot j} y_{jm}. \quad (12)$$

Here, $\eta_1, \eta_2, \ldots, \eta_K$ are the eigenvalues of the decomposed matrix. The eigenvectors should be scaled to have squared weighted lengths of one:

$$\sum f_{i\cdot} x_{im}^2 = \sum f_{\cdot j} y_{jm}^2 = 1 \qquad (m = 1,2,\ldots,K) \quad (13)$$

These results can be quite useful in understanding how CA relates to more standard statistical analyses of two-way contingency tables, which we discuss in more detail in section 5.

3.3. Some Examples

In this section we illustrate correspondence analysis using the two 2-way contingency tables defined earlier: corporations by levels of donations made, Table 2b, and nonprofit agencies by levels of donations received, Table 2a. We used Greenacre's computer program SIMCA (Greenacre 1986) extensively, and recommend it. The program calculates standard CA solutions. We have relied on it for the correspondence analyses reported here.

We look first at the donations made at different levels by the corporations. The first three eigenvalues (η^2) from a correspondence analysis of this table are .282, .198 and .072, and respectively account for 45%, 31.6% and 11.5% (cumulatively, 45%, 76.6%, and 88.1%) of the total inertia. This implies that we can summarize well the relationship between the corporations (rows) and donation levels (columns) by (at most) three sets of scores. The scores for row categories (individual corporations) and column categories (levels of donation) are presented in Table 3. These scores are displayed graphically in Figure 1. From the figure, one can see how the

TABLE 3
Row and column scores from correspondence analysis of corporations and levels of donation

Corporation	Eigenvectors 1	2	3	4
1	−0.128	−0.989	0.875	−0.439
2	−0.083	−0.983	1.147	−0.766
3	1.116	1.265	−0.108	−0.787
4	−0.714	−0.562	0.424	0.060
5	1.305	1.550	0.134	0.328
6	0.301	−1.142	−2.459	−0.873
7	−2.487	1.366	−0.388	0.135
8	0.039	0.382	0.290	−1.120
9	0.209	−0.326	−0.801	2.209
10	0.442	−0.560	0.886	1.255
Donation Level				
1	0.051	−0.634	0.309	−0.226
2	−2.818	1.303	−0.164	0.085
3	0.051	0.296	−0.935	1.672
4	0.533	0.366	0.540	0.450
5	1.843	2.510	−0.123	0.026
6	0.806	1.968	−0.961	−2.788
7	0.569	−2.566	−9.159	−4.686
8	0.480	−1.650	−6.075	3.586
9	1.407	1.283	0.611	1.795
Eigenvalue	0.282	0.198	0.072	0.035
% of Inertia	45.00	31.58	11.49	5.55

corporation variable interacts with the donation variable. Notice not only how the corporations relate to each other and how the donation variable is clustered along the first dimension, but also how corporations #5 and #3 are located in the northeast corner of the plot along with donation levels #5, #6, #9. This implies that these firms are likely to send money to the nonprofits at high levels. These types of interactions are easy to spot from graphs of CA solutions and can be quite valuable analytic tools.

Correspondence analysis of levels of donations received by different nonprofit agencies give eigenvalues (η^2) of .584, .178 and .138, accounting for 51.4%, 15.7% and 12.1% (cumulatively, 51.4%, 67.1%, and 79.2%) of the inertia, respectively. The scores for row categories (nonprofit agencies) and column categories (levels of donation) are presented in Table 4. They are displayed in Figure 2. This CA is more informative than that for the corporations and donation levels. The first axis of this analysis distinguishes both between the highest levels of donation (7, 8 and 9) and medium and low levels of donation, and between nonprofit agencies #11 and #20 and the rest. The most important feature of this CA is that nonprofits #11 and #20, which are supported at the highest donation levels, are both media organizations (public television and radio stations), an attribute of these actors that we will incorporate into later analyses.

Correspondence analysis of these two-way tables is useful for exploring three different kinds of relationships. These include: 1) the tendency for corporations to

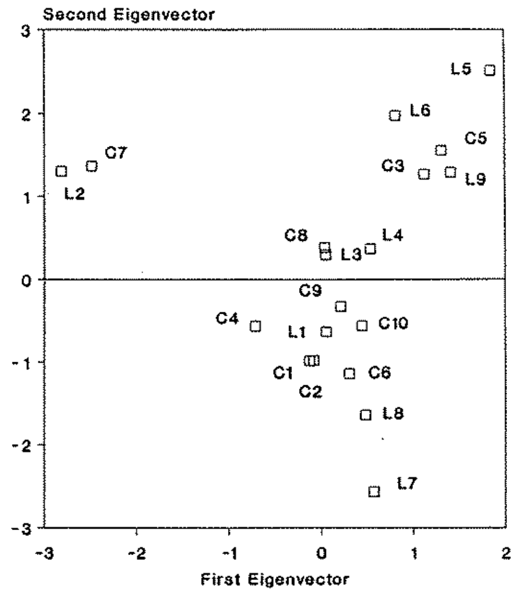

FIGURE 1. Correspondence analysis of corporations (C1,C2,...,C10) and donation levels (L1,L2,..., L9).

make donations at different levels and for nonprofit agencies to receive donations at different levels; 2) displaying similarities among corporations or nonprofit agencies in terms of the levels of donations they make or receive; and 3) displaying similarities among levels of donations with respect to either the corporations which make them or the nonprofit agencies which receive them. However, CA of these two-way tables does not allow us to analyze the relational nature of the data, nor does it allow us to incorporate attributes of either nonprofits or corporations. Both of these problems require multiple correspondence analysis, which we will now discuss. Following this, we will present a discussion of other network methodologies and their relationship to CA.

4. MULTIPLE CORRESPONDENCE ANALYSIS

Up to this point we have focussed on the analysis of two-way data structures. In this section we discuss *multiple correspondence analysis* (or simply MCA) for higher-way tables, a generalization of correspondence analysis. References which discuss the theory and mathematics of multiple correspondence include Greenacre (1984), Nishisato (1980), Tenenhaus and Young (1985), and Deville and Saporta (1983).

Multiple correspondence analysis is often used to explore the relationships among several categorical variables. It is a useful tool for examining the relationship between questionnaire responses and respondent characteristics. For example Hayashi (1980) looked at the difference between Japanese and American responses to ques-

TABLE 4
Row and column scores from correspondence analysis of nonprofit agencies and levels of donation

	Eigenvectors			
Nonprofit	1	2	3	4
1	−0.572	1.614	−0.912	0.326
2	−0.184	−0.756	0.436	−0.816
3	−0.602	2.272	−1.823	0.255
4	−0.145	−0.284	1.262	−2.064
5	−0.304	−0.521	−0.017	0.484
6	−0.016	−2.040	−1.316	0.452
7	−0.212	−1.009	0.027	0.823
8	−0.208	−0.889	0.590	1.278
9	−0.098	−0.419	−1.443	−1.203
10	−0.297	0.491	−0.666	−1.290
11	3.109	1.623	1.391	0.684
12	−0.411	−1.102	0.803	0.736
13	−0.444	0.438	−0.672	0.210
14	−0.411	−0.102	0.803	0.736
15	−0.432	0.654	−0.257	0.600
16	−0.522	0.197	1.057	0.533
17	−0.522	0.197	1.057	0.533
18	−0.228	0.301	1.070	−2.467
19	−0.313	−0.616	0.132	0.552
20	2.810	−1.047	−1.521	0.361
Donation Level				
1	−0.384	0.164	0.282	0.124
2	0.389	−1.498	−0.111	0.962
3	−0.456	−0.242	0.832	0.330
4	0.365	−2.010	−2.206	−0.451
5	−0.613	2.951	−3.096	−0.093
6	0.669	−0.019	0.854	−4.319
7	3.677	−2.484	−4.102	−1.164
8	3.872	0.680	−0.176	0.520
9	3.920	1.472	0.805	0.943
Eigenvalue	0.584	0.178	0.138	0.096
% of Inertia	51.37	15.66	12.09	8.44

tions on traditionalism. Deville and Saporta (1983) looked at women's marital status through time in relation to their socioeconomic background and number of children. Iwatsubo (1978) examined ratings of patient symptoms on several scales across several weeks of treatment. Multiple correspondence analysis has also been used to explore the relationship between sets of categorical variables, for example in Whiting et al.'s (1986) cross-cultural analysis of the relationship between kinship and social structure. It has also been used to explore the relationship between a set of categorical predictor variables and a categorical outcome variable, for example by Greenacre (1981) who looked at climatic conditions such as surface temperature and wind speed and direction as predictors of rainfall in Africa.

4.1. Indicator or Response Pattern Matrices

Discussion of multiple correspondence analysis is facilitated by introducing a data array called a *multiple indicator matrix*, or a *response pattern matrix*. This array

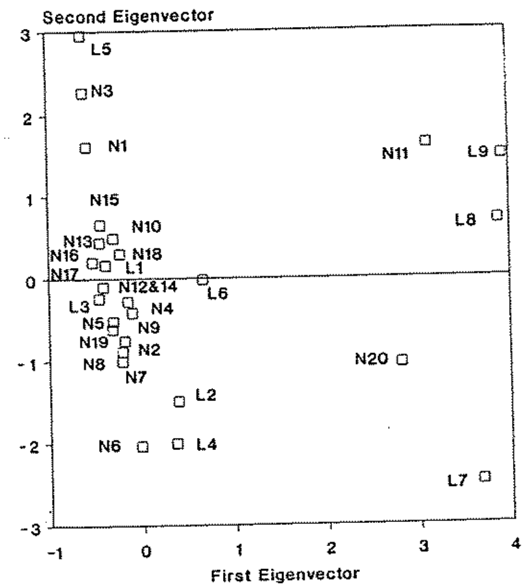

FIGURE 2. Correspondence analysis of nonprofit agencies (N1,N2,...,N20) and levels of donation (L1,L2,...,L9).

was mentioned briefly in connection with the optimal scaling of a two-way array. In this section we describe it in more detail and show how it can be used to represent the multi-way nature of relational data. A multiple indicator matrix, which we will denote **M**, is a subjects by variables array in which categorical variables of a contingency table are dummy-coded. In the multiple indicator matrix each dimension (row–sending actors, column–receiving actors, layer–relation, etc.) of the contingency table **F** is coded as a collection of dummy variables. There is one set of dummy variables for each dimension of **F**, and the number of levels for a dummy variable equals the number of levels of the corresponding categorical variable of **F**. For a two-way table, there are two sets of dummy-coded variables, one set with I variables, and the second set with J variables. Units falling in the (i,j)th cell of **F** would be coded unity on the ith variable in the first set and on the jth variable in the second set, and zero on the other $I+J-2$ variables. The multiple indicator matrix is easily extended to higher-way tables by introducing additional sets of variables, one for each additional dimension of **F**.

To be a bit more precise about the structure of the multiple indicator matrix, **M**, we will let N be the number of observational units (for example the number of dyads or pairs) and let Q be the number of categorical variables in **F**. We will let $a_1, a_2, \ldots, a_Q$ be the number of categories (levels) on each variable, respectively. The total number of categories for all variables is then $L = \sum a_q$. **M** is therefore an $N \times L$ array. We can see that **M** consists of Q submatrices. It can be viewed as the horizontal arrangement $\mathbf{M} = [\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_Q]$, where

$$m_{iql} = \begin{cases} 1 & \text{if subject } i \text{ responded to variable } q \text{ (of } \mathbf{F}\text{) at level } l \\ 0 & \text{otherwise} \end{cases}$$

($i = 1,2,\ldots,N$; $q = 1,2,\ldots,Q$; $l = 1,2,\ldots,a_q$). Note that the subscripts refer to the ith row or unit and the lth dummy variable in the qth (set) of variable(s) of **F**. Since there is a single "one" coding the response for each unit or row on each of the Q variables, the row totals of **M** will all equal Q. The column totals are the marginal frequencies of the response categories. An important note is that for Q = 2 variables, there is an important relationship between **M** and the contingency table, $\mathbf{F} : \mathbf{F} = \mathbf{M}_1'\mathbf{M}_2$. Table 5 gives an example of an **M** array. We discuss this example in more detail in the next section.

Given **M**, one then performs a standard CA on this $N \times L$ multiple indicator matrix. There will be sets of scores for the rows/units and for the columns/variables. The results and interpretation of a multiple correspondence analysis of **M** directly parallel the results of CA, though usually we will be interested only in the scores for the columns. These scores (for a specific eigenvector) are L in number and can be partitioned into Q sets. One usually takes each set, and scales the a_q scores in the qth set to have zero weighted mean and unit weighted variance. The weights are the reciprocals of the row and column profiles. We will denote the scores in the qth set as x_q, which has elements $\{x_{qi}\}$.

The results of multiple correspondence analyses are interpreted in terms of the simultaneous analysis of all two-way relationships. This joint-bivariate interpretation is perhaps more apparent if we note that MCA may be accomplished by analysis of an alternative data array constructed from the response pattern matrix **M**. This new array, termed a *Burt matrix* after Cyril Burt (1950) who first used it, and denoted here by **B**, is defined as $\mathbf{B} = \mathbf{M}'\mathbf{M}$. The **B** array is a square, symmetric matrix with a special form. It is composed of submatrices which are the 2-way cross-classifications of the Q sets of categorical variables constituting the columns of **M**. We can display **B** in terms of the Q submatrices of **M** as:

$$\begin{pmatrix} \mathbf{M}_1'\mathbf{M}_1 & \mathbf{M}_1'\mathbf{M}_2 & \ldots & \mathbf{M}_1'\mathbf{M}_Q \\ \mathbf{M}_2'\mathbf{M}_1 & \mathbf{M}_2'\mathbf{M}_2 & \ldots & \mathbf{M}_2'\mathbf{M}_Q \\ \vdots & \vdots & & \vdots \\ \mathbf{M}_Q'\mathbf{M}_1 & \mathbf{M}_Q'\mathbf{M}_2 & \ldots & \mathbf{M}_Q'\mathbf{M}_Q \end{pmatrix}$$

We see that each submatrix of **B** is a two-way contingency table, of size $a_i \times a_j$. The diagonal submatrices of **B**, $\mathbf{M}_i'\mathbf{M}_i$, are themselves diagonal matrices and contain the frequencies of the response categories: the (j,j)th diagonal entry of the ith submatrix contains the jth column total for the ith response variable.

Correspondence analysis of **B** gives results which are easily related to MCA of the response pattern matrix **M**. Since **B** is symmetric, the row and column scores from a CA of **B** will be equal. These scores, once scaled to have zero weighted mean and unit weighted variance (within each of the Q sets) will be equal to the scaled column scores from an MCA of **M** (see Greenacre, 1984, chapter 6). There will be no scores for subjects, which constitute the rows of **M**, when doing a CA on **B**. The eigenvalues from MCA of **B** are the squares of the eigenvalues from an MCA of **M**.

Analysis of the **B** array is not only helpful for illustrative purposes, it is also computationally efficient since the **B** array is usually considerably smaller than the response pattern matrix **M**. This efficiency is due to the fact that N is usually a good deal larger than L. However, in an analysis of **B**, one sacrifices the scores associated with the N units. Frequently, these units are not of interest, so can be ignored.

Multiple correspondence analysis also provides additional mathematical interpretations, which we have chosen not to mention here (to save some space). We do want to point out one important fact. The column scores maximize the sum of the Pearson ϕ^2 statistics for all of the $Q(Q-1)/2$ two-way tables. This property emphasizes that multiple correspondence analysis is a "joint bivariate" analysis. The method looks at two-way relationships but not at higher-order interactions. We mentioned this fact above, in conjunction with the introduction of Burt matrices. We will need this property in later sections of this paper.

4.2. Applications to Relational Data

Multiple correspondence analysis will allow us to analyze many forms of relational data. The only "tricks" we will need are to define our observational unit and our categorical variables appropriately. We start with a social relational system for a two-mode network with no attribute variables and a single relational variable, as defined in section 1. Recall that our unit of analysis is the dyad and that there are three variables necessary to describe the dyadic relation: the ID of the actor $(1,2,\ldots g)$, the ID of the partner $(1,2,\ldots h)$, and the level of the relational variable $(0,1,\ldots C-1)$. We can therefore construct the multiple indicator matrix as an array with as many rows as there are dyads ($N = gh$) and Q = three sets of categorical variables with $a_1 = g$, $a_2 = h$, and $a_3 = C$, for a total of $L = g + h + C$ categories or columns of **M**. Each row of the **M** array codes the identity of the actor in the dyad, the identity of the partner in the dyad, and the strength of the relation from the actor to the partner. Table 5 shows the multiple indicator matrix for the data on corporate-nonprofit donations from Table 1. This array can be analyzed with a standard correspondence analysis program.

Correspondence analysis of relational data expressed this way will be interpreted in terms of simultaneous two-way relationships, because of the relationship of **M** to Burt matrices. For two-mode networks there are three two-way relationships: actors initiating relations at different levels, partners receiving relations at different levels, and actors relating to partners.

4.3. Actor Attribute Variables

Attributes of actors and partners are easily incorporated into a correspondence analysis of relational data by including additional sets of categorical variables in the multiple indicator matrix. Each categorical attribute variable is coded as a set of "dummy" variables. If the attribute variable q has a_q levels, it gives rise to a set of

TABLE 5
Multiple indicator matrix of three-way relational structure of corporations, nonprofit agencies and levels of donation

Dyad	Corporation ID 1 2 3 … 10	Nonprofit ID 1 2 3 … 20	Contribution 1 2 3 … 9
(1,1)	1 0 0 … 0	1 0 0 … 0	1 0 0 … 0
(1,2)	1 0 0 … 0	0 1 0 … 0	0 0 1 … 0
(1,3)	1 0 0 … 0	0 0 1 … 0	1 0 0 … 0
⋮	⋮	⋮	⋮
(1,20)	1 0 0 … 0	0 0 0 … 1	0 1 0 … 0
(2,1)	0 1 0 … 0	1 0 0 … 0	1 0 0 … 0
⋮	⋮	⋮	⋮
(10,19)	0 0 0 … 1	0 0 0 … 0	1 0 0 … 0
(10,20)	0 0 0 … 1	0 0 0 … 1	0 0 0 … 1

a_q column variables in the multiple indicator matrix, which pertain to the actor (or the partner) in the dyad. An actor (partner) receives a "one" on the lth variable in this qth set of variables if the attribute for that actor (partner) is equal to the lth level of the attribute. We note that we can also include additional relations in a similar manner by including more column variables to code the level of the new relational variable.

4.4. Comparison to Other Methodologies

The many applications of correspondence and multiple correspondence analysis in psychology, ecology, marketing, and other disciplines have demonstrated that CA and MCA have two main types of application: 1) Scaling the rows and columns of a two-way data structure; and 2) Automatic classification of the categories of the row and/or column and/or other variables under study. The methods have two big advantages over similar methods: they are simple to describe mathematically and computationally easy and inexpensive.

Ecologists, who use CA frequently as a scaling device (see Gauch, Whittaker, and Wentworth, 1977) report that the method does well when interest centers on the first non-trivial eigenvector of $\mathbf{F}^*\mathbf{F}^{*\prime}$. As is often the case with such methods, later eigenvectors can be difficult to interpret. Social scientists should be more interested in the use of CA as a classification device. The hope is that if the rows and/or columns of **F** can be placed or partitioned into "blocks", then CA should be able to detect them. This is one reason why we are interested in CA, since this is one of the main goals of relational data analysis.

When the **Z** array is a sociomatrix with row and column categories representing interpretatively different social entities as with a two-mode network, we can study it with correspondence analysis. Using MCA, we can turn a sociomatrix **Z** into a two-way contingency table **B** that can be analyzed by MCA. The results of such an analysis would yield row and column scores $\{x_{im}\}$ and $\{y_{jm}\}$, and eigenvalues, $\{\eta^2_m\}$, which would have clear social network interpretations. Row and column

scores would be associated with actors in the network (rather than with the general categories of an $I \times J$ cross-classification). Row scores would be interpreted as *actor* or *sending* effects, while column scores would be *partner* or *receiver* effects. Actors with similar row (column) scores across the K dimensions of the correspondence analysis solution would have similar patterns of relationships with the other sending (receiving) actors. So, the idea is to use the correspondence score values to group together equivalent or at least similar actors and partners. This goal is quite common in social network analysis. We do not recommend that **Z** be analyzed directly with a CA computer program. This may seem surprising, since **Z** is a two-dimensional array, and CA operates on such arrays. This recommendation is based on the realization that the relational data comprising **Z** are three dimensional, and should be analyzed with *multiple* correspondence analysis.

We want to mention the goals of most relational data methodologies and compare them to the things that can be learned from CA and MCA. Researchers studying relational data usually adopt the primary goal of learning how similar (or dissimilar) their network actors are. This is accomplished by methods which partition actors into subgroups of similar actors. The most widely used substantive notion for justification of actor partitioning (which also assumes that the network is just one-mode) relies on the similarity in patterns of social interaction. The notion, *structural equivalence*, is defined by stating that two actors in a network are structurally equivalent if they have identical ties to and from all other actors in the network (Lorrain and White, 1971). In any actual application it is unlikely that two actors will have exactly identical ties; consequently, this strict notion of structural equivalence is often relaxed to allow for varying "degrees" of structural equivalence. Common alternative measures of this equivalence are based on Euclidean distance between actors represented as points in a multidimensional space and correlations among rows and/or columns in the sociomatrix.

Identification of structurally equivalent rows (columns) allows the researcher to simplify the analysis of a sociomatrix by equating equivalent units. One simplification is provided by a data structure called a *blockmodel* (see Breiger, Boorman, and Arabie, 1975; or White, Boorman, and Breiger, 1976). A blockmodel is a partition of rows and columns of a sociomatrix (or, in substantive terms, the actors) into equivalent *blocks* or actor subsets, and a mapping from the original relations in the sociomatrix to relations between subsets (blocks) in the model. Equivalent individuals are assigned to blocks in the model. If a relation exists between individuals in the sociomatrix, then a relation exists between blocks in the blockmodel. Actors within the same block are assumed to be structurally equivalent. In practice, one applies this theory to one-mode network data. However, there is no reason why these ideas can not be extended to the two-mode, rectangular situation that we have described here. Indeed, such data are certainly social network data, and one can conceive of blockmodels for both the sending and receiving actors. One nice feature of CA, is that similarities among both sets of actors are easy to find, thus giving the researcher a simple algorithm for locating structurally equivalent actors.

We point out these notions of structural equivalence and blockmodels to make the following observation. There is a direct parallel between this data structure and the observation by Hill (1974) that if the matrix **Z** has the block-diagonal form

$$\begin{pmatrix} \mathbf{Z}_a & 0 \\ 0 & \mathbf{Z}_b \end{pmatrix}$$

where $\mathbf{Z}_a$ and $\mathbf{Z}_b$ are sociomatrices for disjoint subsets of the actors, then a permutation of rows and columns of **Z** according to scores **x** and **y** from a correspondence analysis will reveal this pattern. The above block-diagonal form is termed an *exact blockmodel* by White, Boorman, and Breiger (1976).

If the pattern is not exact (distinct subsets of equivalent actors do not exist in the network) then approximate blockmodels may be found by grouping rows or columns on the basis of degree of similarity in scores. The notion is that the existing actor *interactions* in the sociomatrix can be found by a correspondence analysis. Schriever (1983) comments on the use of correspondence analysis for this purpose when the rows and columns of the two-way contingency table both are levels of ordinal discrete variables. To be able to use CA to study the interactions in a two- (or higher-way) structure would be very beneficial to social networkers. Furthermore, there has been much interest recently in the analysis of ordinal categorical variables. Several researchers have shown that if one postulates a special class of log linear models for the data from such variables, then the higher-order interactions can be written as functions of correspondence analysis-derived scores. We will return to this idea in later sections of this paper.

Another goal of social network research is to identify groups of actors who interact frequently with each other, or who relate to each other in preference to others. For a one-mode network, such patterning is revealed by rearranging rows and columns of a sociomatrix so that positive choices among actors are concentrated along the main diagonal of a sociomatrix. This problem was central to early investigations in sociometry (Beum and Brundage, 1950; Coleman and MacRae, 1960; and Katz, 1947). More recently, it has been seen as an instance of the more general problem of seriation or ordination of data arrays (Deutsch and Martin, 1971; Spath 1980). Kendall (1971) was among the first people to recognize the usefulness of CA for seriation. He suggested using CA row and column scores for ordering archaeological sites and artifacts to reveal their relative historical age. For rectangular data arrays, seriation based on CA row and column scores reveals data structures of "block trapezoidal" form. Gradient analysis, a variant of CA, has been used in ecology for seriation purposes, specifically to arrange geographical locations and characteristic plant or animal populations (Greenacre, 1984; Hill, 1974; and Lebart, et al., 1984). As we noted above, the earliest applications of correspondence analysis to social network data had as one goal the reordering of rows and columns of a sociomatrix to make the group structure more apparent.

A theoretically important proposition in the analysis of social structure is that individuals are linked through their participation in social events, and that events in turn are linked by their participants. This notion has been termed by Breiger (1974) the *duality of persons and groups*, and is related to the concept of social circles (Simmel, 1955; and Kadushin, 1966). This gives rise to a special form of two-mode relational data which records actors' participation in social events. One classic example of this form of data is the attendance of Southern club women at

social events (Davis, Gardner, and Gardner, 1941). The dual scaling and reciprocal averaging interpretations of correspondence analysis make CA potentially useful for analyzing such data. However, except for the analysis by Romney and Boyd (1987) of the Davis, Gardner and Gardner data, this remains a relatively unexplored application.

We turn now to an example of multiple correspondence analysis, looking at the three-way relational structure of corporate donations to nonprofit agencies.

4.5. Example Revisited

In this section we use multiple correspondence analysis to look at the multi-way structure of a relational data set. We look first at the three way structure of corporations making donations to nonprofit agencies at one of $C = 9$ different levels. We then incorporate attributes of corporations and nonprofit agencies into the analysis in order to better understand the relationships.

An MCA of the three-way table of corporations ID, nonprofit ID and level of donations gives three sets of scores, which are presented in Table 6 for our simple example of $g = 10$ corporations and $h = 20$ nonprofits. The first two eigenvectors are used to present the results graphically in Figures 3a, 3b, and 3c. Although we present separate figures for corporations, nonprofits, and donation levels, one can examine the relationship between the three sets of scores. Since multiple correspondence analysis is a "joint bivariate" solution (i.e., we analyze simultaneously the interactions between all possible pairs of variables), this analysis attempts to simultaneously quantify the two-way relationships given in Tables 2a and 2b. The solution incorporates aspects of all two-way relationships, and will not in general be identical to any CA of just one of the two-way tables. One can see that the scatter in Figures 3a–3c combines features of both of the two-way analyses presented in Figures 1 and 2. The first axis separates the highest levels of donation (7, 8 and 9) and the nonprofit agencies #11 and #20 (the two media organizations) from the corporations and the remainder of donation levels and nonprofit agencies, as in the analysis of nonprofit agencies and donation levels. The second axis of the three way result resembles the first axis of the two-way analysis of corporations by levels of donations, distinguishing between level 2 donations and corporation #7 on the one hand and donations at level 5 and corporations #3 and #5 on the other.

By referring to the cross-classification of corporations by levels of donation in Table 2b, we can see that these firms are distinctive in the levels of donations they make. These three (#7, #3, #5) are the only corporations which made more than half of their donations to the 20 nonprofit agencies at a level of 2 or higher. Corporation #7 made half of its donations at level 2. Corporations #3 and #5 are similar to each other and different from the rest of the corporations in that they made multiple donations at levels 5, 6 and 9, and made no donations at levels 2, 7, or 8. One can see this association between corporations and donation levels by comparing Figures 3b (for corporations) and 3a (for the donations).

As mentioned above, multiple correspondence analysis of relational data is easily generalized to include attributes of actors and partners, a feature which we now demonstrate. We include 5 categorical attribute variables in an MCA. For nonprofits we look at two attributes: the activity in which they are involved (6 levels) and

TABLE 6
Scores from correspondence analysis of three-way table: corporations, nonprofit agencies and levels of donation

Corporation	Eigenvectors 1	2	3	4
1	1.428	0.029	−0.790	0.007
2	1.279	0.040	−0.707	−0.832
3	−1.347	−1.201	0.980	0.817
4	1.343	0.574	−0.384	1.058
5	−1.504	−1.345	1.465	−0.742
6	−0.336	−0.093	−1.257	−2.325
7	0.093	2.510	1.717	1.076
8	−0.208	−0.076	0.335	0.558
9	−0.382	−0.165	−0.511	0.292
10	−0.365	−0.273	−0.848	0.090
Nonprofit Agency				
1	0.434	−1.649	0.829	0.337
2	0.173	0.377	0.457	−0.179
3	0.365	−2.304	2.072	0.301
4	0.097	0.175	0.715	0.686
5	0.395	0.416	−0.686	−0.403
6	0.071	1.339	1.160	−1.860
7	0.301	1.184	0.249	−0.187
8	0.340	1.834	0.513	0.766
9	0.020	−0.288	1.723	−1.248
10	0.212	−1.090	0.548	−0.530
11	−3.128	−0.071	−1.513	2.515
12	0.494	0.382	−0.929	0.529
13	0.388	−0.921	0.047	−0.461
14	0.494	0.382	−0.929	0.529
15	0.441	−0.315	0.096	0.395
16	0.566	−0.298	−1.436	0.511
17	0.556	−0.298	−1.436	0.511
18	0.206	−0.640	−0.434	0.371
19	0.379	0.460	−0.472	−0.305
20	−2.794	1.324	−0.572	−2.281
Donation Level				
1	0.443	−0.149	−0.515	0.109
2	−0.141	2.643	1.475	0.519
3	0.323	0.021	0.077	0.451
4	−0.426	0.147	0.745	−2.794
5	−0.078	−2.666	2.909	−0.021
6	−0.966	−0.635	1.459	0.209
7	−3.461	1.164	−2.764	−7.859
8	−3.670	0.444	−2.669	−0.987
9	−3.929	−0.485	−0.682	1.756
Eigenvalue (M)	0.598	0.544	0.495	0.466
% of Inertia	4.99	4.53	4.15	3.88
Eigenvalue ($M'M$)	0.358	0.295	0.248	0.217
% of Inertia	8.16	6.73	5.64	4.94

TABLE 7
Scores from multiple correspondence analysis of corporations, nonprofit agencies, levels of donation, and attributes

	Eigenvectors 1	2
Corporation		
1	−1.142	−0.370
2	−1.458	−0.914
3	0.047	−1.243
4	−1.051	−0.300
5	0.937	0.931
6	0.487	0.263
7	2.031	2.424
8	−0.331	−0.721
9	0.254	0.138
10	0.226	−0.209
Nonprofit Agency		
1	−0.998	0.593
2	0.692	−0.784
3	−0.787	0.488
4	0.732	−0.409
5	−0.717	0.456
6	0.163	0.734
7	−0.446	0.430
8	0.105	0.792
9	0.079	0.614
10	−0.715	0.367
11	2.604	−2.889
12	0.220	−0.433
13	−0.755	0.430
14	−0.577	0.338
15	−0.931	0.643
16	−0.314	−0.065
17	−0.429	0.485
18	−0.542	0.152
19	−0.017	0.635
20	2.631	−2.577
Donation Level		
1	−0.605	−0.026
2	1.284	1.767
3	0.250	0.481
4	0.461	0.167
5	−0.422	0.696
6	1.154	−0.877
7	3.366	−2.617
8	3.210	−2.845
9	3.082	−3.624

	Eigenvectors 1	2
Nonprofit Activity		
media	2.732	−2.881
legal	−0.971	0.677
housing & urban	−1.041	0.625
health & welfare	−0.550	−0.318
educational	0.196	−0.447
cultural	0.222	0.498
Nonprofit Essential		
1 (low)	−0.554	0.258
2	−0.964	0.595
3	−0.024	0.784
4 (high)	1.624	−1.705
Corporation Industry		
manufacturing	−0.129	0.142
utilities	2.556	2.586
wholesale/retail	0.059	−1.328
finance/real estate	−0.066	−0.496
other services	−1.836	−0.977
Corporation Prestige		
1 (low)	−1.411	−0.787
2	0.077	−0.845
3 (high)	1.001	1.224
Corporation Pretax Income, 1980		
1 (low)	−1.059	−0.586
2	−0.293	−0.705
3 (high)	1.449	1.526
Eigenvalue ($M'M$)	.172	.152
% of Inertia	11.35	10.01

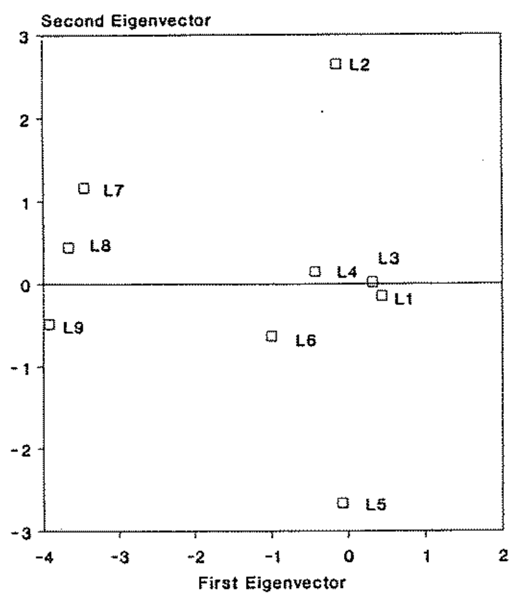

FIGURE 3a. Multiple correspondence analysis of corporations, nonprofits and donation levels: donation levels (L1,L2,...,L9).

the degree to which they are perceived as offering essential services (4 levels). For corporations we look at three attributes: the industry of the corporation (from the seven levels/industries in the full data set, five levels are represented in this small example), the degree to which members of the corporate board of directors are linked to prestigious clubs and associations (3 levels), and the 1980 pretax income of the corporation (3 levels). The attribute variables measuring the degree to which a nonprofit agency is perceived as offering essential services, the prestigious links of corporations, and corporate pretax income were all initially continuous variables, but were categorized (to make them easier to compare to the categorical relational variable) by choosing cutpoints to make the frequencies in the categories as equal as possible.

There are now $Q = 8$ sets of variables: three for the minimal relational structure plus five sets of attribute variables. We present the multiple correspondence analysis scores for these eight variables in Table 7. These scores are displayed in five separate figures (Figures 4a-4e).

Inclusion of attributes of corporations and nonprofits in the multiple correspondence analysis allows us to better interpret the relationships among corporations, nonprofits, and donation levels, represented in Figures 4a to 4e. We first note that the levels of donation are fairly well ordered using the first two eigenvectors from this analysis. The highest levels of donation (levels 7, 8 and 9) have high positive scores on the first eigenvector, and high negative scores on the second eigenvector.

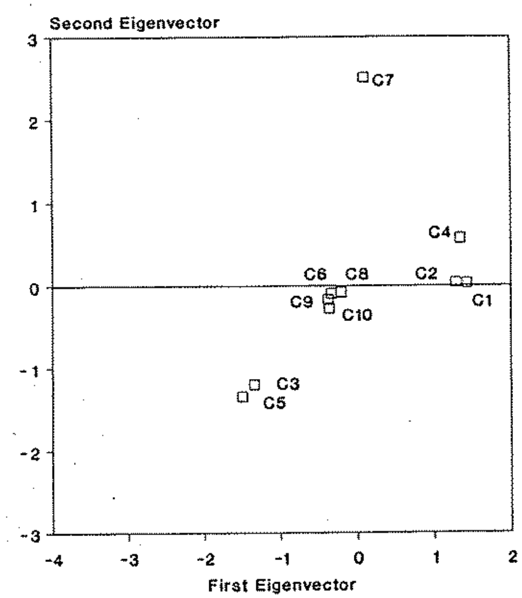

FIGURE 3b. Multiple correspondence analysis of corporations, nonprofits and donation levels: corporations (C1,C2,...,C10).

The donation levels are thus roughly ordered from highest levels in the lower right corner of Figure 4a to low and moderate values in the upper left. We also see nonprofits #11 and #20 are associated with the highest levels of donations, as in the previous results. Inclusion of nonprofit attributes in this analysis confirms our earlier observation that these are media nonprofits. We can also see that the attribute variable based on the ratings by community elites of whether they recognize the nonprofits, and whether they provide essential and outstanding services, is also ordered from highest (4) to lowest (1), with the category for highest ratings associated with higher donations and with the media nonprofits.

Corporations and their attributes are arranged in a somewhat different pattern, and appear to be less clearly related to levels of donations. Two of the corporate attributes (pretax income in 1980 and the categorical variable coding linkages of corporate boards of directors to prestigious individuals) are ordered on the first eigenvector, and this order is apparent in Figure 4c. Highest level of pretax income and links to prestigious individuals have positive scores on both the first and second eigenvectors and are characteristic of corporations involved in utilities. Low levels of pretax income and low and moderate links to prestigious individuals are characteristic of corporations involved in other services, as can be seen from their negative scores on both eigenvectors. The relationship between donation levels and corporation attributes is less clear. Focusing on the first eigenvector shows that high levels of donations are characteristic of utilities. Low levels of donations are characteristic of corporations in other services, with low links to prestigious individuals,

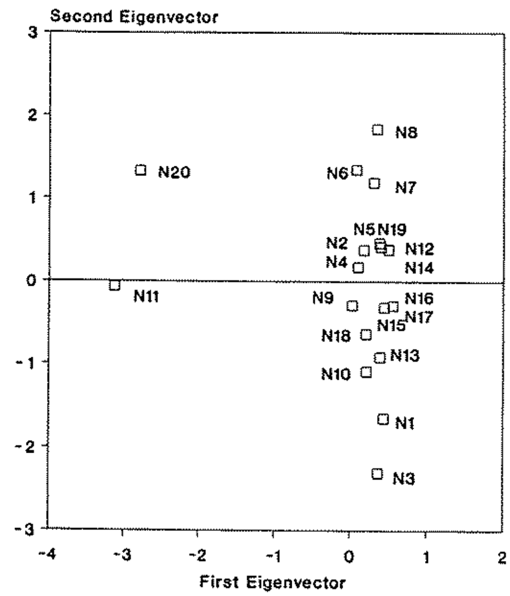

FIGURE 3c. Multiple correspondence analysis of corporations, nonprofits and donation levels: nonprofits (N1,N2,...,N20).

or with lower pretax income in 1980. However, focusing on both eigenvectors reveals that the relationship between corporation attributes and levels of donations is far from simple, with low to moderate donation levels characteristic of most kinds of corporations.

5. CORRESPONDENCE OR CANONICAL ANALYSIS—STATISTICAL MODEL

Over the past 20 years, there has been tremendous growth in research methodology for categorical data. Most of the new methods use log linear models for the probabilities associated with the multinomial or Poisson distributions assumed as sampling models for the data. Goodman, Mosteller, Bishop, Fienberg, Holland, Haberman, and several others are credited with the development of this important new methodology. This research has led to consideration of contingency tables in which one or more of the categorical variables are ordinal; i.e., the levels of the variable(s) can be ordered in some way (as with SES categories—Lower, Middle, Upper classes), and perhaps even have scores attached to them (as with age categories—under 30 years, 30–50 years, etc.). Haberman (1974), Simon (1974), Goodman (1979a, 1979b, 1981a, 1981b, 1981c, 1983, all of which can be found in the 1984 collection), Andersen (1980), Clogg (1982a, 1982b), and especially the review of Agresti (1983), present much of this new research.

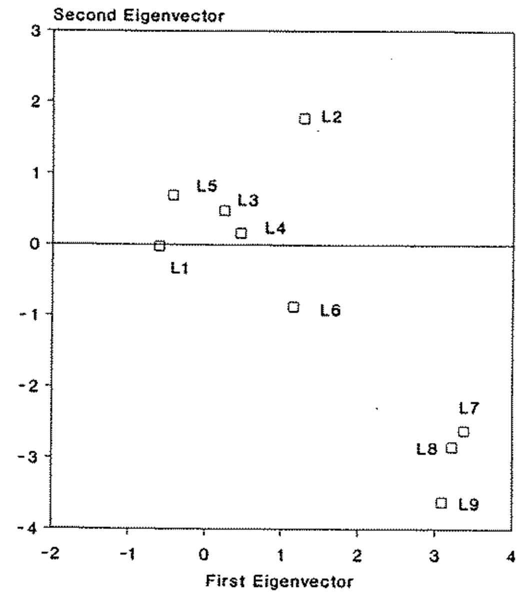

FIGURE 4a. Multiple correspondence analysis including corporation and nonprofit attributes: donation levels (L1,L2,...,L9).

Very recently, several authors have noted that models for ordinal categorical data are related in special ways to methods for correspondence analysis. This connection was first noted by Goodman (see Goodman, 1985, for a thorough review) and further elaborated upon by Goodman (1986), Gilula and Haberman (1986a, 1986b), Gilula (1986), and van der Heijden and de Leeuw (1985). This connection has been very important to the development of CA because it has allowed the extensive statistical machinery of log linear modelling to be applied to CA. This has extended the inferential aspects of CA in much needed ways. We have chosen to call this approach the *correspondence analysis statistical model*, since it is so different from the traditional, more exploratory CA criteria and solutions, resembling a log linear model for **F**, and because it allows for statistical tests on the number of eigenvalues of **F** that need to be examined. Goodman (1986) states that those researchers "sold on" correspondence analysis "...may view the methods described below as a replacement for the usual (CA) approach", while others, who are more "rooted" to the standard log linear models may use these methods to "supplement or augment" their usual methods. We will briefly review the CA statistical model, and show how it can be extended to higher-way tables. After this discussion, we will compare this model to a family of models for relational data, and show how this family can be generalized to include CA.

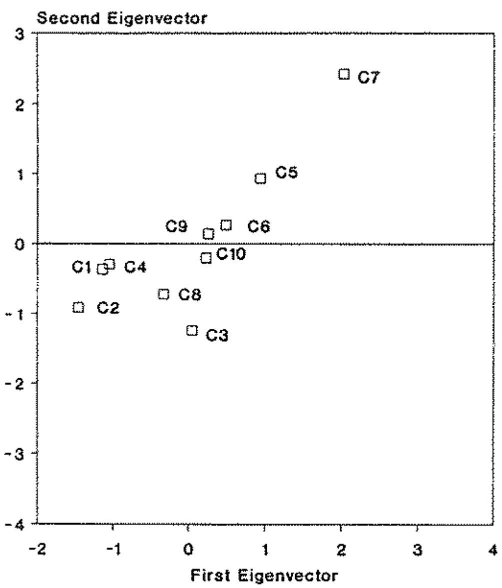

FIGURE 4b. Multiple correspondence analysis including corporation and nonprofit attributes: corporations (C1,C2,...,C10).

5.1. Mathematics

We begin, as we have in earlier sections, with a two-way contingency table **F**, of size $I \times J$, but now define P_{ij} as the probability that an observation falls into the ith row and jth column of the table. The standard, "null" model of association assumes that P_{ij} is equal to the product of $P_{i\cdot}$ and $P_{\cdot j}$, the sums of the $\{P_{ij}\}$ over j and i, respectively. If this model (which can be tested by examining the likelihood ratio test statistic, which, asymptotically, is a χ^2 random variable with $(I-1)(J-1)$ degrees of freedom) does not provide an adequate explanation of the data, there are a number of alternative models that can be tried. All of these alternatives attempt to model the interaction between the variables in a parsimonious fashion. Agresti (1983) reviews these alternatives, many of which originated with Goodman.

One very special generalization, termed the *saturated RC canonical correlation model*, takes the above model

$$P_{ij} = P_{i\cdot}P_{\cdot j} \tag{14}$$

and adds another term to it:

$$P_{ij} = P_{i\cdot}P_{\cdot j}\left(1 + \sum_{m=1}^{K} \eta_m x_{im} y_{jm}\right) \tag{15}$$

where $K = \min(I-1, J-1)$. This additional term is a multiplicative interaction, and is dependent on (up to) K components. The **x**'s and **y**'s in model (15) are con-

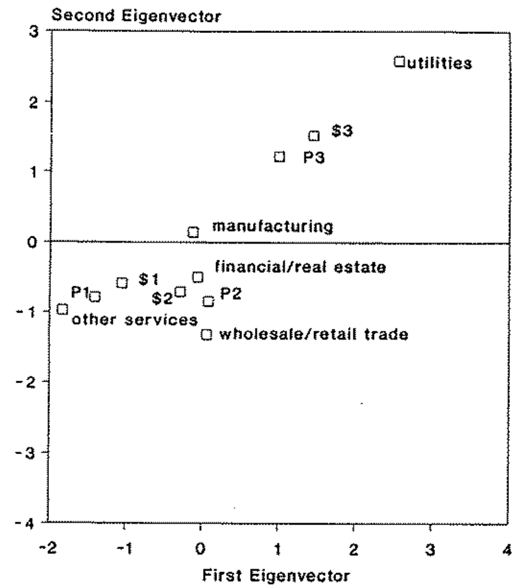

FIGURE 4c. Multiple correspondence analysis including corporation and nonprofit attributes: corporation attributes–corporation industry, links to prestigious individuals (P1,P2,P3) and corporate pre-tax income in 1980 ($1,$2,$3).

strained to have weighted means of zero, weighted variances of unity, and the **x**'s (**y**'s) are uncorrelated among themselves:

$$\begin{aligned} \sum_{i=1}^{I} x_{im} P_{i\cdot} &= \sum_{j=1}^{J} y_{jm} P_{\cdot j} = 0, \\ \sum_{i=1}^{I} x_{im}^2 P_{i\cdot} &= \sum_{j=1}^{J} y_{jm}^2 P_{\cdot j} = 1, \end{aligned} \tag{16}$$

and

$$\sum_{i=1}^{I} x_{im} x_{im'} P_{i\cdot} = \sum_{j=1}^{J} y_{jm} y_{jm'} P_{\cdot j} = 0$$

for $m \neq m'$. The interaction parameters $\{x_{im}\}$ and $\{y_{jm}\}$ are unknown row and column scores (scaled to have weighted means of zero and variances of unity). The parameters $\{\eta_m\}$ measure the correlation between the **x**'s and **y**'s, since

$$\sum_{i=1}^{I}\sum_{j=1}^{J} x_{im} y_{jm} P_{ij} = \eta_m, \qquad m = 1, 2, \ldots, K, \tag{17}$$

and are usually ordered $1 \geq \eta_1 \geq \eta_2 \geq \cdots \geq \eta_K \geq 0$.

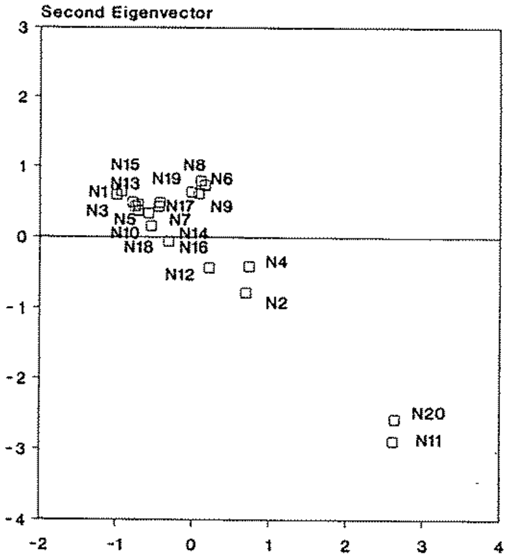

FIGURE 4d. Multiple correspondence analysis including corporation and nonprofit attributes: nonprofits (N1,N2,...,N20).

Equations (15)–(17) highlight why this model is labelled a canonical correlation model. The x's and y's, with $m = 1$, are the standardized row and column scores that maximize the correlation η_1 defined by (17). Further, picking any m between 2 and K, the parameters x_{im} and y_{jm} are standardized row and column scores, uncorrelated with all row and column scores with $m' < m$, that maximize the correlation η_m between them. This interpretation is virtually identical to the solution of the standard canonical correlation problem in multivariate analysis.

An equivalent model to (15)–(17), which Goodman terms the *saturated RC correspondence analysis model* uses scaled **x**'s and **y**'s,

$$x'_{im} = \eta_m x_{im} \qquad \text{and} \qquad y'_{jm} = \eta_m y_{jm}$$

so that equation (15) becomes equivalent to

$$P_{ij} = P_{i\cdot}P_{\cdot j}\left(1 + \sum_{m=1}^{K} x'_{im}y'_{jm}/\eta_m\right) \tag{18}$$

where the weighted variances of the x's and y's are now the η^2's.

To understand how these models compare to a standard correspondence analysis of **F**, one can show that both the saturated RC canonical correlation model and the

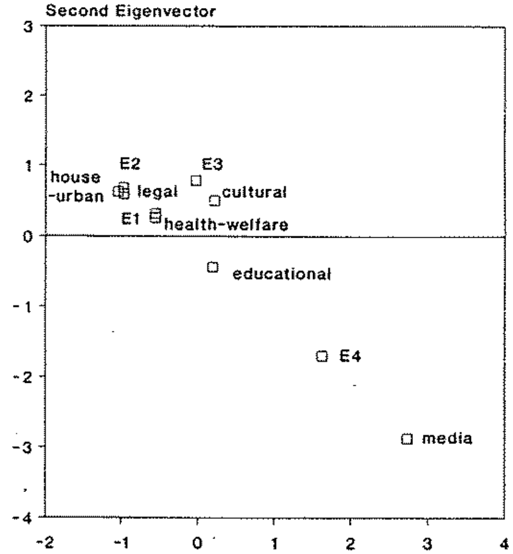

FIGURE 4e. Multiple correspondence analysis including corporation and nonprofit attributes: nonprofit attributes–nonprofit activity, and nonprofit recognition and esseential rating (E1,E2,E3,E4).

saturated RC correspondence analysis model partition Pearson's X^2 as follows:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(P_{ij} - P_{i\cdot}P_{\cdot j})^2/P_{i\cdot}P_{\cdot j} = N\sum_{m=1}^{K}\eta_m^2 \tag{19}$$

Thus, the statistic can be partitioned into K components which are the squares of the canonical correlations. Compare (19) to (11)–(13), and it becomes apparent that the canonical correlations are simply the square roots of the eigenvalues of our favorite matrix $\mathbf{F^*F^{*\prime}}$.

In fact, one can rearrange model (15) to:

$$\frac{P_{ij} - P_{i\cdot}P_{\cdot j}}{P_{i\cdot}P_{\cdot j}} = \sum_{m=1}^{K}\eta_m x_{im} y_{jm} \tag{20}$$

which makes it clear that the saturated RC correspondence analysis model is simply a "parametric" spectral decomposition of the residuals from the standard independence log-linear model. This equation should help the reader realize the equivalence between the models discussed here and the more standard or classical CA approach. Furthermore, the unknown x's and y's in these models are simply rescalings of the eigenvectors of $\mathbf{F^*F^{*\prime}}$ (remember that $P_{ij} = f_{ij}/f_{\cdot\cdot}$).

The research of van der Heijden and de Leeuw (1985) also provides an interesting characterization of correspondence analysis modelling. These authors contrast

CA with standard log linear models, and show how one can use CA to decompose the difference between two sets of fitted values, each from a different log linear model, but where one is nested within the other. The difference between the two fitted models is a function of the interaction terms that are in the larger, more comprehensive model, but not the smaller, nested model. Thus, a CA of the difference allows one to study in detail the interactions present in the bigger model, but not the smaller.

Goodman's models demonstrate that the CA statistical models are equivalent to the CA classical models, in the sense that the CA statistical models contain interaction terms that depend on the standard CA-derived scores. Since the canonical correlation models are equivalent to the correspondence analysis models, Goodman refers to both as simply *correlation models*, although Gilula (1986) and Gilula and Haberman (1986) refer to these models as *canonical models*, and this approach, as a *canonical analysis* of the contingency table **F**. The $\mathbf{x}_m$'s and $\mathbf{y}_m$'s are called *mth-order canonical scores*, and η_m, the *m-th order canonical correlation*.

The basic model (15) can be simplified by including fewer **x**'s and **y**'s. The *unsaturated* RC correspondence analysis model replaces the symbol K with K^*, where $1 \leq K^* < K$. The special unsaturated case with $K^* = 1$ has been considered by Goodman (1985), and there are some simplifications of the parameters (depending on the spacing and/or equality of the **x**'s and **y**'s) that make interpretation easier and relate this model to some of Goodman's other models for ordinal categorical data. The RC correspondence analysis models are quite similar to another class of models discussed by Goodman (1986)—the saturated/unsaturated RC association models. Clogg (1986) shows that even though the association and CA models look similar, there are some important differences. The association models are sometimes referred to as *log-bilinear models*. Further statistical comments on these models can be found in the discussion following Goodman's (1986) paper; for lack of space, we will not elaborate on these comments here.

The saturated and unsaturated RC correspondence analysis models can be fit using maximum-likelihood (ML) estimation, and goodness-of-fit techniques can be used to determine how unsaturated the model should be. Such tests are quite important, and basically determine the dimensionality of the CA problem. Gilula and Haberman (1986a) and Goodman (1985) discuss estimation. These authors show that the ML estimates are quite close to the usual CA calculated scores, so that the maximum likelihood estimates $\hat{\mathbf{x}}$'s and $\hat{\mathbf{y}}$'s, from model (15), can be approximated well by the scaled eigenvectors of $\mathbf{F}^*\mathbf{F}^{*\prime}$, found via a standard correspondence analysis.

We now consider how to fit RC correspondence analysis models to relational data. Rather than taking Goodman's ideas and applying them directly to relational data, we have found that the best approach is to take existing log-linear models for relational data, and then modify them so that model parameters can be interpreted as CA scores.

5.2. Independent Dyadic Choice Models—p_1 and relatives

We turn our attention to recent methodology for relational data, with the goal of demonstrating how correspondence and canonical analysis can be used to complement these methods. The relationship between these independent dyadic choice models and correspondence and canonical analysis will be detailed in the next section. Since these dyadic models have been described in other places (beginning in 1981) and because the emphasis in this paper is on correspondence and canonical analysis, the discussion in this section is intentionally brief.

In 1977, Holland and Leinhardt introduced a model that they termed p_1 for the statistical analysis of a single, binary relational variable. Their paradigm was developed further by Holland and Leinhardt (1981) and Fienberg and Wasserman (1981), and since then, has been the basis of much research activity. Here we present one extension of p_1, first described by Wasserman and Iacobucci (1986), and show how it includes Holland and Leinhardt's model as a special case. We then generalize the model to handle two-mode or rectangular single relational data, and briefly describe other generalizations of it that have appeared in the literature since 1981. We conclude this section with a discussion of how this model can be generalized even further to include standard correspondence analysis as a special case.

Once again, consider a single relational variable, defined for all pairs of actors in a one-mode network of size g. We assume that the variable is valued and can take on C values. The situation can be generalized if we assume that the relational variable is actually ordinal and we will comment on this generalization later. It is usually the case that relational strengths or values are measured with positive integers. We define, as we did earlier,

$$Z_{ij} = \text{value of relation from actor } i \text{ to actor } j$$

and assume that the $g(g-1)$ Z variables $(i \neq j)$ take on one of the possible values. The model to be described here has been termed by Frank, Hallinan, and Nowicki (1985), and Frank, Komanska, and Widaman (1985) a *dyad independence model* since it assumes that the $g(g-1)/2$ dyads, the basic modelling units, are independent. We first must define

$$D_{ij} = (Z_{ij}, Z_{ji})$$

as the dyad involving the pair of actors i and j. These D_{ij} are bivariate discrete-valued random variables, and have C^2 possible realizations. It will be convenient for us to define a new random variable, based on D, that reflects the state of the dyad. We can construct a four-dimensional array $\mathbf{Z}^* = \{Z^*_{ijkl}\}$ from $\mathbf{Z}$, where the subscripts i and j refer to the two actors in the dyad, and the subscripts k and l refer to the strengths of the relation between the actors. The array indicator variable $\mathbf{Z}^*$ has entries defined as

$$Z^*_{ijkl} = \begin{matrix} 1 & \text{if} \quad D_{ij} = (k,l) \\ 0 & \text{otherwise.} \end{matrix}$$

In our earlier research, we labelled this array **Y** (see Fienberg and Wasserman, 1981), but have decided to relabel it here so that it would not be confused with the correspondence analysis scores, **y**. Thus, for a specific dyad, we have a $C \times C$ submatrix, a table of indicator variables. The location of the single 1 in this submatrix indicates the state of the dyad. For example, suppose that $C = 5$, and i relates to j at strength 3 and j relates to i at strength 4. Then the 5×5 submatrix will have a

1 in the (4,5) cell, and zeros elsewhere. The introduction of the $\mathbf{Z}^*$ array facilitates the statistical modelling of the variable.

Define π_{ijkl} as the probability that the observed dyad is in state (k,l), and let $\mu_{ijkl} = \log\pi_{ijkl}$ be the elements of a table of log probabilities corresponding to the Z^*_{ijkl}. We postulate the following log-linear model

$$\mu_{ijkl} = \lambda_{ij} + \theta_k + \theta_l + \alpha_{i(k)} + \beta_{j(k)} + \alpha_{j(l)} + \beta_{i(l)} + (\alpha\beta)_{(kl)} \tag{21}$$

subject to the constraints that insure proper probability distributions,

$$\sum_{k=0}^{C-1}\sum_{l=0}^{C-1} \exp\{\mu_{ijkl}\} = 1$$

for all dyads. Model (21) contains four types of parameters. A thorough discussion of the types can be found in Wasserman and Iacobucci (1986). We will note here that the α and β parameters are main effects for choosing (actors, expansiveness) and receiving (partners, popularity), respectively. They are constrained as follows:

$$\alpha_{i(0)} = 0, \qquad i = 1,2,\ldots,g,$$

$$\sum_{i=1}^{g} \alpha_{i(k)} = 0, \qquad k = 0,1,\ldots,C-1,$$

$$\beta_{j(0)} = 0, \qquad j = 1,2,\ldots,g,$$

$$\sum_{j=1}^{g} \beta_{j(k)} = 0, \qquad k = 0,1,\ldots,C-1.$$

These constraints allow the model (21) to reduce to the special case of p_1 when $C = 2$.

The $(\alpha\beta)$ interaction parameters measure how likely it is that two actors (such as i and j) interact at strengths k and l, given that actors i and j have specific tendencies to choose at strengths k (for i) and l (for j), and tendencies to receive at strengths l (for i) and k (for j). These interactions can also be interpreted as measuring how likely it is that i "chooses" j at strength k, given that j "chooses" i at strength l. This collection of interactions occupy the cells of a $C \times C$ array, and are constrained in such a way that the first row and first column contain zero entries.

The θ parameters in (21) are general "strength" effects, and are constrained to sum to zero. The remaining parameters in model (21), the λ_{ij} quantities, are present simply to insure that the probabilities sum to unity for each dyad.

When $C = 2$, there are $2g$ α parameters, half of which are zero. Those that are non-zero are equivalent to the α parameters in p_1. The same is true for the β parameters. There is just one non-zero $\alpha\beta$ interaction, $(\alpha\beta)_{11}$, which is equal to the ρ reciprocity parameter in p_1. It is important to remember these constraints are different from standard ANOVA constraints when interpreting the parameters. We note that another reason for adopting these constraints is that they are equivalent to

those associated with generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983) which are fit by the statistical computer package GLIM 3.77 (Payne, 1985). GLIM offers several features not found in comparable packages, and we have used it extensively in our research.

Model (21) can be simplified by using the ordinal nature of the relational variable. Since Z_{ij} can take on any integer value in the range $0,1,\ldots,C-1$, we can easily allow the parameters in (21) to depend on the value attained by D_{ij}. We let

$$\begin{aligned} \theta_k &= k\theta \\ \alpha_{i(k)} &= k\alpha_i \\ \beta_{j(k)} &= k\beta_j \\ (\alpha\beta)_{kl} &= kl(\alpha\beta) \end{aligned} \tag{22}$$

where $D_{ij} = (k,l)$. Models which incorporate such parameter restrictions are sometimes termed *homogeneous linear effect* log-linear models, and have been used by Goodman (1979), Clogg (1982b), and others (see Agresti, 1983, 1984) when modelling categorical variables with ordinal categories.

Model (21), incorporating the parameter constraints given by (22), has exactly $2g$ parameters, one θ, $(g-1)$ α's, $(g-1)$ β's, and a single interaction, $(\alpha\beta)$. This simplified model is very similar to p_1, and is easily fit with GLIM 3.77. We now turn our attention to parameter estimation and model goodness-of-fit.

If we assume that the dyads $\{D_{ij}\}$ are statistically independent, then the log likelihood function for model (21) given data $\mathbf{z}^*$, depends on the following sufficient statistics, which are simple two-dimensional margins:

$$z^*_{++kl}/2 = \text{number of dyads at strength } (k,l), \qquad (k,l = 0,1,\ldots,C-1)$$

$$z^*_{i+k+} = \text{number of choices made by actor } i \text{ at strength } k \qquad (i = 1,2,\ldots,g;\ k = 0,1,\ldots,C-1)$$

$$z^*_{+jk+} = \text{number of choices made of actor } j \text{ at strength } k \qquad (j = 1,2,\ldots,g;\ k = 0,1,\ldots,C-1).$$

When $C = 2$, as with p_1, these sufficient statistics are three of the four two-dimensional margins of the $\mathbf{z}^*$ array (in-degrees, out-degrees, number of mutual dyads). The fourth two-dimensional margin, the [12] margin, $\{z^*_{ij++}\}$, has all entries equal to unity (for $i \neq j$), and is the sufficient statistic associated with the λ parameters. We note that we have seen arrays such as these statistics before, in the context of the Burt matrices that arise during the multiple correspondence analysis of relational data.

We use standard theory for maximum likelihood estimation of unknown parameters in log-linear models for cross-classified, categorical data (see Fienberg, 1980, Appendix II). This leads us to set these sufficient statistics equal to their expected values under the model. Further, as noted by Meyer (1981, 1982), Wasserman and Weaver (1985), and Wasserman and Iacobucci (1986), fitting model (21) to $\mathbf{z}$ can be accomplished simply by fitting the "no-three-factor interaction" log linear model to the $g \times g \times C \times C$ $\mathbf{z}^*$ array. This model, in Fienberg (1980) notation, is [12] [13] [14] [23] [24] [34]. This implies that fitting p_1 is equivalent to "controlling" simultaneously for all two-way interactions that arise from the four variables (the actor and partner variables, and the sending and receiving variables). General comments on this model-fitting approach, and details on how to calculate maximum likelihood parameter estimates (which are simple functions of the estimated log linear model u-terms), can be found in Wasserman and Weaver (1985). We will return to our discussion of this model-fitting process in the next section.

If we simplify model (21) by utilizing the reduced set of parameters given by (22), then standard iterative proportional fitting can not be used. However, as mentioned, other algorithms can be used, particularly the approach based on Fisher's scoring method as implemented by GLIM 3.77 (see McCullagh and Nelder, 1983, section 2.5). Specific details on how to fit this model and special cases of it (found by setting various parameters equal to zero or by incorporating other kinds of restrictions) can be found in Wasserman and Iacobucci (1986; see Table 3).

Define $\hat{\pi}_{ijkl}$ as the expected value of z^*_{ijkl} calculated by assuming some log linear model for the π's. The likelihood-ratio statistic for the model is

$$G^2 = 2\sum_{i<j}\sum_{k,l} z^*_{ijkl}\log(z^*_{ijkl}/\hat{\pi}_{ijkl}) \tag{23}$$

which has degrees of freedom equal to $(C-1)g(g-1)$, which is the number of independent estimated model parameters. One can test a wide variety of null hypotheses (concerning restrictions on parameters) by calculating conditional likelihood-ratio test statistics, which are found by subtracting G^2's. The null model must be a special case of the alternative model, derived by placing the restrictions on the alternative model parameters.

As mentioned above, there are many special cases of model (21), found by allowing some of the parameters to equal zero or by placing restrictions on the parameters, such as the restrictions in equations (22). There are also many ways of generalizing this model, most of which are designed for data sets containing either additional relational variables or information on the attributes of the actors. We now briefly describe these generalizations, with the goal of eventually being able to relate this methodology to correspondence and canonical analysis.

We have described methods to analyze single relational data consisting of information on the sociometric relation that exists between the actors. Suppose that we also have attribute information consisting of individual measurements on the actors. For example, we may know their sex, place of residence, financial characteristics, or age. We can assume that all actors that have identical values on the actor attribute variables also have equal parameter values. If we use sex of actor as an attribute variable, then if we number the male actors $1, 2, \ldots, g_1$, and the female actors $g_1+1, g_1+2, \ldots, g$, then we can use this additional information in our model by assuming that

$$\begin{aligned} \alpha_{1(k)} &= \alpha_{2(k)} = \cdots = \alpha_{g_1(k)} = \alpha^{(1)}_{(k)} \\ \beta_{1(k)} &= \beta_{2(k)} = \cdots = \beta_{g_1(k)} = \beta^{(1)}_{(k)} \end{aligned} \tag{24a}$$

for the first subgroup of actors, the males, and

$$\begin{aligned} \alpha_{g_1+1(k)} &= \alpha_{g_1+2(k)} = \cdots = \alpha_{g(k)} = \alpha^{(2)}_{(k)} \\ \beta_{g_1+1(k)} &= \beta_{g_1+2(k)} = \cdots = \beta_{g(k)} = \beta^{(2)}_{(k)} \end{aligned} \tag{24b}$$

for the second subgroup, the females. Equalities (24a) and (24b) hold for all strengths k, $k = 0, 1, \ldots, C-1$. Attribute variables have been called *structural parameters* by Blau (1977). In a generalization of this methodology to ego-centric networks, Frank, Lundquist, Wellman, and Wilson (1986) refer to parameters that depend on actors or actor attributes as *compositional* and parameters that depend solely on the relational information as *structural*. In brief, through the use of attribute variables, a researcher can partition the network actors exhaustively into S mutually exclusive subgroups by placing all actors with equal values on the attribute variables into the same subgroup. Thus, all actors in the sth subgroup have equal compositional parameters, $\{\alpha^{(s)}_{(k)};\ k = 0, 1, \ldots, C-1\}$ and $\{\beta^{(s)}_{(k)};\ k = 0, 1, \ldots, C-1\}$. Remember, that if the relational variable takes on values from the set of the first C positive integers, then we can simplify the compositional parameters so that they no longer depend on the strength of the choice k (see equation 22).

Wasserman and Anderson (1987) discuss the substantive significance of the use of compositional parameters, and define the concept of *stochastic equivalence* as a generalization of Lorrain and White's classic notion of *structural equivalence* (see Lorrain and White, 1971; White, Boorman, and Breiger, 1976; Arabie, Boorman, and Levitt, 1978; and especially, Holland, Laskey, and Leinhardt, 1983). Through the use of stochastic or structural equivalence, a researcher tries to place g actors in S blocks, which is the standard classification problem. The connection between models for actor equivalence (such as blockmodels) and correspondence analysis was first noted by Wasserman and Anderson (1987). We comment more on this connection later in this section.

An equally important, but alternative approach to stochastic blockmodelling can be found in Wang and Wong (1987), who have generalized the ideas of Breiger (1981) and Fienberg, Meyer, and Wasserman (1985). These authors have taken the basic p_1 model for binary data and have added blockmodel parameters that measure how likely it is that actors choose others inside or outside their particular subgroup or block. These models contain both actor and subgroup parameters (an idea first discussed and used by Fienberg, Meyer, and Wasserman, 1985) and allow a researcher to test for equal subgroup choice tendencies. We feel that this approach is of great value for social network analysis.

This model of discrete-valued, single relational data has been generalized in other ways. The interested reader can consult Wasserman and Galaskiewicz (1984), Fienberg, Meyer, and Wasserman (1985), Iacobucci and Wasserman (1987), Wasserman (1987), and Weaver and Wasserman (1986). Recently, there has been interest in methods for the study of dyadic interactions over time. Most of this interest has surfaced in the psychology literature, particularly in **Psychological Bulletin**. Noteworthy is the research by Gottman (1979a, 1979b), Allison and Liker (1982), Wampold and Margolin (1982), Dillon, Madden and Kumar (1983), Budescu (1984), Wampold

(1984), and Feick and Novak (1985). Wasserman and Iacobucci (1987) and Iacobucci and Wasserman (1988) show how generalizations of the models described here can be applied to this study of *sequential*, or relational data observed over time. These authors show how one can posit predictive models that allow a researcher to model one relation as a function of other relations and actor attributes.

The extension of these ideas to rectangular relational data or two-mode networks is not difficult to present. We generalize the **Z** array to be of size $g \times h$, where g is the number of actors in the first or sending set (the corporations) and h is the number of actors in the second or receiving set (the nonprofits). Note that when subscripting the entries of **Z**, $i \in$ sending set and $j \in$ receiving set. Consequently, there is no direct connection between the (i,j)th and (j,i)th elements of the array. Since actors of the second set can not *relate to* actors of the first—they can only *be related to*—we do not need to study the bivariate $\{D_{ij}\}$ random variables. Our models can be placed directly on the probabilities of the Z_{ij}'s.

Again assuming that the elements of **Z** can take on any value from 0 to $C-1$, we define $\nu_{ijk} = \log P\{Z_{ij} = k\}$, for $k = 0,1,2,\ldots,C-1$. We postulate the following log linear model

$$\nu_{ijk} = \lambda_{ij} + \theta_k + \alpha_{i(k)} + \beta_{j(k)} \tag{25}$$

subject to the constraints that insure probability distributions,

$$\sum_{k=0}^{C-1} \exp\{\nu_{ijk}\} = 1$$

for all dyads. We have insurance that these constraints hold by the presence of the $\{\lambda_{ij}\}$ parameters in model (25). Note that this model is considerably simpler than model (21), since actor j can not "choose" actor i; consequently, the notion of reciprocity is not meaningful. To fit this model, we rearrange the entries in **z** into a three-dimensional, $g \times h \times C$ array in which the first two variables refer to the two actors in the dyad, and the third to the strength of the relationship.

5.3. Relationship between p_1 and Correspondence Analysis

We have presented the mathematics of corrrespondence analysis, multiple correspondence analysis, and log linear models for relational data. The question before us is how are these methodologies related? Let us remember that (at a minimum) we have three (with two-mode networks) or four (with one-mode networks) variables which must be analyzed simultaneously. Thus, the simplest network data set requires MCA, rather than CA. In general, we will assume that we have Q variables, which include both relational variables (Q_R in number) and attribute and actor variables (Q_A) where $Q_R + Q_A = Q$. We have restricted attention to single relational networks, so that Q_R is one for two-mode networks and two for one-mode networks (since we must code both relational variables in the dyad).

There are two key points. The first is that MCA can be performed by applying CA to a Burt matrix, **B**. As we have mentioned, Burt matrices contain all two-way margins among the network relational and attribute variables. These two-way margins are exactly the sufficient statistics for the standard log linear model for relational data, p_1. This connection is very interesting and will be exploited here. The second point is that CA can be viewed statistically, by adopting an appropriate model for the logs of the probabilities of a two-way contingency table. These models should be applicable to two-way tables formed by manipulations of the multivariate network data sets. We comment on these two points here.

Consider first a Burt matrix for a multivariate relational data set. This matrix contains all two-way interactions among the Q variables in the data set. If we were to fit a p_1-type model to this network, and were estimating our model parameters with the maximum likelihood technique, then we would be calculating a set of fitted values from the model that satisfied a set of constraints determined by the maximum likelihood equations. It is well known (see Fienberg, 1980) that these constraints are based on the sufficient statistics for the model parameters, and require that the fitted values have margins that are exactly equal to the margins represented by the sufficient statistics. For p_1, these margins are the two-dimensional margins of the original $\mathbf{Z}^*$ matrix. But; these margins are also present in the Burt matrix for the data set, and as we have pointed out, a MCA of the data set is equivalent to a simultaneous bivariate analysis of all pairs of variables. This is one way (indeed, a very useful way) of viewing the relationship between p_1 and correspondence analysis.

We can also state that a CA of **B** yields sets of scores (the number of such sets is equal to one less than the smallest a_q) for each variable. Further, because of Goodman's research, we can view these scores as the components of approximate multiplicative interactions between pairs of variables. For example, consider one of the actor attribute variables (with a_1 levels) and a relational variable (with C categories, as usual). For simplicity, we will number the attribute variable 1, and the relational variable, 2. Models (24) and (25) state that the parameters $\alpha_{(k)}^{(s)}$ reflect the tendency for actors in subgroup s ($s = 1,2,\ldots,a_1$) to choose partners at strength k ($k = 0,1,\ldots,C-1$). CA tries to approximate these interactions with multiplicative interactions of the form $x_{1s}x_{2k}$, where $\mathbf{x}_{1s}$ is the first CA score for the sth subgroup of attribute variable #1 and $\mathbf{x}_{2k}$ is the first CA score for level k of the relational variable (as can be clearly seen in model 15). More saturated CA models would approximate the log linear model interactions with sums of such products.

Thus, one can take CA scores from a CA of **B**, or an MCA of **M**, and multiply them together to obtain approximate multiplicative interactions. We will demonstrate and illustrate this idea with our example, which follows shortly.

The second point that we want to make concerns the variables in the relational data set and the fact that they can be arranged in many different ways. At the beginning of this section, we stated that these variables are of two types: *compositional*, referring to actors and their attributes, and *structural*, the variables giving the relations among the actors. We take the $\mathbf{Z}^*$ matrix, which contains Q variables and hence, is Q dimensional, and rearrange it into a two-way matrix where the row variable consists of the products or the cross-classification of the Q_A compositional variables, and the column variable, the Q_R structural variable(s). Following Gilula and Haberman (1986b), we can correspondence analyze this rearrangement by fitting RC correspondence analysis models, and use contrasts of the parameters from these models to obtain CA scores for the original variables. This is the approach taken by Faust and Wasserman (1987).

The table to be correspondence analyzed, which we will call $\mathbf{Z}^{**}$, is two-way, of size $L_1 \times L_2$, where the L's are the products of the number of levels of the compositional and structural variables, respectively. A correspondence analysis of $\mathbf{Z}^{**}$ is equivalent to fitting a saturated RC correspondence analysis model to the original data that statistically controls for all the interactions among the compositional variables and all the interactions among the relational variables. The interactions between these two sets are then modeled using model (15). One can take the scores from the rows and the scores from the columns and plot them, take contrasts of them, and make tables of them to better understand the important relationships between the two types of variables. Our continued analysis of our example will illustrate this further.

5.4. Example Revisited Yet Again

To conclude this paper, we return to the larger set of data consisting of 75 corporations and the 67 nonprofits which received contributions from them. We look first at a model using a single corporation attribute and a single nonprofit attribute, and then consider a model with more than one attribute for both corporations and nonprofits. In each case we compare three different approaches: multiple correspondence analysis, as described in section 4, statistical models for relational data (p_1 and its relatives), and regular (two-way) correspondence analysis focusing on the relationship between the compositional and the relational structure in the network. The arrays for these analyses are $\mathbf{B} = \mathbf{M}'\mathbf{M}$, $\mathbf{Z}^*$, and $\mathbf{Z}^{**}$, respectively.

The variables we use are identical to those used in the multiple correspondence analyses described above, except we have collapsed some of the categories of the relational variable (to make the computations easier). Our donation variable now has three levels: 1 = no donation (previously level 1), 2 = previous levels 2–6, and 3 = previous levels 7–9. We also (in some analyses) consider models which collapse the 9 nonprofit activity categories to just two categories, primarily to distinguish between media nonprofits and others (2 levels). We note that these simplifications were suggested by the multiple correspondence analyses of these data reported earlier; e.g., see Figures 4a and 4e. These simplifications help in fitting some of the larger statistical models by reducing the sizes of the tables.

For the multiple correspondence analyses reported here, we analyzed Burt matrices, as we did in the earlier MCAs. We focussed on two tables gleaned from the entire set of variables in this data set. First, we looked at a table that included one corporate and one nonprofit attribute, and the single relational variable. For these MCAs, we obtained three sets of scores (since we are using a Burt matrix there are no scores for dyads). These scores are scaled to have mean of zero and unit variance (weighted by the reciprocal of the variable relative frequencies) within each set. We next looked at a table that had two attributes for the corporations (pretax income and the degree to which corporate boards are linked to prestigious individuals) and two attributes for the nonprofit agencies (the activity of the nonprofit agency and the degree to which it is recognized, and preceived as offering essential and outstanding services). This MCA gave us five sets of scores. We chose not

to look at the other corporate attribute variable (corporate industry) since earlier analyses showed it was less important.

The next type of analyses we performed were correspondence analyses of the tables highlighting compositional and relational structure. These two-way tables are $\mathbf{Z}^{**}$ arrays. A $\mathbf{Z}^{**}$ array is constructed so that the columns are the levels of the relational variable (the three levels of donations) and the rows are the combinations or cells in the cross-classification of the two (or four) attribute variables. Correspondence analysis of this array gives two sets of scores, one set for the levels of the relational variable (the structure) and one set for the composition, the $a_1 \times a_2$ combinations of categories of the attribute variables (or $a_1 \times a_2 \times a_3 \times a_4$ for the analysis including four attribute variables). The two sets of scores can be scaled to have means of zero and unit variances. We can examine the row scores in a two- or four-way table, and using contrasts, remove main effects for the two (or four) variables. We demonstrate below.

Lastly, we fit independent dyadic choice models. Since our network is two-mode, we fit model (25) using assumptions in 24a and 24b that parameters are constant within subgroups of actors and partners defined by the attribute variables. The alpha parameters reflect the tendencies of the corporations to send donations at various levels, and the betas, the tendencies of nonprofits to receive donations at various levels. For the first table, we looked at a single attribute for the actors (corporations) and a single attribute for the partners (nonprofits). There are $g_1 = a_1$ corporation subgroups and $g_2 = a_2$ nonprofit subgroups. For the table with 4 attributes, we have $g_1 = a_1a_2$ corporation subgroups, and $g_2 = a_3a_4$ nonprofit subgroups. We fit the no three-factor interaction log linear model [12] [13] [23] to each $\mathbf{Z}^*$ array to obtain the fitted values and estimated parameters associated with these p_1-type models.

5.4.1. The Example With One Corporate and One Nonprofit Attribute

The first table we studied included one of the corporate attribute variables and one of the nonprofit attribute variables. The most interesting data set with just two attribute variables had the corporate prestige attribute variable ($a_1 = 3$ levels) and the nonprofit activity attribute variable ($a_2 = 9$ levels), and of course, the corporate donation relational variable (which was recoded to have 3 levels). We first fit model (25) to the $3 \times 9 \times 3$ $\mathbf{Z}^*$ array, and concluded that the nine activities fell neatly into two or three groups which distinguished the media organizations from the other nonprofits. So, we recoded this attribute variable just to make this distinction. Thus, $a_2 = 2$ in subsequent analyses.

The α and β parameter estimates are given in Table 8. These two-way tables of estimates are constrained to have zero row and column sums. One can see that low prestige corporations tend to donate at low levels and high prestige corporations at high levels. Non-media nonprofits are more likely to get low levels of donations while media organizations (as we have already noted) are likely to get high levels. None of these findings are particularly surprising. To understand the interaction between the actors and the relational variable better, we can turn to our correspondence and multiple correspondence analyses.

TABLE 8
p_1 parameter estimates for prestige × activity × donation data set

		Donation 1	2	3
$\hat{\alpha}$'s	Prestige 1	3.213	2.050	−5.263
	2	−0.899	−0.776	1.675
	3	−2.315	−1.274	3.589
		Donation 1	**2**	**3**
$\hat{\beta}$'s	not medial	1.765	−0.404	−1.361
	medial	−1.765	0.404	1.361

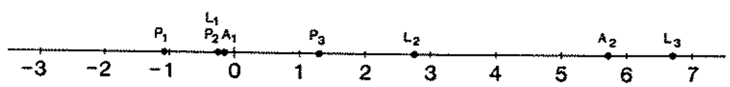

FIGURE 5. MCA scores for prestige (P_1, P_2, P_3) × activity (A_1, A_2) × donation level (L_1, L_2, L_3) data set.

We note that the number of sets of CA or MCA scores must be no greater than one less than the smallest number of categories of the variables. Since $a_2 = 2$, there is only one set of scores in the multiple correspondence analyses of this table. This is advantageous, since the smaller the dimension, the easier it is to study the results. We can place the scores on a one-dimensional axis and easily see how they relate to each other.

We constructed a Burt matrix and do a MCA on the prestige × activity × donation data set. The matrix was square, with $L = 3 + 2 + 3$ (the total number of levels) rows and columns. The scaled scores are: Prestige = −1.088, −0.311, 1.303; Activity = −0.175 (Not media), 5.701 (Media); and Donation = −0.305, 2.735, 6.701. These scores are plotted in Figure 5. One can quickly see that media organizations (A_2) receive donations at the highest levels (L_3) and the new information that high prestige corporations make donations at medium (L_2) or high (L_3) levels. The other categories are all associated and lumped together near 0 or −1. These findings support the p_1 model results *and* have the advantage that they are much easier to calculate.

Lastly for this table, we constructed a composition by structure $\mathbf{Z}^{**}$ array and obtained CA scores for the cross-classification of the attribute variables ($3 \times 2 = 6$ levels) and the three levels of the structure variable. Unlike the MCA just discussed, this CA yields two sets of scores, since there are three columns of the matrix. We obtain the first set of scores −0.281, 2.411, and 8.599, which are remarkably close to the MCA donation variable scores. The first eigenvalue explains roughly 80% of the inertia so that the second set of scores (which appears to basically reflect just the third donation level) is not very important.

For the composition variables in this CA, we present the first set of scores in Table 9. We see the strong effect of the media organizations, particularly at the highest prestige level. To better understand this two-way array, we can extract main effects for the two attribute variables (by finding row and column sums, centered to

TABLE 9
Correspondence analysis scores from the composition × structure analysis of the prestige × activity × donation data set

	Activity: Not media	Media
Prestige 1	−0.246	1.386
2	−0.124	1.869
3	0.128	4.003

TABLE 10
Approximate p_1 parameter estimates based on CA scores

Approximate $\hat{\alpha}$'s		Donation 1	2	3
	Prestige 1	2.311	0.698	−3.009
	2	1.146	0.346	−1.492
	3	−3.456	−1.044	4.501
Approximate $\hat{\beta}$'s		**Donation 1**	**2**	**3**
	not media	5.328	2.279	−7.606
	media	−5.328	−2.279	7.606

have zero means). We calculate the prestige main effect scores as −0.599, −0.297, and 0.896, and the activity main effect scores as −1.250 and 1.250. It is difficult to compare these scores directly to the MCA scores since the two approaches use different scalings. However, the overall impression one obtains from the CA is very much the same as that from the MCA. The advantages of the CA are two: first, one can usually extract more eigenvectors since one combines the categories of the attribute variables; and secondly, one controls for the interactions among the attribute/composition variables with this approach.

The last thing that we did was to approximate the alpha and beta p_1 parameter estimates using the CA scores. If we take the rescaled donation relational variable scores, and the prestige main effects reported above, we can take their cross-product to obtain a 3×3 table. We then centered this table to have zero row and column sums. This recentered table should approximate the $\hat{\alpha}$'s shown in Table 8. We can do similar calculations using the donation scores and activity main effects to approximate the $\hat{\beta}$'s. The approximations are shown in Table 10. The approximated $\hat{\alpha}$ estimates in Table 10 are quite close to the estimates in Table 8. The approximated $\hat{\beta}$'s have too much variation, which may be due to the fact that the activity attribute variable has only two levels. It appears that approximate parameter estimates based on CA scores could serve as proxies for estimates obtained by fitting dyadic independence models. However, the accuracy of this approximation will have to be determined by future research efforts.

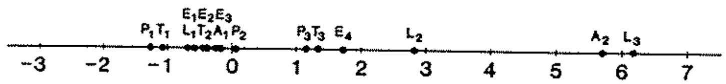

FIGURE 6. MCA scores for prestige (P_1, P_2, P_3) × pretax income (T_1, T_2, T_3) × nonprofit essential variable (E_1, E_2, E_3, E_4) × nonprofit activity (A_1, A_2) and donation level (L_1, L_2, L_3) data set.

5.4.2. The Example With Two Corporate and Two Nonprofit Attributes

We conclude this section with an analysis of the five variable data set, which includes four attribute variables and the donation relation. Remember that the four attributes are corporate prestige ($a_1 = 3$ levels), corporate pretax income ($a_2 = 3$ levels), elite opinions of how essential the nonprofit is ($a_3 = 4$ levels), and nonprofit activity (recoded to have $a_4 = 2$ levels). Since a_4 is equal to 2, there will be just a single set of MCA scores. These scores are shown in Figure 6. One immediately sees that the pretax income and prestige variable are not nearly as interesting as the other two attributes. All scores are monotonically increasing with their indices, which is a desirable property since the categories of the variables are ordered. Once again, the media organizations stand out. In addition, the nonprofits viewed as essential have a high score, near the second score for the donation variable. This figure clarifies the relationships among the 5 variables.

Other comparisons of these three approaches using different combinations of attribute variables confirms our observations that p_1 parameter estimates can be well approximated. These approximations are cross-products of scores for attribute variables with scores for the relational variables.

6. CONCLUSIONS

We have shown in this paper how correspondence analysis and its generalization, multiple correspondence analysis, can be used to understand the structure of relational data. It appears that even complex network data sets, containing many attribute variables, can be successfully studied with these new techniques. We also discussed statistical approaches to correspondence analysis, and compared these approaches to models specifically designed for relational data. There are many similarities between CA and these log linear models which should help network researchers in their statistical studies of social networks.

Although in this paper we have primarily examined two-mode network data, social network data often consist of a single set of actors. The approach we have outlined in this paper is also applicable to data gathered from such systems. As mentioned, the **Z** sociomatrix for such data is square. An example of a one-mode network, which is part of data set gathered by Galaskiewicz on the Minneapolis/St. Paul local grants economy, is based on acquaintanceship among donation officers in corporations. These data were collected by asking the donation officer in each of the corporations to examine a list of other firms in the area and indicate whether he or she knew anyone in the firm who was responsible for corporate giving. These data form a four-way structure: ID of the corporation of the donation officer making a choice, ID of the corporation of the donation officer receiving a choice, the value of the acquaintance relation from the first officer to the second (z_{ij}, which is

either 0 indicating no acquaintances or 1, in the event that there are acquaintances) and the choice from the second officer to the first (z_{ji}). Note that this relation is not symmetric, since actor i may report the acquaintance but not actor j, and vice versa. Thus, z_{ij} need not equal z_{ji}. In general, symmetric relations are not common in sociometric studies, particularly those reporting friendship. However, data are occasionally "symmetrized" to force off-diagonal elements of **Z** to equal their transposes. For example, a researcher may define a link only if both actors report that it is present.

There are six two-way margins that can be constructed from the four-way system. To illustrate, we use the example introduced above, although these tables arise from any one-mode, single relational study. The six are:

1. The ID of the donation officer expressing acquaintanceship by the ID of the donation officer receiving a choice; in general a $g \times g$ table.
2. The ID of the donation officers making choices by the strength of choices they make, in general a $g \times C$ table; with binary relations (as above), this table is $g \times 2$.
3. The ID of the donations officers making choices by the strength of choices they receive; a $g \times C$ table.
4. The ID of the donations officers receiving choices by the strength of choices they receive; a $g \times C$ table.
5. The ID of the donations officers receiving choices by the strength of choices they make; a $g \times C$ table.
6. The strength of choices given by the strength of choices received, in general a $C \times C$ table; with binary data, a 2×2 table.

Notice that there is some redundancy in this set of tables. Since the identities of the actors (donation officers) are the same whether they are initiating or receiving choices, tables 2 and 5 are identical, as are tables 3 and 4. These six tables are the two-way margins of the special four-dimensional contingency table which is used to fit the models described in section 5 of this paper.

When we have a one-mode network, our relational system, and thus our multiple indicator matrix, requires four variables to unambiguously describe the dyadic relation: the ID of the actor, the ID of the partner, the level of the relation from actor to partner, and the level of the relation from partner to actor. Since actors and partners are from the same set there are $g(g-1)/2$ dyads or $g(g-1)$ pairs. The multiple indicator matrix for a one-mode network will have $g(g-1)$ rows, since it is necessary to include each dyad twice (once from the perspective of the actor and once from the perspective of the partner), and $g + g + C$ columns.

Perhaps the most interesting aspect of this methodology is the ability to understand both the composition and structural/relational aspects of either one-mode or two-mode network data. Extending the ideas described here to multiple relational data will allow the researcher to study how similar the relational variables are, while controlling for either the dyads themselves or the attributes of the actors. We look forward to such extensions.

ACKNOWLEDGMENT

Research supported by NSF Grants #SES84-08626 (to the University of Illinois) and #SES83-19364 (to the University of Minnesota), and a postdoctoral traineeship awarded to the second author by the Quantitative Methods Program of the Department of Psychology, University of Illinois, funded by ADAHMA, National Research Service Award #MH14257.

We are grateful to Dawn Iacobucci for her research assistance and comments on this paper, to John T. Daws for research assistance, and to David R. Holtgrave and to Lawrence Hubert for comments on earlier versions. Several referees and editors made valuable suggestions. We also want to thank Zvi Gilula and Shelby J. Haberman for providing us with FORTRAN correspondence analysis computer programs.

Requests for copies of this paper should be addressed to the first author at the Department of Psychology and Department of Statistics, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820.

REFERENCES

Agresti, A. (1983) A survey of strategies for modelling cross-classifications having ordinal variables. *Journal of the American Statistical Association* **78:** 184–198.

Agresti, A. (1984) *Analysis of Ordered Categorical Data.* New York: Academic Press.

Allison, P. D., and Liker, J. K. (1982) Analyzing sequential categorical data on dyadic interaction: a comment on Gottman. *Psychological Bulletin* **91:** 393–403.

Andersen, E. (1980) *Discrete Statistical Models with Social Science Applications.* Amsterdam: North-Holland.

Arabie, P., Boorman, S. A., and Levitt, P. R. (1978) Constructing blockmodels: how and why. *Journal of Mathematical Psychology* **17:** 21–63.

Benzecri, J.-P. (1969) Statistical analysis as a tool to make patterns emerge from data. In: S. Watanabe (ed.): *Methodologies of Pattern Recognition,* 35–74. New York: Academic Press.

Benzecri, J.-P. (1973) *L'Analyse des Donnees. Tome 2: L'Analyse des Correspondances.* Paris: Dunod.

Berkowitz, S. D. (1982) *An Introduction to Structural Analysis: The Network Approach to Social Research.* Toronto: Butterworths.

Beum, C. O., and Brundage, E. G. (1950) A method for analyzing the sociomatrix. *Sociometry* **13:** 141–145.

Blau, P. (1977) A macrosociological theory of social structure. *American Journal of Sociology* **83:** 26–54.

Breiger, R. L. (1974) The duality of persons and events. *Social Forces* **53:** 181–190.

Breiger, R. L. (1981) Discussion of Holland and Leinhardt. *Journal of the American Statistical Association* **76:** 51–53.

Breiger, R. L., Boorman, S. A., and Arabie, P. (1975) An algorithm for clustering relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology* **12:** 328–283.

Budescu, D. V. (1984) Tests of laggeed dominance in sequential dyadic interaction. *Psychological Bulletin* **96:** 402–414.

Burt, C. (1950) The factorial analysis of qualitative data. *British Journal of Psychology, Statistical Section* **3:** 166–185.

Burt, R. S. (1980) Models of network structure. *Annual Review of Sociology* **6:** 79–141.

Burt, R. S., and Minor, M. J. (eds.) (1983) *Applied Network Analysis.* Beverly Hills: Sage Publications.

Clogg, C. C. (1982a) Using association models in sociological research: some examples. *American Journal of Sociology* **88:** 114–134.

Clogg, C. C. (1982b) Some models for the analysis of association in multiway cross-classifications having ordered categories. *Journal of the American Statistical Association* **77:** 803–815.

Clogg, C. C. (1986) Statistical modelling versus singular value decomposition. Comment on Goodman. *International Statistical Review* **54:** 284–288.

Coleman, J. S., and MacRae, D. (1960) Electronic processing of sociometric data for groups up to 1,000 in size. *American Sociological Review* **25:** 722–727.

Davis, A., Gardner, B. G., and Gardner, M. R. (1941) *Deep South: A Social Anthropological Study of Caste and Class.* Chicago: University of Chicago Press.

de Leeuw, J. (1973) *Canonical Analysis of Categorical Data.* Unpublished doctoral dissertation, University of Leiden, Leiden, The Netherlands. (Published in 1984 by the DWSO Press, Leiden.)

Deutsch, S. B., and Martin, J. J. (1971) An ordering algorithm for analysis of data arrays. *Operations Research* **19:** 1350–1362.

Deville, J.-C., and Saporta, G. (1983) Correspondence analysis, with an extension towards nominal time series. *Journal of Econometrics* **22:** 169–189.

Dillon, W. R., Madden, T. J., and Kumar, A. (1983) Analyzing sequential categorical data on dyadic interaction: a latent structure approach. *Psychological Bulletin* **94:** 564–583.

Faust, K., and Wasserman, S. (1987) Correspondence analysis and log linear models for relational data. Paper presented at the Sunbelt VII Social Network Conference, Clearwater, Florida, February, 1987.

Feick, L. F., and Novak, J. A. (1985) Analyzing sequential categorical data on dyadic interaction: log-linear models exploiting the order in variables. *Psychological Bulletin* **98:** 600–611.

Fienberg, S. E. (1980) *The Analysis of Cross-Classified Categorical Data.* 2nd Ed. Cambridge, MA: The MIT Press.

Fienberg, S. E., and Wasserman, S. (1981) Categorical data analysis of single sociometric relations. In: S. Leinhardt (ed.): *Sociological Methodology 1981.* San Francisco: Jossey-Bass.

Fienberg, S. E., Meyer, M. M., and Wasserman, S. (1985) Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association* **80:** 51–67.

Fisher, R. A. (1940) The precision of discriminant functions. *Annals of Eugenics* **10:** 422–429.

Frank, O. (1981) A survey of statistical methods for graph analysis. In: S. Leinhardt (ed.): *Sociological Methodology 1981,* 110–155. Edited by S. Leinhardt. San Francisco: Jossey-Bass.

Frank, O., Hallinan, M., and Nowicki, K. (1985) Clustering of dyad distributions as a tool in network modeling. *Journal of Mathematical Sociology* **11:** 47–64.

Frank, O., Komanska, H., and Widaman, K. F. (1985) Cluster analysis of dyad distributions in networks. *Journal of Classification* **2:** 219–238.

Frank, O., Lundquist, S., Wellman, B., and Wilson, C. (1986) Analysis of Composition and Structure of Social Networks. Unpublished Manuscript.

Freeman, L. C., White, D. R., and Romney, A. K. (1989) *Research Methods in Social Network Analysis.* Fairfax, VA: George Mason University Press.

Galaskiewicz, J. (1985) *Social Organization of an Urban Grants Economy: A Study of Business Philanthropy and Nonprofit Organizations.* Orlando, FL: Academic Press.

Galaskiewicz, J. (1987) The study of a business elite and corporate philanthropy in an United States metropolitan area. In: G. Moyser and M. Wagstaffe (ed.): *Research Methods for Elite Studies,* 147–165. London: Allen & Unwin.

Galaskiewicz, J., and Bielefeld, W. (1986) Organizational Non-Profit Response to Shifting Resource Markets: 1980–1984. Paper read at the XI World Congress of Sociology, August 18–22, New Delhi, India.

Galaskiewicz, J., and Wasserman, S. (1986) An approach to the study of organizational change. Paper read at the 1986 Annual American Sociological Association Meeting, August, 1986, New York City.

Galaskiewicz, J., and Wasserman, S. (1989) Mimetic and normative processes within an interorganizational field: An empirical test. *Administrative Science Quarterly* **34:** 454–480.

Gauch, H. G., Whittaker, R. H., and Wentworth, T. R. (1977) A comparative study of reciprocal averaging and other ordination techniques. *Journal of Ecology* **65:** 157–174.

Gilula, Z. (1986) Grouping and association in contingency tables: an exploratory canonical correlation approach. *Journal of the American Statistical Association* **81:** 773–779.

Gilula, Z., and Haberman, S. J. (1986a) Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association* **81:** 780–788.

Gilula, Z., and Haberman, S. J. (1986b) Analysis of multiway contingency tables by restricted canonical and restricted association models. Technical Report No. 48, Statistics Research Center, Graduate School of Business, University of Chicago.

Goodman, L. A. (1979a) Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* **74:** 537–552.

Goodman, L. A. (1979b) Multiplicative models for square contingency tables with ordered categories. *Biometrika* **66:** 413–418.

Goodman, L. A. (1981a) Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika* **68:** 347–355.
Goodman, L. A. (1981b) Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association* **76:** 320–334.
Goodman, L. A. (1981c) Three elementary views of log-linear models for the analysis of cross-classifications having ordered categories. In: S. Leinhardt (ed.): *Sociological Methodology 1981* San Francisco: Jossey-Bass.
Goodman, L. A. (1983) The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics* **39:** 149–160.
Goodman, L. A. (1984) *The Analysis of Cross-Classified Data Having Ordered Categories*. Cambridge, MA: Harvard University Press.
Goodman, L. A. (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.
Goodman, L. A. (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.
Gottman, J. (1979a) *Marital Interaction: Experimental Investigations*. New York: Academic Press.
Gottman, J. (1979b) Detecting cyclicity in social interaction. *Psychological Bulletin* **86:** 338–348.
Greenacre, M. J. (1981) Practical correspondence analysis. In: Vic Barnett (ed.): *Interpreting Multivariate Data* New York: John Wiley and Sons.
Greenacre, M. J. (1984) *Theory and Application of Correspondence Analysis*. New York: Academic Press.
Greenacre, M. J. (1986) SIMCA: A program to perform simple correspondence analysis. *Psychometrika* **51:** 172–173.
Guttman, L. (1941) The quantification of a class of attributes. In: P. Horst et al. (eds.): *The Prediction of Personal Adjustment*, 319–348. New York: Social Science Research Council.
Guttman, L. (1946) An approach for quantifying paired comparisons and rank order. *Annals of Mathematical Statistics* **17:** 144–163.
Guttman, L. (1959) Metricizing rank-ordered and unordered data for a linear factor analysis. *Sankhya: The Indian Journal of Statistics* **21:** 257–268.
Haberman, S. J. (1974) Log-linear models for frequency tables with ordered classifications. *Biometrics* **30:** 589–600.
Hayashi, C. (1980) Data analysis in a comparative study. In: E. Diday, L. Lebart, J. P. Pages and R. Tomassone (eds.): *Data Analysis and Informatics*, 31–52. Amsterdam: North Holland.
Heiser, W. J., and Meulman, J. (1983) Analyzing rectangular tables by joint and constrained multidimensional scaling. *Journal of Econometrics* **22:** 139–167.
Hill, M. O. (1974) Correspondence analysis: a neglected multivariate method. *Applied Statistics* **23:** 340–354.
Hill, M. O. (1982) Correspondence analysis. In: S. Kotz and N. L. Johnson (ed.): *Encyclopedia of Statistical Sciences*, 204–210. New York: Wiley.
Hirschfeld, H. O. (1935) A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society* **31:** 520–524.
Holland, P. W., and Leinhardt, S. (1977) Notes on the statistical analysis of social network data. Presented at the Advanced Research Symposium on Stochastic Process Models of Social Structure. Carnegie-Mellon University, Pittsburgh, PA.
Holland, P. W., and Leinhardt, S. (1979) *Perspectives in Social Network Research*. New York: Academic Press.
Holland, P. W., and Leinhardt, S. (1981) An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76:** 33–50.
Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983) Stochastic blockmodels: first steps. *Social Networks* **5:** 109–137.
Iacobucci, D., and Wasserman, S. (1987) Dyadic social interactions. *Psychological Bulletin* **102:** 293–306.
Iacobucci, D., and Wasserman, S. (1988) A general framework for the statistical analysis of sequential dyadic interaction data. *Psychological Bulletin* **103:** 379–390.
Iwatsubo, S. (1978) An optimal scoring method for detecting clusters and interrelations from multi-way qualitative data. *Behaviormetrika* **5:** 1–22.

Kadushin, C. (1966) The friends and supporters of psychotherapy: on social circles in urban life. *American Sociological Review* **31:** 786–802.
Katz, L. (1947) On the matric analysis of sociometric data. *Sociometry* **10:** 233–241.
Kendall, D. G. (1971) Seriation from abundance matrices. In: C. R. Hodson, D. G. Kendall, and P. Tautu (eds.): *Mathematics in the Archaeological and Historical Sciences*, 215–222. Edinburgh: Edinburgh Press.
Kendall, M., and Stuart, A. (1973) *The Advanced Theory of Statistics*, Volume 2, 3rd Edition. London: Griffin.
Knoke, D., and Kuklinski, J. H. (1982) *Network Analysis*. Beverly Hills, CA: Sage Publications.
Lancaster, H. O. (1966) Kolmogorov's remark on the Hotelling canonical correlations. *Biometrika* **53:** 585–588.
Lebart, L., Morineau, A., and Warwick, K. M. (1984) *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: John Wiley and Sons.
Levine, J. H. (1979) Joint-space analysis of 'pick-any' data: Analysis of choices from an unconstrained set of alternatives. *Psychometrika* **44:** 85–92.
Levine, J. H. (1984) *Atlas of Corporate Interlocks Volume I*. Hanover, NH: Worldnet.
Lorrain, F., and White, H. C. (1971) Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology* **1:** 49–80.
Marsden, P. V., and Lin, N. (1982) *Social Structure and Network Analysis*. Beverly Hills: Sage Publications.
Maung, K. (1941) Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children. *Annals of Eugenics* **11:** 189–223.
McCullagh, P., and Nelder, J. A. (1983) *Generalized Linear Models*. London: Chapman and Hall.
Meyer, M. M. (1981) *Applications and Generalizations of the Iterative Proportional Fitting Procedure*. Ph.D. Thesis, School of Statistics, University of Minnesota.
Meyer, M. M. (1982) Transforming contingency tables. *Annals of Statistics* **10:** 1172–1181.
Nelder, J. A., and Wedderburn, R. W. (1972) Generalized linear models. *Journal of the Royal Statistical Society*, Series A, **135:** 370–384.
Nishisato, S. (1978) Optimal scaling of paired comparison and rank order data: an alternative to Guttman's formulation. *Psychometrika* **43:** 263–271.
Nishisato, S. (1980) *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto: University of Toronto Press.
Noma, E. (1982a) Untangling citation networks. *Information Processing and Management* **18:** 43–53.
Noma, E. (1982b) The simultaneous scaling of cited and citing articles in a common space. *Scientometrics* **4:** 205–231.
Noma, E., and Smith, D. R. (1985) Scaling sociomatrices by optimizing an explicit function: correspondence analysis of binary single response sociomatrices. *Multivariate Behavioral Research* **20:** 179–197.
Payne, C. D. (1985) *The GLIM System Release 3.77: Generalized Linear Interactive Modelling Manual*. Oxford: The Numerical Algorithms Group.
Romney, A. K., and Boyd, J. P. (1986) Optimal dual scaling of the Sampson data. Paper presented at Sunbelt VI Social Network Conference, Santa Barbara, California, February, 1986.
Schriever, B. F. (1983) Scaling of order dependent categorical variables with correspondence analysis. *International Statistical Review* **51:** 225–238.
Simmel, G. (1955) *Conflict and the Web of Group-Affiliations*. New York: Free Press.
Simon, G. (1974) Alternative analyses for the singly ordered contingency table. *Journal of the American Statistical Association* **69:** 971–976.
Spath, H. (1980) *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. Chichester, ENG: Ellis Harwood.
Tenenhaus, M., and Young, F. W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* **50:** 91–119.
Torgerson, W. S. (1958) *Theory and Methods of Scaling*. New York: Wiley.
Tucker, L. R. (1964) The extension of factor analysis to three-dimensional matrices. In: N. Frederiksen and H. Gulliksen (eds.): *Contributions to Mathematical Psychology* New York: Holt, Rinehart, & Winston.
van der Heijden, P. G. M., and de Leeuw, J. (1985) Correspondence analysis used complementary to loglinear analysis. *Psychometrika* **50:** 429–447.

Wampold, B. E. (1984) Tests of dominance in sequential categorical data. *Psychological Bulletin* **96:** 424–429.

Wampold, B. E., and Margolin, G. (1982) Nonparametric strategies to test the independence of behavioral states in sequential data. *Psychological Bulletin* **92:** 755–765.

Wang, Y. J., and Wong, G. Y. (1987) Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association* **82:** 8–19.

Wasserman, S. (1987) Conformity of two sociometric relations. *Psychometrika* **52:** 3–18.

Wasserman, S., and Anderson, C. (1987) Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks* **9:** 1–36.

Wasserman, S., and Galaskiewicz, J. (1984) Some generalizations of p_1: External constraints, interactions, and non-binary relations. *Social Networks* **6:** 177–192.

Wasserman, S., and Iacobucci, D. (1986) Statistical analysis of discrete relational data. *British Journal of Mathematical and Statistical Psychology* **39:** 41–64.

Wasserman, S., and Iacobucci, D. (1988) Sequential social network data. *Psychometrika* **53:** 261–282.

Wasserman, S., and Weaver, S. O. (1985) Statistical analysis of binary relational data: parameter estimation. *Journal of Mathematical Psychology* **29:** 406–427.

Weaver, S., and Wasserman, S. (1986) RELTWO: interactive log linear model fitting for pairs of sociometric relations. *Connections: Bulletin of the International Network for Social Network Analysis* **9:** 38–46.

White, H. C., Boorman, S. A., and Breiger, R. L. (1976) Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology* **81:** 730–780.

Whiting, J. W. M., Romney, A. K., White, D. R., Burton, M., and Moore, C. (1986) Reanalysis of kinship and social structure. Paper presented at the 85th meeting of the American Anthropological Association, Philadelphia, PA.

NOTES FOR CONTRIBUTORS

Manuscripts should be typewritten on high quality bond paper with wide margins, double spaced throughout. Three copies should be submitted; each copy should include all figures and tables. Every effort will be made to maintain anonymity during the refereeing process. Therefore, the first page of the manuscript should contain the title, the author's name and institutional affiliation. At the bottom of the first page include the name and address of the author with whom the editor should correspond about the manuscript, and to whom proofs are to be sent. The second page should contain the title of the article and an abstract no longer than 120 words in length; the second page should not contain the author's name. Authors should include at the end of the abstract a list of six key words to be used for database word searches. The third page should contain the title and the beginning of the text; it too, should not contain the author's name.

In matters of style authors should follow the instructions of the *Publication Manual of the American Psychological Association* with the following exceptions: (1) The caution not to use equations is obviously inappropriate in this journal; (2) Do not use the abbreviations S, O, and E for subject, observer, and experimenter; (3) Do not underline symbols for they will normally be set in italics; (4) Please identify any unusual symbols in the margin at their first occurrence.

Figures submitted should be clear enough for direct reproduction. Redrawing, and author's alterations in excess of 10%, will be charged.

Authors should not submit articles to this journal if they are being considered elsewhere. It is a condition of the acceptance by the editor of a typescript for publication that the publishers acquire automatically the copyright in the typescript throughout the world. Every attempt will be made to render a prompt decision about publication.

Authors will receive page proofs for correction by air mail. Reprints can be ordered when proofs are returned.

Manuscripts should be submitted to the editor.

Patrick Doreian
Department of Sociology
University of Pittsburgh
Pittsburgh, PA 15260