SCISPACE
formerly Typeset

Open access • Journal Article • DOI:10.2307/270947

Canonical Analysis of the Composition and Structure of Social Networks

— Source link

Stanley Wasserman, Katherine Faust

Published on: 01 Jan 1989 - Sociological Methodology

Topics: Relational database, Correspondence analysis, Categorical variable and Network model

Related papers:

- Social Network Analysis: Methods and Applications
- Correspondence and canonical analysis of relational data
- Theory and applications of correspondence analysis
- An Exponential Family of Probability Distributions for Directed Graphs
- A Comparison of χ 2 Metrics for the Assessment of Relational Similarities in Affiliation Networks

Share this paper:

View more about this paper here: https://typeset.io/papers/canonical-analysis-of-the-composition-and-structure-of-392es1y7b7

Canonical Analysis of the Composition and Structure of Social Networks
Author(s): Stanley Wasserman and Katherine Faust
Source: *Sociological Methodology*, Vol. 19, (1989), pp. 1-42
Published by: American Sociological Association
Stable URL: http://www.jstor.org/stable/270947
Accessed: 11/06/2008 13:06

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=asa.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We enable the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

http://www.jstor.org

1

Canonical Analysis of the Composition and Structure of Social Networks

*Stanley Wasserman** *and Katherine Faust†*

Developing network models that allow for simultaneous analysis of actor attributes and network relational structure provides a challenge for network researchers. Such models would allow one to look at the characteristics of actors and partners in a network and at the patterns of social relations at the same time. In this paper, we show how recent developments in the statistical analysis of categorical data can be used to analyze the structure and composition of a wide variety of network or relational data. The techniques we use are canonical analysis and its relative, correspondence analysis. Both are methods for studying two-way cross-classifications

This paper was presented at the Sunbelt Social Network Conference, sponsored by the International Network for Social Network Analysis, San Diego, February 1988, and at the annual meetings of the Psychometric Society, Los Angeles, June 1988. Research was supported by NSF grant SES84-08626 to the University of Illinois. We are grateful to John T. Daws and Michael E. Walker for computing assistance, Ferdinanda Marcic for research assistance, Carolyn Anderson, Clifford Clogg, Lawrence J. Hubert, Peter van der Heijden, and an anonymous reviewer for comments on the paper, and Zvi Gilula and Shelby Haberman for providing us with FORTRAN canonical and association analysis computer programs. We also thank Joseph Galaskiewicz, Linton Freeman, and Jeffrey Johnson for allowing us to analyze social network data that they gathered.

*University of Illinois
†University of South Carolina

that several researchers have recently applied to network data. We have found that these techniques allow network analysts to easily study aspects of social structure that, until recently, have been too difficult because of the reliance on "heavy" computational methods. We begin with an overview of network data sets and then show how such data can be organized into a special two-dimensional patterned matrix. We then present the canonical analysis model and show how it can be applied to the network pattern matrices, which highlight the relationship between actor and dyad attributes (the network composition) and relational structure. Canonical analysis allows us to test hypotheses about several interesting types of relationships, including associations among actor and partner characteristics, actor/partner attributes (i.e. network composition) and relational patterns (i.e. network structure), and multiple relations.

The focus of this paper is on social network data, which give measurements on the actors and relations of a network. We will present models designed to study several interesting social network hypotheses that are not easily investigated using other approaches. Our models focus on the association between actor characteristics (network composition) and the strengths and patterns of social relationships (network structure). Hypotheses about the association between these two kinds of variables are of great interest to social networkers looking either at the impact of network structure on outcome measures pertaining to individuals or at the influence of individual characteristics on the development of network structure. These research questions require the incorporation of variables on both network structure and composition into a single model. The models that we will describe here also allow the researcher to test statistical hypotheses about network structure.

We will discuss a variety of network data sets and will present a unified approach to the analysis of network data. The approach is based on a canonical analysis of a special two-way array constructed from the network data. First, we give an overview of the variety of network data sets, so that the reader can appreciate how diverse network studies can be. Following this overview, we will discuss how to construct a *network pattern matrix*, the two-dimensional matrix that highlights the relationship between the dyads, or pairs of network actors, the information that may be

measured on the actors themselves, and the relational information that is measured on the dyads. The statistical technique that we will then apply to these design matrices will be described at length, and we will illustrate it with several examples.

1. NETWORK DATA

Network data sets can contain a variety of information, but at a minimum, the data set must consist of a set of actors or *nodes* (to use the graph-theoretic network perspective) who send and receive relational information to and from the other actors. The classic network data set contains a square *sociomatrix*, which gives information on whether each of the g actors in the network "send relational information to" (or simply "choose") the other $(g-1)$ actors. For example, a network may exist among a set of major publicly held corporations in a large metropolitan area. We may have information on whether a specific corporation purchases goods or services from any of the other corporations, and we may record this transaction as a binary variable: $X_{ij} = 1$, if corporation i has a business transaction with corporation j, and 0, otherwise. Such a sociomatrix is a $g \times g$ binary array, with zeros along the diagonal (since it is conventional to ignore intra-actor information). Many methodologists have developed mathematical and statistical techniques to study such data sets. There is a very rich literature of methods for square, binary sociomatrices. Much of this literature has appeared in the serial *Sociological Methodology* (Holland and Leinhardt 1975; Schwartz 1976; Bonacich 1977; Erickson 1978; Bonacich and McConaghy 1979; Wasserman 1979; Fienberg and Wasserman 1981; Frank 1981; Winship and Mandel 1983; Wu 1983; Mizruchi et al. 1986). It will be convenient to view these methods as basically of three types: graph-theoretic (Harary, Norman, and Cartwright 1965), relational or algebraic (see Wu 1983 and Winship and Mandel 1983 for thorough reviews), and statistical (e.g., Holland and Leinhardt 1975; Fienberg and Wasserman 1981; see also the review by Frank 1981). Our focus here is on statistical methods, but our goals are quite similar to those researchers employing either the algebraic or the graph-theoretic approaches: Specifically, we search for the *social structure* of the network by

looking for cliques or blocks of equivalent actors and by understanding the associations among the measured relations. Our approach moves considerably beyond the description of social structure by incorporating characteristics of network actors into derived data condensations designed to summarize the inherent structure of the network. Further, our approach works in the reverse too: These methods allow social structural patterns to explain variation in actor characteristics.

We are primarily interested in network data sets that contain more information than just a single relation existing among a single set of g actors. We might have more than a single relational variable. Or we may have information about the actors themselves. For example, our corporations may not only have business transactions with each other but also send information or exchange officers to serve on each other's board of directors. Each of these relations may be coded into a sociomatrix. If there are R measured binary relations, then we will let $\mathbf{X}_i$ denote the ith sociomatrix, and we will define $\boldsymbol{X}$ as the *super-sociomatrix*, of size $g \times g \times R$, containing the R two-dimensional sociomatrices as its layers. Multirelational data sets, consisting of the information in $\boldsymbol{X}$, can be effectively analyzed with the blockmodeling techniques of White, Boorman, and Breiger (1976) and Breiger, Boorman, and Arabie (1975), with the statistical methods of Fienberg, Meyer, and Wasserman (1985), Wasserman (1987), and Iacobucci and Wasserman (1987), or perhaps even with the *stochastic blockmodeling* methods of Holland, Laskey, and Leinhardt (1983), Wasserman and Anderson (1987), and Wang and Wong (1987), which combine the first two approaches. One interesting type of multirelational network data set considers the relations as different realizations of a longitudinal network process, evolving over time. For example, we could record money transactions among the corporations for 1980, 1984, and 1988 and define $\boldsymbol{X}$ as the $g \times g \times 3$ super-sociomatrix containing this sequentially measured relation (see Iacobucci and Wasserman 1988; Wasserman and Iacobucci 1988).

Network data sets may contain information on transactions or flows from one set of actors to a second, completely different set. The first set of actors, the *senders*, which we will label $\mathcal{G}$, consists of actors that are defined to send relational information to a second set of actors, $\mathcal{H}$, containing the *receivers*. The receivers, because of the nature of the relational variables measured in the

data set, cannot send information to either the senders or other receivers. Wasserman, Faust, and Galaskiewicz (1989) refer to such networks as *two-mode*, since the row actors (the first mode) in the sociomatrix are not the same as the column actors (the second mode). Specifically, with a single relational variable, we define X_{ij} = 1 if actor $i \in \mathcal{G}$ "relates to" actor $j \in \mathcal{H}$, and 0 otherwise. Assuming that set $\mathcal{G}$ contains g actors and that set $\mathcal{H}$ contains h actors, the sociomatrix $\mathbf{X}$ is of size $g \times h$. It will be *rectangular*, rather than square, if $g \neq h$. Such matrices and their super-sociomatrix generalizations (if more than one relational variable is measured from the actors in $\mathcal{G}$ to the actors in $\mathcal{H}$) require sophisticated, nonstandard analytical methods. A good example can be found in Galaskiewicz (1987) and Galaskiewicz and Wasserman (1988), who studied how the corporations in the Minneapolis/St. Paul metropolitan area supported the nonprofit organizations in the area. The relational variables in their study (gathered by Galaskiewicz 1985) were the levels of monetary donations (coded to be discrete, on a scale from 1 to 9, rather than binary) for 1980 and 1984. The first mode of their two sociomatrices consisted of 96 corporations, and the second mode, 175 nonprofit organizations. This example also illustrates that one can study nonbinary, but still discrete (noncontinuous), relational variables using methods first proposed by Wasserman and Iacobucci (1986). Clearly, one or more of the relational variables can be discrete-valued in multirelational network data sets.

Lastly, one can have information consisting of actor or dyadic characteristics, i.e., nonrelational information in network data sets. In fact, such variables are standard social and behavioral science fare and are usually organized into familiar n (actors or subjects or units) by p (variables) arrays. For example, we may have information about the corporations, such as their size, influence in the community, profitability (as measured by reported pre-tax income), etc. Or we may be studying a collection of married couples and how well the spouses in a couple communicate with each other. If we also study couples that are in marital "difficulty," then we could easily measure a variable for dyadic marital satisfaction. Note that such a variable is a function of the couple rather than one of the spouses. This would produce a binary *dyadic* rather than *actor* attribute variable.

Social network analysis differs from the standard actor-by-

attribute approach of usual social and behavioral science, since the focus is primarily on the relational variables in network data sets. The introduction of actor or dyadic attribute information is important to network analysis, since researchers can use such attributes to understand the results of their network analyses. For example, why does this subset of actors choose that subset of actors? Is there something unique to these subsets that can be captured by studying how well the actor or dyadic characteristics model the partition of the actors into subsets? How associated are the attributes with the social structure found by analyzing all the relational variables? We have focused much of our statistical research in the past decade on methods to integrate relational and attribute information in analyses (see Fienberg and Wasserman 1981; Wasserman and Anderson 1987). The techniques described in this paper should further this integration.

Regardless of the variables measured on the actors and included in the network data set, the primary unit of the data set is the *dyad*. Relational variables, or simply *ties*, are defined for dyads. Network data sets certainly may include attribute variables, defined for either the individual actors or the dyads, but these variables are not required for many network analytic techniques. We have borrowed a very nice terminology from Frank et al. (1986) and Frank et al. (1988), who refer to the collection of relational variables as the *structure* of the network and to the individual attributes as the *composition*. Thus, the goal of a network analysis must be to understand the relationship between the network structure and the composition of the dyads and actors. Frank et al. (1986), who present one statistical method for such an analysis, state this goal quite well:

> Ideally, [statistical] models would explain how the attributes of individuals affect their likelihood of being linked or how the attributes are affected by existing ties. [The models] would relate this interaction to the kinds of network patterns which are prevalent in a social system. [P. 3]

To use our continuing example as an illustration, we would like to be able to predict why donative ties, of the various types and

perhaps for different periods of time, exist between the corporate actors and the nonprofit actors, using the composition variables such as size of corporation, the composition of corporate and nonprofit boards of directors, the prestige of the corporations and nonprofits within the community, etc. (see Galaskiewicz and Wasserman 1988 for such an analysis). Another example, taken from Wellman et al. (1987), focuses on the well-known East York (a suburb of Toronto, Ontario) study of community and support and shows how a multirelational network data set, containing many different types of ties (such as companionship, emotional aid, services, financial aid, and information) and even more attribute variables (age, marital status, sex, employment status, and education, to name just a few), can be modeled *in full* using sophisticated statistical methods. This dichotomy of network variables into these two types, structure and composition, is crucial to the analyses presented here and to the network pattern matrices that we will introduce below.

Complete, rather than piecemeal, analyses of complicated data sets such as these are certainly welcome in the network literature. Unfortunately, the methods these authors employ are quite complicated, utilizing logistic regressions and independent dyadic choice models (as described in the appendix of Galaskiewicz and Wasserman 1988). The primary purpose of this paper is to present an alternative method for the analysis of such complicated data sets and to illustrate how easily it can be applied to study the composition and structure of network data. This technique, *canonical analysis*, which is quite similar to the more well known method of *correspondence analysis*, will be described after we show how to organize complicated network data sets into two-way structures that highlight the relationship of the network composition to its structure.

2. NETWORK PATTERN MATRICES

We begin with $Q+R$ variables divided into two sets of variables: Q composition variables and R structure variables. A dichotomy of variables (here, into composition and structure) is quite common in data analysis, in which one frequently tries to predict one or more response variables as functions of a collection

of explanatory variables. We let $\mathscr{C}$ refer to the composition variables and $\mathscr{S}$ to the structure variables.

Following the terminology of Wasserman et al. (1989), we want to consider the variables measured on each of the $g(g-1)/2$ dyads (in the case of a square or one-mode network) or the gh dyads (in the case of a rectangular or two-mode network). We will view the N pairs of possible inter-actor relationships (where N is either $g(g-1)$ or gh) as the rows of a matrix and consider the variables that are measured on the N rows. To make this discussion easier, we will treat the two different types of network data separately, first describing the situation for two-mode networks, then for one-mode networks.

2.1. *Two-Mode Network Data*

We will assume that the Q attribute variables measured on the actors in a two-mode network can be split into Q_1 variables for the actors in $\mathscr{G}$ constituting the first mode and Q_2 variables for the actors in $\mathscr{H}$, the second mode. We first discuss the various states that the actors can fall into and then define $\mathscr{K}$ as the set containing the possible states for the dyads. These states depend on the variables in $\mathscr{C}$.

Consider the first mode, i.e., the actors in $\mathscr{G}$. We have Q_1 variables measured for them, so these actors have a value for each of these variables (which we implicitly assume are all categorical or discrete). We define k_q as the number of levels of the qth categorical attribute variable. For example, if we have a single attribute variable for actors in the first mode, then $Q_1 = 1$. If $k_1 = 2$ levels, then actors in the first mode will fall into one of two subgroups defined by levels of the variable. If we also have a second attribute variable for these actors, then $Q_1 = 2$. If $k_2 = 3$ levels (for example high, medium, and low prestige), then there are $k_1\ k_2 = 6$ subgroups defined by the combinations of the levels of the two actor attribute variables. We let K_{c1} be the product of the levels of these categorical attribute variables, across all Q_1 variables. To summarize the above example, $Q_1 = 2$, and $k_1 = 2$ and $k_2 = 3$. Thus, K_{c1}, the product of the number of levels associated with the sender composition variables, is 6. More generally, $K_{c1} = \Pi_q\ k_q$. We do the same for the second mode to obtain K_{c2}, the product of the number of levels of the attribute

variables associated with the receivers, the second mode of the network. In simpler terms, K_{c1} is the number of subgroups for the actors in the first mode as defined by the levels of the attribute composition variables, and K_{c2} is the number of subgroups for the second mode.

Of most interest to us are the states of the dyads. Each sending actor falls into one of the K_{c1} actor subgroups defined by the cross-classification of the Q_1 relevant sending attribute variables, and each receiving actor falls into one of the K_{c2} actor subgroups defined by the cross-classification of the Q_2 relevant receiving attribute variables. For example, suppose that $\mathcal{C}$ has four variables, two for each mode, so that $Q_1 = Q_2 = 2$. If each variable has two levels, then there are $2^2 = 4$ possible subgroups for the sending actors and $2^2 = 4$ for the receiving actors. Thus, the dyads themselves can fall into $4 \times 4 = 16$ possible dyadic states. We define $\mathcal{K}$ as the set consisting of all possible dyadic states, which for this example has 16 elements, denoted by states (1,1) . . ., (4,4), where the notation (r,s) implies that for the dyad (i,j), sending actor $i \in$ subgroup r, and receiving actor $j \in$ subgroup s. Clearly, in general, we will have $K = K_{c1}K_{c2}$ elements in $\mathcal{K}$ for two-mode network data. It is important to remember that the elements in $\mathcal{K}$ represent states for ordered pairs. The elements are all possible *pairs* of subgroups of the sending and receiving actors in the dyad.

Consider now the structure or relational variables. We will let $\mathcal{L}$ refer to the levels of the structure variables. For example, if we have a pair of structure variables, with l_1 and l_2 levels, respectively, then $\mathcal{L}$ can take on one of $L = l_1l_2$ values, all possible combinations of the two structure variables. In general, we will assume that there are R relations measured on our N units and that associated with relation r is a $g \times h$ sociomatrix $\mathbf{X}_r$, whose elements, measuring the strength of this type of tie from actor i to partner j, take on values from 0 to $l_r - 1$. With l_r denoting the number of levels of the rth structural variable, we define $L = \Pi_r l_r$ as the total number of levels associated with the structural variables. For binary relations, $l_r = 2$.

2.2. *One-Mode Network Data*

As with two-mode networks, we will assume that we have Q discrete attribute variables, with levels k_1, k_2, etc. The product

of these levels gives us the total number of possible subgroups for the actors. For a one-mode network, actors fall into one of $K_c = \Pi_q k_q$ actor subgroups defined by the combined levels of the Q attribute variables.

From these actor subgroups, we can form dyadic states. This task is simple, since the two actors in the dyad are from the same set and hence fall into the same collection of subgroups. We will let K denote the total number of states for the dyads, all possible pairs of actor subgroups defined by the composition variable(s). As mentioned above, these states are the elements of the set $\mathcal{K}$. For an example of a single attribute variable with three levels, $K = 3\times3 = 9$, and $\mathcal{K}$ has elements (1,1), (1,2), (1,3), (2,1), . . ., (3,3), where the notation (r,s) implies that for dyad (i,j), actor $i \in$ subgroup r, and $j \in$ subgroup s. Dyads in general can fall into $K = K_c^2$ states.

For the structure variables, we will let $\mathcal{L}$ refer to the levels of the structure variables. Each pair of actors constitutes a dyad. We will have two pieces of structural information for each relation: how the first actor in the dyad relates to the second, and how the second relates to the first. For one-mode networks, we will consider these two dyadic pieces of information (per relation) separately, so that we work with $N=g(g-1)$ pairs, on which the structure variables are defined. The structure variables are therefore treated exactly the same as with two-mode relational data sets. To reiterate, we will assume that there are R relations measured on our N units and that associated with relation r is a $g \times g$ sociomatrix $\mathbf{X}_r$, whose elements, measuring the strength of this type of tie from actor i to actor j, take on values from 0 to $l_r - 1$. With l_r denoting the number of levels of the rth structure variable, we define $L = \Pi_r l_r$ as the total number of levels associated with the structure variables.

Thus, in brief, we assume that we have Q composition and R structure variables. The composition variables define actor subgroupings, which are then paired to generate dyadic states, given by the elements of $\mathcal{K}$. The structure variables produce a cross-classified set of L possible structural states. These states are entries of the set $\mathcal{L}$, with elements $\ell \in \mathcal{L}$. Clearly, the elements of $\mathcal{L}$ are the entries in an R-dimensional contingency table that cross-classifies the structure variables. Fienberg et al. (1985) and Iacobucci and Wasserman (1987) give several examples of such contingency tables.

We will refer to these sets $\mathcal{K}$ and $\mathcal{L}$ frequently throughout this paper and will give several examples of them later in this section.

We should note that if we are interested in estimating individual actor effects, then the Q composition variables may include variables in the network data set that give the identification code of the individual actors. For example, in a one-mode network, the first composition variable may be simply the label associated with the individual actors. In a two-mode network, the first two composition variables could label the actors in $\mathcal{G}$ and the actors in $\mathcal{H}$. This is exactly the approach taken by Wasserman et al. (1989).

We now take the network data sets and the sets $\mathcal{K}$ and $\mathcal{L}$ defined by the composition and structure variables, respectively, and consider how to reorganize the data to facilitate canonical analyses. To begin, we describe one such reorganization advocated by earlier researchers interested in correspondence analysis.

2.3. *Response Pattern Matrices*

Wasserman et al. (1989) employ a *response pattern matrix* with N rows and several sets of columns. This matrix is an indicator array, with one set of columns for each of the composition and structure variables. Indicator matrices such as this are frequently used in correspondence analysis (see Greenacre 1984), a technique closely related to canonical analysis. For example, we can examine the network of donative transfers from corporations to nonprofit agencies. This example is a subset of the network data set analyzed by Galaskiewicz and Wasserman (1988) and Wasserman et al. (1989). The subset consists of 75 corporations and 67 nonprofit agencies in the Minneapolis/St. Paul area. So, there are $N = 75 \times 67 = 5{,}025$ dyads, which form the rows of the response pattern matrix for this example. The single structure variable is the amount of a donative transfer from each corporation to each nonprofit agency, coded on a scale from 1 to 3, so that $L = l_1 = 3$. We will consider two composition variables. For the corporations, the composition variable is the extent to which members of the various corporate boards of directors have ties to prestigious individuals in the community. This variable takes on one of three levels (high,

TABLE 1
Composition and Structure Pattern Matrices for the Example of Donative Transfers from Corporations to Nonprofit Agencies

	Composition									Structure		
	Corporation Prestige											
	Low			Medium			High					
	Nonprofit Essential Rating									Donation Level		
Dyad	Low	Med	High	Low	Med	High	Low	Med	High	1	2	3
(1,1)	0	0	0	0	0	0	1	0	0	1	0	0
(1,2)	0	0	0	0	0	0	0	0	1	0	0	1
.		.						.			.	
.		.						.			.	
.		.						.			.	
(1,67)	0	0	0	0	0	0	0	1	0	1	0	0
(2,1)	1	0	0	0	0	0	0	0	0	1	0	0
.		.						.			.	
.		.						.			.	
.		.						.			.	
(75,67)	0	0	0	0	1	0	0	0	0	0	1	0

medium, or low). For the second mode, the nonprofit agencies, we code the extent (high, medium, or low) to which the agency is thought to offer essential and outstanding services. Thus, $Q_1 = Q_2 = 1$, and $Q = 2$. This gives $K = 3 \times 3 = 9$ possible dyadic states. The set $\mathcal{H}$ then has nine elements: (1,1), (1,2), (1,3), . . ., (3,3), where (r,s) indicates that for a specific dyad (i,j), the corporation i is in level r of the corporate prestige variable, and nonprofit j is in activity s.

A portion of the response pattern matrix for this example is given in Table 1. The entire array contains $N = 5{,}025$ rows and $K_c + L_s = 9 + 3 = 12$ columns. Notice that the matrix is composed of two submatrices. The first submatrix consists of the first nine columns and contains the coded network composition variables. The second submatrix consists of the next three columns and contains the network structure variable. We will call these submatrices $\mathbf{M}_c$ and $\mathbf{M}_s$ for the composition and structure pattern matrices, respectively. We will refer to the entire array as $\mathbf{M}$. We now examine these two pattern matrices in more detail.

2.4. *Composition Pattern Matrices*

Consider first the pattern matrix for the network composition. We start with a set of Q categorical attribute variables. Recall that for one-mode networks, actors fall into one of $K_c = \Pi_q k_q$ subgroups defined by the combined levels of the Q attribute variables. Dyads then can fall into $K = K_c^2$ states.

For a one-mode network, the composition pattern matrix, $\mathbf{M}_c$, has one column for each entry in $\mathcal{H}$. There are $g(g-1)$ ordered pairs constituting the rows. Each row has a single 1, which indicates the compositional state of the pair of actors associated with that row. Thus, $\mathbf{M}_c$ is a dummy or indicator matrix coding the state of each dyadic pair; i.e., the subgroup to which actor i belongs *and* the subgroup to which actor j belongs are indicated by the row of $\mathbf{M}_c$ associated with the ordered pair (i,j).

Now consider a two-mode network. Senders and receivers will be described by different attribute variables. However, the composition pattern matrix is constructed in a way that is logically identical to the construction of the pattern matrix for a one-mode network. Recall that with Q_1 categorical sender attribute variables

and Q_2 categorical receiver attribute variables, there are K_{c1} sender subgroups and K_{c2} receiver subgroups. This implies that there are $K = K_{c1} \times K_{c2}$ states for the dyads defined by the composition variables and therefore K columns for the composition pattern matrix coding the dyadic states. These are gh dyads constituting the rows. As above, a single 1 in each row codes the compositional state for each dyad in the matrix, $\mathbf{M}_c$.

2.5. *Structure Pattern Matrices*

The example discussed above, a network of donations from corporations to nonprofit agencies, is a two-mode network. For this example involving a single structure variable, as for other two-mode networks, the structure pattern matrix, $\mathbf{M}_s$, requires only a single structure variable, since the direction of the relational link in a two-mode network is unambiguous. Therefore, $L = l$, and the network structure pattern matrix has as many columns as there are levels of the structure variable. A single 1 in each row of this matrix indicates the level of the relation from actor i to partner j in the dyad (i,j).

If we have more than one structure variable in a two-mode network, then we simply consider the cross-classification of the R structure variables. Such a cross-classification has $L = \Pi_r l_r$ cells, and we define L dummy variables to be the columns of the structure pattern matrix. These variables code which cell (in the R-dimensional contingency table) a specific dyad falls into. In this way, the structure pattern matrix can easily accommodate multiple relations, just as the composition pattern matrix could be generalized to networks with several categorical attribute variables. This will allow us to examine interesting higher-order network properties, such as multiplexity and exchange (see Fienberg et al. 1985; Iacobucci and Wasserman 1987). We note that this approach is exactly the same when we have a one-mode network.

In a one-mode network, however, actors are both senders and recipients of relational ties. For example, computer users may send messages to each other over a computer network, or corporations may purchase goods and services from each other. To code the structure in a one-mode network, we must consider actors as both initiators and recipients of relations. This leads us to analyze

ordered pairs of actors. There are $N = g(g-1)$ such pairs in a one-mode network. By coding relations for ordered pairs of actors, we can construct the structure pattern matrix for a one-mode network in the same way as for a two-mode network. Assuming just one structure variable, the structure pattern matrix for a one-mode network, then, has $L = l$ columns and $g(g-1)$ rows. We follow the rules given above for two-mode networks if we have more than one structure variable in a one-mode network.

The advantage of the response pattern matrix is that it calls attention to the fact that we are modeling dyads. Consequently, it also allows us to estimate parameters for individual dyads. However, the size of the matrix places a large computational constraint on analyses. We do not recommend using these matrices unless one is interested in individual actor or dyad effects. To estimate the parameters of the canonical analysis models we describe below, and to completely study the composition and structure of network data, we will need to introduce a different and more parsimonious array: a two-way *composition-by-structure response pattern matrix*, or simply, the *network pattern matrix*.

2.6. *The Composition-by-Structure Network Pattern Matrix*

Since canonical analysis is designed for two-way contingency tables, and since we are interested in the relationship between network composition, described by the set of Q composition variables, and network structure, described by the set of R structure variables, we define another two-way array to examine this relationship. We will let $\mathbf{Z}$ be the cross-classification of the K dyadic states defined by the combined composition variables and the L levels of the combined structure variables. The rows of this $K \times L$ array are the dyadic states defined by the levels of the network composition variables, and the columns are the cross-classification of the network structure variables. Each pair in the network accounts for one observation in this array, so the total frequency of the table is either $g(g-1)$ or gh for one- or two-mode networks, respectively.

The $\mathbf{Z} = (z_{kl})$ array, as we will label the network pattern matrix, is easily constructed from the composition and structure pattern matrices. If we let $\mathbf{M}_c$ be the $N \times K$ composition pattern matrix and $\mathbf{M}_s$ be the $N \times L$ structure pattern matrix, then we

TABLE 2
Composition-by-Structure Network Pattern Matrix for the Example of Donations from Corporations to Nonprofit Agencies

Composition		Structure		
	Nonprofit Essential	Donation Level		
Corporation Prestige	Rating	1	2	3
Low	Low	575	1	0
	Medium	501	2	1
	High	480	41	7
Medium	Low	585	9	6
	Medium	507	10	8
	High	450	53	47
High	Low	583	8	33
	Medium	486	23	37
	High	393	37	142

have $\mathbf{Z} = \mathbf{M}_c^t \mathbf{M}_s$ (where the superscript t indicates matrix transposition). This is the array we analyze in canonical analysis. Table 2 presents the $\mathbf{Z}$ array for the example of donations from corporations to nonprofit agencies.

This matrix is an "ordinary" contingency table and should be more familiar to readers than the response pattern matrices of Table 1. It is also small and thus easier to analyze. However, we feel that an examination of response pattern matrices is quite informative, since these arrays are based on the basic unit of a social network data set — the dyad. We should mention that a simple canonical analysis of $\mathbf{Z}$ is identical to a canonical analysis of an $\mathbf{M}$, except for a rescaling of the derived scores (see Gifi 1981; van der Heijden and de Leeuw 1985).

3. *CANONICAL ANALYSIS*

3.1. *The Model*

Canonical analysis, first used by several researchers in the 1930s (see Nishisato 1980; Gifi 1981; Greenacre 1984; Wasserman

et al. 1989), uses correlation coefficients to analyze the relationship between two polytomous variables, whose levels are arranged as the rows and columns of a two-way contingency table. The technique has recently become popular, primarily because of its relationship to correspondence analysis and to Goodman's association and correlation models for contingency tables. Goodman has called the canonical analysis model the *RC correlation model* (Goodman 1985). Gilula and Haberman (1986, 1988) and Goodman (1985) were the first to give a thorough treatment of maximum likelihood (ML) estimation of the parameters of the canonical analysis models, standard errors, and likelihood-ratio tests. Gilula and Haberman (1988) has the most relevance to our research, since these authors were the first to consider canonical analyses of multiway tables.

As we stated earlier, the array that we analyze is the $K \times L$ composition-by-structure (or simply, network) pattern matrix $\mathbf{Z} = \mathbf{M}_c'\mathbf{M}_s$. An examination of the network pattern matrix clearly shows that the elements of $\mathcal{K}$ (the dyadic states) and the possible values of $\mathcal{L}$ (the structure variables) are in one-to-one correspondence with the rows and columns of $\mathbf{Z}$.

To define a canonical decomposition of the network data, as viewed via $\mathbf{Z}$, we will let $P(k,\ell)$ be the probability that a dyad, chosen at random, belongs to level $k \in \mathcal{K}$ of the dyadic states (defined by the composition variables) and that the structure variables, defined for this dyad, take on value $\ell \in \mathcal{L}$. Further, we will let $P_c(k)$ be the probability that the dyad falls into state k, and we will let $P_s(\ell)$ be the probability that the structure variables take on value ℓ. We now define t as the smaller of $K-1$ and $L-1$. The *canonical decomposition* of $\mathbf{Z}$ is defined for constants $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_t \geq 0$, which are the canonical correlations themselves, and for functions $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$, defined on $\mathcal{K}$, and functions $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t$, defined on $\mathcal{L}$:

$$P(k,\ell) = P_c(k)\, P_s(\ell) \left[1 + \sum_{u=1}^{t} \rho_u\, x_u(k)\, y_u(\ell) \right]. \tag{1}$$

The elements of $\mathbf{x}_u$ and $\mathbf{y}_u$ have zero means (for all u), variances of unity (for all u), and

$$E[x_u(k)\, x_v(k)] = E[y_u(\ell)\, y_v(\ell)] = 0 \quad (\text{for } u \neq v). \tag{2}$$

The most important fact about these functions is that the ρ's are the correlations between these canonical functions:

$$E[x_u(k)\, y_u(\ell)] = \rho_u \quad \text{(for all } u\text{)}. \tag{3}$$

The correlations between the elements of $\mathbf{x}_u$ and $\mathbf{y}_v$ (if $u \neq v$) are zero. As long as the correlations are distinct and nonzero, then the $\mathbf{x}$'s and $\mathbf{y}$'s are unique, except for possible changes in sign. Replacing $x_u(k)$ by $-x_u(k)$ and $y_u(\ell)$ by $-y_u(\ell)$ clearly does not change (1).

This canonical decomposition of a two-way cross-classification has been known since the work of Hirschfeld (1935), Lancaster (1957, 1958) and Guttman (1959). We can view this canonical decomposition as a canonical correlation analysis of a two-way contingency table; specifically, if we ask what scores should be allocated to the rows and columns of the table to maximize the correlation coefficient between the linear combinations of the variables defined by the scores, then the answer is exactly $\mathbf{x}_1$ and $\mathbf{y}_1$. The maximal *canonical correlation* is ρ_1, and the remaining scores are orthogonal to (uncorrelated with) all the scores with lower indices, while still maximizing correlations. There can be only t canonical correlations, since the rank of the $\mathbf{Z}$ array is at most t. The linear combinations, defined by the $\mathbf{x}$'s and $\mathbf{y}$'s, are *canonical variables*. Lancaster (1957) proves the existence of this decomposition, where the table arises by grouping bivariate Gaussian frequencies. The best exposition of canonical decompositions and related theorems (that we have found) is in Kendall and Stuart (1973, ch. 33). A related result following from (1) is that Pearson's X^2 statistic is N (the total of the frequencies in the table) times the sum of squares of the canonical correlations. Thus, the squared canonical correlations are sometimes referred to as *components* or *partitions* of X^2. We also want to note that this decomposition has been termed the saturated RC canonical correlation model by Goodman (1985).

This canonical decomposition (1) uses all the available degrees of freedom in a two-way table and hence does not provide a very parsimonious model for the table. The most logical way to obtain special cases of (1) is to assume that the correlations for $u > w$, where w is less than t, are all zero; i.e., we assume that the smallest $t-w$ correlations are zero. In this case, we have only w nonzero correlations. We will refer to this special case as the $CA(w)$ *model*.

Goodman (1985) labels this model the *RC(w) canonical correlation model*, and Gilula and Haberman (1986, 1988) label it simply C_w. $CA(t)$ is the full, saturated canonical decomposition, and $CA(0)$ is complete independence of the composition and structure variables. Specifically, we have

$$P(k, \ell) = P_c(k)\, P_s(\ell) \left[1 + \sum_{u=1}^{w} \rho_u\, x_u(k)\, y_u(\ell) \right] (w < t), \qquad (4)$$

where the $\mathbf{x}$'s and $\mathbf{y}$'s still have all the properties of (2) and (3). Given N dyads and the composition and structure variable measurements for each, if we assume independent dyads (as is commonly done in the dyadic independence models of Holland, Leinhardt, Fienberg, and Wasserman et al. [see also Frank and Strauss 1986]), then we can estimate the parameters in $CA(w)$: (a) the w canonical correlations, (b) the w sets of K scores for the dyadic composition variables ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_w$, each of which has one score for each element of $\mathcal{K}$), and (c) the w sets of L scores for the structure variables ($\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_w$, each of which has one score for each level $\ell \in \mathcal{L}$, the levels of the structure variables). These parameters have considerable substantive interest. We will discuss how to interpret them in terms of standard network concepts later in this section. Clearly, there are many parameters in a $CA(w)$ model if w is large. One should choose the value for w carefully; fortunately, it is straightforward to test the goodness of fit of $CA(w)$ versus $CA(w^*)$, where $w < w^*$. Goodman (1985) and Gilula and Haberman (1986) discuss ML estimation of these parameters, and Gilula and Haberman (1986, sect. 4) give a thorough description of an ML algorithm. Gilula and Haberman justify significance tests for these models by the large-sample theory given in section 5 and Appendices A and B of their important paper.

3.2. *Related Models*

Before discussing estimation and testing of $CA(w)$, we want to point out the relationship between this class of canonical correlation models and related models for two-way cross-classified categorical data. First, we should mention that there is another class of models, known as *association models*, that is quite

similar to the canonical models discussed here. An association decomposition assumes that

$$P(k, \ell) = \exp\left[\theta + \alpha_r(k) + \beta_s(\ell) + \sum_{u=1}^{t} \lambda_u \mu_u(k) \nu_u(\ell)\right], \qquad (5)$$

where all the parameters are unknown. Conditions for the existence of these parameters are given in Goodman (1985) and Gilula and Haberman (1986). Unsaturated or parsimonious versions of this model arise by including fewer than all t terms in the sum in (5). Goodman (1985) refers to this unsaturated model as the *RC(w) association model*, where w is the number of terms in the sum, i.e., the number of nonzero lambdas and hence the number of components in the multiplicative interaction. Association models have been more popular than canonical models, perhaps because of their similarity to standard loglinear models, but also because the logarithms of the cross-product ratios arising from the two-way array depend on multiplicative differences of the μ's and ν's. Of course, the canonical models would appeal to researchers more interested in expressing relationships (in this case, between the composition and structure variables) using correlations. Many authors have discussed association models, and versions of these models appear in the well-known categorical data analysis texts of Fienberg (1980) and Agresti (1984) in the context of methods for discrete variables with ordered categories. Goodman (1981) and Haberman (1981) give the first thorough treatment of models such as (5), and Becker and Clogg (1989) present a very readable discussion of the relationship between the association and canonical models. Several researchers have described association models for multiway tables (Clogg 1982*a*,*b*; Agresti and Kezouh 1983; Becker 1985; Goodman 1986; and Becker and Clogg 1989).

The association models coincide with the $CA(w)$ models when w is either 0 or t. We can show that the association model for a specific w (denoted by A_w in Gilula and Haberman 1986) is approximately the same as $CA(w)$ when ρ_1 is small. Thus, we can think of the association models as approximations to the canonical models, and vice versa (see Goodman 1985). The models may give quite different fitted values, however, as several researchers, particularly Clogg (1986), have pointed out.

We prefer the canonical correlation models over the association models because of their similarity to *correspondence analysis*. Evidence of this similarity, and a good sociological discussion of these models, can be found in Goodman (1987). Further, the canonical models (and correspondence analysis, too) contain parameters that have straightforward sociological interpretations (more on this later). Goodman (1985, 1986) takes $CA(w)$ and reparameterizes it as

$$P(k, \ell) = P_c(k)\, P_s(\ell) \left[1 + \sum_{u=1}^{w} x'_u(k)\, y'_u(\ell) / \rho_u \right] \tag{6}$$

to obtain the $RC(w)$ *correspondence analysis model*. With the definitions

$$\begin{aligned} x'_u(k) &= \rho_u\, x_u(k)\ , \\ y'_u(\ell) &= \rho_u\, y_u(\ell), \end{aligned} \tag{7}$$

we can easily see the equivalence of (4) and (6). Further, either of these models can be written

$$\frac{P(k,\ell) - P_c(k)P_s(\ell)}{P_c(k)\, P_s(\ell)} = \sum_{u=1}^{w} x'_u(k)\, y'_u(\ell) / \rho_u\ , \tag{8}$$

which looks surprisingly like a spectral decomposition of a matrix consisting of the elements given by the left-hand side of (8).

Let us work with the full model in which $w = t$. From (8), simple algebra (see Goodman 1986, eqs. (2.9) – (2.10)) shows that (a) we can decompose X^2 into the sum of squares of the correlations (as we mentioned earlier), (b) each squared correlation can be partitioned further into the subcomponents $x'_u(k)\ P_c(k)$ or the subcomponents $y'_u(\ell)\ P_s(\ell)$, and (c) the x' parameters are weighted averages of the y parameters, with weights $P(k, \ell)/P_c(k)$, and the y' parameters are weighted averages of the x parameters, with weights $P(k, \ell)/P_s(\ell)$. These facts are exactly the objectives achieved by a correspondence analysis of a two-way contingency table, as described by many authors, including Wasserman et al. (1989). Correspondence analysis has been applied to sociometric data by several researchers in recent years. As the reader can see, the

$CA(t)$ model is equivalent to a complete correspondence analysis of the **Z** array. This is exactly why we prefer canonical models over association models. We comment further on this relationship below.

3.3. *Estimation, Testing, and Approximations*

As described in Goodman (1985, 1986) and in Gilula and Haberman (1986, 1988), the parameters of the $CA(w)$ model can be estimated using ML. We assume that we are given independent observations (k_n, ℓ_n), $n = 1, 2, \ldots, N$, i.e., a dyadic composition state and a value of the structure variables for each pair of actors. Each observation is recorded in the network pattern matrix, **Z**. We also assume that these independent observations follow a multinomial distribution, with probabilities governed by model (4). ML estimates of the parameters of $CA(w)$ [ρ_u, $u = 1, 2, \ldots, w$; $x_u(k)$, $u = 1, 2, \ldots, w$, and $k \in \mathcal{K}$; and $y_u(\ell)$, $u = 1, 2, \ldots, w$, and $\ell \in \mathcal{L}$] can be calculated using standard algorithms. We used the scoring algorithm described in Gilula and Haberman (1986), which can also give restricted ML estimates if constraints are placed on the model parameters (see section 3.4). We refer the reader to section 4 of Gilula and Haberman (1986) for algorithmic details.

Estimated asymptotic standard errors can also be calculated, allowing the researcher to place confidence intervals on the ρ's, x's, and y's (using normal distribution approximations). Likelihood-ratio tests are also possible, as shown by the large-sample theory presented in Gilula and Haberman (1986). Specifically, Pearson's chi-squared statistic,

$$X^2 = \sum \sum (z_{kl} - N\hat{P}_{kl})^2 / (N\hat{P}_{kl}), \tag{9}$$

and the likelihood-ratio chi-squared statistic,

$$G^2 = \sum \sum z_{kl} \log [z_{kl} / (N\hat{P}_{kl})], \tag{10}$$

where $\hat{P}_{kl}$ is the ML estimate of $P(k,\ell)$ under $CA(w)$, are asymptotically equivalent and have approximate χ^2 distributions with $(K-w-1)$ $(L-w-1)$ degrees of freedom. It is important to note that if any of the w ρ's is zero, then customary approximate distributions are not correct.

Of most interest to us are conditional likelihood-ratio tests that allow us to test goodness of fit of $CA(w)$ for specific values of

w. For example, we can fit both $CA(2)$ and $CA(1)$, and if $CA(2)$ provides an adequate description of the data (i.e., if the statistic (10) is not statistically large), then we can examine the difference of the likelihood-ratio chi-square statistics. This difference is approximately χ^2 with $(K-2)(L-2) - (K-3)(L-3) = K + L - 3$ degrees of freedom. Such conditional tests are very useful in data analysis.

We have alluded to the relationship between correspondence analysis and canonical analysis throughout this paper. Correspondence analysis (see Wasserman et al. 1989, and references therein) is a technique for the decomposition of a two-way contingency table. Computer programs for correspondence analyses of contingency tables are becoming available, such as Nishisato's (1986) DUAL3 and Greenacre's (1986) SIMCA. The technique is quite important in Europe, particularly in France, but it has made slower "in-roads" in the U.S.

Escoufier and Junca (1986) show that the scores from correspondence analysis, which yields sets of all t eigenvalues and associated eigenvectors for both the rows and columns of a two-way table, can be viewed as least squares approximations to the ML estimates from the $CA(w)$ model (for any w less than t). Goodman (1986) shows, as we have discussed in this section, that the saturated $CA(t)$ model is equivalent to a complete correspondence analysis of the two-way table. If we focus on fewer than all t eigenvalues arising from a correspondence analysis (say, just w of them), then we can view these w eigenvalues and their associated eigenvectors as approximate canonical analysis model parameter estimates. The importance of Escoufier and Junca's statement is that the results of a correspondence analysis can be used as approximate canonical analysis parameters. The availability of correspondence analysis software should allow network researchers to utilize the models presented here more frequently. One simply must keep in mind that the scores derived from correspondence analyses are just *approximate* canonical analysis x's and y's.

We should note that canonical analysis of social network data can be a valuable exploratory tool even when the assumptions necessary for ML estimation are not valid. In such circumstances, one can simply use the eigenvalues and eigenvectors mentioned above as least squares approximations to the ML estimates. One

should refrain from significance tests and use canonical analysis simply to summarize the data. In our specific application, the assumption of independent dyads could be questioned. If one is concerned substantively that this assumption might not hold, then "least squares canonical analysis" is recommended. Further research into the instances when "ML canonical analysis" is inferior to least squares canonical analysis is certainly needed. Stability of parameter estimates could be studied, using techniques such as jackknifing and bootstrapping.

3.4. *Restricted Canonical Analysis*

As we have mentioned, $CA(w)$ has $w + wK + wL$ (unconstrained) parameters: w canonical correlations, w sets of $\mathbf{x}$'s (each $\mathbf{x}$ is a vector of K components), and w sets of $\mathbf{y}$'s (each $\mathbf{y}$ is a vector of L components). Even with small w's, the number of parameters could be large. We can place linear constraints on these scores to obtain restricted canonical models and thereby simultaneously reduce the number of parameters to be estimated and increase the degrees of freedom for goodness-of-fit tests. Gilula and Haberman (1988) refer to such models as *restricted canonical models*, and we will denote them by $CA'(w)$.

For example, we might have a single structure variable. If the $L = l$ levels of this variable correspond to ordinal, equally spaced categories, then we might want to impose the constraint on the first $\mathbf{y}$ (the scores associated with this structure variable) that

$$y_1(\ell+1) - y_1(\ell) = y_1(\ell+2) - y_1(\ell+1), (\ell = 1, 2, \ldots, L-2). \tag{11}$$

In general, we can place constraints on any of the sets of scores. Further, these constraints are in the form of linear combinations of the $\mathbf{x}$'s and $\mathbf{y}$'s, which are then set equal to zero. For the above example, if $L=6$, then the constraints (11) consist of four linear combinations defined on the six scores for $\mathbf{y}_1$ (the first set of scores for the structure variable). The coefficients for the linear combinations constrain the spacings between levels ℓ and $\ell + 1$ to be equal to the spacings between levels $\ell + 1$ and $\ell + 2$. The coefficients are 1, -2, 1, 0, 0, 0 (for $\ell = 1$); 0, 1, -2, 1, 0, 0 (for

$\ell = 2$); 0, 0, 1, −2, 1, 0 (for $\ell = 3$); and 0, 0, 0, 1, −2, 1 (for $\ell = 4$).

The model $CA'(w)$ assumes that the model $CA(w)$, defined by (4), holds and that the **x**'s and **y**'s are constrained as in equations (2) and (3). We further assume that

$$\sum_{k=1}^{K} a_{knu} x_u(k) = 0 \quad (n = 1, 2, \ldots, n_{xu}; u = 1, 2, \ldots, w) \quad (12)$$

and that

$$\sum_{l=1}^{L} b_{lnu} y_u(l) = 0 \quad (n = 1, 2, \ldots, n_{yu}; u = 1, 2, \ldots, w). \quad (13)$$

The a's define the linear combination constraints placed on the **x**'s, and the b's define those placed on the **y**'s. There are linear combinations defined for all w **x**'s and **y**'s. Note that there are n_{xu} constraints for the uth set of scores for the **x**'s (and n_{yu} for the uth set of scores for the **y**'s) so that the number of constraints may depend on which set of scores we are studying. In the above example, there are $n_{yu} = 4$ constraints for $u = 1$, and $n_{y2} = n_{y3} = \ldots = n_{yw} = 0$. Remember that u denotes which set of scores is being studied, so that u can be anywhere from 1 to w, where w is the dimension of the model. Also remember that w must be between 1 and t. The coefficients $\{a_{knu}\}$ and $\{b_{lnu}\}$ must sum to zero for each of the constraints and for each of the sets of scores:

$$\sum_{k=1}^{K} a_{knu} = 0 \, (n = 1, 2, \ldots, n_{xu}; u = 1, 2, \ldots, w) \quad (14)$$

and

$$\sum_{l=1}^{L} b_{lnu} = 0 \, (n = 1, 2, \ldots, n_{yu}; u = 1, 2, \ldots, w) \, . \quad (15)$$

For each set of scores u, the n_{xu} vectors of coefficients $(a_{1nu}, a_{2nu}, \ldots, a_{Knu})'$ must be linearly independent (i.e., orthogonal), and the n_{yu} vectors of coefficients $(b_{1nu}, b_{2nu}, \ldots, b_{Lnu})'$ must also be linearly independent.

Gilula and Haberman (1986) discuss how to estimate the

parameters of $CA'(w)$, which are subject to the constraints (12) and (13). There are $\gamma = \sum_{u=1}^{w} [n_{xu} + n_{yu}]$ total constraints. If we make the same assumptions about the data and the ρ's as we do with $CA(w)$, then ML estimates of the parameters are asymptotically normal and X^2 and G^2 are asymptotically distributed as χ^2 random variables with $(K-w-1)(L-w-1) + \gamma$ degrees of freedom; i.e., we gain an additional γ degrees of freedom for testing goodness of fit because of the addition of γ constraints placed on our parameters. We can test whether the constraints are important by comparing the likelihood-ratio statistic for $CA(w)$ with the statistic for $CA'(w)$. As Gilula and Haberman (1986) show, the difference in these statistics is asymptotically distributed as a χ^2_γ random variable. Goodman (1985) also discusses estimation of RC canonical correlation model parameter subject to linear constraints.

One could argue that restricted canonical analysis models are not appropriate if ML canonical analysis is invalid because of restrictive statistical assumptions. We believe that the benefits of ML estimates greatly outweigh the costs here. The advantage of furthering our understanding of network composition and structure is very real; furthermore, the evidence (see Goodman 1986, 1987) shows that ML estimates and least squares estimates are virtually equal in most circumstances.

After a short discussion about why these canonical models should be useful to social networkers, we will illustrate their usefulness with two examples.

3.5. *Why Canonical Analysis for Social Network Data?*

In the last decade or so, interest in *social network* (sometimes called *structural*) approaches to social and behavioral science has grown in the mainline research community. However, the application of network concepts and models to more standard substantive and theoretical questions has been hindered by several problems. Two of the more important problems are that (a) there are few easily computable models that allow the researcher to incorporate network properties into more standard actor attribute studies, and (b) standard network methods are often limited to descriptive conclusions. Moving beyond mere description of structural properties is a necessary advance if network concepts are to continue to be

an important and influential part of social and behavioral science thinking. Canonical analysis using the models described here provides one resolution.

There are also several interesting social network hypotheses that may be addressed using the models we have described. We note the following as being of special interest to social network researchers. First, the canonical analysis model gives as standard output the canonical correlations (or square roots of the eigenvalues) $\rho_1, \rho_2, \ldots, \rho_t$. These parameters may be interpreted as the canonical correlations (or approximate correlations) between combinations of the row and column variables. In the analysis of network structure and composition, ρ_1 may be interpreted as the overall association between actor characteristics (composition) and relational strength (structure). Second, the canonical scores themselves have interesting network interpretations. Scores for values of the structural variables (**y**'s) give information about relational strength. Restricted canonical analysis models (which we have just described and which we illustrate in section 4.3) allow us to test hypotheses about spacing, linearity, and homogeneity of relational response categories (something other network approaches do only in an ad hoc way, if at all). Third, inspection of compositional and structural scores together gives insight into the relative strength of relations among pairs of actors with different combinations of characteristics or attributes. This interpretation may be used either in exploratory analyses (as advocated by Frank et al. 1986) or to make significance tests.

While canonical analysis provides a great advance in the kinds of network data and substantive questions that may be addressed, there are some limitations that should also be mentioned. First, canonical analysis relies on a number of assumptions that the researcher may be unwilling (or unable) to make. The model is assumed to be correct, and the dyads are assumed to be independent and identically distributed. But assumptions of models such as (1), (4), and (6) are usually considerably less severe than those of normal theory linear models. Second, since we have presented ML estimates of canonical analysis parameters, the parameter estimates are dependent upon good (i.e., computationally stable and efficient) ML algorithms. It is well known that these algorithms may misbehave when data arrays are sparse (when many cells in contingency tables such as **Z** are near zero). This may be a problem

for some network pattern matrices. In such instances, we recommend using correspondence analysis, which is not subject to this problem and which many have found to give good approximations to canonical analysis scores. On the whole, we believe these canonical analysis models will be quite useful to network researchers.

4. EXAMPLES

In this section we use the canonical analysis models described above to analyze two quite different network data sets. First, we look at a two-mode network consisting of donations from corporations to nonprofit agencies. We then turn to a one-mode network of electronic mail computer communications among researchers in an emerging scientific specialty.

4.1. *A Two-Mode Network of Donative Transfers*

In this example we look at a single relation and two attribute variables defined for a two-mode network. The example consists of information on contributions made from major publicly held corporations in the Minneapolis/St. Paul area to not-for-profit agencies. The data come from a long-term study conducted by Galaskiewicz (1985) of an urban grants economy. These data are described in detail in Galaskiewicz and Wasserman (1988) and in Galaskiewicz (1985). In the current example, we focus on a subset of 75 corporations and 67 nonprofit agencies. We look at a single relational variable: the level of the donation from each corporation to each nonprofit agency. This variable takes on three levels: (1) no donation, (2) a donation of up to $1,000, and (3) a donation of $1,000 or more.

In addition, we look at two attribute variables, one for corporations and one for nonprofit agencies. For corporations, we consider the degree to which corporations are linked to prestigious members of the local business elite. This variable is a combination of the degree to which prestigious community members know corporate officers or board members and the degree to which chief executive officers belong to the same clubs and cultural boards as the business elite. Details on how this variable was constructed can

be found in Galaskiewicz and Wasserman (1987). This variable takes on three levels: (roughly) low, medium, and high. For nonprofit agencies we code the degree to which corporate giving officers recognize the nonprofit agencies and perceive them as providing essential and outstanding services. This variable takes three levels: low, medium, and high. Again, details on how this measure was constructed can be found in Galaskiewicz and Wasserman (1988). So, using our notational system, $R = 1$ structure variable and $Q_1 = Q_2 = 1$ attribute variable for each mode, and $L = 3$ levels for the structure variable and $K = 3 \times 3 = 9$ dyadic states. We note that the form of the $\mathbf{M}_c$ and $\mathbf{M}_s$ matrices is shown in Table 1.

The network pattern matrix ($\mathbf{Z}$) for these data is presented in Table 2. Canonical analysis of this two-way array gives the scores presented in Table 3. These results show that donations at medium

TABLE 3
Canonical Analysis of Corporation Prestige, Nonprofit Essential Rating, and Donation Level

A. Structure

Donation Level		
1	2	3
−0.329	2.320	3.436

B. Composition

Corporation Prestige	Nonprofit Essential Rating		
	Low	Medium	High
Low	−0.876	−0.837	−0.187
Medium	−0.677	−0.581	0.693
High	−0.251	0.117	2.329

Canonical correlation = 0.364
$X^2 = 59.767$
7 *df*

and high levels are most likely to be made from high-prestige corporations to those nonprofit agencies perceived as offering a medium or high level of essential services and from medium-prestige corporations to those nonprofit agencies perceived as offering the most essential serivces. Absence of donations is most likely from corporations with low prestige to nonprofit agencies that are not perceived as offering essential services. An additional observation is that scores for both network composition variables are ordered, each within the levels of the other composition variable. Scores for levels of the structure variable are also ordered.

The interpretations of the estimates in Table 3 are made by simultaneously examining the scores for the levels of the composition variables (the **x**'s) and scores for the levels of the structure variables (the **y**'s). A glance at equations (1) and (8) shows that canonical analysis scores can be viewed as a multiplicative decomposition of the residuals from $CA(0)$, the model of complete independence. Using this interpretation and the unsaturated canonical analysis model (4), we can see that the estimated probability that a dyad falls into level ℓ of the structure variables and level k of the composition variables depends loglinearly on the product $x(k)y(\ell)$. When this product is positive, the estimated probability exceeds the expected probability calculated under an assumption of independence.

We note that the $CA(1)$ model does not fully account for the data, since $X^2 = 59.767$ with 7 degrees of freedom. One could consider fitting $CA(2)$ to these data, but this is a saturated model and would fit perfectly (remember that $t=2$ for these data). We also note that $CA(2)$ is identical to a full correspondence analysis. The best-fitting model might be a restricted version of $CA(2)$, $CA'(2)$. However, we have not fit such a restricted model to this network because the **Z** matrix is so sparse that the likelihood-ratio test statistics may not be distributed asymptotically as χ^2 random variables. Even though **Z** is sparse, the ML estimated scores from $CA(1)$ given in Table 3 appear to be good estimates because they closely match scores obtained from correspondence analysis of the same data. As we have emphasized, the correspondence analysis scores are approximations to ML estimates of the $CA(1)$ scores.

4.2. *A One-Mode Network of a Computer Conference among Scientific Specialists*

Our second example comes from a computer conference among researchers working in the emerging scientific speciality of social network research. These data were collected as part of a study of the impact of the Electronic Information Exchange System (EIES) housed at the New Jersey Institute of Technology. Fifty researchers interested in social network research participated. We focus here on the 32 people who completed the study. These researchers included a wide range of social scientists, among them, sociologists, anthropologists, and mathematicians. As part of the conference, a computer network was set up and participants were given computer terminals and access to a network for sending electronic mail messages to other participants. We note that this study was done prior to the widespread use of BITNET and other popular computer networks that are widely available to academics today; consequently, this study involved a novel way for researchers to communicate. For more details of this study, see Freeman and Freeman (1979, 1980) and Freeman (1986).

Of particular interest to us are network data arising from the study. As part of this project, the computer system recorded all message transactions, specifically, the origin and destination of the message, the day and time, and the number of lines in the message. Records were kept for several months. We therefore have a record of the number of messages sent from each participant to every other participant. An examination of these data reveal that while people differed in the degree of message exchange, most pairs who exchanged messages did so repeatedly. We therefore chose not to consider the number of messages but rather to define a binary variable indicating whether or not a message-sending relationship existed between each participant and every other participant. We use this sending relational variable as our first structure variable and note that it has $l_1 = 2$ levels. We constructed a second relational variable from a sociometric question. At the beginning and at the end of the project, participants were asked to fill out a questionnaire that included, among other things, the following network question. Each participant was asked to indicate, for every other participant, whether she/he (1) did not know the other, (2) had heard of the

TABLE 4
Network Pattern Matrix for the Example of Recognition (Citations), Messages, and Acquaintanceship on a Computer Network

		Messages							
		No				Yes			
Citations		Acquaintanceship							
Actor	Partner	Unknown	Heard of	Met	Friend	Unknown	Heard of	Met	Friend
Low	Low	67	11	38	15	6	4	58	41
	High	50	29	52	13	10	5	65	32
High	Low	59	20	47	11	20	6	63	30
	High	18	25	71	26	3	6	55	36

other but had not met him/her, (3) had met the other, (4) was a friend, or (5) was a close personal friend. We use the results of the second administration of the question, and because of the sparseness of responses in the last two categories, we combine levels (4) and (5) to create a category "friend." Our second structure variable has four levels ($l_2 = 4$).

The attribute variable we used indicates how well each participant is recognized by the social science community. This is measured by recording the number of citations of the researcher's work in the *Social Science Citation Index* for the year 1978 (when the research started). This variable is coded as binary, distinguishing between the most cited half and the least cited half of the participants. Thus, in this example, there are two relations ($R=2$) and one attribute variable ($Q=1$), and $L = 2\times4 = 8$ cells in the cross-classification of the two structure variables and $K = 2\times2 = 4$ dyadic states defined by the single binary composition variable.

The network pattern matrix for this example is presented in Table 4. Since the interpretation of the entire table is complicated, we first present analyses of several of the *marginal* tables. We note that the composition of the two members of a dyad is independent of the structural variable of sending messages, since $CA(0)$ yields $G^2 = 1.322$ and $X^2 = 1.321$ with 3 degrees of freedom. Participants with higher or lower recognition are neither more likely nor less likely to send messages to others of high or low recognition. However, from the second marginal table, we see that there is a relationship between the recognition of pair members and their acquaintanceship: $CA(0)$ fit to these data yields $G^2 = 59.254$, and $X^2 = 52.616$ with 9 degrees of freedom. Since $CA(0)$ does not adequately describe the association in this table, we fit $CA(1)$. Table 5 presents the $CA(1)$ canonical analysis of the $K = 2\times2$ composition by $L = 4$ structure marginal table of recognition by acquaintanceship. All participants, regardless of recognition, are likely to be unaware of others with relatively low recognition. Furthermore, participants with relatively high recognition are more likely to have heard of, to have met, or to be friends with other participants of high recognition.

We next looked at the association between just the two relations, ignoring the single composition variable. Acquaintanceship and sending messages are (clearly) not independent, since $CA(0)$

TABLE 5
Canonical Analysis of Recognition (Citations) and Acquaintanceship on a Computer Network

A. Structure

Acquaintanceship			
Unknown	Heard of	Met	Friend
−1.7862	0.9506	0.4656	0.5214

B. Composition

Citations		
Actor	Partner	
	Low	High
Low	−0.8595	0.0552
High	−0.7949	1.6484

Canonical correlation = 0.2121
G^2 = 9.577
X^2 = 9.577
4 *df*

has X^2 = 160.421, and G^2 = 171.149 with 3 degrees of freedom. The $CA(1)$ scores for analysis of acquaintanceship and recognition are presented in Table 6. Not surprisingly, people are more likely to send messages to others they have met or with whom they are friends.

Finally, a canonical analysis of the entire composition by structure network pattern matrix shows aspects of all of these relationships. First note that the independence model, $CA(0)$, does not fit these data (G^2 = 80.484 and X^2 = 77.269, with 21 degrees of freedom, $CA(1)$ does provide a reasonable model for these data (G^2 = 19.382 and X^2 = 19.056, with 12 degrees of freedom). The results are presented in Table 7.

TABLE 6
Canonical Analysis of Messages and Acquaintanceship on a Computer Network

Acquaintanceship			
Unknown	Heard of	Met	Friend
−1.3823	−1.2285	0.4665	1.1904

Messages	
No	Yes
−0.8928	1.1201

Canonical correlation = 0.4021
0 *df*

4.3. *Restricted Canonical Analysis of a Computer Network of Scientific Specialists*

In this section we briefly illustrate the use of restricted canonical analysis models to test substantively interesting hypotheses about the structure and composition of the computer network analyzed in the previous section. Looking at the scores for the $CA(1)$ model of these data (presented in Table 7), we see that we may be able to understand patterns of message sending without including all four levels of the structure variable of acquaintanceship; i.e., it might be possible to fit a more parsimonious model than $CA(1)$. Note that the scores for three levels of acquaintanceship, "heard of," "met," and "friend," within each level of message sending are quite close. Apart from whether one person is unknown to another, distinctions among levels of acquaintanceship may not be important.

From this observation, we considered a class of restricted models which assume that $CA(1)$ was appropriate, as well as various restrictions on the scores. These restrictions imply that except for actor and partner recognition (i.e., citations), there are few

TABLE 7
Canonical Analysis of Recognition (Citations), Messages, and Acquaintanceship on a Computer Network

A. Structure

	Acquaintanceship			
Messages	Unknown	Heard of	Met	Friend
No	−1.5572	0.9665	1.0076	1.3111
Yes	−1.7007	0.4004	−0.1042	0.0590

B. Composition

	Citations	
Actor	Partner	
	Low	High
Low	−0.8614	−0.0048
High	−0.7527	1.6694

Canonical correlation = 0.2378
G^2 = 19.382
X^2 = 19.056
12 *df*

distinctions among the levels of acquaintanceship. These restricted models can be tested by comparing $CA(1)$ to a $CA'(1)$ restricted canonical analysis model that sets some of the canonical scores equal to each other.

We examined the restricted model which assumes that apart from the scores for the "unknown" category, there is no distinction among the three levels of acquaintanceship. This model places the following restrictions on $CA(1)$:

$$y_1(i,2) = y_1(i,3) = y_1(i,4), \quad \text{for } i=1,2. \tag{16}$$

The coefficients for the linear combinations constraining the equality of these three levels of the structure variable are 0, −1, 1, 0, 0, 0,

0, 0; 0, 0, −1, 1, 0, 0, 0, 0; 0, 0, 0, 0, 0, −1, 1, 0; and 0, 0, 0, 0, 0, 0, −1, 1. There are two sets of coefficients (with eight elements in each combination), since we constrain acquaintanceship scores within each level of the structure variable "message sending." We can test this restricted model against the unrestricted $CA(1)$ model, which has $X^2 = 19.056$ and $G^2 = 19.382$, with 12 degrees of freedom. The restricted model yields $X^2 = 19.543$ and $G^2 = 19.958$, with 16 degrees of freedom. Clearly, this model fits the data. Further, the difference between $CA(1)$ and the restricted model $CA'(1)$, which contains these four restrictions, is statistically small; thus, it is valid to conclude that the categories of this variable can be collapsed into two.

This restricted model makes no distinction among the acquaintanceship variable categories "heard of," "met," or "friend" of a communication partner. Table 8 presents the $CA'(1)$ scores for this model. Other restricted models—for example, those equating levels of the other structure variable "message sending" or levels of the composition variable "citations"—do not fit well. Therefore, the levels of these variables should not be collapsed, as indicated by the restrictions shown in the table. These other variables, with all their unrestricted levels, are important in understanding the structure and composition of this communications network.

Other restricted models, which we do not present here, could be used to test more complicated network relationships. For example, we could examine the slope of the levels of one or more ordered categorical variables. We leave applications such as this for future research.

5. CONCLUSION

We have discussed the wide variety of social network data and have described how such data sets can be organized into special pattern matrices. These matrices consist of collections of indicator variables and code the values of the dyads, for which we have data, on two types of variables: composition variables and structure variables. From these matrices, one can easily construct a composition-by-structure network pattern matrix, which can then be analyzed with either canonical or correspondence analysis. One of the most fascinating features of relational data is that they can be viewed in

TABLE 8
Restricted Canonical Analysis of Recognition (Citations) Messages and Acquaintanceship on a Computer Network

A. Structure

Messages	Acquaintanceship	
	Unknown	Heard of & Met & Friend
No	−1.5678	1.0607
Yes	−1.6957	−0.0236

B. Composition

	Citations	
	Partner	
Actor	Low	High
Low	−0.9128	0.0318
High	−0.7267	1.6540

Canonical correlation = 0.2380
G^2 = 19.958
X^2 = 19.543
16 *df*

many different ways and modeled with many different strategies. Our network pattern matrices and the canonical analysis models that can be applied to them provide a novel, and in some ways unique, approach to social network analysis.

We have actually described two kinds of canonical analysis models: unrestricted $CA(w)$ models, very similar to correspondence analysis models, and restricted canonical analysis models, which we labeled $CA'(w)$. Because of the expositional tone that we have adopted in this paper, we have not devoted enough space to the latter models, nor have we shown in depth how they can be applied to specific examples. We hope that other researchers will see how

useful both classes of canonical analysis models can be and will devote their resources to a further exposition of the restricted models. We feel that the $CA'(w)$ models have tremendous utility in social network analysis, particularly when network data sets include multiple relational variables. With these models, one should be able to study the higher-order relational effects, such as exchange and multiplexity, that are so important to a complete understanding of the social structure of a closed group. At present, the analysis of such effects (as described by Fienberg et al. 1985; Iacobucci and Wasserman 1987) is hindered by the complexity of the independent dyadic choice models that contain such effects and the overpowering notational system of these models. However, canonical models are so straightforward and intuitive that the problems inherent in dyadic choice models may now be easily overcome.

REFERENCES

Agresti, A. 1984. *Analysis of Ordinal Categorical Data.* New York: Wiley.

Agresti, A., and A. Kezouh. 1983. "Association Models for Multi-Dimensional Cross-Classifications of Ordinal Variables." *Communications in Statistics, Series A: Theory and Methods* 12: 1261–76.

Becker, M. P. 1985. "Analysis of Discrete Data Using Log-Multiplicative Models and Other Log-Linear Models." Ph.D. diss., Department of Statistics, Penn State University.

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

Bonacich, P. 1977. "Using Boolean Algebra to Analyze Overlapping Memberships." Pp. 101–15 in *Sociological Methodology 1978*, edited by K. F. Schuessler. San Francisco: Jossey-Bass.

Bonacich, P., and M. J. McConaghy. 1979. "The Algebra of Blockmodelling." Pp. 489–532 in *Sociological Methodology 1980*, edited by K. F. Schuessler. San Francisco: Jossey-Bass.

Breiger, R. L., S. A. Boorman, and P. Arabie. 1975. "An Algorithm for Clustering Relational Data with Applications to Social Network Analysis and Comparison with Multidimensional Scaling." *Journal of Mathematical Psychology* 12: 328–83.

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

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

———. 1986. "Statistical Modelling Versus Singular Value Decompo-

sition." Comment on Goodman. *International Statistical Review* 54: 284–88.

Erickson, B. H. 1978. "Some Problems of Inference from Chain Data." Pp. 276–302 in *Sociological Methodology 1979*, edited by K. F. Schuessler. San Francisco: Jossey-Bass.

Escoufier, Y., and S. Junca. 1986. "Least-Squares Approximation of Frequencies of their Logarithms." Comment on Goodman. *International Statistical Review* 54: 279–83.

Fienberg, S. E. 1980. *The Analysis of Cross-Classified Categorical Data*. 2d ed. Cambridge, MA: MIT Press.

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

Fienberg, S. E., M. M. Meyer, and S. Wasserman. 1985. "Statistical Analysis of Multiple Sociometric Relations." *Journal of the American Statistical Association* 80: 51–67.

Frank, O. 1981. "A Survey of Statistical Methods for Graph Analysis." Pp. 110–55 in *Sociological Methodology 1981*, edited by S. Leinhardt. San Francisco: Jossey-Bass.

Frank, O., S. Lundquist, B. Wellman, and C. Wilson. 1986. "Integrating the Analysis of Structures and Attributes." Working Paper No. 86–11. Stockholm: University of Stockholm, Statistical Institute.

Frank, O., and D. Strauss. 1986. "Markov Graphs." *Journal of the American Statistical Association* 81: 832–42.

Frank, O., B. Wellman, C. Wilson, S. Lundquist, and V. Espinoza. 1988. "Integrating Individual, Relational, and Structural Analysis." Unpublished manuscript.

Freeman, L. C. 1986. "The Impact of Computer Based Communication on the Social Structure of an Emerging Scientific Specialty." *Social Networks* 6: 201–21.

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

———. 1980. "A Semi-Visible College: Structural Effects of Seven Months of EIES Participation by a Social Networks Community." Pp. 77–85 in *Electronic Communication: Technology and Impacts*, edited by M. M. Henderson and M. J. MacNaughton. AAAS Symposium 52. Washington, DC: American Association for the Advancement of Science.

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

———. 1987. "The Study of a Business Elite in an American Metropolitan Area." Pp. 147–65 in *Research Methods for Elite Studies*, edited by G. Moyser and M. Wagstaffe. London: Allen and Unwin.

Galaskiewicz, J., and S. Wasserman. 1988. "Social Action Models for the Study of Change in Organizational Fields." In *Social Networks Through*

Time, edited by H. Flap and J. Weesie. Utrecht, Netherlands: University of Utrecht Press.

Gifi, A. 1981. "Nonlinear Multivariate Analysis." Unpublished manuscript.

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

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

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

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

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

————. 1987. "New Methods for Analyzing Qualitative Variables." *American Journal of Sociology* 93: 529–83.

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

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

Guttman, L. 1959. "Metricizing Rank-Ordered and Unordered Data for a Linear Factor Analysis." *Sankhya: The Indian Journal of Statistics* 21: 257–68.

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

Harary, F., R. Z. Norman, and D. Cartwright. 1965. *Structural Models: An Introduction to the Theory of Directed Graphs*. New York: Wiley.

Hirschfeld, H. O. 1935. "A Connection Between Correlation and Contingency." *Proceedings of the Cambridge Philosophical Society* 31: 520–24.

Holland, P. W., K. B. Laskey, and S. Leinhardt. 1983. "Stochastic Blockmodels: First Steps." *Social Networks* 5: 109–37.

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

Iacobucci, D., and S. Wasserman. 1987. "Dyadic Social Interactions." *Psychological Bulletin* 101: 292–306.

————. 1988. "A General Framework for the Statistical Analysis of Sequential Dyadic Interaction Data." *Psychological Bulletin* 103: 379–90.

Kendall, M., and A. Stuart. 1973. *The Advanced Theory of Statistics*. Vol. 2. 3d ed. London: Griffin.

Lancaster, H. O. 1957. "Some Properties of the Bivariate Normal Distribution Considered in the Form of a Contingency Table." *Biometrika* 44: 289–92.

————. 1958. "The Structure of Bivariate Observations." *Annals of Mathematical Statistics* 29: 719–36.

Mizruchi, M. S., P. Mariolis, M. Schwartz, and B. Mintz. 1986. "Techniques for Disaggregating Centrality Scores in Social Networks." Pp. 26–48 in *Sociological Methodology 1986*, edited by N. B. Tuma. Washington, DC: American Sociological Association.

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

————. 1986. "The DUAL3 Statistical Software Series: A Dual Scaling Package for the IBM-PC and the NEC-APC." *Psychometrika* 51: 614–15.

Schwartz, J. E. 1976. "An Examination of CONCOR and Related Methods for Blocking Sociometric Data." Pp. 255–82 in *Sociological Methodology 1977*, edited by D. R. Heise. San Francisco: Jossey-Bass.

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

Wang, Y. J., and G. Y. Wong. 1987. "Stochastic Blockmodels for Directed Graphs." *Journal of the American Statistical Association* 82: 8–19.

Wasserman, S. 1979. "A Stochastic Model for Directed Graphs with Transition Rates Determined by Reciprocity." Pp. 392–412 in *Sociological Methodology 1980*, edited by K. F. Schuessler. San Francisco: Jossey-Bass.

————. 1987. "Conformity of Two Sociometric Relations." *Psychometrika* 52: 3–18.

Wasserman, S., and C. Anderson. 1987. "Stochastic A Posteriori Blockmodels: Construction and Assessment." *Social Networks* 9: 1–36.

Wasserman, S., K. Faust, and J. Galaskiewicz. 1989. "Correspondence and Canonical Analysis of Relational Data." *Journal of Mathematical Sociology* 15: forthcoming.

Wasserman, S., and D. Iacobucci. 1986. "Statistical Analysis of Discrete Relational Data." *British Journal of Mathematical and Statistical Psychology* 39: 41–64.

————. 1988. "Sequential Social Network Data." *Psychometrika* 53: 261–82.

Wellman, B., C. Mosher, C. Rottenberg, and V. Espinoza. 1987. "Different Strokes from Different Folks: Which Ties Provide What Kinds of Social Support." Unpublished manuscript.

White, H. C., S. A. Boorman, and R. L. Breiger. 1976. "Social Structure from Multiple Networks I: Blockmodels of Roles and Positions." *American Journal of Sociology* 81: 730–80.

Winship, C., and M. J. Mandel. 1983. "Roles and Positions: A Critique and Extension of the Blockmodeling Approach." Pp. 314–44 in *Sociological Methodology 1983–84*, edited by S. Leinhardt. San Francisco: Jossey-Bass.

Wu, L. L. 1983. "Local Blockmodel Algebras for Analyzing Social Networks." Pp. 272–313 in *Sociological Methodology 1983–84*, edited by S. Leinhardt. San Francisco: Jossey-Bass.