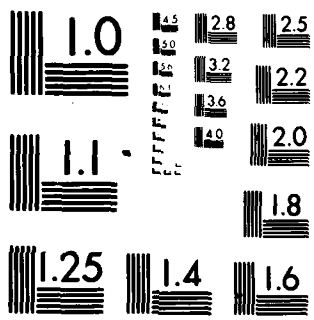
AD-A089 194 CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS F/G 12/1
ANALYZING DATA FROM MULTIVARIATE DIRECTED GRAPHS: AN APPLICATIO--ETC(U)
JUL 80 S E FIENBERG, M M MEYER, S S WASSERMAN N00014-80-C-0637
UNCLASSIFIED TR-185 NL

END
DATE
FILMED
10 80
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

12

DEPARTMENT

OF

STATISTICS

Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

July 1980

(18)

ANALYZING DATA FROM MULTIVARIATE DIRECTED GRAPHS: AN APPLICATION TO SOCIAL NETWORKS[1]

by

Stephen E. Fienberg[2]
Michael M. Meyer
Stanley S. Wasserman

Department of Applied Statistics
School of Statistics
University of Minnesota
St. Paul, Minnesota 55108

Technical Report No. 185
Carnegie-Mellon University
Pittsburgh, PA 15213

To appear in Vic Barnett, ed. (1981).
Looking at Multivariate Data
John Wiley & Sons, Chichester, England

[1]The preparation to this paper was supported in part by Grants SOC78-26075 and SES80-08573 from the National Science Foundation, and Contract N00014-78-C-0151 from the Office of Naval Research, to the University of Minnesota.

[2]Present Address: Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

Summary

A multivariate directed graph consists of a set of g nodes, and a family of directed arcs (one for each relation) connecting pairs of nodes. Such multivariate directed graphs provide natural representations for social networks. In this paper ~~we consider~~ methods to analyse a network of 73 organizations in a Midwest American community linked by three types of relations is considered: information, money, and support. The resulting data set, described by Galaskiewicz and Marsden (1978), involves 3 × 73 × 72 = 15,768 possible arcs or "observations". ~~We~~ THE REPORT describes a class of stochastic loglinear models for multivariate directed graphs, demonstrate how they can be fit to the data using generalized iterative scaling of Darroch and Ratcliff (1972), and explain the connection between these models and variants on standard loglinear models for multidimensional contingency tables discussed by Bishop, Fienberg, and Holland (1975). ~~We~~ IT also considers a disaggregation of the organizations into sub-groups, and demonstrate how to adapt ~~our~~ THE models to explore the intra- and inter-group relationships. These methods generalize research of Holland and Leinhardt (1980), who develop a model for dyadic relationships in univariate directed graph data. The paper includes a detailed analysis of the Galaskiewicz-Marsden data.

1. Introduction

Although this conference is entitled "Looking at Multivariate Data", most attendees and authors have interpreted this to mean looking at multivariate data "using graphical methods". Despite the fact that the title of the present paper contains the words "data", "multivariate", and "graphs", we shall break step with these other authors and describe a class of multivariate network problems. We would have liked to address these problems using graphical methods but, for the moment, we have been forced to settle for a more traditional multivariate model-based approach. This may seem even more surprising since the network problems we address begin with data that correspond to a picture or graph.

-- Figure 1 goes about here --

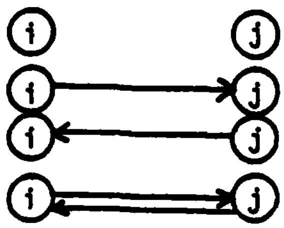
Figure 1 contains an example of a univariate directed graph, a graphical representation of a network involving $g = 6$ individuals. There are $g(g-1) = 30$ possible arrows or directed arcs linking these 6 individuals in pairs, only 12 of which are present in Figure 1. The information in a univariate graph for g individuals can be summarized by means of a $g \times g$ adjacency matrix $\underset{\sim}{x}$, with elements

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ relates to } j \\ 0 & \text{otherwise,} \end{cases}$$

where, by convention, the diagonal terms, x_{ii}, are set equal to zero. The adjacency matrix for Figure 1 is:

$$\underset{\sim}{x} = \begin{array}{cc} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

There are several approaches that we might adopt to model the data in the adjacency matrix, $\underset{\sim}{x}$. For example, we might focus on the 6 individuals and assume that individual i makes 5 possible independent choices (corresponding to arrows), with some unknown Bernoulli parameter, p_i ($i = 1,2,\ldots,6$). Then a suitable data summary would be the row totals of $\underset{\sim}{x}$, i.e., (3,3,2,1,3,0). The assumption of independence of choices is not likely to be satisfied in practice, however. Alternatively, we might focus on the $6 \times 5 = 30$ pairs of individuals, and assume that the data for the pairs are independent and identically distributed. In effect, then, we would choose to focus on relationships, and would observe three different types:



Null,

Asymmetric,

Mutual.

Thus the observed data would be summarized in the following 2×2 table:

		Receive Choice		
		Yes	No	
Send Choice	Yes	4	8	12
	No	8	10	18
		12	18	30

Note that each pair has been counted twice, once for "sending" and once for "receiving", thus merging the asymmetric relationships.

The approach involving pairs essentially uses the g(g-1) permutations of the g individuals, two at a time, and thus leads to a doublecounting of each pair. By focussing on the $\binom{g}{2} = g(g-1)/2$ combinations or <u>dyads</u>, we can eliminate the doublecounting and obtain the following summary:

	No. of Dyads
$i \qquad j$	5
$i \longrightarrow j$	8
$i \rightleftarrows j$	2

In this paper, we consider stochastic models of multivariate directed graphs, involving several types of arrows or relationships, that treat the $\binom{g}{2}$ dyads as independent random variables. We do this in the full knowledge that for most network problems dyads are constructs. We do not sample them. Rather, if we sample at all, we take a sample of individuals and we measure information on dyadic relationships. The independence of dyadic information is an assumption which in practice is in need of some verification. We do not address this issue in this paper. For population directed graph data, consisting of the dyad information for all of the individuals in a network, the use of stochastic models leans for support on (a) randomization arguments, (b) superpopulation ideas, or (c) it simply provides a convenient framework for exploratory data analysis.

In the next section we describe a set of network data involving organizations and three types of organizational relations. Then, in Section 3, we describe a class of models and multivariate methods for the analysis of such data, which treats the organizations as a single group. After fitting these models to the data in Section 4, we further develop the models in Section 5 to allow for disaggregation of the organizations into subgroups. We conclude by returning to the graphical theme of this conference, and suggest some extensions of our modelling approach which might lead to interesting graphical summaries.

2. A Specific Network: Towertown, U.S.A.

The data that have motivated our work on this topic come from a study of 109 formal organizations (with more than 20 employees) in a small midwest United States community of 32,000 persons, referred to by the pseudonym "Towertown". Galaskiewicz (1979) described the survey of Towertown, Galaskiewicz and Marsden (1978) report on the data considered here, and we have described the data elsewhere in detail (Fienberg and Wasserman, 1981). For the present, it will suffice to note that we are concerned with the results of questionnaire data for a subset of 73 organizations, representing the ties between pairs of organizations for three types of relations: (i) information, (ii) money, and (iii) support. This data set can then be represented as a multivariate directed graph, summarized by three adjacency matrices defined for the same 73 organizations,

$(\underset{\sim}{x}_1, \underset{\sim}{x}_2, \underset{\sim}{x}_3)$. Each matrix is of dimension 73 × 73 and represents 73(73-1) = 5256 possible directed arcs using 0's and 1's.* Given the size of these matrices, it should not be surprising that graphical representations of even the univariate links are too complex to comprehend.

Thus, we still need a way to look at and, perhaps more importantly, summarize the data. Table 1 contains one such summary of the data given by Galaskiewicz and Marsden (1978), in the form of a 2^6 table of counts of pairs of organizations. This table gives the direct multivariate generalization of the 2 × 2 representation for a single relation given in Section 1. Each pair of organizations is counted twice, once from the perspective of each member. Thus, the total of the counts in the table is 5256, twice the number of pairs, $\binom{73}{2} = 2628$. Henceforth, we refer to Table 1 as the $\underset{\sim}{w}$-table with entries $\{w_{ii'jj'kk'}\}$.

-- Table 1 goes about here --

The 2^6 cells of Table 1 consist of (a) 8 cells whose counts are doubled, and (b) 28 cells whose counts are duplicated. If we eliminate the duplication and doubling of counts, we get an arrangement of 36 cells, whose counts correctly total 2628. In Table 2 we give one possible representation of these 36 cells in a form resembling a three-dimensional 4 × 4 × 4 cross-classification, where the three "variables" correspond to the three relations (1) information, (2) money, and (3) support.

-- Table 2 goes about here --

---

*Throughout this paper we work with summaries of this data set. The full data set, consisting of three adjacency matrices and pseudonyms for the organizations, is available on request from the authors.

When the dyadic structure for a single relation is asymmetric, the "direction" of the corresponding arc does not matter. We use a single subscript, A, to denote the relation in such situations. When the dyadic links for two or more relations are both asymmetric, we need to distinguish between situations where the arcs for a pair of relations go in the same or different directions. Thus, for these situations, we use two different subscripts, A and $\bar{A}$, with identical subscripts for those relations whose asymmetric directed arcs go in the same direction. We arbitrarily assign the subscript A to the lowest numbered asymmetric generator. (Note that interchanging the subscripts A and $\bar{A}$ yields the same dyadic structural relationship.) We denote the observed counts in Table 2 by z_{abc}, for a, b, c = M, A, $\bar{A}$, N (for Mutual, Asymmetric, $\bar{A}$symmetric, and Null), where the convention for the use of the subscripts A and $\bar{A}$ is as described above. These observed counts can be thought of as realizations of a set of random variables, $\{Z_{abc}\}$, whose probability structure we wish to model.

3. Loglinear Models for Multivariate Directed Graphs

We wish to model the probability p_{abc} that a randomly selected dyad would be assigned to cell (a,b,c) in Table 1, where

$$(3.1) \qquad \sum_{\text{all cells}} p_{abc} = 1.$$

Although we might think of using loglinear models directly for the $\{p_{abc}\}$, such an approach leads to difficulties of interpretation (see Fienberg and Wasserman, 1980, for further details). Instead, we define

$$(3.2) \qquad \xi_{abc} = \begin{cases} \log p_{abc} & \text{if a, b, and c are each equal to M or N,} \\ \log \left[\frac{p_{abc}}{2}\right] & \text{if one of a, b, and c equals A.} \end{cases}$$

Our plan is to develop a class of linear models for the $\{\xi_{abc}\}$, which for $\{p_{abc}\}$ yields an affine translation of a class of loglinear models (see Chapter 9 of both Haberman, 1974 and Haberman, 1979). This approach

(a) treats dyads involving asymmetric ties as having been produced with an orientation and then pooled. (This also accounts for the divisor of 2 for counts involving asymmetric ties.)

(b) includes as a special case the model of independent individual choices (see the discussion of Section 1).

(c) is directly related to an approach of Holland and Leinhardt (1980) which allows for parameters associated with the individuals in the dyad (see also Fienberg and Wasserman, 1981).

We plan to consider models for the $\{\xi_{abc}\}$ which are linear in parameters that reflect the 13 distinct types of dyadic patterns depicted in Figure 2. Note that the patterns have a hierarchical structure. For example, the six-arrow full symmetry pattern, (xiii), contains all the other patterns as special cases, and the conditional multiplex mutuality pattern, (xii), contains patterns (i) through (xi) as special cases. We consider a class of increasingly complex loglinear models for the $\{\xi_{abc}\}$ with parameters based on the patterns in Figure 2.

-- Figure 2 goes about here --

(I) The null model corresponding to Figure 2(i) depicts the probabilities $\{p_{abc}\}$ as being constant, and could be represented as

$$\xi_{abc} = \theta,$$

where $\theta = \log(1/36)$. This is an individual, independent Bernoulli choice model. For subsequent models we use θ as a <u>normalizing constant</u>.

(II) At the next level, we add choice parameters $\{\theta_1, \theta_2, \theta_3\}$ for the relations (Figure 2(ii)), one for each directed arc. For example,

$$\xi_{MAN} = \theta + 2\theta_1 + \theta_2$$

$$\xi_{MAA} = \theta + 2\theta_1 + \theta_2 + \theta_3$$

$$\xi_{MA\bar{A}} = \theta + 2\theta_1 + \theta_2 + \theta_3.$$

(III) Next, we add sets of parameters corresponding to heightened or diminished effects related to pairs of directed arcs:

(a) ρ_{11}, ρ_{12}, ρ_{33} for <u>mutuality</u> effects (see Figure 2(iii)),

(b) ρ_{12}, ρ_{13}, ρ_{23} for <u>exchange</u> effects (see Figure 2(iv)),

(c) θ_{12}, θ_{13}, θ_{23} for <u>multiplexity</u> effects (see Figure 2(v)),

For example:

$$\xi_{MA\bar{A}} = \theta + 2\theta_1 + \theta_2 + \theta_3 + \rho_{11} + \rho_{12} + \rho_{13} + \rho_{23} + \theta_{12} + \theta_{13},$$

$$\xi_{MAM} = \theta + 2\theta_1 + \theta_2 + 2\theta_3 + \rho_{11} + \rho_{33} + \rho_{12} + 2\rho_{13} + \rho_{23} + \theta_{12} + 2\theta_{13} + \theta_{23}.$$

There are additional sets of parameters corresponding to the remaining 4 levels in Figure 2. At level IV, one of these parameters involves only multiplexity and thus is denoted by a triple subscripted θ, i.e., θ_{123}. The remaining parameters involve mixtures of mutuality, exchange, and multiplexity, and are denoted by subscripted $(\rho\theta)$'s. Overbars on subscripts are used to distinguish asymmetric directed arcs going in opposite directions, e.g., $(\rho\theta)_{1\overline{23}}$.

The parameters in this class of models are GLIM-like in structure (e.g., see Nelder and Wedderburn, 1972), in that a parameter is included in the model if and only if the corresponding effect is present. The entries of the resulting "design matrix" for the parameter structure for any given model will be 0's, 1's, and 2's. This particular problem could be handled in GLIM directly only through the explicit construction of this design matrix, which is a formidable task.

The parameters have a hierarchical structure, i.e., if we set some parameters equal to zero, all related higher-order terms are also zero. For example,

$$\begin{aligned}\theta_{12} = 0 \Rightarrow \theta_{123} &= (\rho\theta)_{112} = (\rho\theta)_{221} = (\rho\theta)_{3\overline{12}} \\ &= (\rho\theta)_{1123} = (\rho\theta)_{11\overline{23}} = (\rho\theta)_{2213} \\ &= (\rho\theta)_{22\overline{13}} = (\rho\theta)_{1122} = (\rho\theta)_{11223} \\ &= (\rho\theta)_{11332} = (\rho\theta)_{22331} \\ &= (\rho\theta)_{112233} = 0,\end{aligned}$$

and

$$\begin{aligned}\rho_{11} = 0 \Rightarrow (\rho\theta)_{112} &= (\rho\theta)_{113} = (\rho\theta)_{1123} = (\rho\theta)_{11\overline{23}} \\ &= (\rho\theta)_{1122} = (\rho\theta)_{1133} = (\rho\theta)_{11223} \\ &= (\rho\theta)_{11332} = (\rho\theta)_{112233} = 0.\end{aligned}$$

In the next section we discuss how to fit these models to social network data.

4. Fitting the Models to Data

Fitting the loglinear models of the preceding section to data in Table 2 follows, in principle, directly from the general results for loglinear models in Haberman (1974) or Appendix II of Fienberg (1980). The minimal sufficient statistics (MSS's) take the form of linear combinations of the $\{z_{abc}\}$,

(4.1) $$\sum_{\text{all cells}} \alpha_{abc}\, z_{abc},$$

where for a MSS corresponding to "generic" parameter, β,

(4.2) $$\alpha_{abc} = \text{multiple of } \beta \text{ in } \xi_{abc}.$$

The multiples of all parameters are either 0, 1, or 2, and thus all of the α's are *either* 0, 1, or 2.

If we let the expected value for the (a,b,c) cell be $m_{abc} = N \cdot p_{abc}$ where $N = \binom{g}{2}$, then the likelihood equations are found by setting the MSS's equal to their estimated expected values, i.e., for a generic parameter the likelihood equation is:

(4.3) $$\sum_{\text{all cells}} \alpha_{abc}\, \hat{m}_{abc} = \sum_{\text{all cells}} \alpha_{abc}\, z_{abc}.$$

We can solve a set of likelihood equations, each of the form (4.3), by using a version of the generalized iterative scaling algorithm due to Darroch and Ratcliff (1972), with starting values as follows:

(4.4) $$\hat{m}^{(0)}_{abc} = \begin{cases} 1 & \text{if a, b, and c are each equal to M or N} \\ \tfrac{1}{2} & \text{if one or more of a, b, and c equals A.} \end{cases}$$

There are two drawbacks to this approach. First, one needs to work with data arrays of the irregular shape of Table 2. Second, the convergence of generalized iterative scaling can be excruciatingly slow.

All, however, is not lost. Two results, one simple and one relatively complex, lead us to a very straightforward alternative approach for computing the $\{\hat{m}_{abc}\}$.

<u>Result 1</u>: For the class of affine translations of hierarchical loglinear models described in Section 3, each set of MSS's is equivalent to a set of marginal totals for the 2^6 table (i.e., the $\underset{\sim}{w}$-table) with doubled and duplicated counts.

For example, the simple model with only a choice parameter, θ_1, and a mutuality parameter, ρ_{11}, for the first relation has MSS's $\{z_{M++}, z_{A++}, z_{N++}\}$, and

(4.5)
$$z_{M++} = \tfrac{1}{2} w_{11++++},$$
$$z_{A++} = w_{10++++} = w_{01++++},$$
$$z_{N++} = \tfrac{1}{2} w_{00++++}.$$

<u>Result 2</u>: For each affine translation of a loglinear model for the $\underset{\sim}{z}$-table, there is a corresponding loglinear model for the $\underset{\sim}{w}$-table, with equivalent estimated expected values, once we take account of the duplication and doubling.

For example, for the model with choice and mutuality parameters, i.e.,

(4.6)
$$\{\theta, \theta_1, \theta_2, \theta_3, \rho_{11}, \rho_{22}, \rho_{33}\},$$

the corresponding loglinear model for the $\underset{\sim}{w}$-table that yields equivalent MLE's is, in GLIM-like notation:

$$(4.7)\qquad \log m_{ii'jj'kk'} = \lambda + \lambda_1\delta_i + \lambda_{1'}\delta_{i'} + \lambda_2\delta_j + \lambda_{2'}\delta_{j'} + \lambda_3\delta_k + \lambda_{3'}\delta_{k'} + \lambda_{11}\delta_i\delta_{i'} + \lambda_{22}\delta_j\delta_{j'} + \lambda_{33}\delta_k\delta_{k'}.$$

Here $m_{ii'jj'kk'}$ is the expected value for the (i,i',j,j',k,k') cell, and each δ-term equals 1 if the subscript takes the value 1, and is zero otherwise.

To understand Result 2 we need to note the following correspondences between the $\underset{\sim}{w}$-table and the $\underset{\sim}{z}$-table:

	$\underset{\sim}{w}$-table	$\underset{\sim}{z}$-table
Cell:	(i,i',j,j',k,k')	(a,b,c)
Symmetric flows:	$i = i'$, $j = j'$, $k = k'$	a,b,c = M or N

Because of the doubling of the counts in Table 1, we have:

$$(4.8)\quad \log m_{ii'jj'kk'} = \begin{cases} \log\,(2\, m_{abc}) & \text{for symmetric flows,} \\ \log\,(m_{abc}) & \text{for asymmetric flows.} \end{cases}$$

Substituting expression (3.2) into (4.8) and noting that $m_{abc} = \binom{g}{2} p_{abc}$, we get

(4.9) $$\log m_{ii'jj'kk'} = \left[2\binom{g}{2}\right] + \xi_{abc}.$$

Thus the models for $\log m_{ii'jj'kk'}$ and ξ_{abc} differ by only a constant.

A direct consequence of these two results is that we can compute MLE's for the expected values under the models of Section 3 using standard iterative methods for contingency tables. (This is in fact what Galaskiewicz and Marsden (1978) did in their original analyses of Table 1.). For example, for the model with parameters given by (4.6), the MSS's are equivalently given by the two-way marginal totals of the $\underset{\sim}{w}$-table:

$$\{w_{ii'++++}\},\{w_{++jj'++}\},\{w_{++++kk'}\}$$

These marginals can be fit to the 2^6 table using the standard iterative proportional fitting procedure (or some other program such as GLIM). Because of symmetries in marginal totals, e.g.,

$$w_{10++++} = w_{01++++},$$

$$w_{++10++} = w_{++01++},$$

$$w_{++++10} = w_{++++01},$$

the resulting parameter estimates are such that

$$\hat{\lambda}_1 = \hat{\lambda}_{1'}, \quad \hat{\lambda}_2 = \hat{\lambda}_{2'}, \quad \hat{\lambda}_3 = \hat{\lambda}_{3'}.$$

The estimated parameters for the models for ξ_{abc} can be computed directly from these parameters:

$$\hat{\theta} = \hat{\lambda} - \log\left[2\binom{g}{2}\right]$$

$$\hat{\theta}_i = \hat{\lambda}_i \qquad i = 1,2,3$$

$$\hat{\rho}_{ii} = \hat{\lambda}_{ii} \qquad i = 1,2,3.$$

We note that the d.f. for any model <u>must</u> be calculated using the model for the $\underset{\sim}{z}$-table, not the one for the $\underset{\sim}{w}$-table, and the value of any standard goodness-of-fit statistic computed directly on the fitted $\underset{\sim}{w}$-table must be divided by 2.

5. Initial Analyses of the Towertown Data

In Table 3 we list a set of seven loglinear models that we have fit to the Galaskiewicz-Marsden data of Table 1 (some of these models correspond to ones fit by Galaskiewicz and Marsden). The first six models are of increasing complexity, and only the most complex of these models, (6), provides a fit which it not significant at the 0.05 or even 0.01 level. Model (7) is a compromise between models (5) and (6) that drops one of the conditional mutuality and two of the multiplex mutuality effects but still provides an acceptable fit to the data. Its parameter estimates are listed in Table 4.

-- Tables 3 and 4 go about here --

The most substantial estimated effects (in terms of magnitude) are those associated with choice ($\hat{\theta}_i$'s), mutuality ($\hat{\rho}_{ii}$'s), conditional mutuality $(\widehat{\rho\theta})_{331} = -2.15$ and multiplex mutuality $(\widehat{\rho\theta})_{1133} = 2.88$. Interpreting these effects is complicated. For all hierarchical models, with nonorthogonal designs, the parameters that are easiest to interpret are those associated with the highest-order effects. Here the multiplex mutuality parmaeter estimate implies a heightened likelihood of simultaneous reciprocation of both information and support, relative to what we would expect in a model without the multiplex mutuality parameter.

One of the major difficulties with the models of Section 3 is that dyads are considered to be homogeneous and thus do not allow for the inherent differences among the organizations. Without some allowance for this heterogeneity, further interpretation of fitted models makes little sense. In Table 5 we list pseudonyms for each of the 73 organizations, and provide a partition of them into four sub-groups:

1. Business $g_1 = 16$,
2. Political $g_2 = 24$,
3. Nonprofit voluntary associations $g_3 = 21$,
4. Nonprofit service associations $g_4 = 12$.

We postulate that the sociological factors affecting interaction should be relatively homogeneous within these groups. Thus, we can categorize the original $\binom{g}{2} = \binom{73}{2} = 2628$ dyads into the cells of an upper triangular 4×4 array:

No. of Dyads:

G_1	G_2	G_3	G_4	
120	384	336	192	G_1
	276	504	288	G_2
		210	252	G_3
			66	G_4

For each cell in this array there is a 2^6 table.

-- Table 5 goes about here --

Within each of the four groups we can analyze flows using a 2^6 table and the models from Section 3. These 2^6 tables have the same doublings and duplications as the aggregated 2^6 table. The flows between groups (in pairs) now have an orientation and there are corresponding 2^6 tables describing these flows which contain no doubling and no duplication. We can analyze each of these tables with standard loglinear models that parallel those models for within group flows. The total number of cells in the full table is $(4 \times 36) + (6 \times 64) = 528$.

In Table 6, we report the result of fitting separate multiplex mutuality models (model (6) of Table 3) to each of the 10 2^6 arrays. While this model fits extremely well (G^2 is less than the d.f.), this is in large part the result of fitting 352 parameters. An alternative modelling approach links the within and between group models. For example, we might take a common "interaction structure" for all 10 2^6 tables, but allow only the choice parameters (the θ_i's) to depend on groups. The result is model (2) in Table 5, whose fit is not horrid but is still significant at the 0.005 level. A compromise between models (1) and (2) of Table 3 would have a common model for within-group flows and a separate variant on model (2) for between-group flows. We report the fit of two such models in Table 6. Model (3b) fits extremely well, and provides a convenient starting point for further analyses of the data.

-- Table 6 goes about here --

6. A Possible Graphical Display for Multivariate Directed Graphs

The second set of analyses of the preceding section leads quite naturally to analyses involving a further disaggregation of organizations. Indeed we could carry the disaggregation to the limit, with each organization forming its own group of one. We could postulate models with different choice parameters for each organization and a common higher-order parametric structure. Actually, we would end up with individual sending and receiving parameters for each organization and each relation. The resulting model is in the same spirit as the bivariate models suggested by Holland and Leinhardt (1980).

The attractive feature of this fully-disaggregated approach is that we can examine the estimated higher order structure in a tabular form similar to that of Table 4, and look separately at the estimated individual parameters. The latter can be displayed in a set of three overlayed "correspondence-like" plots of the 73 organizations. The sending and receiving parameter estimates for an organization could be used as the abscissa and ordinate for a corresponding point, and the three points for different relations could be linked to form a triangle. This plot should show not only the clustering of organizations but also the similarities of their behavior with regard to the three different relations being considered. We have stopped short of producing the plot for the Towertown data for computational reasons. The iterative methods used here, and in Fienberg and Wasserman (1981) for the univariate version of the disaggregated model, when applied to the Towertown data simply take up too much computing storage. We hope, however, that alternative computational methods currently under development might make possible some graphical displays for multivariate directed graphs in the not-too-distant future.

REFERENCES

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). Discrete Multivariate Analysis, Cambridge, MA: The MIT Press.

Darroch, J.N. and Ratcliff, D. (1972). "Generalized iterative scaling of loglinear models," The Annals of Mathematical Statistics, 43:1470-80.

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

Fienberg, S.E. and Wasserman, S. (1980). "Methods for the analysis of data from multivariate directed graphs," Proceedings of the Conference on Recent Developments in Statistical Methods and Applications. Taipei, Taiwan: Institute of Mathematics, Academia Sinica.

Fienberg, S.E. and Wasserman, S. (1981). "Categorical data analysis of single sociometric relations," to appear in Sociological Methodology 1981, edited by S. Leinhardt. San Francisco: Jossey-Bass.

Galaskiewicz, J. (1979). Exchange Networks and Community Politics. Beverly Hills: Sage.

Galaskiewicz, J. and Marsden, P.V. (1978). "Interorganizational resource networks: Formal patterns of overlap," Social Science Research, 7:89-107.

Haberman, S. (1974). The Analysis of Frequency Data. Chicago: University of Chicago Press.

Haberman, S. (1979). Analysis of Qualitative Data, Volume 2: New Developments. New York: Academic Press.

Holland, P.W. and Leinhardt, S. (1975). "Local structure in social networks." In Sociological Methodology 1976, edited by D.R. Heise. San Francisco: Jossey-Bass, 1-45.

Holland, P.W. and Leinhardt, S. (1980). "An exponential family of probability densities for directed graphs," Journal of the American Statistical Association, to appear.

Nelder, J.A. and Wedderburn, R.W. (1972). "Generalized linear models," Journal of the Royal Statistical Society, Series A, 135:370-384.

Figure 1: Example of a Univariate Directed Graph Involving g = 6 Individuals

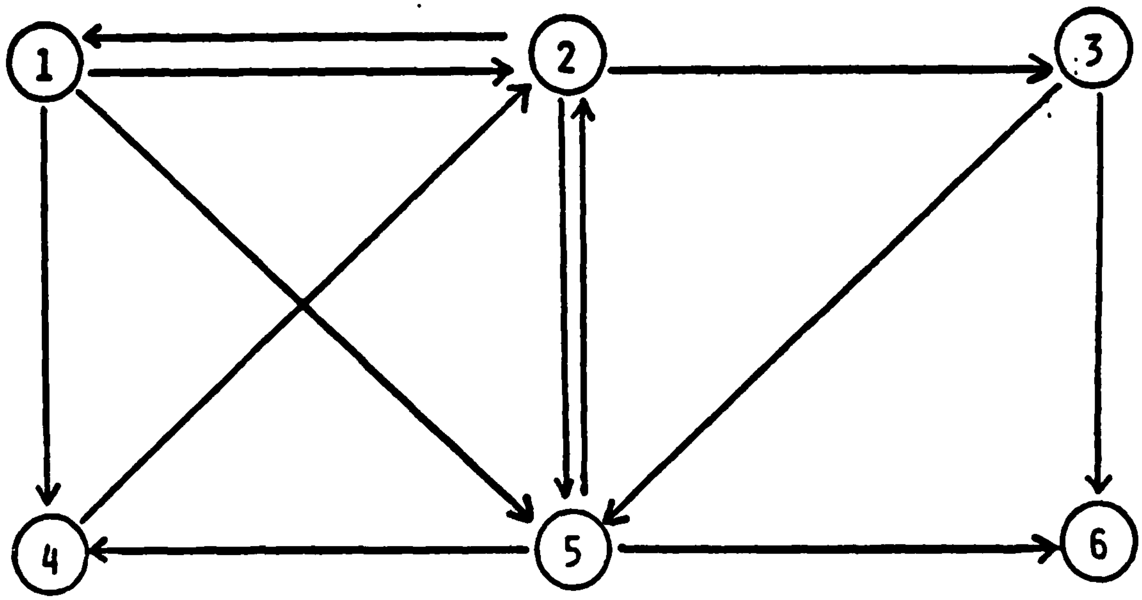


FIGURE 2. PATTERNS OF FLOW DEPENDENCY IN DYADIC PATTERNS

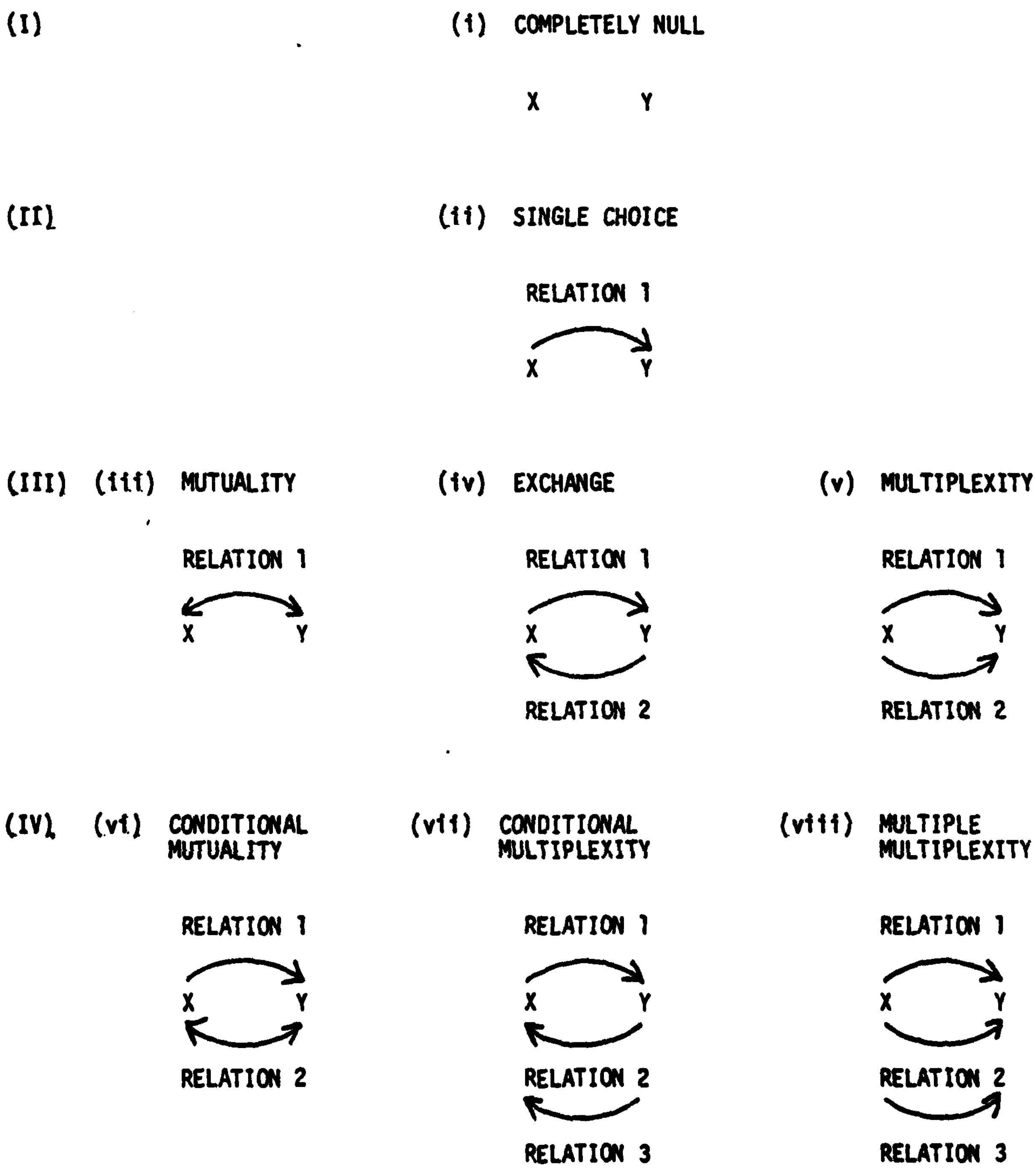


FIGURE 2 (CONTINUED)

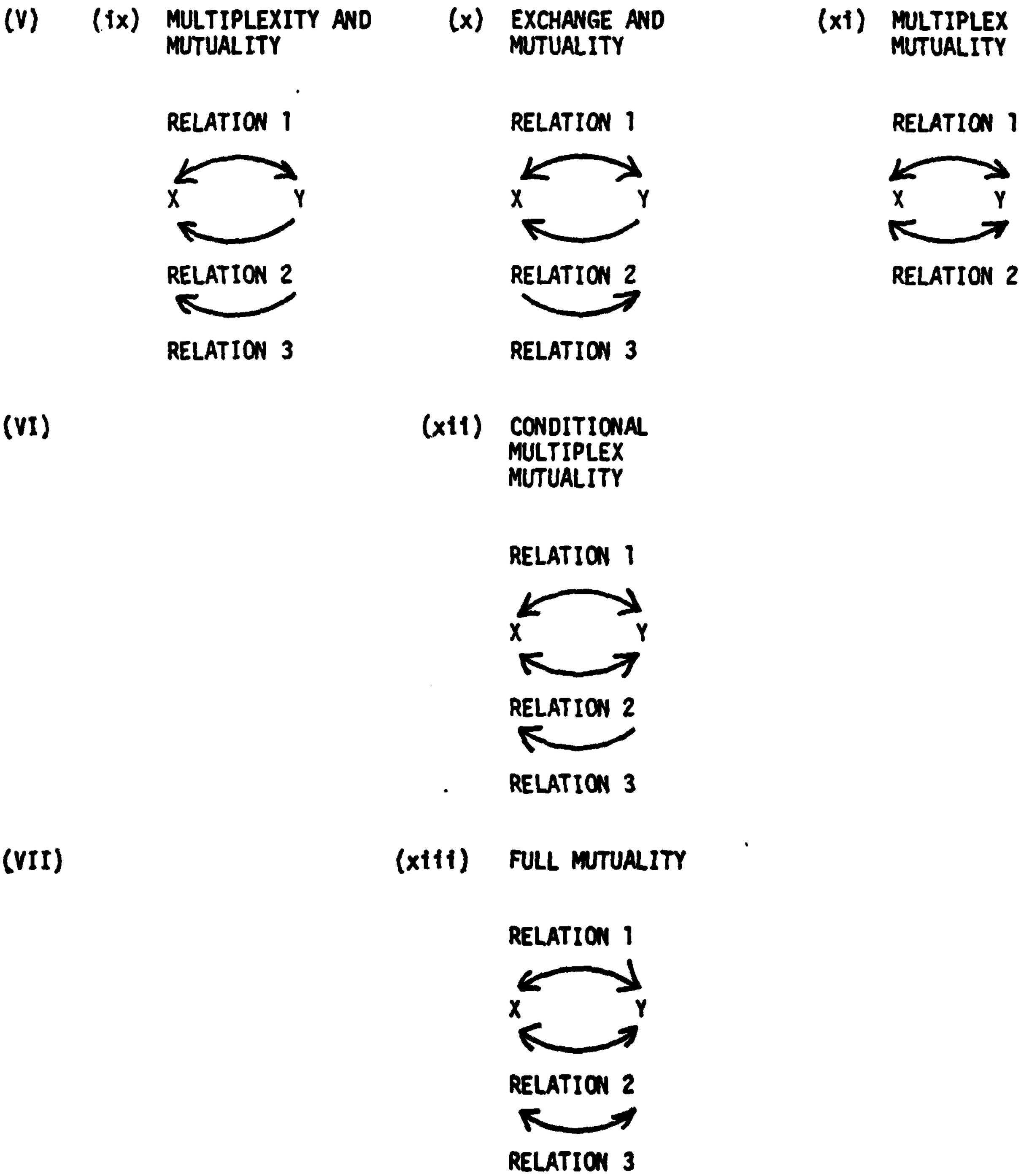


TABLE 1. OBSERVED DISTRIBUTION OF INTERORGANIZATIONAL TRANSACTIONS INVOLVING THREE RELATIONS AND 73 ORGANIZATIONS[A]

1. INFORMATION SENT		-								+							
1'. INFORMATION RECEIVED		-				+				-				+			
2. MONEY SENT		-		+		-		+		-		+		-		+	
2'. MONEY RECEIVED		-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+
3. SUPPORT SENT	3'. SUPPORT RECEIVED																
-	-	3042	100	100	20	145	15	11	7	145	11	15	7	322	47	47	16
	+	111	18	12	3	20	9	3	4	32	20	2	1	77	38	15	25
+	-	111	12	18	3	32	2	20	1	20	3	9	4	77	15	38	25
	+	116	14	14	4	15	4	6	0	15	6	4	0	100	50	50	28

TABLE TOTAL = 5256

A"+" INDICATES THAT A DIRECTED FLOW IS PRESENT, "-" INDICATES THAT A DIRECTED FLOW IS ABSENT. A SOMEWHAT DIFFERENT VERSION OF THESE DATA APPEARED IN GALASKIEWICZ AND MARSDEN (1978).

TABLE 2. STRUCTURE FOR ACTUAL TABLE OF 36 COUNTS

<table>
<tr><td colspan="4"></td><td colspan="4">Support</td></tr>
<tr><td colspan="4"></td><td>M</td><td colspan="2">A</td><td>N</td></tr>
<tr><td rowspan="10">Information</td><td rowspan="3">M</td><td rowspan="3">Money</td><td>M</td><td>z_{MMM} 14</td><td colspan="2">z_{MMA} 25</td><td>z_{MMN} 8</td></tr>
<tr><td>A</td><td>z_{MAM} 50</td><td>z_{MAA} 38</td><td>$z_{MA\bar{A}}$ 15</td><td>z_{MAN} 47</td></tr>
<tr><td>N</td><td>z_{MNM} 50</td><td colspan="2">z_{MNA} 77</td><td>z_{MNN} 161</td></tr>
<tr><td rowspan="4">A</td><td rowspan="4">Money</td><td>M</td><td>z_{AMM} 0</td><td>z_{AMA} 4</td><td>$z_{AM\bar{A}}$ 1</td><td>z_{AMN} 7</td></tr>
<tr><td rowspan="2">A</td><td>z_{AAM} 4</td><td>z_{AAA} 9</td><td>$z_{AA\bar{A}}$ 2</td><td>z_{AAN} 15</td></tr>
<tr><td>$z_{A\bar{A}M}$ 6</td><td>$z_{A\bar{A}A}$ 3</td><td>$z_{A\bar{A}\bar{A}}$ 20</td><td>$z_{A\bar{A}N}$ 11</td></tr>
<tr><td>N</td><td>z_{ANM} 15</td><td>z_{ANA} 20</td><td>$z_{AN\bar{A}}$ 32</td><td>z_{ANN} 145</td></tr>
<tr><td rowspan="3">N</td><td rowspan="3">Money</td><td>M</td><td>z_{NMM} 2</td><td colspan="2">z_{NMA} 3</td><td>$z_{A\bar{A}N}$ 10</td></tr>
<tr><td>A</td><td>z_{NAM} 14</td><td>z_{NAA} 18</td><td>$z_{NA\bar{A}}$ 12</td><td>z_{NAN} 10</td></tr>
<tr><td>N</td><td>z_{NNM} 58</td><td colspan="2">z_{NNA} 111</td><td>z_{NNN} 1521</td></tr>
</table>

TABLE 3. VARIOUS LOGLINEAR MODELS FITTED TO DATA IN TABLE 1

Model	D.F.	G^{2*}
(1) $\theta, \theta_1, \theta_2, \theta_3$	32	2528.5
(2) $\theta, \theta_1, \theta_2, \theta_3, \rho_{11}, \rho_{22}, \rho_{33}$	29	895.0
(3) $\theta, \theta_1, \theta_2, \theta_3, \rho_{11}, \rho_{22}, \rho_{33}, \rho_{12}, \rho_{13}, \rho_{23}$	26	224.1
(4) $\theta, \theta_1, \theta_2, \theta_3, \rho_{11}, \rho_{22}, \rho_{33}, \rho_{12}, \rho_{13}, \rho_{23}, \theta_{12}, \theta_{13}, \theta_{23}$	23	122.15
(5) $(\rho\theta)_{112}, (\rho\theta)_{113}, (\rho\theta)_{221}, (\rho\theta)_{223}, (\rho\theta)_{331}, (\rho\theta)_{332}$, plus all implied lower-order terms	17	40.315
(6) $(\rho\theta)_{1122}, (\rho\theta)_{1133}, (\rho\theta)_{2233}$, plus all implied lower-order terms	14	20.73
(7) parameters from model (4) plus $(\rho\theta)_{113}, (\rho\theta)_{331}, (\rho\theta)_{112}, (\rho\theta)_{223}, (\rho\theta)_{332}$, and $(\rho\theta)_{1133}$	17	22.24

*G^2 is the log-likelihoodratio chi-square goodness-of-fit statistics.

TABLE 4. PARAMETER ESTIMATES FOR MODEL (7) FITTED TO THE DATA FROM TABLE 1

Parameter	Estimate	
$\hat{\theta}$	-0.55	Normalization Constant
$\hat{\theta}_1$	-3.02	Choice
$\hat{\theta}_2$	-3.35	
$\hat{\theta}_3$	-3.28	
$\hat{\rho}_{11}$	3.82	Mutuality
$\hat{\rho}_{22}$	1.52	
$\hat{\rho}_{33}$	3.28	
$\hat{\rho}_{12}$	1.01	Exchange
$\hat{\rho}_{13}$	1.73	
$\hat{\rho}_{23}$	0.60	
$\hat{\theta}_{12}$	0.78	Multiplex
$\hat{\theta}_{13}$	1.34	
$\hat{\theta}_{23}$	1.57	
$(\widehat{\rho\theta})_{112}$	-0.52	Conditional Mutuality
$(\widehat{\rho\theta})_{113}$	-1.30	
$(\widehat{\rho\theta})_{223}$	-0.70	
$(\widehat{\rho\theta})_{331}$	-2.15	
$(\widehat{\rho\theta})_{332}$	-0.83	
$(\widehat{\rho\theta})_{1133}$	2.88	Multiplex Mutuality

TABLE 5. PARTITION OF 73 ORGANIZATIONS INTO 4 GROUPS

G_1		G_2		G_3		G_4	
Business		Political		Nonprofit Voluntary Associations		Nonprofit Service Organizations	
2.	Farm Equipment Co.	25.	City Council	1.	Farm Bureau	46.	Health Services Center
3.	Clothing Mfg. Co.	26.	City Manager	9.	Chamber of Commerce	52.	United Fund
4.	Farm Supply Co.	27.	County Board	10.	Banker's Association	60.	St. Hilary's Catholic Church
5.	Mechanical Co.	28.	Fire Department	18.	Bar Association	61.	1st Baptist Church
6.	Electrical Equip. Co.	29.	Human Relations Dept.	19.	Board of Realtors	62.	1st Church of the Light
7.	Metal Products Co.	30.	Mayor's Office	20.	Small Business Assoc.	63.	1st Congregational Church
8.	Music Equip. Co.	31.	Police Dept.	21.	Music Employee Union #1	64.	1st Methodist Church
11.	1st Towertown Bank	32.	Sanation Dept.	22.	Music Employee Union #2	65.	Unity Lutheran Church
12.	Towertown Savings & Loan	33.	Streets and Sanitation	23.	Teacher's Union	66.	University Methodist Church
13.	Bank of Towertown	34.	Park District	24.	Central Labor Union	69.	Family Services
14.	2nd Towertown Bank	35.	Zoning Board	36.	Democratic Committee	71.	YMCA
15.	Brinkman Law Firm	41.	Hospital Board	37.	Republican Committee	72.	Towertown Mental Health Center
16.	Cater Law Firm	42.	Public Hospital	38.	League of Women Voters		
17.	Knapp Law Firm	44.	Board of Mental Health	43.	Medical Society		
39.	Towertown News	45.	County Board of Health	48.	1st Kiwanis Club		
40.	WTWR Radio	47.	Highway Authority	49.	2nd Kiwanis Club		
		53.	School Board	50.	Rotary Club		
		54.	High School	51.	Lions Club		
		56.	Community College	55.	Parent-Teacher Assc.		
		57.	State University	58.	1st Assoc. of Churches		
		67.	Dept. of Public Aid	59.	2nd Assoc. of Churches		
		68.	Housing Authority				
		70.	Employment Services				
		73.	Youth Services				
	(g_1 = 16)		(g_2 = 24)		(g_3 = 21)		(g_4 = 12)

TABLE 6. MODELS FIT TO THE 10 2^6 TABLES FORMED BY THE PARTITION OF THE 73 ORGANIZATIONS INTO THE 4 GROUPS GIVEN IN TABLE 8 (528 CELLS)

Model	G^2	D.F.
(1) Separate models for each 2^6 table, each based on all multiplex mutuality and implied lower-order terms	136.0	176
(2) A common interaction structure for all 2^6 tables, based on all multiplex mutuality and implied lower-order terms, but one-factor choice parameters (θ_i's) depending on the groups	629.0	482
(3a) A common multiplex mutuality model for within group flows plus a between group model similar to model 2	409.0	352
(3b) Model (3a) plus a set of "information" multiplex parameters (θ_{11}) for between groups that depend on the groups	355.7	343

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) 14 TR-185

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 9 Technical Paper #185	2. GOVT ACCESSION NO. AD-A089194	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) 6 ANALYZING DATA FROM MULTIVARIATE DIRECTED GRAPHS: AN APPLICATION TO SOCIAL NETWORKS.		5. TYPE OF REPORT & PERIOD COVERED TR, to July 1980 6. PERFORMING ORG. REPORT NUMBER TP #185
7. AUTHOR(s) 10 Stephen E. Fienberg Michael M. Meyer Stanley S. Wasserman		8. CONTRACT OR GRANT NUMBER(s) 15 N00014-80-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Carnegie-Mellon University Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Contracts Office Carnegie-Mellon University Pittsburgh, PA 15213		12. REPORT DATE 11 Jul[redacted]80 13. NUMBER OF PAGES 29
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) 12 32		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

multivariate directed graph, social networks, stochastic loglinear models

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A multivariate directed graph consists of a set of g nodes, and a family of directed arcs (one for each relation) connecting pairs of nodes. Such multivariate directed graphs provide natural representations for social networks. In this paper we consider methods to analyse a network of 73 organizations in a Midwest American community linked by three types of

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

391190

relations: information, money, and support. The resulting data set, described by Galaskiewicz and Marsden (1978), involves 3 x 73 x 72 = 15,768 possible arcs or "observations". We describe a class of stochastic loglinear models for multivariate directed graphs, demonstrate how they can be fit to the data using generalized iterative scaling of Darroch and Ratcliff (1972), and explain the connection between these models and variants on standard loglinear models for multidimensional contingency tables discussed by Bishop, Fienberg, and Holland (1975). We also consider a disaggregation of the organizations into sub-groups, and demonstrate how to adapt our models to explore the intra- and inter-group relationships. These methods generalize research of Holland and Leinhardt (1980), who develop a model for dyadic relationships in univariate directed graph data. The paper includes a detailed analysis of the Galaskiewicz-Marsden data.