ELSEVIER

Social Networks 19 (1997) 243-269

SOCIAL
NETWORKS

Network analysis of 2-mode data

Stephen P. Borgatti [a,*], Martin G. Everett [b]

[a] *Dept. of Organizational Studies, Carroll School of Management, Boston College, Chestnut Hill, Boston, MA 02167, USA*
[b] *University of Greenwich, Wellington Street, London SE18 6PF, UK*

Abstract

Network analysis is distinguished from traditional social science by the dyadic nature of the standard data set. Whereas in traditional social science we study monadic attributes of individuals, in network analysis we study dyadic attributes of pairs of individuals. These dyadic attributes (e.g. social relations) may be represented in matrix form by a square 1-mode matrix. In contrast, the data in traditional social science are represented as 2-mode matrices. However, network analysis is not completely divorced from traditional social science, and often has occasion to collect and analyze 2-mode matrices. Furthermore, some of the methods developed in network analysis have uses in analysing non-network data. This paper presents and discusses ways of applying and interpreting traditional network analytic techniques to 2-mode data, as well as developing new techniques. Three areas are covered in detail: displaying 2-mode data as networks, detecting clusters and measuring centrality.

Keywords: *Networks; Clusters; Centrality*

1. Introduction

According to Wellman (1988), social network analysis differs from traditional social science in that traditional social science studies personal attributes whereas network analysis studies social relations. While we agree with this distinction, as methodologists, we would prefer to put it another way. In our view, it is better to say that traditional social science studies attributes of INDIVIDUALS (call these *monadic attributes*) whereas network analysis studies attributes of PAIRS OF INDIVIDUALS (call these *dyadic attributes*). Social relations are just one type of dyadic attribute. Other members of this set are distances (such as miles between cities), and similarities (such as correlations among respondents' responses across a set of questionnaire items).

Even in the case of social relations, the data that are actually collected are attributes of the relation, not the relation itself. For example, for the case of the friendship relation,

[*] Corresponding author. E-mail: steve_borgatti@msn.com

PII S0378-8733(96)00301-2

researchers might measure (for each pair of actors) the *strength* of the friendship (Krackhardt, 1990), which is one particular aspect of the friendship relation. It is also possible to measure other aspects of the relationship, such as its duration, the reason for its existence, etc.. In sexual networks, what is actually measured is typically the frequency of sexual contact with each partner. In communication networks, researchers may record both the frequency and the length of individual communications between pairs of actors.

The differential focus on monadic and dyadic attributes results in traditional social science and network analysis having different canonical data sets. In the traditional case, the canonical data set is a person-by-attribute matrix in which the persons are seen as cases and the monadic attributes are seen as variables. In the network case, the canonical data set is a person-by-person matrix, which is conceived of as recording a single social relation (or other dyadic attribute) among a set of actors. Here, the cases are the (ordered or unordered) pairs of actors, and the entire relation or dyadic attribute is a single variable.

Both data matrices may be described as having two dimensions or *ways*, which means simply that they have more than one row and more than one column [1]. The number of ways in a matrix is just the number of subscripts needed to identify each individual datum, as in x_{ij}. The data sets differ in the number of *modes*, which are distinct sets of entities pointed to by the subscripts. In the traditional data set, the two subscripts refer to two different sets of entities, persons and attributes. Hence, a methodologist might describe traditional social science as the study of 2-way 2-mode matrices. In contrast, in the network data set, both subscripts refer to the same set of entities, persons. The same methodologist might describe network analysis as the study of 2-way 1-mode matrices.

In practice, however, the distinction between traditional social science and network analysis is not nearly so neat. One reason is that there are ways of converting 2-mode data sets into 1-mode matrices [2], to which we can then apply the techniques (if not the theories) of network analysis. Another reason is that some 2-mode data are clearly relational in spirit and arise naturally in network research. For example, the assignment of faculty to courses may be seen as a relation between the set of faculty and the set of courses. Similarly the membership of individuals in voluntary organizations may be seen as a relation between two equally interesting sets. Still another reason is that some data can be recorded either as 1-mode or 2-mode, at the convenience of the researcher. For example, in some universities, faculty are asked to name one or more graduate students that they would like as research assistants, while, simultaneously, students are asked to name faculty that they would like to work with. If faculty and graduate students are regarded as separate sets of entities, the data are 2-mode, but if they are seen as a single set of entities (persons), the data are 1-mode.

This paper considers new methods of visualizing and analysing 2-mode data using

[1] Typically, 1-way matrices are simply called ‘arrays’.

[2] For example, we can compute correlations or other measures of similarity among the rows or the columns of the 2-mode matrix yielding a 1-mode correlation matrix.

network analytic techniques. In particular, we consider the following topics: visual representations of 2-mode data, clustering, centrality and structural similarity. Methods for dealing with 2-mode data have been suggested by a number of authors. Freeman (1980) and Doreian (1980) used Q-Analysis. Seidman (1981) proposed using hypergraphs. The idea we develop is based upon bipartite graphs and was first suggested by Wilson (1982). See also Breiger (1974) and McPherson (1982) for relevant discussions of 2-mode data.

2. Two or three views of 2-mode data in network analysis

Some 2-mode data sets are not seen as particularly relevant to network analysis. For example, if we randomly sample 1500 Americans and ask their opinion on a 100 attitude questions, we get a person-by-question 2-mode matrix that we do not normally regard as a candidate for network analysis. However, we could in fact apply network methods to the analysis of these data by deriving a dyadic variable from the data. For example, we can compute correlations among all pairs of respondents across all 100 attitudes to get a 1500-by-1500 person-by-person matrix which records the degree to which each pair of persons in the sample has similar attitudes. This 1-mode matrix can then be analyzed like any other dyadic attribute. We could then use cohesive subgroup algorithms to find groups of respondents with similar attitudes, or use centrality measures to identify respondents whose views are more ‘middle-of-the-road’ and less ‘fringy’ than others.

In other cases, 2-mode data sets are collected or constructed explicitly as an intermediary step toward the construction of a 1-mode network data set. For example, one may record, as Davis et al. (1941) did, the guest lists of a series of social events attended by society women. The data are arranged as a woman-by-event matrix X in which $x_{ij} = 1$ if the ith woman attended the jth event, and $x_{ij} = 0$ otherwise. The data matrix is shown in Fig. 1. Given matrix X, it is possible to construct the product of matrix X and its transpose XX′, whose ijth cell gives the number of events that *both* woman I and woman j attended. This value is interpreted as an index of the strength of social proximity of the two women.

```
                                      1 1 1 1 1
                1 2 3 4 5 6 7 8 9 0 1 2 3 4
                E1E2E3E4E5E6E7E8E9E1E1E1E1E1
                - - - - - - - - - - - - - -
 1     EVELYN   1 1 1 1 1 1 0 1 1 0 0 0 0 0
 2      LAURA   1 1 1 0 1 1 1 1 0 0 0 0 0 0
 3    THERESA   0 1 1 1 1 1 1 1 1 0 0 0 0 0
 4     BRENDA   1 0 1 1 1 1 1 1 0 0 0 0 0 0
 5 CHARLOTTE    0 0 1 1 1 0 1 0 0 0 0 0 0 0
 6    FRANCES   0 0 1 0 1 1 0 1 0 0 0 0 0 0
 7    ELEANOR   0 0 0 0 1 1 1 1 0 0 0 0 0 0
 8      PEARL   0 0 0 0 0 1 0 1 1 0 0 0 0 0
 9       RUTH   0 0 0 0 1 0 1 1 1 0 0 0 0 0
10      VERNE   0 0 0 0 0 0 1 1 1 0 0 1 0 0
11      MYRNA   0 0 0 0 0 0 0 1 1 1 0 1 0 0
12 KATHERINE    0 0 0 0 0 0 0 1 1 1 0 1 1 1
13     SYLVIA   0 0 0 0 0 0 1 1 1 1 0 1 1 1
14       NORA   0 0 0 0 0 1 1 0 1 1 1 1 1 1
15      HELEN   0 0 0 0 0 0 1 1 0 1 1 1 1 1
16    DOROTHY   0 0 0 0 0 0 0 1 1 1 0 1 0 0
17     OLIVIA   0 0 0 0 0 0 0 0 1 0 1 0 0 0
18      FLORA   0 0 0 0 0 0 0 0 1 0 1 0 0 0
```

Fig. 1. Davis, Gardner and Gardner data (DGG).

What this index of social proximity means exactly is not always clear. In some cases we would be willing to assume that strong proximity reflects a positive affective tie. In other cases, we would recognize that certain pairs of highly proximate women might not like each other at all (e.g., have a competitive relationship), but are still closely familiar with and influenced by each other. In still other cases, we would recognize the possibility that two women could co-attend a series of the same (large) events and not ever have even met each other, in which case we might regard the large value of XX'_{ij} as an index of the potential for some kind of tie to develop between a pair. In all of these cases, XX′ is regarded as representing the valued graph of a social network which could not be measured directly and was instead constructed from an intermediate data set X.

In still other cases, 2-mode data sets may be collected with the explicit intention that they will remain 2-mode data sets, but the spirit of the analysis is still relational in character. That is, the data consist of relations between two equally important sets of entities. For example, in a classic assignment problem, we may ask faculty to indicate which courses they would like to teach, or ask fraternity members which rooms they would like to occupy. The result is a matrix X in which $x_{ij} > 0$ if person i chooses item (e.g., room) j and $x_{ij} = 0$ otherwise. Here, what is of primary interest is which person is connected to which room or course, not how the persons are connected via the items, nor how the items are connected via the persons. Yet the latter two issues are not unimportant, and could still play an important part in the analysis.

In the first two cases there is no need to develop any new techniques to analyze 2-mode data, since the data are immediately converted to 1-mode data, for which the full range of network analytic methods are available. In contrast, for the last case, we need techniques that can work with the 2-mode data directly, without reducing it to 1-mode data first.

3. Visual representations of 2-mode data

One technique that is specifically designed for the analysis of relations between two modes is correspondence analysis. In simplified terms, correspondence analysis can be seen as a method for representing both the rows and columns of a 2-mode matrix as points in a metric space such that distances between the points are meaningful. Applied to the Davis, Gardner and Gardner data, a correspondence analysis results in a map in which (a) points representing the women are placed close together if the women attended mostly the same events, (b) points representing the social events are placed near each other if they were attended by mostly the same women, and (c) women-points are placed near event-points if those women attended those events [3]. A correspondence analysis map of the Davis, Gardner and Gardner data is given in Fig. 2.

[3] Actually, correspondence analysis includes an adjustment for marginal effects with the result that women are placed close to events to the extent that (a) those events were attended by few other women, and (b) those women attended few other events.

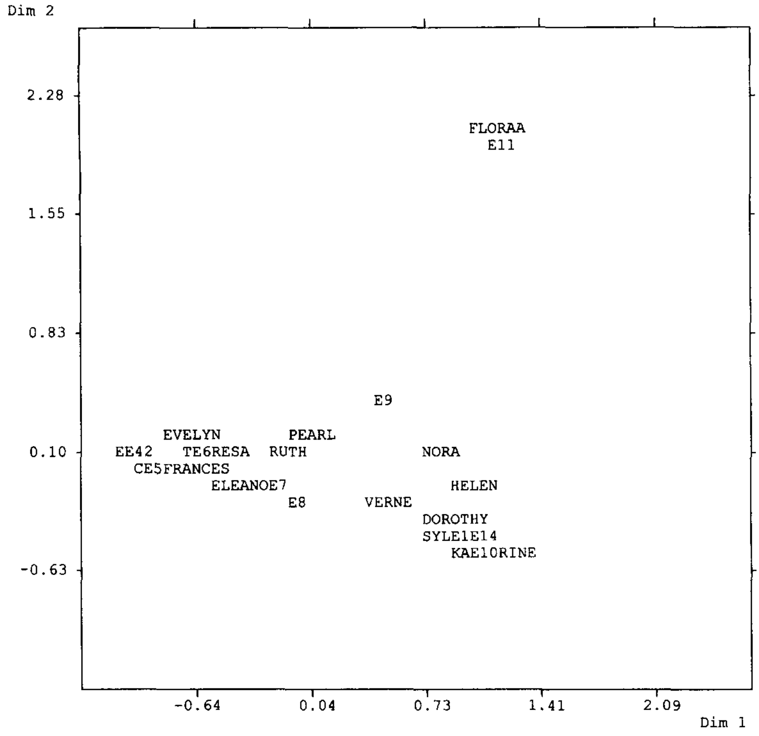

Fig. 2. Correspondence analysis of the Davis data. Note: some points are obscured by others.

Three problems should be noted with respect to correspondence analysis representations of these kind of data. First, because the data values in relational data sets have a severely limited range (all zeros and ones), they are difficult to fit using a continuous distance model of low dimensionality. This means that 2-dimensional maps will almost always be severely inaccurate and misleading. Second, correspondence analysis is designed to model frequency data, such as might arise from a Poisson, multinomial or product-multinomial sampling process. Yet the data values in the Davis, Gardner and Gardner data set are very different — they are not frequencies that just happen to have a restricted range. They are not, for example, the outcome of a set of Bernoulli processes with one trial. Rather, they are an arbitrary numerical representation of a discrete binary relation. Therefore, a modeled value of 0.7 (approximating 1.0) for any data cell may be meaningless, and the distances on the map difficult to interpret. In fact, there is no way, based on the 2-dimensional map, to determine which women attended what events, which would seem like an elementary criterion for representation adequacy. Third, the distances in correspondence analysis are not Euclidean, yet human users of the technique find it very difficult to comprehend the maps in any other way.

An alternative approach begins by treating the data as a bipartite graph. A graph is bipartite if the vertices may be partitioned in exactly two mutually exclusive sets such that there are no ties wholly within either set — i.e., the endpoints of every tie come

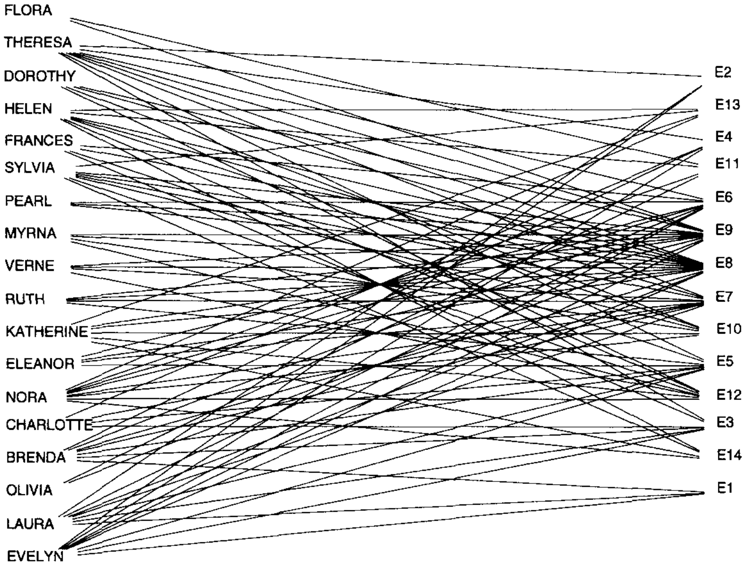

Fig. 3. Simple bipartite graph representation of the DGG dataset.

from different sets [4]. The advantage of the bipartite representation is that no data is lost: we always know which women attended which events.

Fig. 3 shows the Davis, Gardner and Gardner data represented as a classic bipartite graph. Armed with the knowledge that what is being represented is a mathematical graph, we can remind ourselves that the information in the figure is contained solely in the pattern of connections, not in the spatial positioning of the nodes or the length of lines. However, the picture is so complex that it is difficult for the mind to get a sense for the structure of the data.

A compromise approach is to combine the correspondence analysis and classical graph representations so that nodes are positioned spatially according to the coordinates from the correspondence analysis, but lines are drawn between women and events to show membership linkages. The result of this is shown in Fig. 4. The advantage of this approach is that the complexity of the figure is considerably reduced by the non-random placement of nodes in the space.

While more successful than either approach alone, the combination still has faults. For one thing, all the flaws of applying correspondence analysis to these kind of data are still present. We are tempted to interpret the distances between points, but this would be a mistake for the reasons mentioned earlier. Instead, we must interpret the figure as a representation of the bipartite graph drawn in such a way as to improve the aesthetic quality of the representation and consequently its readability. Yet the figure is clearly not optimal with respect to this last criterion. This is to be expected given that

[4] In addition, in this paper, we assume that the graph is connected and undirected.

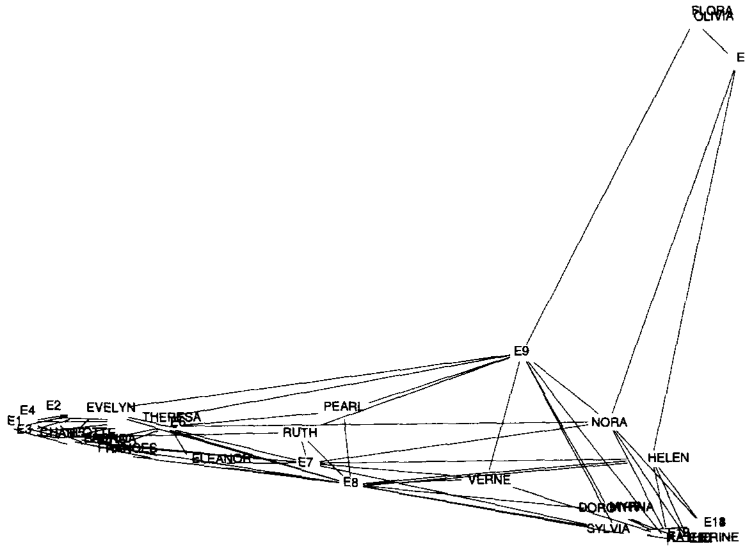

Fig. 4. DGG bipartite graph with points located according to coordinates derived from correspondence analysis.

correspondence analysis was not designed as an algorithm for drawing bipartite (or any other) graphs. If the objective is to represent the data in such as way that the mind can readily absorb its structure, then there are other methods which are better designed to deliver that benefit.

A more direct approach is as follows. First, compute geodesic distances between all pairs of nodes in the bipartite graph (see Fig. 5). Note that the geodesic distances among women (and among events) cannot be less than two (nor odd-valued) because women are not directly connected to other women (the same goes for events). Second, submit this geodesic distance matrix to ordinary multidimensional scaling. Fig. 6 shows the results of non-metric scaling, with ties between women and events superimposed. The map is quite revealing, making it easy to draw rough conclusions at a glance. For example, we can readily see that there are two groups of women and corresponding events. Even more interesting, however, is that we can see that there are some central women and events that are connected to both groups and serve to bring these groups together.

Davis, Gardner and Gardner also give a description of the social structure of the women. They identify two groups and within each group they have levels of participation. The first group consists of the women labelled 1 to 8 in Fig. 1. They identify a core within this consisting of the first four women, namely, Evelyn, Laura, Theresa and Brenda. The second group consists of the women 10 to 18; the core of this group is Sylvia, Nora and Helen. They place Ruth in both groups but at the lowest level. The map is in close agreement with this analysis.

A still more direct approach is to define what figure qualities contribute to aesthetic

```
                                     1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3
                   1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2
                   EVLATHBRCHFRELPERUVEMYKASYNOHEDOOLFLE1E2E3E4E5E6E7E8E9E1E1E1E1E1
                   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 1       EVELYN    0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 3 1 1 3 3 3 3 3
 2        LAURA    2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 4 1 1 1 3 1 1 1 1 3 3 3 3 3 3
 3      THERESA    2 2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 1 1 1 1 1 1 1 3 3 3 3 3
 4       BRENDA    2 2 2 0 2 2 2 2 2 2 2 2 2 2 2 2 4 4 1 3 1 1 1 1 1 1 3 3 3 3 3 3
 5    CHARLOTTE    2 2 2 2 0 2 2 4 2 2 4 4 2 2 2 4 4 4 3 3 1 1 1 3 1 3 3 3 3 3 3 3
 6      FRANCES    2 2 2 2 2 0 2 2 2 2 2 2 2 2 2 2 4 4 3 3 1 3 1 1 3 1 3 3 3 3 3 3
 7      ELEANOR    2 2 2 2 2 2 0 2 2 2 2 2 2 2 2 2 4 4 3 3 3 3 1 1 1 1 3 3 3 3 3 3
 8        PEARL    2 2 2 2 4 2 2 0 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 1 3 1 1 3 3 3 3 3
 9         RUTH    2 2 2 2 2 2 2 2 0 2 2 2 2 2 2 2 2 2 3 3 3 3 1 3 1 1 1 3 3 3 3 3
10        VERNE    2 2 2 2 2 2 2 2 2 0 2 2 2 2 2 2 2 2 3 3 3 3 3 3 1 1 1 3 3 1 3 3
11        MYRNA    2 2 2 2 4 2 2 2 2 2 0 2 2 2 2 2 2 2 3 3 3 3 3 3 3 1 1 1 3 1 3 3
12    KATHERINE    2 2 2 2 4 2 2 2 2 2 2 0 2 2 2 2 2 2 3 3 3 3 3 3 3 1 1 1 3 1 1 1
13       SYLVIA    2 2 2 2 2 2 2 2 2 2 2 2 0 2 2 2 2 2 3 3 3 3 3 3 1 1 1 1 3 1 1 1
14         NORA    2 2 2 2 2 2 2 2 2 2 2 2 2 0 2 2 2 2 3 3 3 3 3 1 1 3 1 1 1 1 1 1
15        HELEN    2 2 2 2 2 2 2 2 2 2 2 2 2 2 0 2 2 2 3 3 3 3 3 3 1 1 3 1 1 1 1 1
16      DOROTHY    2 2 2 2 4 2 2 2 2 2 2 2 2 2 2 0 2 2 3 3 3 3 3 3 3 1 1 1 3 1 3 3
17       OLIVIA    2 4 2 4 4 4 4 2 2 2 2 2 2 2 2 2 0 2 3 3 3 3 3 3 3 3 1 3 1 3 3 3
18        FLORA    2 4 2 4 4 4 4 2 2 2 2 2 2 2 2 2 2 0 3 3 3 3 3 3 3 3 1 3 1 3 3 3
19           E1    1 1 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 0 2 2 2 2 2 2 2 2 4 4 4 4 4
20           E2    1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 0 2 2 2 2 2 2 2 4 4 4 4 4
21           E3    1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 2 2 0 2 2 2 2 2 2 4 4 4 4 4
22           E4    1 3 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 0 2 2 2 2 2 4 4 4 4 4
23           E5    1 1 1 1 1 1 1 3 1 3 3 3 3 3 3 3 3 3 2 2 2 2 0 2 2 2 2 4 4 4 4 4
24           E6    1 1 1 1 3 1 1 1 3 3 3 3 3 1 3 3 3 3 2 2 2 2 2 0 2 2 2 2 2 2 2 2
25           E7    3 1 1 1 1 3 1 3 1 1 3 3 1 1 1 3 3 3 2 2 2 2 2 2 0 2 2 2 2 2 2 2
26           E8    1 1 1 1 3 1 1 1 1 1 1 1 1 3 1 1 3 3 2 2 2 2 2 2 2 0 2 2 2 2 2 2
27           E9    1 3 1 3 3 3 3 1 1 1 1 1 1 1 3 1 1 1 2 2 2 2 2 2 2 2 0 2 2 2 2 2
28          E10    3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 3 3 4 4 4 4 4 2 2 2 2 0 2 2 2 2
29          E11    3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 3 1 1 4 4 4 4 4 2 2 2 2 2 0 2 2 2
30          E12    3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 3 3 4 4 4 4 4 2 2 2 2 2 2 0 2 2
31          E13    3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 3 3 3 4 4 4 4 4 2 2 2 2 2 2 2 0 2
32          E14    3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 3 3 3 4 4 4 4 4 2 2 2 2 2 2 2 2 0
```

Fig. 5. Geodesic distances among nodes in the bipartite DGG graph.

approval and human readability, and design an optimization algorithm to minimize an appropriate cost function. Such definitions have already appeared in the network analytic literature (Chung and Borgatti, 1994, Krackhardt et al., 1994). Perhaps the most obvious

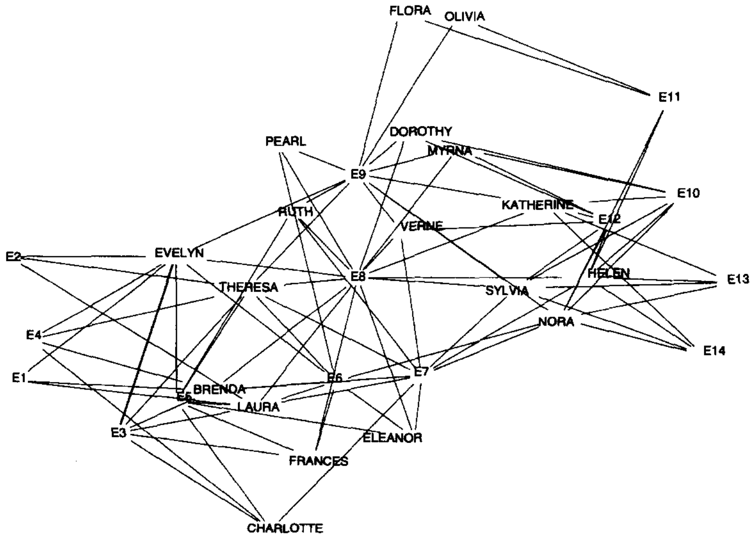

Fig. 6. Bipartite DGG data with nodes located by MDS of geodesic distances.

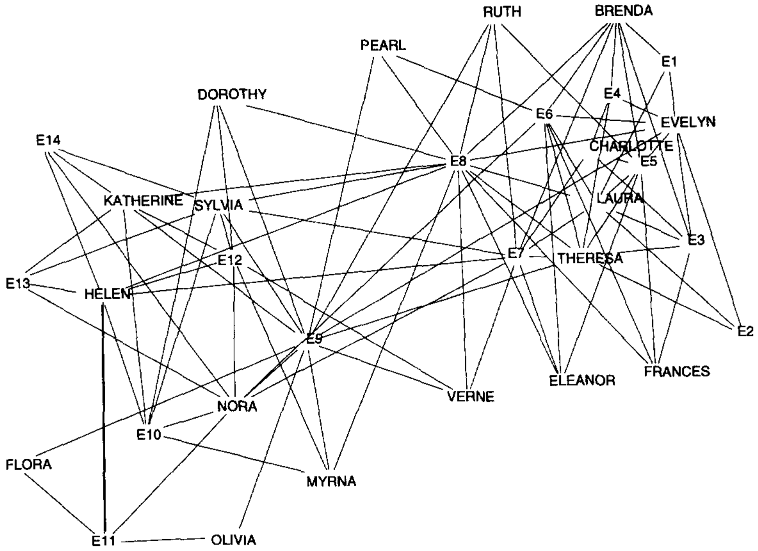

Fig. 7. MINLEN drawing of DGG bipartite graph.

criteria for readability are a minimum number of crossed lines, the preservation of some space around each node (that is, preventing nodes from being placed right on top of each other), and the placement of nodes near each other that are adjacent (directly connected by a tie). We have written a PASCAL program called MINCROSS based on a standard (Fletcher-Powell) function minimization algorithm (Press et al., 1989). The procedure tries to minimize a cost function containing terms for the number of crossed lines (Davidson and Harel, 1989), the spatial proximity of vertices, and the sum of Euclidean lengths of all lines. Unfortunately, calculating the number of crossed lines is expensive, and the MINCROSS algorithm is unacceptably slow for graphs of more than 20 points. However, we have found that if we eliminate the term for the number of crossed lines, the routine runs quite rapidly and the picture quality deteriorates only marginally. This is because minimizing the sum of lengths of all lines tends to reduce the number of crossed lines anyway, and particularly reduces the number of crossed *long* lines, which are the most visually disturbing. We call this new routine MINLEN [5].

Applied to the bipartite Davis, Gardner and Gardner data, MINLEN yielded the map in Fig. 7. The results are very similar to those of the MDS map in Fig. 6, but the groups are more clearly distinguished.

Both the MINCROSS and MINLEN programs accept a set of starting coordinates for each node as input. By starting the programs with nearly optimal initial coordinates, the user can greatly reduce the processing time. This feature enables the programs to be used to improve the visual output from correspondence analysis or MDS.

[5] MINLEN and other programs discussed in this paper are available via the INTERNET from the worldwide web site presently located at http://www.analytictech.com/download.htm.

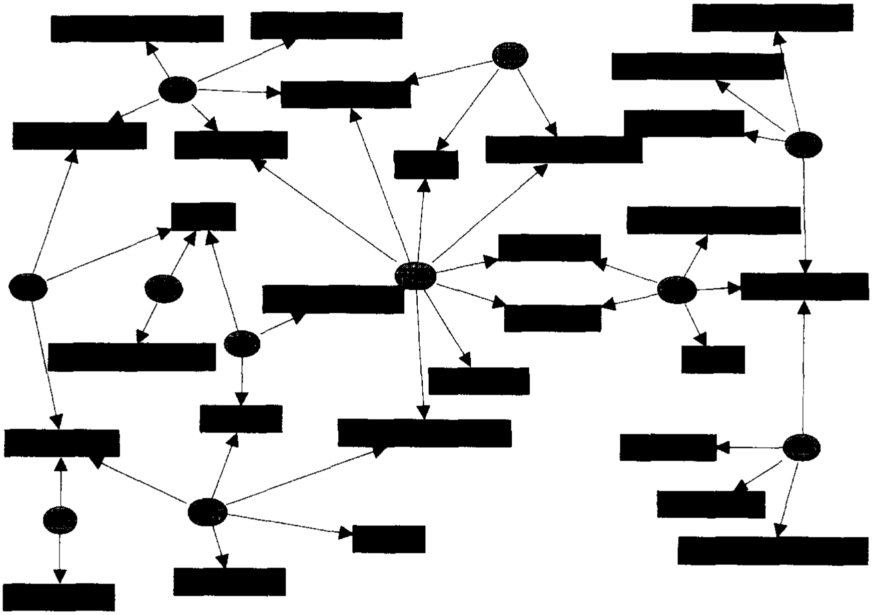

Fig. 8. Circles represent faculty, squares are courses. Arrows indicate which faculty chose which courses.

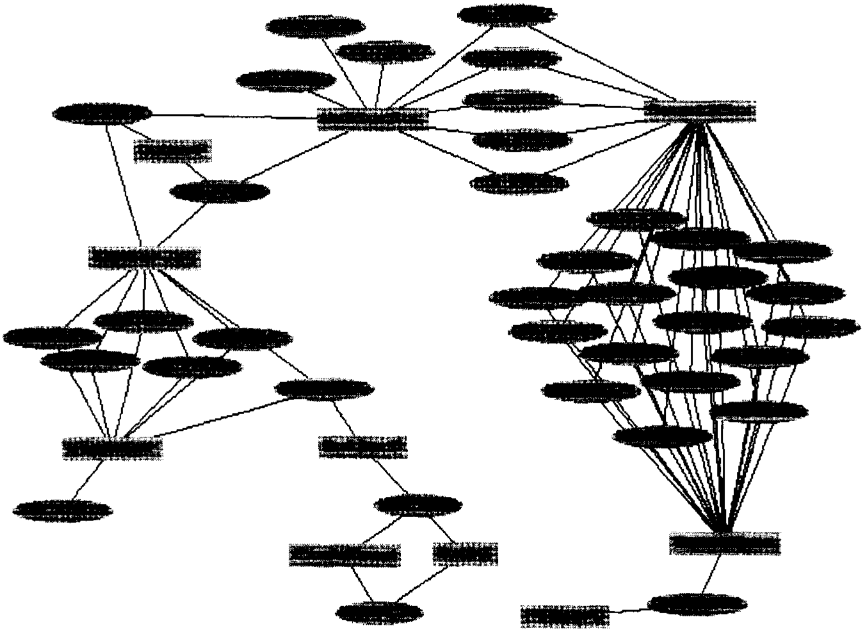

Fig. 9. Film data courtesy of Candace Jones, Boston College. Squares are project team members, ellipses are film projects.

An application of this methodology to data that is not normally thought of as network data is provided by the familiar situation in which an academic department asks its members which courses they would like to teach. Often, the results of such a questionnaire are presented as a list of courses with the names of persons choosing that course alongside, or as a list of persons followed by the courses they have selected. Such representations can be difficult to grasp overall, and require choosing a dominant mode of interest. In contrast, representing the data as a graph makes it easier to see global patterns, and permits the focus to be on persons, courses or relations between the two (see Fig. 8). For example, consider the structural similarity evident in the choices by TB and BJ on the far right of the graph. They compete for one course, and both choose three other courses which no others compete for.

Another example is provided by data drawn from film credits (Jones et al., 1996). Films are made by project teams that are assembled for just that one film, and then are dissolved when the film is finished. The members then go on to form part of other film teams. A partial graph of films made in 1977–1979 is shown in Fig. 9.

4. Density

One of the most basic attributes of a social network is its density — essentially a count of the number of ties present. To improve the interpretability of this value, it is standard practice to compute a normalized value by dividing the raw count by the maximum possible in a graph of the same size. A directed graph without self-loops has at most $n(n - l)$ possible edges and an undirected graph has half this value. These are the standard denominators used to divide into the number of observed ties. For the Davis, Gardner and Gardner data set (which is undirected) the standard density measure gives a value of 0.18, which would normally suggest a fairly sparse network.

However, the standard denominators are clearly not appropriate for our 2-mode data, since no ties are possible within vertex sets. The maximum number of ties possible occurs when every vertex in one set is connected to every vertex in the other. If the vertex sets are of size n_i and n_o then this amounts to $n_i n_o$ edges in the undirected case and $2n_i n_o$ in the directed case (again assuming self-loops are not allowed). Using this denominator, the Davis, Gardner and Gardner dataset has density 0.37, which is twice as high as previously calculated.

5. Centrality

Previous attempts (Bonacich, 1991) to measure centrality in 2-mode data sets have concentrated on methods that can be applied to the 2-mode matrix representation of the data. However, by representing a 2-mode data set as a bipartite graph instead, it is clear that we can mechanically if not sensibly utilize any standard measure of centrality. The main question is whether any shifts in interpretation are necessitated by the unusual nature of the data. In this section, we consider applying four standard measures of centrality, including degree, closeness, betweenness and eigenvector centrality.

5.1. Degree

The degree centrality of a node is defined as the number of edges incident upon that node. Applied to the Davis, Gardner and Gardner data, this means that the degree of a woman is the number of events she attended, and the degree of an event is the number of women who attended it. Thus, degree has a clear and simple interpretation in the 2-mode case.

It should be noted, however, that the normalization of degree recommended by Freeman (1979) and computed by programs like UCINET (Borgatti et al., 1990) is not necessarily the most appropriate for the 2-mode case. Freeman recommends dividing by the number of nodes in the network minus one, which is the theoretical maximum in an ordinary graph. In the case of a bipartite graph, however, the maximum degree of a node is given by the number of nodes in the opposing set. Hence, the maximum degree for a woman in the Davis, Gardner and Gardner data is the total number of events, and the maximum degree for an event is the total number of women. The only way that a node in a bipartite graph can achieve maximum degree in Freeman's terms is the case where one vertex set contains just one node and the other set contains all other nodes.

Consequently, if we regard the number of nodes in each vertex set as fixed, then we might prefer an alternative normalization in which we divide each score by the size of the opposite vertex set (i.e., the vertex set to which a given node does NOT belong). However, this means that the normalized scores would not be a linear transformation of the raw scores, unlike the Freeman normalization.

This is not necessarily a disadvantage, however, since we are after all considering different entities. While we expect a duality between the centrality scores across the modes it is not surprising that special adjustments are needed to make the centralities of, say, actors and events directly comparable.

Raw, normalized and 2-mode normalized degree measures for the Davis, Gardner and Gardner data set are given in the first three columns of Table 1. Note that the nonlinear normalization has changed the rank order of the centralities. For example event 14 and Dorothy both have the same degree and normalized degree value, but the 2-mode normalization gives a higher centrality score to Dorothy reflecting the fact there are fewer events than women. Note that here, and elsewhere in this paper, we shall express all normalized scores as a percentage.

5.2. Closeness

The closeness centrality of a node was defined by Freeman (1979) and is inversely proportional to the total geodesic distance from the node to all other nodes in the network. Geodesic distance is defined as the length (number of edges) of the shortest path linking two nodes. In a bipartite graph, all paths consist of an alternating series of nodes and edges of the form u–v–u^*–v^* where the u's represent nodes from one vertex set, the dashes represent edges, and the v's represent nodes from the other vertex set. The minimum distance of a node to another node in the same vertex set is 2.

As with degree centrality, the normalization of closeness proposed by Freeman is not appropriate for the 2-mode case, if one regards the size of the vertex sets as fixed. In

Table 1
Degree and closeness centrality

		1 Deg	2 NDeg	3 2mNDeg	4 Farness	5 Close	6 2mClose
1	EVELYN	8.00	25.81	57.14	60.00	51.67	80.00
2	LAURA	7.00	22.58	50.00	66.00	46.97	72.73
3	THERESA	8.00	25.81	57.14	60.00	51.67	80.00
4	BRENDA	7.00	22.58	50.00	66.00	46.97	72.73
5	CHARLOTTE	4.00	12.90	28.57	80.00	38.75	60.00
6	FRANCES	4.00	12.90	28.57	72.00	43.06	66.67
7	ELEANOR	4.00	12.90	28.57	72.00	43.06	66.67
8	PEARL	3.00	9.68	21.43	72.00	43.06	66.67
9	RUTH	4.00	12.90	28.57	68.00	45.59	70.59
10	VERNE	4.00	12.90	28.57	68.00	45.59	70.59
11	MYRNA	4.00	12.90	28.57	70.00	44.29	68.57
12	KATHERINE	6.00	19.35	42.86	66.00	46.57	72.73
13	SYLVIA	7.00	22.58	50.00	62.00	50.00	77.42
14	NORA	8.00	25.81	57.14	60.00	51.67	80.00
15	HELEN	7.00	22.58	50.00	62.00	50.00	77.42
16	DOROTHY	4.00	12.90	28.57	70.00	44.29	68.57
17	OLIVIA	2.00	6.45	14.29	82.00	37.80	58.54
18	FLORA	2.00	6.45	14.29	82.00	37.80	58.54
19	E1	3.00	9.68	16.67	84.00	36.90	52.38
20	E2	3.00	9.68	16.67	84.00	36.90	52.38
21	E3	6.00	19.38	33.33	78.00	39.74	56.41
22	E4	4.00	12.90	22.22	82.00	37.80	53.66
23	E5	8.00	25.81	44.44	74.00	41.89	59.46
24	E6	8.00	25.81	44.44	64.00	48.44	68.75
25	E7	10.00	32.26	55.56	60.00	51.67	73.33
26	E8	14.00	45.16	77.78	52.00	59.62	84.62
27	E9	12.00	38.71	66.67	56.00	55.36	78.57
28	E10	6.00	19.35	33.33	78.00	39.74	56.42
29	E11	4.00	12.90	22.22	82.00	37.80	53.66
30	E12	7.00	22.58	38.89	76.00	40.79	57.89
31	E13	4.00	12.90	22.22	82.00	37.80	53.66
32	E14	4.00	12.90	22.22	82.00	37.00	53.66

Freeman's normalization, the total distance score is divided into the quantity $n-1$, which represents the minimum score possible for a node in an ordinary graph. This is because it is possible for a node to be directly connected to all $n-1$ others, placing it a distance of 1 from each of these. Summing all of these unitary distances we get a total score of $n-1$. The beauty of this normalization is that it can be interpreted as the reciprocal of the average distance to all other nodes.

However, in the bipartite case, it is not possible for any node to be a distance of 1 from all other nodes. Instead, a node may be distance 1 from all nodes in the opposite vertex set, and distance 2 from all nodes in its own vertex set. (Actually, for our connected bipartite graphs, any node which is at a distance 1 from all nodes in the opposite vertex set must be a distance of 2 from all the nodes in its own vertex set.)

Therefore, the theoretical minimum raw score for a node is $n_i + 2n_o - 2$ where n_o is the size of the node's own vertex set and n_i is the size of the other vertex set. This formula, therefore, generates two different values for any bipartite graph with unequal vertex sets. Hence, if we regard the size of each vertex set as fixed, we should normalize closeness by dividing the raw score into one of these quantities, as appropriate. Hence, as in the degree case, we again obtain a nonlinear normalization.

Raw, normalized and 2-mode normalized closeness measures for the Davis, Gardner and Gardner data set are given in the last three columns of Table 1. The nonlinear normalization has had an even more dramatic effect on the closeness centrality. In all the degree and closeness centrality measures, except 2-mode normalized closeness, the second-most central node after E8 has been E9. However, in the 2-mode normalized closeness, the second-highest value of 80 is given to three women.

5.3. Betweenness

Betweenness may be roughly defined as the number of geodesic paths that pass through a given node, weighted inversely by the total number of equivalent paths between the same two nodes, including those that do not pass through the given node. In a bipartite graph, paths can originate and terminate at a node from either vertex set. In the case of the Davis, Gardner and Gardner data, this means that the betweenness of a woman (or an event) is a function of paths from women to women, from women to events (or vice versa), and from events to events.

In a bipartite graph, the only way that a node can achieve the theoretical maximum given by Freeman is if it is the only member of its vertex set. If we consider the size of each vertex set to be fixed, then it can be proved that the maximum is given by

$$\begin{array}{cc} 2(n_o - 1)(n_i - 1) & n_o > n_i \\ \frac{1}{2}n_i(n_i - 1) + \frac{1}{2}(n_o - 1)(n_o - 2) + (n_o - 1)(n_i - 1) & n_o \le n_i \end{array}$$

where n_o is the size of the node's own vertex set and n_i is the size of the other vertex set. (A proof of this and some of the other more mathematically involved results will be the subject of a separate paper.) The graph which gives these values consists of a node connected to all nodes in the opposite set, the remaining nodes are then connected pairwise so as to avoid concentrating ties on a single opposing node (see Fig. 7). This again produces a nonlinear normalization. As can readily be seen, the maximum centrality is a descending function of the relative size of a node's own vertex set. When $n_o = 1$, the equation (case $n_o \le n_i$) reduces to Freeman's absolute maximum, which is

$$\frac{n^2 - 3n + 2}{2}$$

where $n = n_i + n_o$.

Unlike closeness, betweenness can be said to have a built-in sense of exclusivity or competitiveness, such that a node is only central to the extent that it is the only node in its vertex set. Furthermore, if the vertex set contains two or more nodes, adding any ties beyond the minimum required to ensure connectedness can only reduce the centrality of the most central node.

5.4. Eigenvector centrality

Eigenvector centrality (Bonacich, 1972) is defined as the principal eigenvector of the adjacency matrix of a graph. It may be thought of as a weighted degree measure in which the centrality of a node is proportional to the sum of centralities of the nodes it is adjacent to. In the Davis, Gardner and Gardner case, this means that a woman's centrality is determined by the sum of the centralities of the events she attended, and, simultaneously, an event's centrality is determined by the sum of centralities of the women who attended it. This interpretation is identical (and the scores are proportional) to that of the first factor resulting from a singular value decomposition (SVD) of the raw 2-mode incidence matrix, which is the approach taken by Bonacich (1991). As he points out, this approach is also equivalent to computing eigenvectors of XX' and $X'X$, where X is again the raw 2-mode incidence matrix. Thus, all three approaches yield the same scores and two of them have the same interpretations.

Bonacich (1972) does not provide a normalization of eigenvector centrality. However, in the UCINET program (Borgatti et al., 1990), eigenvector centrality is normalized by dividing each raw eigenvector score [6] by the square root of one half, which is the maximum score attainable in any graph. Since it appears that the maximum occurs only in the center of a star graph, we can derive this maximum as a special case of the general principle that for any *complete* bipartite graph, the eigenvector score for any node is equal to

$$\sqrt{\frac{1}{2n_o}}$$

where n_o is the size of the vertex set the node belongs to. In the star graph, $n_o = 1$. This equation also gives the minimum score by substituting $n_i = n - n_o$ for n_o in the equation.

For our purposes we require the maximum value obtainable among all connected bipartite graphs in which the vertex sets have fixed sizes. This is precisely the same problem as for the betweenness case. It is interesting to note that eigenvector centrality resembles betweenness in that reducing the size of a node's vertex set generally improves its centrality score, and never worsens it, all else being equal. In fact, we conjecture that the relationship between the two concepts is closer than this and that the maximum eigenvector score occurs on precisely the same graphs that maximize the betweenness scores. We have performed a number of trials and this does seem to be the case. It is unlikely that a proof for this can easily be found. In fact, even for the 1-mode case there is no published proof that the star maximizes the denominator in the

[6] This terminology is unfortunate as the 'raw eigenvectors' in this paper correspond to what in the mathematical literature would be called 'normalized eigenvectors' since their Euclidean norm is unity.

Table 2
Betweenness and eigenvector centrality

		1	2	3	4	5	6
		Bet	NBet	2mNBet	Eig	NEig	2mNEig
1	EVELYN	42.76	9.20	9.67	0.22	31.27	32.71
2	LAURA	22.86	4.92	5.17	0.20	28.81	30.14
3	THERESA	38.74	8.33	8.76	0.25	34.84	36.44
4	BRENDA	22.01	4.73	4.98	0.21	29.15	30.49
5	CHARLOTTE	4.73	1.02	1.07	0.11	15.48	16.19
6	FRANCES	4.75	1.02	1.08	0.14	19.45	20.39
7	ELEANOR	4.14	0.89	0.94	0.15	21.54	22.53
8	PEARL	2.98	0.64	0.67	0.12	17.35	18.15
9	RUTH	7.36	1.58	1.67	0.16	22.65	23.69
10	VERNE	6.37	1.37	1.44	0.15	21.91	22.91
11	MYRNA	5.94	1.28	1.34	0.14	19.55	20.45
12	KATHERINE	16.29	3.50	3.69	0.17	24.09	25.20
13	SYLVIA	25.30	5.44	5.72	0.21	29.55	30.91
14	NORA	43.94	9.45	9.94	0.20	28.13	29.42
15	HELEN	30.73	6.61	6.95	0.18	25.41	26.58
16	DOROTHY	5.94	1.28	1.34	0.14	19.55	20.45
17	OLIVIA	2.09	0.45	0.47	0.05	7.00	7.33
18	FLORA	2.09	0.45	0.47	0.05	7.00	7.35
19	E1	0.97	0.21	0.22	0.09	12.99	13.25
20	E2	0.94	0.20	0.21	0.10	13.82	14.10
21	E3	8.20	1.76	1.81	0.16	23.15	23.62
22	E4	3.45	0.74	0.76	0.11	16.12	16.45
23	E5	16.98	3.65	3.76	0.21	29.58	30.18
24	E6	28.01	6.02	6.20	0.22	30.65	31.27
25	E7	58.10	12.49	12.85	0.27	37.48	38.24
26	E8	108.26	23.28	23.95	0.36	50.24	51.26
27	E9	96.23	20.69	21.29	0.27	38.27	39.05
28	E10	6.82	1.47	1.51	0.15	21.30	21.73
29	E11	9.02	1.94	2.00	0.07	9.83	10.03
30	E12	10.24	2.20	2.26	0.17	24.48	24.98
31	E13	1.89	0.41	0.42	0.11	15.60	15.92
32	E14	1.89	0.41	0.42	0.11	15.60	15.92

centralization formula[7]. Under the assumption that the conjecture is correct, we can proceed with normalization. However, we have not been able to derive a closed formula and must resort to numerical calculations. Essentially, we find the graph that maximizes betweenness centralization and then compute eigenvectors to obtain the needed denominator.

Raw, normalized and 2-mode normalized eigenvector and betweenness centrality scores for the Davis, Gardner and Gardner data set are given in Table 2. We note that in this table the centrality of E8 and E9 is not a factor of the size of the groups but is a real phenomenon. This is demonstrated by the fact that they still dominate after 2-mode

[7] The authors have a partial proof for this case but it still needs additional work.

normalization has been applied. If we look across all the tables and try and identify a group of most central women we see that Evelyn, Laura, Theresa, Brenda, Sylvia, Nora and Helen are the most central. These women are precisely those which Davis, Gardner and Gardner identified as core members of the two groups.

6. Centralization

In 1-mode data, an important concept is graph centralization. This measures the extent to which a particular network has a highly central actor around which highly peripheral actors collect. In all the examples of centrality mentioned above this is equivalent to providing a formal measure of the extent to which the network resembles a star. The general principle under which all centralization measures are now computed was proposed by Freeman (1979). In his formulation, to compute centralization we begin by summing the differences between the most central vertex and all other vertices. Then we normalize by dividing by the maximum possible, which is the value attained by a star graph. This is summarized by the formula

$$C_{\mathrm{G}} = \frac{\Sigma[C(p^*) - C(p_i)]}{\max\Sigma[C(p^*) - C(p_i)]}$$

where C_{G} is the centralization of the network G, C is any centrality measure, and the maximum is taken over all possible graphs of the same size.

With 2-mode data represented as a bipartite graph we could simply apply the centralization methods directly. But we again come across a problem of interpretation since we assume the two modes are fixed in size. If we could find the bipartite connected graphs with specified vertex sizes which give the maximum for the centralization formula then we could use these as the basis for our normalization. We therefore replace the denominator with a new maximum which is taken over all connected bipartite graphs of specified vertex sizes. It seems reasonable to suspect that these are the same graphs which give us the normalization for the centrality measures. This is indeed the case.

For degree centralization the denominator required in the centralization formula is

$$(n_{\mathrm{i}} + n_{\mathrm{o}})n_{\mathrm{i}} - 2(n_{\mathrm{i}} + n_{\mathrm{o}} - 1)$$

It is interesting to note that this is independent of the distribution of edges required to make the graph connected. That is, the maximum is achieved by any tree that has $K_{1,n_{\mathrm{i}}}$ as a subgraph. This is because the degree centrality measure is local: it is concerned only with the degree of the most central vertex which must be as high as possible and then we must apply the connectivity constraint. The distribution of degrees in the opposite vertex set is irrelevant. This local nature is not apparent in the non-bipartite case as there is only one connected graph which satisfies these constraints.

The formula above assumes that we have not normalized our degree centrality scores by the method suggested in the previous section. In the normal (non-bipartite) case this is not an issue since the normalization formula is a linear transformation (division by a constant). This means that the normalization divisor cancels out (this is not the case for closeness) and it does not matter whether you apply the formula to the normalized or un-normalized scores. However, in the bipartite case the normalization is not linear and this will in turn affect the centralization score. Clearly if we decide that the nonlinear normalization makes sense then we should base our centralization formula on it. In that case, we obtain the following denominator using the same notation:

$$(n_o + n_i - 1) - \frac{n_o - 1}{n_i} - \frac{n_i + n_o - 1}{n_o}$$

For our 2-mode data, each of these formulae can potentially provide us with two measures of centralization. We could measure to what extent the actors and events are centralized around a particular actor and to what extent actors and events are centered around a particular event. To obtain these measures we simply have to compute the numerator around the most central actor in each group. Note that it is possible for one of these values to be negative and this could provide difficulties in interpretation. Clearly we can avoid this problem if we just base our centralization on the most central actor or event in the graph as a whole and just have one measure of centralization.

Another approach to centralization in bipartite graphs is to develop what we shall call *single mode centralizations*. A single mode centralization measures the extent to which nodes in one vertex are central relative only to other nodes in the same vertex set. The nodes in the other vertex set are not ignored, however, as they are included in the computation of each node's centrality score. It is quite possible for there to be two very different structures internal to each mode of the dataset. It could happen that in one mode there are a lot of actors with a similar centrality score whereas in the other mode there may be a highly central event with very peripheral other events. We therefore have a network in which there is no centralization among the actors but a high degree of centralization among the events. The centralization of the whole network will give an artificial measure which would be an average of the two extremes. Since these are actually different modes, it would seem sensible to obtain values for each mode separately.

The single mode centralizations represent a significant advance over the traditional procedure of converting the 2-mode data to 1-mode data, then computing centrality and centralization. One reason is that most centrality measures, such as closeness and betweenness, are defined only for binary data, so after converting to 1-mode co-occurrence frequencies, the data must be dichotomized before computing centralization, which destroys information that is not lost in the single mode centralization approach. Another reason is that even without dichotomization, the conversion to 1-mode data destroys information about the pattern of overlaps. For example, if Evelyn and Laura shared 6 events, and Laura and Charlotte shared 3 events, and Charlotte and Evelyn shared 3 events, there is no way to know if the events that Charlotte and Evelyn shared are the same ones that Charlotte and Laura shared, even though we know that Evelyn and Laura shared many events. By calculating the centrality scores on the original 2-mode data and then restricting our attention to a single mode, we avoid losing that information.

For degree centrality we obtain the following denominator for the single mode centralization formula, we note that the sum in the numerator is now taken over just one of the modes and the centrality scores for the nodes in the other mode are ignored.

$$(n_i-1)(n_o-1)$$

Note that we do not have to worry about whether the centralities have been normalized, since with only one mode, we again have a linear normalization. We therefore give only the un-normalized formula. This formula requires more than one vertex in the mode we are computing our centralization on; therefore we cannot discuss centralization of a single vertex.

We can apply these general ideas to provide centralization and single mode centralization formulae for closeness and betweenness. Note that since closeness is a normalized measure there is no unnormalized version. (It would be possible to have a farness centralization but as this is not used in the non-bipartite case it is not presented here either.) Formulae for the denominators for closeness and betweenness centralization are given below.

Closeness centralization denominator:

$$2(n_i-1)\frac{n_i+n_o-2}{3n_i+4n_o-8}+2(n_o-n_i)\frac{2n_i-1}{5n_i+2n_o-6}$$
$$+2(n_i-1)\frac{n_o-2}{2n_i+3n_o-6}+2\frac{n_i-1}{n_o+4n_i-4} \qquad n_o>n_i$$
$$2(n_o-1)\frac{n_i+n_o-4}{3n_i+4n_o-8}+2(n_o-1)\frac{n_o-2}{2n_i+3n_o-6}$$
$$+2(n_o-1)\frac{n_i-n_o+1}{2n_i+3n_o-4} \qquad n_o\le n_i$$

Closeness single mode denominator:

$$\frac{(n_i-1)(n_o-2)}{2n_o-3}+\frac{(n_i-1)(n_o-n_i)}{n_o+n_i-2} \qquad n_o>n_i$$
$$\frac{(n_o-2)(n_o-1)}{2n_o-3} \qquad n_o\le n_i$$

It is interesting to note that the single mode closeness denominator for n_o less than n_i is independent of n_i. This shows that once n gets to a certain size it has no effect on the centralization of the vertices in the other set.

Betweenness centralization denominator (un-normalized):

$$2(n_o-1)(n_i-1)(n_o+n_i-1)-(n_i-1)(n_o+n_i-2)$$
$$-\frac{1}{2}(n_o-n_i)(n_o+3n_i-3) \qquad n_o>n_i$$
$$\left[\frac{1}{2}n_i(n_i-1)+\frac{1}{2}(n_o-1)(n_o-2)+(n_o-1)(n_i-2)\right]$$
$$(n_o+n_i-1)+(n_o-1) \qquad n_o\le n_i$$

Table 3
Centralization of DGG Data

2-Mode Degree Centralization

	Raw	Normalized	Single Mode
Actor	18.13	23.07	23.08
Event	50.97	46.61	46.61
Network Degree Centralization = 28.17			

2-Mode Closeness Centralization

	Normalized	Single Mode
Actor	28.43	21.35
Event	44.20	52.86
Network Closeness Centralization = 31.89		

2-Mode Betweenness Centralization

	Raw	Normalized	Single Mode
Actor	5.80	5.86	6.68
Event	20.73	20.70	19.82
Network Betweenness Centralization = 19.59			

Betweenness normalized centralization denominator:

$$(n_{\mathrm{o}} + n_{\mathrm{i}} - 1) - \frac{(n_{\mathrm{i}} - 1)(n_{\mathrm{o}} + n_{\mathrm{i}} - 2) + \frac{1}{2}(n_{\mathrm{o}} - n_{\mathrm{i}})(n_{\mathrm{o}} + 3n_{\mathrm{i}} - 3)}{\frac{1}{2}(n_{\mathrm{o}}(n_{\mathrm{o}} - 1) + \frac{1}{2}(n_{\mathrm{i}} - 1)(n_{\mathrm{i}} - 2) + (n_{\mathrm{o}} - 1)(n_{\mathrm{i}} - 1)} \qquad n_{\mathrm{o}} > n_{\mathrm{i}}$$

$$(n_{\mathrm{o}} + n_{\mathrm{i}} - 1) - \frac{(n_{\mathrm{o}} - 1)(n_{\mathrm{o}} + n_{\mathrm{i}} - 2)}{2(n_{\mathrm{o}} - 1)(n_{\mathrm{i}} - 1)} \qquad n_{\mathrm{o}} \leq n_{\mathrm{i}}$$

Betweenness centralization single mode (un-normalized version):

$$2(n_{\mathrm{o}} - 1)^2(n_{\mathrm{i}} - 1) \qquad n_{\mathrm{o}} > n_{\mathrm{i}}$$

$$(n_{\mathrm{o}} - 1)\left[\frac{1}{2}n_{\mathrm{i}}(n_{\mathrm{i}} - 1) + \frac{1}{2}(n_{\mathrm{o}} - 1)(n_{\mathrm{o}} - 2) + (n_{\mathrm{o}} - 1)(n_{\mathrm{i}} - 1)\right] \qquad n_{\mathrm{o}} \leq n_{\mathrm{i}}$$

Table 3 gives the various centralization scores for the Davis, Gardner and Gardner data. When we simply use the raw scores in the formula we obtain a bias against the larger data set, in this case the women. The normalized version always increases the actor centralization and decreases the event centralization. Since, in all the measures of centrality, events have the highest value, it makes the most sense to consider the normalized event centralization measure as the overall measure of centralization. Clearly this value must be higher than the network centralization reported in the table (the network centralization is simply the common centralization of the bipartite graph). It is interesting to note that it is significantly higher for degree and closeness but only marginally higher for betweenness. The reason for this is that the graph which maximizes the betweenness centrality score is very highly centralized itself, with a betweenness centralization of 94.5%. We also note the different consequences of the single mode centralization. For the degree case the results are similar to the normalized 2-mode; for closeness, actor centralization increased and event decreased, whereas the reverse occurs for the betweenness case. The reason for this is the greater dominance of the events over the actors in the betweenness centrality and this effect is ameliorated when single mode centralization is used. Overall, the picture is one in which there is a reasonable amount of centralization among the events and very little centralization among the actors. Clearly the greatest potential of these techniques (like their 1-mode counterparts) is to compare networks with each other.

We have not presented the results for eigenvector centralization but clearly the same methods could be applied; again we would need to perform all the computations numerically as we do not have formulae for the denominators.

7. Subgroups

One obvious feature of the centrality scores presented above is the disparity between the betweenness scores of events E8 and E9 and their scores on all the other centrality measures. This occurrence is characteristic of situations in which the nodes of a graph fall into two or more groups with some nodes acting as links between the groups. Indeed, glancing at the map in Fig. 7, it does appear that there is a left and a right group of nodes which are joined by E8 and E9. Note that these two groups do not correspond to women and events. Rather, each group consists of a set of women together with a set of events that they attended.

The obvious next step is to try to identify these subgroups using one of the standard approaches in the social networks toolkit. For example, we can search for cliques (Luce and Perry, 1949), n-cliques (Luce, 1950; Alba, 1973), n-clans, n-clubs (Mokken, 1979), k-plexes (Seidman and Foster, 1978), lambda sets (Borgatti et al., 1990) and ls-sets (Seidman, 1983). Unfortunately, these methods are not well suited for analysing a bipartite graph. In fact, bipartite graphs contain no cliques, as strictly defined by Luce and Perry. In contrast, bipartite graphs contain too many 2-cliques and 2-clans. The Davis, Gardner and Gardner graph contains 70 2-cliques, 65 2-clans, and 438 k-plexes ($k = 2$).

One of the problems is that, in the bipartite graph, all nodes of the same type are necessarily two links distant. While nodes of different types may be adjacent (e.g., a woman and an event she attended), thereby forming the kernel of a subgroup, but it is difficult to add a third node to the group because no matter what it is it can only be adjacent to one of the previous members: if it is an event it can only be adjacent to the woman, and if it is a woman it can only be adjacent to the event. Another problem is the existence of strong bridging nodes, like events E9 and E8. A meta analysis of the k-plex analysis, in which we count (for each pair of nodes) the number of k-plexes they both are members of, suggests a core/periphery structure in which the bridging nodes are at the core, and the nodes adjacent to them are in the semiperiphery, and so on.

Clearly we need to consider special types of subgraphs which are more appropriate

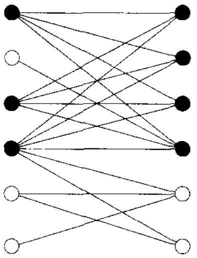

Fig. 10. Dark nodes form a biclique.

for 2-mode data. As with our centrality scores we do not expect comparisons between the 2 modes but a duality. To achieve this we need to take account of features such as the relative sizes of the two vertex sets. A maximally dense subgraph of a bipartite graph would simply be a complete bipartite graph. We define a biclique as a maximal complete bipartite subgraph of a given bipartite graph (see Fig. 10). For cliques we normally only consider cliques greater than size 2 and it would seem reasonable to adopt the same criteria for our 2-mode data. However, in this case, as each of the modes should form dual cohesive structures then it would seem reasonable to insist that we only consider bicliques of the form $K_{m,n}$ where m and n are greater than or equal to 3. Of course in analysing real data we may wish to increase or decrease these values depending on our data. Of the 70 2-cliques found in the Davis, Gardner and Gardner data only 24 are bicliques with minimum size (3,3) and only 4 have minimum size (4,4). We can use these bicliques to reveal some of the structure in the data which was beginning to be revealed by the visualization methods.

As already mentioned, one commonly employed technique for analysing cliques is to look at clusters of the clique overlap matrix. This operation is performed automatically on all cohesive subgroup methods contained within UCINET (Borgatti et al., 1990). We can perform exactly the same technique for bicliques. Fig. 11 contains this analysis of the biclique structure for the (3,3) bicliques. An examination of the hierarchical clustering of the (3,3) bicliques reveals two basic groups together with some outsiders. In the first group we have: Charlotte, Ruth, Frances, Eleanor, Evelyn, Laura, Brenda and Theresa together with events 4, 7, 3, 6, 5 and 8. The second group consists of Verne, Myrna, Dorothy, Helen, Nora, Katherine and Sylvia together with events 9, 10, 12, 13 and 14. This leaves us with Olivia, Flora and Pearl together with events 11, 2 and 1 as outsiders. These groupings have a remarkable amount of agreement with the plot shown in Fig. 7. We see for example the two major groups identified and the outsiders do all appear at the edge of the diagram. This is also in close agreement with the description given by Davis, Gardner and Gardner and the lattice analysis performed by Freeman and White (1993). If we now look at the (4,4) bicliques in Fig. 12 we see that we again have a similar structure but with a few important differences. We see that Ruth has been placed in a variety of cliques and this suggests she is a waverer between the two groups. Davis, Gardner and Gardner assigned Ruth to both major groups. The prevalence of event 8 throughout all the bicliques suggests that this event occupies a very central

```
                      C                                         K
                      H                                         A
                      A     F E             T         D         T
          O           R     R L     E     B H         O         H S
          L F   P     L     A E     V L   R E     V M R H       E Y
          I L   E     O   R N A     E A   E R     E Y O E N     R L
          V O e A     T   U C N     L U   N E     R R T L O   e I V e e e
          I R 1 R e e T e T E O e e Y R e D S e e N N H E R e 1 N I 1 1 1
          A A 1 L 2 1 E 4 H S R 7 3 N A 6 A A 5 8 E A Y N A 9 0 E A 2 3 4

          1 1 2   2 1   2       2 2     2     2 2 1 1 1 1 1 2 2 1 1 3 3 3
  Level   7 8 9 8 0 9 5 2 9 6 7 5 1 1 2 4 4 3 3 6 0 1 6 5 4 7 8 2 3 0 1 2
-------   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
11.0000   . . . . . . . . . . . . . . . . . XXX . . . . . . . . . . . . .
10.0000   . . . . . . . . . . . . . . . . . XXX . . . . . . . . . XXX . .
 9.6667   . . . . . . . . . . . . . . . . XXXXX . . . . . . . . . XXX . .
 8.7500   . . . . . . . . . . . . . . . . XXXXXXX . . . . . . . . XXX . .
 7.4643   . . . . . . . . . . . . . . . XXXXXXXXX . . . . . . . . XXX . .
 7.0119   . . . . . . . . . . . . . . XXXXXXXXXXX . . . . . . . . XXX . .
 7.0000   . . . . . . . . . . . . . . XXXXXXXXXXX . . . . . . . XXXXX . .
 6.7500   . . . . . . . . . . . . . . XXXXXXXXXXX . . . . . . XXXXXXX . .
 6.2755   . . . . . . . . . . . . . XXXXXXXXXXXXX . . . . . . XXXXXXX . .
 5.6696   . . . . . . . . . . . . XXXXXXXXXXXXXXX . . . . . . XXXXXXX . .
 4.6000   . . . . . . . . . . . . XXXXXXXXXXXXXXX . . . . . XXXXXXXXX . .
 4.1667   . . . . . . . . . . . . XXXXXXXXXXXXXXX . . . . XXXXXXXXXXX . .
 4.0000   . . . . . . . . . . . . XXXXXXXXXXXXXXX . XXX . XXXXXXXXXXX XXX
 3.8373   . . . . . . . . . . . XXXXXXXXXXXXXXXXX . XXX . XXXXXXXXXXX XXX
 3.4286   . . . . . . . . . . . XXXXXXXXXXXXXXXXX . XXX XXXXXXXXXXXXX XXX
 3.2500   . . . . . . . . . . . XXXXXXXXXXXXXXXXX . XXX XXXXXXXXXXXXXXXXX
 2.3000   . . . . . . . . . . XXXXXXXXXXXXXXXXXXX . XXX XXXXXXXXXXXXXXXXX
 2.0000   . . . . . . XXX . . XXXXXXXXXXXXXXXXXXX . XXX XXXXXXXXXXXXXXXXX
 1.6364   . . . . . . XXX . XXXXXXXXXXXXXXXXXXXXX . XXX XXXXXXXXXXXXXXXXX
 1.4881   . . . . . . XXX XXXXXXXXXXXXXXXXXXXXXXX . XXX XXXXXXXXXXXXXXXXX
 1.3125   . . . . . . XXX XXXXXXXXXXXXXXXXXXXXXXX . XXXXXXXXXXXXXXXXXXXXX
 1.1795   . . . . . . XXXXXXXXXXXXXXXXXXXXXXXXXXX . XXXXXXXXXXXXXXXXXXXXX
 0.8942   . . . . . . XXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXX
 0.5939   . . . . . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 0.1994   . . . . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 0.1923   . . . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 0.1441   . . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 0.0000   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Fig. 11. Clustering of (3,3) bicliques.

position; again this conclusion is matched by both our visual representation and all our centrality measures given in previous sections.

Clearly, we can define extensions of *n*-cliques, *n*-clubs and *n*-clans to *n*-bicliques, *n*-biclubs and *n*-biclans. The extensions are in many senses unnatural since *n* would need to be odd. One relaxation of the standard clique is the notion of a *k*-plex (Seidman and Foster, 1978). This can be used as a basis for relaxing the biclique condition. We define an $(1,m)$ biplex as a maximal bipartite graph with vertex sets V_1 and V_2 of sizes p and q, respectively, where every member of V_1 is connected to $(q-m)$ vertices in V_2 and every member of V_2 is connected to $(p-l)$ members of V_1 (see Fig. 13). Clearly a biclique is a (0,0) biplex. This relaxation has the advantage that we can take account of the different sizes of the vertex sets of the original graph. We can retain a strong condition for one vertex set and a much weaker condition for the other. We shall pursue the properties and applications of biplexes in a subsequent paper.

An alternative approach to finding subgroups is based on block modelling and traditional clustering techniques. The FACTIONS routine in UCINET takes the bipartite

```
1: MYRNA RUTH SYLVIA DOROTHY E8 E9 E10 E12
2: EVELYN LAURA THERESA BRENDA FRANCES E3 E5 E6 E8
3: LAURA THERESA BRENDA ELEANOR E5 E6 E7 E8
4: RUTH SYLVIA NORA HELEN E10 E12 E13 E14
```

Fig. 12. List of (4,4) bicliques.

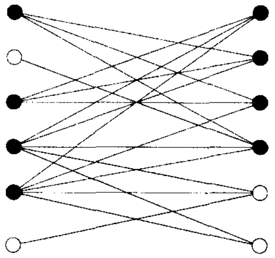

Fig. 13. Example of a (1,1) *biplex*.

graph as input and uses a combinatorial optimization algorithm called Tabu Search (Glover, 1989) to assign nodes to as many clusters as hypothesized by the user so as to maximize a fit criterion. The fit criterion is a correlation between the observed data and an idealized pattern in which the density of ties within groups is 100% and the density of ties between groups is 0%. The results of this routine (Fig. 14) further confirm our visual representation of the data. The revealed groups correspond to both the visual interpretation of the map in Fig. 7 and the clusters found using biclique overlap. There is also a near perfect agreement with the conclusions of Davis, Gardner and Gardner. The only difference is that Ruth is only placed in one group; this is of course inevitable using this routine as every actor must be assigned to one and only one group.

Note, however, that while the correlation between the blocked data and idealized pattern matrix is adequate, it can never be really high because the bipartite structure prevents ties of the same type from being adjacent. Hence, the maximum fit possible is

```
Correlation:   0.375

                1 1 1 1 1 1 1 1 2 1 2 2 3 3 3                     2 2 1 2 2 2 2 2
                7 8 1 2 3 4 5 6 9 0 7 8 1 0 2   9 1 2 3 4 5 6 7 8 5 6 9 0 1 2 3 4
                O F M K S N H D E V E E E E E   R E L T B C F E P E E E E E E E E
              ------------------------------------------------------------------
17    OLIVIA | 1               1   1         |
18     FLORA |   1             1   1         |
11     MYRNA |     1               1 1   1   |                     1
12 KATHERINE |       1             1 1 1 1 1 |                     1
13    SYLVIA |         1           1 1 1 1 1 |                   1 1
14      NORA |           1     1   1 1 1 1 1 |                   1             1
15     HELEN |             1   1     1 1 1 1 |                   1 1
16  DOROTHY |               1     1 1   1   |                     1
29       E11 | 1 1       1 1   1             |
10     VERNE |                   1 1     1   |                   1 1
27        E9 | 1 1 1 1 1 1   1   1 1         | 1 1   1         1
28       E10 |     1 1 1 1 1 1       1       |
31       E13 |       1 1 1 1           1     |
30       E12 |     1 1 1 1 1 1   1       1   |
32       E14 |       1 1 1 1               1 |
              ------------------------------------------------------------------
 9      RUTH |                     1         | 1                 1 1         1
 1    EVELYN |                     1         |   1                 1 1 1 1 1 1 1
 2     LAURA |                               |     1             1 1 1 1 1   1 1
 3   THERESA |                     1         |       1           1 1   1 1 1 1 1
 4    BRENDA |                               |         1         1 1 1   1 1 1 1
 5 CHARLOTTE |                               |           1       1       1 1 1
 6   FRANCES |                               |             1       1     1   1 1
 7   ELEANOR |                               |               1   1 1         1 1
 8     PEARL |                     1         |                 1   1           1
25        E7 |         1 1 1     1           | 1   1 1 1 1   1   1
26        E8 |     1 1 1   1 1   1           | 1 1 1 1 1   1 1 1   1
19        E1 |                               |   1 1   1             1
20        E2 |                               |   1 1 1                 1
21        E3 |                               |   1 1 1 1 1 1             1
22        E4 |                               |   1   1 1 1                 1
23        E5 |                               | 1 1 1 1 1 1 1 1               1
24        E6 |           1                   |   1 1 1 1   1 1 1               1
              ------------------------------------------------------------------
```

Fig. 14. Output of FACTIONS procedure with bipartite graph as input.

Fit: 0.518

		1 E	2 E	3 E	4 E	5 E	6 E	7 E	8 E	9 E	10 E	11 E	12 E	13 E	14 E
1	EVELYN	1	1	1	1	1	1		1	1					
2	LAURA	1	1	1		1	1	1	1						
3	THERESA		1	1	1	1	1	1	1	1					
4	BRENDA	1		1	1	1	1	1	1						
5	CHARLOTTE			1	1	1		1							
6	FRANCES			1		1	1		1						
7	ELEANOR					1	1	1	1						
8	PEARL						1		1	1					
9	RUTH					1		1	1	1					
10	VERNE							1	1	1			1		
11	MYRNA								1	1	1		1		
12	KATHERINE								1	1	1		1	1	1
13	SYLVIA							1	1	1	1		1	1	1
14	NORA						1	1		1	1	1	1	1	1
15	HELEN							1	1		1	1	1	1	1
16	DOROTHY								1	1	1		1		
17	OLIVIA									1		1			
18	FLORA									1		1			

Fig. 15. GENFAC2 routine results.

considerably less than 1.0. In cases where the data have clear enough structure, this is not a problem — we simply remember to adjust our expectations of the fit criterion downward. However, in cases where the structure is not terribly clear, the lack of ties

		V	H	B	C	D	A	S
1	P						1	
10	UF	1	1					
11	G	1	1	1				
4	C		1					
5	K		1	1				
6	Q			1				
7	B	1					1	
8	D	1			1		1	
26	BETA	1					1	
31	ZP	1	1		1			1
32	J	1	1					
12	GG	1		1				
34	UR	1		1				1
33	Y	1	1	1				
23	CC		1					
24	X		1	1				
13	TS				1	1	1	1
2	T				1		1	
3	TT				1			
18	F						1	1
21	INT				1			1
22	SP		1		1			1
19	THETA				1		1	
20	S				1		1	1
25	CHI			1				1
9	DD	1			1			
27	V	1					1	1
28	DC	1			1		1	
29	Z	1			1		1	1
30	ZZ	1			1			1
14	DZ	1			1	1	1	1
15	CV				1	1		1
16	JV	1			1	1		1
17	PHI						1	

Fig. 16. Presence/absence of features (columns) of language sounds (consonants only). Feature legend: V = voiced, H = high, B = back, A = anterior, S = strident, C = coronal, D = delay. Source: Crystal (1987).

within vertex sets could interfere with the algorithm's ability to find the best groupings.

Therefore, it seems advisable to develop a special procedure specifically designed for finding subgroups in 2-mode data. Taking a block modelling and combinatorial optimization approach, this is not a difficult task. We have written a new routine called GENFAC2 (see Footnote 5 for availability) which takes as input the raw 2-mode data matrix (women-by-events) and uses a genetic algorithm (Goldberg, 1989) to find a PAIR of partitions (one of the rows and one of the columns) that maximizes the same fit criterion described above. The results are given in Fig. 15. Note that one group consists of women 1–9 and events 1–8, while the other group consists of women 10–18 and events 9–14. These are the same groups found by the previous method, but these were found faster because the data matrix is smaller (18 × 14 rather than 32 × 32).

GENFAC2 is a general-purpose 2-mode clustering routine that can be used in a wide variety of settings. For example, linguists classify language sounds according to a series of features including the locations in the mouth where the sounds are formed. We can use an algorithm like GENFAC2 to simultaneously cluster the language sounds and the features into collections that go together. See Fig. 16 for the results.

8. Positions

Positions in 2-mode data have been discussed by Borgatti and Everett (1992). In essence we are able to apply our normal methods to the incidence matrix and this is equivalent to applying the techniques to the bipartite graph. Routines able to work directly on incidence data are available on the Internet (see Footnote 5), using algorithms similar to the 2-mode fractions presented earlier.

9. Discussion

The purpose of this paper was to develop network techniques for 2-mode data. We have shown how a number of standard methods can be applied to 2-mode data and briefly demonstrated how they work on one particular data set. There are a variety of potential applications for this type of analysis. For example, medical anthropologists frequently work with 1/0 illness-by-treatment matrices in which native informants indicate which treatments are used with which illnesses. Similarly, other anthropologists work with item-by-use matrices, frame substitution data, and other what-goes-with-what kinds of data. Taxonomists also use binary 2-mode species-by-characteristic matrices. Linguists characterize sounds in terms of a set of features that describe how the sounds are formed by the mouth. Marketing researchers describe product brands in terms of collections of binary features. Freeman and White (1993) also suggest a number of potential application areas for lattice representations; clearly any data which can make use of the lattice representation can make use of the methods outlined here (and vice versa). In addition, by dichotomizing data it is possible to use the techniques outlined here to analyze most social science data from contingency tables to informant-by-variable matrices.

References

R. Alba, A graph-theoretic definition of a sociometric clique, *Journal of Mathematical Sociology* 3 (1973) 113–126.

P. Bonacich, Factoring and weighting approaches to status scores and clique identification, *Journal of Mathematical Sociology* 2 (1972) 113–120.

P. Bonacich, Simultaneous group and individual centralities, *Social Networks* 13(2) (1991) 155–168.

S.P. Borgatti and M.G. Everett, Regular blockmodels of multiway, multimode matrices, *Social Networks* 14 (1992) 91–120.

S.P. Borgatti, M.G. Everett and P.R. Shirey, LS sets, lambda sets and other cohesive subsets, *Social Networks* 12 (1990) 337–357.

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

C.Y. Chung and S.P. Borgatti, A genetic algorithm for drawing graphs and lattices, *Presentation at the International Social Networks Conference*, New Orleans, LO, (Feb. 1994).

D. Crystal, *The Cambridge Encyclopedia of Language* (Cambridge University Press, New York, NY, 1987).

R. Davidson and D. Harel, Drawing graphs nicely using simulated annealing, *Technical Report CS 89/13* (Dept. of Applied Maths and Comp. Science, Weizmann Inst. of Science, Israel, 1989).

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

P. Doreian, On the evolution of group and network structure, *Social Networks* 2 (1980) 235–252.

L.C. Freeman, Centrality in social networks: I. Conceptual clarification, *Social Networks* 1 (1979) 215–239.

L.C. Freeman, Q-Analysis and the structure of friendship networks, *International Journal of Man-Machine Studies* 12 (1980) 367–378.

L.C. Freeman and D.R. White, Using Galois lattices to represent network data, *Sociological Methodology* 23 (1993) 127–146.

F. Glover, Tabu search — Part 1, *ORSA Journal on Computing* 1 (1989) 190–206.

D.E. Goldberg, *Genetic Algorithms* (Addison Wesley, New York, NY, 1989).

C. Jones, W.S. Hesterly, B. Lichtenstein, S.P. Borgatti and S.B. Tallman, *Intangible Assets of Teams: How Human, Social and Team Capital Influence Project Performance in the Film Industry*, Unpublished Manuscript (1996)

D. Krackhardt, Assessing the political landscape: structure, cognition, and power in organizations,*Admin. Science Quarterly* 35 (1990) 342–369.

D. Krackhardt, J. Blyth and C. McGrath, Krackplot 3.0: An improved network drawing program, *Connections* 17(2) (1994) 53–55.

D. Luce, Connectivity and generalized cliques in sociometric group structure, *Psychometrika* 15 (1950) 169–190.

D. Luce and A. Perry, A method of matrix analysis of group structure, *Psychometrika* 14 (1949) 95–116.

J.M. McPherson, Hypernetwork sampling: Duality and differentiation among voluntary organizations, *Social Networks* 3 (1982) 225–249.

R. Mokken, Cliques, clubs and clans, *Quality and Quantity* 13 (1979) 161–173.

W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vertterling, *Numerical Recipes in Pascal Cambridge* (Cambridge University Press, 1989).

S. Seidman, Structures induced by collections of subsets: A hypergraph approach, *Mathematical Social Sciences* 1 (1989) 381–396.

S. Seidman and B. Foster, A graph–theoretic generalization of the clique concept, *Journal of Mathematical Sociology* 6 (1978) 139–154.

S. Seidman, Internal cohesion of LS sets in graphs, *Social Networks* 5 (1983) 97–107.

B. Wellman, Thinking structurally, in B. Wellman and S.D. Berkowitz (eds.), *Social Structures: A Network Approach* (Cambridge University Press, Cambridge, 1988).

T.P. Wilson, Relational networks: An extension of sociometric concepts, *Social Networks* 4 (1982) 105–116.