Social Networks 11 (1989) 1–37
North-Holland

RETHINKING CENTRALITY: METHODS AND EXAMPLES *

Karen STEPHENSON **
Marvin ZELEN ***
Harvard University

A new model of centrality is proposed for networks. The centrality measure is based on the "information" contained in all possible paths between pairs of points. The method does not require path enumeration and is not limited to the shortest paths or geodesics. We apply this measure to two examples: a network of homosexual men diagnosed with AIDS, and observations on a colony of baboons. Comparisons are made with "betweenness" and "closeness" centrality measures. The processes by which structural changes in networks occur over time are also discussed.

1. Introduction

Networks are implicit in a wide range of social phenomena. The guiding assumption in the logic of network applications is that the overall structure of a network has consequences not only for individual members but for the group as a whole extending well beyond individual behaviors and social roles (cf. de Sola Pool and Kochen 1979; Kapferer 1969, 1973; Bott 1957; Boissevain 1974; Mitchell 1969, 1974; Barnes 1969, 1972). Assessing the quality of relations between people and understanding the pattern of connections has generated much interest and research in various disciplines (cf. Boissevain and Mitchell 1973; Fombrun 1986; Tichy and Fombrun 1979; Hage 1979; Harary 1959, 1970; T. Allen 1977).

* This paper was partially supported by grants MH-18006 and CA-23415 from the National Institutes of Mental Health and the National Cancer Institute. We are grateful to Dr. Brent James who wrote the computer program for calculating the closeness and betweenness measures used in our AIDS example.
** Department of Anthropology, Faculty of Arts and Sciences, Harvard University, Cambridge, MA 02138, U.S.A.
*** Department of Biostatistics, School of Public Health, Harvard University, and the Dana Farber Cancer Institute, Boston, MA 02115, U.S.A.

Our aim in this paper is to propose a technical measure of centrality for the analysis of networks which makes use of all paths between pairs of points. It has a rationale and is readily interpretable. The calculations are relatively straightforward and can be done for large networks. This paper is divided into six sections. Section 2 describes previous theoretical developments as background for rethinking the centrality concept. Section 3 of the paper outlines our basic ideas and applies them to a simple network (both with and without weights). Section 4 consists of the application of our methods to two examples having subject matter interest, i.e., one on a network of homosexuals diagnosed with AIDS and the other on observations of social behavior in a baboon colony. Section 5 discusses our main conclusions and an Appendix contains the mathematics underlying our methods. The discussion in Section 3 suffices for those who wish to understand the method and apply it.

2. Background

The study of networks has generated a large literature. Conceptual clarification, statistical and methodological organization has been provided by Frank 1981; Burt 1978, 1980; Freeman 1979a,b, 1980; Hage and Harary 1983; White *et al.* 1976; Boorman 1976; Mizruchi and Bunting 1981; Cook *et al.* 1983; Bonacich 1987; Johannisson 1987. While recognizing the merits of structural approaches, network studies remain largely ad hoc in nature. In the early history of the discipline, technical ideas often raced ahead of application leading to a criticism that network analysis provided nothing more than a superfluous language that served to recognize what was patently obvious. Such criticism failed to recognize that these techniques yielded results that added substantially to our understanding of social and cultural processes and could not have been obtained by simple common sense notions for large or complex networks.

A review of key centrality concepts can be found in the papers by Freeman (1979a,b). His work has significantly contributed to the conceptual clarification and theoretical application of centrality. Motivated by the work of Nieminen (1974), Sabidussi (1966), and Bavelas (1948), he provides three general measures of centrality termed

"degree", "closeness", and "betweenness". His development is partially motivated by the structural properties of the center of a star graph. The most basic idea of point centrality in a graph is the adjacency count of its constituent points. "The degree of a point, p_i, is simply the count of the number of encounters with other points $p_j (i \neq j)$, that are adjacent to it and with it is, therefore in direct contact" (Freeman 1979a: 219). The second measure relates to the closeness or distance between points. It is based upon the extent to which a point is close to all other points using the shortest path or geodesic (Freeman 1979a: 225). The third measure is called betweenness and is the frequency at which a point occurs on the geodesic that connects pairs of points. Thus, any point that falls on the shortest path between other points can potentially control the transmission of information or effect exchange by being an intermediary. "It is this potential for control that defines the centrality of these points" (Freeman 1979a: 221).

In a later article, Freeman (1980) concludes that "betweenness and closeness based measures of point centrality are determined by the same structural elements of a communication network" (p. 591). Since both are functions of local pair dependency, all measures have in common the same structural element: the geodetic pathway. If one assumes that communication only occurs long the shortest possible path, then communication channels are by default the geodesics. This has always been a fundamental assumption in his development of betweenness, closeness and pair dependence. This theoretical choice neglects measuring communication occurring along reachable, non-geodetic pathways.

It is quite possible that information will take a more circuitous route either by random communication or may be intentionally channeled through many intermediaries in order to "hide" or "shield" information in a way not captured by geodesic paths. These considerations raise questions as to how to include all possible paths in a centrality measure. The measures of centrality discusses above each have their limitations. In studying very large networks, such limitations can have significant ramifications for understanding *total* network processes. For example, it may become necessary to decompose large networks into component networks to avoid computational problems. Decomposition of networks for computational purposes is limiting because it can sacrifice or compromise subtle network infrastructures.

Another area of research on centrality which does not make use of geodesic paths is the work of Bonacich (1972a, b, 1987). In his earlier work he proposes as a centrality measure the eigenvector associated with the largest characteristic eigenvalue of the adjacency matrix. However, this approach neglects multiple shared paths between points in a network (see Section 3.1). More recently he has proposed an important modification which uses a scalar parameter β which weights indirect paths having $(r+1)$ lines by β^r. The measure of centrality for a point is the sum of the weighted paths emanating from that point. An additional difficulty with this approach is that the value of β is arbitrary.

Putting aside the problem of including all possible paths, Donninger (1986) and others articulate another critical aspect of network analysis —that of graph comparability. How can a centrality measure be interpreted for a network of different sizes? What implications does centrality have for clique formation (cf. Alba 1973; M. Allen 1982; Mizruchi 1984), dense or sparse networks (cf. Granovetter 1973; Salzinger 1982; Mariolis 1982), and a changing pattern of linkages over time with the addition and deletion of new members (cf. Barnett and Rice 1985; Tutzauer 1985; Doreian 1980). We discuss some of these notions in our examples.

Our proposed measure of centrality uses all paths, but gives them relative weighting as a function of the "information" they contain. Information is technically defined in our development. The method does not require the enumeration of paths and the calculations are relatively simple. It only requires the inversion of a symmetric matrix having dimension equal to the number of nodes or points in the network. We believe that the ease of calculation of our method without the use of combinatorics or graph-theoretic techniques and the fact that it accounts for all paths will accelerate the practical applications of networks with regard to investigations of: emergent social structures, temporal changes in networks, and changes in networks when nodes and/or communication links are added or deleted.

3. Main ideas and computations

This section discusses and outlines the computations for a proposed new measure of centrality. The general theory and statistical ideas on

which the method is based are in the Appendix. Essentially our ideas are motivated from the theory of the statistical design of experiments and the general ideas of statistical estimation. There is a one-to-one correspondence between networks and these topics. In this section we present our basic ideas with a minimum of technicalities. Section 3.1 is a heuristic development which discusses the consequences of the theory; Section 3.2 illustrates how to directly carry out the necessary calculations; and Section 3.3 discusses how to extend the calculations when there are weights. The new method for measuring centrality departs from several other proposed methods in that it makes use of all paths between points rather than geodesic paths. Furthermore, the calculations are easily organized and are feasible for large networks. This measure of centrality is interpretable and in a certain mathematical sense, cannot be improved on. It is motivated by very different theoretical considerations than in the works described above.

3.1. Heuristic development

Consider a network with n points or nodes and m lines or edges. We shall limit ourselves to non-directed networks in which all pairs of points are reachable although later we shall discuss the implications then there exist pairs of points which are not reachable. To motivate ideas we shall refer to the network of five points and six lines shown in Figure 1*a*.

If two points are connected by the same line, they are said to be incident and the path is referred to as an incident path. The distance

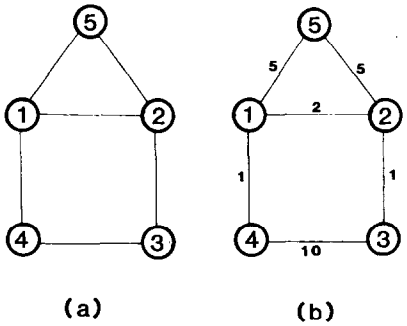

Fig. 1. Network with 5 points and 6 lines. Network (*a*) is without weights; (*b*) is the same topological network with weights for the communication links.

between them will be taken as one unit. In our network (Figure 1*a*) the pairs of points (1, 2), (2, 3), (3, 4), (4, 1), (5, 1), (5, 2) are all incident. In general, however, if (i, j) refer to a pair of points there may be paths other than incident paths that connect them. Suppose for points (i, j) there are k_{ij} paths connecting i and j. These paths will be denoted by $P_{ij}(1), P_{ij}(2), \ldots, P_{ij}(k_{ij})$. Note that the use of this notation "symbolically" represents the specific paths.

For example, for the pair (1, 2), there are three paths which are described by $P_{12}(1) = 1\text{–}2$, $P_{12}(2) = 1\text{–}5\text{–}2$, $P_{12}(3) = 1\text{–}4\text{–}3\text{–}2$. There is only one geodesic path $P_{12}(1)$, but in our measure of centrality all paths are taken into consideration as influencing communication and information. We define a distance measure of a path as the number of lines in the path. Let $D_{ij}(s)$ be defined as the number of lines in path $P_{ij}(s)$, then for our example $D_{12}(1) = 1$, $D_{12}(2) = 2$, $D_{12}(3) = 3$. However all paths do not have the same information and we define an information measure $I_{ij}(s)$ to be the reciprocal of the distance measure, i.e. $I_{ij}(s) = 1/D_{ij}(s)$.

There are essentially two definitions of "information" which are widely used today. One is in the theory of communication and the other is used in the theory of statistical estimation. Our ideas were motivated from the theory of statistical estimation. The statistical definition of the information of a single observation from a normal or Gaussian distribution is the reciprocal of the variance of the observation. Thus, the path $P_{12}(1) = 1\text{–}2$ can be envisioned as a "signal" from point 1 to point 2. The "noise" in the transmission of the signal is measured by the variance of the signal in going from point 1 to point 2. The path $P_{12}(2) = 1\text{–}5\text{–}2$ may be interpreted as the transmission of two independent signals, e.g. 1 to 5 followed by 5 to 2. Since two transmissions are required for a signal from point 1 to reach point 2 through point 5, we can rewrite the path as the sum of two incident paths, $(1\text{–}5) + (5\text{–}2)$. Since the variance of a sum of independent signals is additive we have;

$$\text{Variance}[(1\text{–}5) + (5\text{–}2)] = \text{Variance}(1\text{–}5) + \text{Variance}(5\text{–}2).$$

If the variance of any signal is unity, then the variance of a path simply counts the number of incident paths. Thus the variance $[(1\text{–}5) + (5\text{–}2)] = 2$. The increased variance reflects the condition that there is more "noise" in two transmissions than in one transmission. The length of

any path is simply the variance of transmitting a signal from the first point of a path to the last point in the path. The measure of information of this transmission (using the definition of information from the theory of statistical estimation) would be the reciprocal of the variance.

Since there may be several paths going from i to j (or j to i), we introduce the idea of a combined path which is a weighting function of the individual paths. The "weighted" function of the combined paths from i to j will be denoted by P_{ij} and is defined by

$$P_{ij} = \sum_{s=1}^{k_{ij}} w_s P_{ij}(s), \qquad \sum_{s=1}^{k_{ij}} w_s = 1, \tag{1}$$

where the weights $\{w_s\}$ have to be determined. These weights are numeric quantities and are chosen so that the information in the combined path (denoted by I_{ij}) is maximized. Referring to our example, the optimal combined path for points (1, 2) is obtained by weighting each path proportional to its information. This results in

$$P_{12} = \frac{1P_{12}(1) + (1/2)P_{12}(2) + (1/3)P_{12}(3)}{1 + (1/2) + (1/3)}$$

$$= \frac{6P_{12}(1) + 3P_{12}(2) + 2P_{12}(3)}{11},$$

Note that the incident path has a weight of 6/11 which is greater than the combined weight of the other two paths. The formula for the weights for this example is

$$w_s = \frac{I_{ij}(s)}{I_{ij}(1) + I_{ij}(2) + \cdots + I_{ij}(k_{ij})}, \qquad s = 1, 2, \ldots, k_{ij},$$

which refers to the proportion of the information associated with each path. The information in the combined path P_{12} is given by the sum of the information of the component paths and is the denominator of w_s. Hence for this example $I_{12} = \sum_{s=1}^{3} I_{12}(s) = 1 + \frac{1}{2} + \frac{1}{3} = 1.83$. The interpretation of this number is that the combined path has 83 percent more information than the incident or geodesic pair alone. The numerical result of $I_{12} = 1.83$, based on the given weights, represents the maxi-

mum information for combining the paths. Any other value for the weighting function will result in a smaller amount of information. When one only considers geodesic paths, this would be equivalent to giving the non-geodesic paths a weight of zero in equation (1). It is clear using our definition of information that this will result in non-optimal information.

A combined path for i to j may be interpreted as a signal originating with point i and targeted for point j. The signal at i is transmitted to all of the incident paths associated with point i and will reach point j from each possible path. The estimate of the signal at point j is obtained by weighting the signal from each path proportional to its information. (This weighting procedure is optimum in the sense that any other choice of weights which sum to unity will have less information.) The optimal weighting also has the property that the information of the combined path is equal to the sum of the information from each individual path. The idea of a combined path also arises from the theory of statistical estimation. If one has several estimates of a common parameter, each with a different variance, then the pooled or combined estimate having minimum variance is obtained by weighting each estimate proportional to the inverse of its variance. This is exactly how we are proceeding with the combined path where each path is weighted by its information (reciprocal of its variance).

Our interpretation of "information" is that every path can be evaluated for its information content. Generally the information is inversely proportion to the distance of a path and the information in a combined path is equal to the sum of the information of the individual paths. (This is true when the component paths are independent and contain no common incident paths. The more general case is discussed below.)

The proposal for determining the centrality of any point (i) is to first determine the information of point (i) with all other points; i.e., $I_{i1}, I_{i2}, \ldots, I_{in}$. The information on centrality of point (i) will be defined as the harmonic average of the information associated with the path from (i) to the other points. Specifically, if I_i refers to the centrality or information of (i), then

$$I_i = \frac{n}{\sum_{j=1}^{n} 1/I_{ij}}, \tag{2}$$

where we define $I_{ii} = \infty$. Equation (2) is used to determine the information of (i) in the network.

The three paths going from 1 to 2 had component paths which were independent of one another. By independent, we mean that every path made use of different lines. No pair of paths had any lines in common. In general this is not true and one has to modify the measure of information when component paths share common lines. We shall illustrate the problem of non-independent paths and its solution with the pair of points (1, 3) of Figure 1a. There are three paths which can be distinguished, i.e.

$$P_{13}(1) = 1\text{–}4\text{–}3, \qquad P_{13}(2) = 1\text{–}2\text{–}3, \qquad P_{13}(3) = 1\text{–}5\text{–}2\text{–}3.$$

Note that $P_{13}(2)$ and $P_{13}(3)$ have the incident pair 2–3 in common. It is necessary to take this feature into account in finding the appropriate weights for the combined path and the resulting information. In order to carry this out we define:

$D_{ij}(s)$ = number of lines in path $P_{ij}(s)$,

$D_{ij}(r, s)$ = number of lines in common between $P_{ij}(r)$ and $P_{ij}(s)$.

When $r = s$ we define $D_{ij}(s) = D_{ij}(s, s)$.
The matrix of these quantities for (1, 3) in our example is

$$D_{13} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix}.$$

($P_{13}(1)$ has no lines in common with the other two and hence has zero in the off diagonal elements in the first row and column.)

In order to obtain the combined path and the calculation of the information, it is necessary to invert the matrix D_{ij}. For our example the inverse of D_{13} is

$$D_{13}^{-1} = \tfrac{1}{10}\begin{bmatrix} 5 & 0 & 0 \\ 0 & 6 & -2 \\ 0 & -2 & 4 \end{bmatrix}.$$

The information in the combined path P_{13} is the sum of all the

elements of the D_{13}^{-1} matrix; i.e., $I_{13} = 11/10 = 1.1$. The weights for the combined path are the sum of each row of D_{13}^{-1} divided by the total information, e.g.

$$w_1 = \frac{5/10}{11/10} = 5/11, \quad w_2 = \frac{4/10}{11/10} = 4/11, \quad w_3 = \frac{2/10}{11/10} = 2/11.$$

Hence the combined path can be written

$$P_{13} = [5P_{13}(1) + 4P_{13}(2) + 2P_{13}(3)]/11.$$

The information ($I_{13} = 1.1$) in path P_{13} is 10 percent more than if the

Table 1
Enumeration of paths and information

Point pairs (i, j)	Paths: $P_{ij}(s)$	Information: I_{ij}
1, 2	1–2	1.8333
	1–5–2	
	1–4–3–2	
1, 3	1–4–3	1.1000
	1–2–3	
	1–5–2–3	
1, 4	1–4	1.3750
	1–2–3–4	
	1–5–2–3–4	
1, 5	1–5	1.5714
	1–2–5	
	1–4–3–2–5	
2, 3	2–3	1.3750
	2–1–4–3	
	2–5–1–4–3	
2, 4	2–1–4	1.1000
	2–3–4	
	2–5–1–4	
2, 5	2–5	1.5714
	2–1–5	
	2–3–4–1–5	
3, 4	3–4	1.3750
	3–2–1–4	
	3–2–5–1–4	
3, 5	3–2–5	0.8462
	3–4–1–5	
	3–2–1–5	
4, 5	4–1–5	0.8462
	4–3–2–5	
	4–3–2–1–5	

points 1 and 3 were incident. In general, if the elements of D_{ij}^{-1} are denoted by $I_{ij}(r, s)$, then the information for the combined path is

$$I_{ij} = \sum_{r=1}^{n} \sum_{s=1}^{n} I_{ij}(r, s). \tag{3}$$

To complete our illustration, Table 1 summarizes all paths between pairs of points and their information. It is of interest that the paths (3, 5) and (4, 5) contain less information than if the points were incident. Finally equation (2) is used to compute the centralities.

Carrying out these calculations for our example results in the numerical values summarized in Table 2. Also included in this table is the "betweenness" and "closeness" measure of centrality discussed by Freeman (1979a). Both the information and betweenness measure of centrality rank the points in the same order as to be expected for this simple network. The interpretation of the information I_i for each point is relative to the information of unity corresponding to two points which are incident. Also summing the information results in the total information in the network; i.e., $I = \Sigma_i I_i = 7.73$. Hence one can calculate the relative information associated with each point (0.23, 0.23, 0.18, 0.18, 0.18).

This example and the calculations have been discussed in detail in order to illustrate the main ideas of our proposed method for calculating centrality. In the next section we shall show that the values in Table 2 can be calculated directly without going through the details illustrated here.

Table 2
Information for points in the network of Figure 1a [a]

Point (i)	$I_i = \dfrac{5}{\sum_j' / I_{ij}}$	*Closeness*	$C_B(i)$ (*betweenness*)
1	1.77	5	1.5
2	1.77	5	1.5
3	1.41	6	0.5
4	1.41	6	0.5
5	1.37	6	0.0

[a] The formulas for computing closeness and betweenness are those given by Freeman (1979).

3.2. Direct calculation of information

It is not necessary to carry out the calculations for our measure of centrality as illustrated in the preceding section. It can be done by inverting a simple matrix. This section illustrates how these calculations can be done. The proof is in the Appendix.

Consider a network with n points where every pair of points is reachable. Define the $n \times n$ matrix $B = (b_{ij})$ by:

$$b_{ij} = \begin{cases} 0 & \text{if points } i \text{ and } j \text{ are incident} \\ 1 & \text{otherwise;} \end{cases}$$

$$b_{ii} = \begin{cases} 1 + \text{degree of point } (i) \text{ (number of lines} \\ \text{intersecting point } (i) \text{ plus one).} \end{cases}$$

Our measure of centrality (information) is calculated by inverting the matrix B. Define the matrix $C = (c_{ij}) = B^{-1}$. The values of I_{ij} (the information in the combined path P_{ij}) is given explicitly by

$$I_{ij} = \left(c_{ii} + c_{jj} - 2c_{ij}\right)^{-1}.$$

We can write

$$\sum_{j=1}^{n} 1/I_{ij} = \sum_{j=1}^{n} \left(c_{ii} + c_{jj} - 2c_{ij}\right) = nc_{ii} + T - 2R,$$

where

$$T = \sum_{j=1}^{n} c_{jj} \quad \text{and} \quad R = \sum_{j=1}^{n} c_{ij}.$$

(The row sum R is written without a subscript as it will be identical for all rows.) Therefore the centrality for point (i) can be explicitly written as

$$I_i = \frac{n}{nc_{ii} + T - 2R} = \frac{1}{c_{ii} + (T - 2R)/n}. \tag{4}$$

These calculations will be illustrated by the network in Figure 1. The matrix B is easily constructed and is

$$B = \begin{bmatrix} 4 & 0 & 1 & 0 & 0 \\ 0 & 4 & 0 & 1 & 0 \\ 1 & 0 & 3 & 0 & 1 \\ 0 & 1 & 0 & 3 & 1 \\ 0 & 0 & 1 & 1 & 3 \end{bmatrix}.$$

Its inverse is

$$C = B^{-1} = \tfrac{1}{275} \begin{pmatrix} 76 & 1 & -29 & -4 & 11 \\ 1 & 76 & -4 & 29 & 11 \\ -29 & -4 & 116 & 16 & -44 \\ -4 & 29 & 16 & 116 & -44 \\ 11 & 11 & -44 & -44 & 121 \end{pmatrix},$$

and thus $T = 505/275$ and $R = 55/275$. Since $n = 5$, and $(T - 2R)/5 = 0.2873$, we have the following calculations:

Point (i)	($c_{ii} + 0.2873$)	I_i
1	0.564	1.77
2	0.564	1.77
3	0.709	1.41
4	0.709	1.41
5	0.727	1.37

The calculation procedure only involves inverting a matrix. Hence no special software is needed to carry out the numerical procedure. In our illustration, we immediately constructed the matrix B from the network because it was so simple. However, in more complicated networks it might be worthwhile to first calculate the adjacency matrix. If A represents the adjacency matrix then $B = D(r) - A + J$ where $D(r)$ is a diagonal matrix of the degree for each point and J is a matrix having all elements unity.

Another way to calculate the information is to invert a matrix of dimension $(n + 1)$ which is constructed by bordering the matrix

$D(r) - A$ with row and column elements of unity. If $\mathbf{1}' = (1, 1, \ldots, 1)$ is a row vector with n elements, then the bordered matrix is

$$B_0 = \begin{bmatrix} D(r) - A & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix}.$$

The inverse of B_0 is of the form

$$C_0 = B_0^{-1} = \begin{bmatrix} C & \mathbf{1}/n \\ \mathbf{1}'/n & 0 \end{bmatrix},$$

where C is the $n \times n$ matrix needed to calculate the information. In this case the row sums of C are zero and hence the information for point (i) is $I_i = (c_{ii} + T/n)^{-1}$.

3.3. Calculations of centrality with weights

In some networks the incidence relations between points may be differentiated in a quantitative fashion. The lines between points may carry weights. We use the convention that the higher the weight the more important the communication between the incident points on the line. Our method can be extended easily to the general case of having arbitrary weights for the lines joining the points. We shall illustrate the calculations for the same network as in Figure 1a but with weights. The weighted network is shown in Figure 1b.

The centrality calculation for each point will be illustrated using the method which does not require examining all possible paths. For this purpose define

$b_{ii} = \{1 +$ sum of weights for all lines intersecting point i,

$$b_{ij} = \begin{cases} 1 & \text{if points } i \text{ and } j \text{ are not incident} \\ (1 - \text{weight associated with line connecting } i \text{ and } j). \end{cases}$$

The matrix $B = (b_{ij})$ is calculated readily for the network of Figure 1*b*, i.e.

$$B = \begin{pmatrix} 9 & -1 & 1 & 0 & -4 \\ -1 & 9 & 0 & 1 & -4 \\ 1 & 0 & 12 & -9 & 1 \\ 0 & 1 & -9 & 12 & 1 \\ -4 & -4 & 1 & 1 & 11 \end{pmatrix}.$$

Its inverse matrix is:

$$C = B^{-1} = \frac{1}{26125}\begin{pmatrix} 4552 & 1927 & -2048 & -1923 & 2717 \\ 1927 & 4552 & -1923 & -2048 & 2717 \\ -2048 & -1923 & 6477 & 5277 & -2508 \\ -1923 & -2048 & 5227 & 6477 & -2508 \\ 2717 & 2717 & -2508 & -2508 & 4807 \end{pmatrix}.$$

The calculations for I_i using equation (4) result in the centralities:

$$I_1 = 3.33, \quad I_2 = 3.33, \quad I_3 = 2.68, \quad I_4 = 2.68, \quad I_5 = 3.23.$$

Note that with the weights, point 5 is third ranked compared to being ranked last without weights.

Weights may be important for the calculation of the centralities. In practical situations they may require careful assessment. For example, the weights could correspond to the frequency of communication between incident pairs of points. In some situations the choice of such weights may be subjective.

4. Applications

This section illustrates how our method may be used on actual networks. Two examples are discussed. Our first example is a group of 40 homosexual men who have been diagnosed with AIDS. They have been sexual partners with each other and are thereby linked to each other through sexual contact. We have calculated our measures of centrality (information) for this network and compare the distributions of centrality and ranking with those suggested by Freeman (1979b). The example shows how the measure of information is a more sensitive diagnostic of group structure. Our second example is concerned with changes in the social dynamics of a Gelada baboon colony when a new female adult member is accepted by the group. In this process, we employ a weighted measure of centrality to discern structural changes in the network over time.

4.1. AIDS network

The data described by Auerbach *et al.* (1984) and Klovdahl (1985) consists of information on 40 homosexual men diagnosed with AIDS.

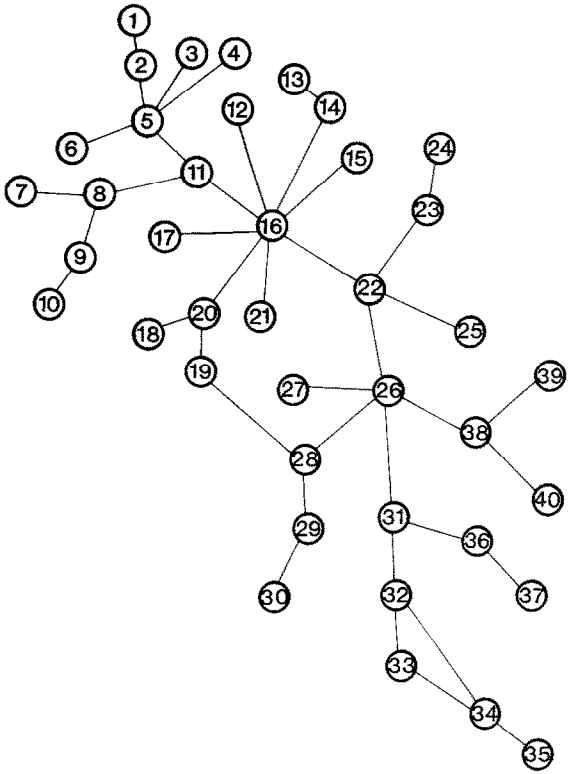

Fig. 2. Network of 40 homosexual men diagnosed with AIDS (from Figure 1 in Klovdahl 1985: 1204).

Initially, 19 men residing in the Los Angeles and Orange County area were interviewed about their previous sexual contacts. This information led to the subsequent identification of an additional 21 sexual partners in San Francisco, New York and other parts of the United States. All 40 homosexual men were linked to each other through sexual contact. Figure 2 depicts the network for these men.

Auerbach *et al.* (1984) concluded that the existence of AIDS cases linked by sexual contact was consistent with an infectious-agent hypothesis. (This was before the HIV virus was shown to be spread by bodily fluid exchange.) Subsequently, Klovdahl demonstrated that personal contact could provide an effective opportunity for transmission of the disease. We will use this network as a way to profile the differences in centrality measures. Table 3 summarizes our measure of

Table 3
Comparison of centrality measures and rankings [a]

Overall rank	Information	Betweenness	Distance	Degree
1	16 (0.417)	16 (17.7)	16 (0.351)	16 (0.205)
2	22 (0.392)	26 (14.0)	22 (0.345)	26 (0.128)
3	26 (0.388)	22 (13.5)	26 (0.322)	5 (0.128)
4	20 (0.357)	11 (11.6)	11 (0.302)	22 (0.103)
5	11 (0.351)	31 (7.6)	20 (0.281)	8 (0.077)
6	28 (0.348)	5 (6.6)	14 (0.265)	11 (0.077)
7	19 (0.336)	8 (4.1)	19 (0.265)	20 (0.077)
8	31 (0.310)	32 (4.0)	31 (0.265)	28 (0.077)
9	14 (0.303)	28 (3.7)	12 (0.262)	31 (0.077)
10	15 (0.299)	20 (3.1)	15 (0.262)	32 (0.077)
11	12 (0.299)	38 (2.8)	17 (0.262)	34 (0.077)
12	21 (0.299)	19 (1.7)	21 (0.262)	38 (0.077)
13	17 (0.299)	2 (1.4)	23 (0.262)	2 (0.051)
14	38 (0.292)	9 (1.4)	28 (0.262)	9 (0.051)
15	23 (0.290)	14 (1.4)	25 (0.258)	14 (0.051)
16	25 (0.285)	23 (1.4)	38 (0.252)	19 (0.051)
17	27 (0.283)	29 (1.4)	5 (0.248)	23 (0.051)
18	5 (0.282)	34 (1.4)	27 (0.245)	29 (0.051)
19	8 (0.274)	36 (1.4)	8 (0.242)	33 (0.051)
20	18 (0.266)	1 (0.0)	18 (0.220)	36 (0.051)
21	29 (0.265)	3 (0.0)	32 (0.218)	1 (0.026)
22	32 (0.248)	4 (0.0)	36 (0.213)	3 (0.026)
23	36 (0.242)	6 (0.0)	13 (0.211)	4 (0.026)
24	13 (0.236)	7 (0.0)	29 (0.211)	6 (0.026)
25	40 (0.228)	10 (0.0)	24 (0.209)	7 (0.026)
26	39 (0.228)	12 (0.0)	2 (0.202)	10 (0.026)
27	24 (0.227)	13 (0.0)	39 (0.202)	12 (0.026)
28	2 (0.225)	15 (0.0)	40 (0.202)	13 (0.026)
29	4 (0.222)	17 (0.0)	3 (0.200)	15 (0.026)
30	3 (0.222)	18 (0.0)	4 (0.200)	17 (0.026)
31	6 (0.222)	21 (0.0)	6 (0.200)	18 (0.026)
32	9 (0.220)	24 (0.0)	9 (0.198)	21 (0.026)
33	7 (0.218)	25 (0.0)	7 (0.196)	24 (0.026)
34	34 (0.217)	27 (0.0)	34 (0.182)	25 (0.026)
35	33 (0.216)	30 (0.0)	33 (0.181)	27 (0.026)
36	30 (0.212)	33 (0.0)	37 (0.176)	30 (0.026)
37	37 (0.197)	35 (0.0)	30 (0.175)	35 (0.026)
38	1 (0.185)	37 (0.0)	1 (0.169)	37 (0.026)
39	10 (0.182)	39 (0.0)	10 (0.166)	39 (0.026)
40	35 (0.180)	40 (0.0)	35 (0.155)	40 (0.026)

[a] The men are assigned numbers 1–40 and are listed in rank order according to their relative centrality. Centralities are given in parenthesis.

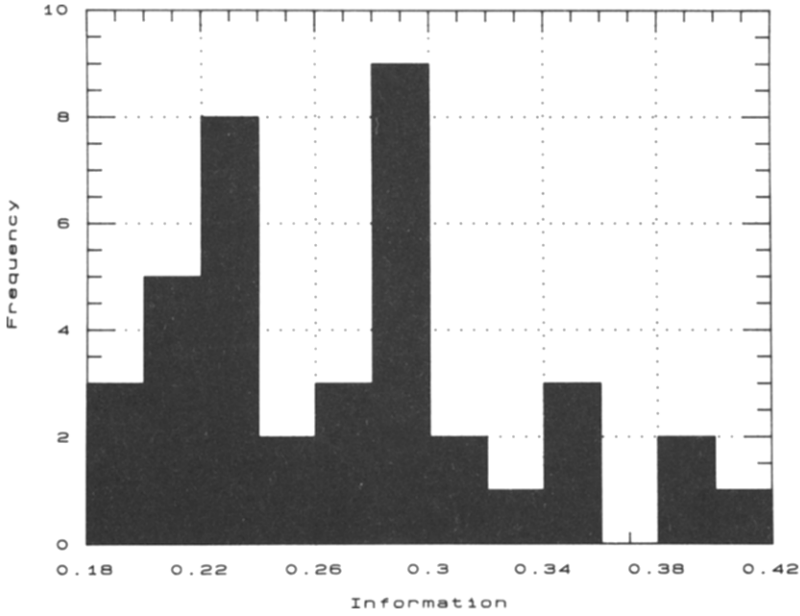

Fig. 3. Histogram of the information measure for 40 homosexual men diagnosed with AIDS.

centrality (information) and compares it with the other proposed centrality measures (degree, closeness and betweenness).

Referring to the information measure of centrality, it is clear that the points which are at the periphery of the network are ranked lowest. (Points having only one path are regarded as peripheral points.) However, even with these points there is an ordering. The highest ranking peripheral points are 15 and 12 (having ranks of 10 and 11); point 35 is the lowest ranked peripheral point. Both points 12 and 15 are incident to the first ranked point 16 and rank above the remaining peripheral points. Among the most central points we can distinguish three levels among the top seven rankings. Point 16 is the most central; a second tier is comprised of 22 and 26; a third tier of points is 20, 11, 28, and 19. The remaining 33 points have information measures which become smaller in a uniform pattern with decreasing rank. Figure 3 is a frequency distribution of our information centrality rankings. A biomodal distribution is indicated with a "cut point" at about 0.25. Without information on date of first contact and diagnosis of AIDS, the biomodality characteristics cannot be readily interpreted. We have information on location, but the biomodality is not correlated with this

factor. However, one interpretation is suggested from the recent evidence that transmission of AIDS by sexual contact may be highly variable (Peterman *et al.* 1988). The central individuals in the network may represent those who can transmit the human immunodeficiency virus (HIV) relatively easily, whereas those on the periphery of the network have a smaller probability of transmission. A less speculative interpretation would require knowledge of frequency of contact and dates. However these data were not available to us.

Comparing our measure of centrality with the others in Table 3 we note that the rankings using the distance measure are closest to our information rankings. One can obtain some insight into the discrepancies between the two centrality measures by investigating the differences in ranks given by these two methods. The top three rankings for both methods are identical. However the remaining four rankings for the information measure are (20, 11, 28, and 19) and for the distance measure are (11, 20, 14, and 19). The principal discrepancy is that the information measure ranks 28 ahead of 14 whereas the distance measure ranks 14 ahead of 28. We shall show that the ranking of 14 ahead of 28 is internally inconsistent by examining the paths for points which are incident to 28 and 14 respectively. From Figure 2 note that points 28 and 14 are of degree 3 and 2 respectively. Furthermore, using the distance ranks, point 28 is incident with 26 (rank 3), 19 (rank 7) and 29 (rank 23.5) whereas point 14 is incident with 16 (rank 1) and 13 (rank 23.5). (We have used average ranks for tied distance scores.) It seems intuitive that, based on the distance measure, point 28 should be ranked ahead of 14. On a more objective basis one can compare the information associated with paths 28 and 14 to the points 13, 16, 19, 26, and 29, which comprise the set of all points incident with 14 and 28. This comparison is shown in Table 4. As one can see from Table 4, point 28 has more information contained in its paths and should therefore rank above point 14. From the point of view of infectivity, AIDS is known to have started with point 16. In this case, if it spreads to 14 through sexual contact, it will go to the peripheral point 13 and stop. However, it is obvious from studying Figure 2 that point 28 is a much more important contact for infiltrating the network with an infectious disease.

Thus using either degree, distance, or information, point 18 should be ranked higher than 14. One can investigate in detail other discrepancies between the ranks given by the information and distance

Table 4

Paths (28)	Path information	Paths (14)	Path information
28–13	0.286	14–13	1.000
28–16	0.667	14–16	1.000
28–19	1.200	14–19	0.429
28–26	1.200	14–26	0.429
28–29	1.000	14–29	0.286
Sum	4.353	Sum	3.144
Harmonic average	0.653		0.492

measures and will find in every case that the information ranking is internally consistent, whereas the distance measure will not be.

The relative rankings in Freeman's betweenness measure indicate three groupings. A highly central group consisting of 4 people, (points 16, 26, 22, and 11), a middle group of 15 individuals and a remaining group of 21 individuals with a zero betweenness measure. When we compare Freeman's betweenness ranking to the relative ranking generated by our information measure, we observe several differences. First, our measure results in point centralities that are more uniformly distributed throughout the entire network. Second, while peripheral individuals to the network may rank low, any contact, i.e., any path connecting a peripheral point can potentially have significant implications throughout the network. This is an important characteristic for studying infectious diseases. Much the same can be said for the notion of communicating in a general sense.

The betweenness measure is not measuring centrality, but ranks points according to the "control" the point exerts in the network. A direct way to evaluate the control of a point is to "remove" it from the network and evaluate the resulting network. If a point exerts control in a network, removing it could generate two sub-networks such that points in one network are not reachable from the other. (This is equivalent to examining "cut points".) One way to measure control is to count the number of points in the smaller subnetwork. The betweenness measures of the AIDS network in Table 3 list 19 points which exert some control. (These are the first 19 ranked points.) Table 5 lists the points in rank order according to the betweenness measure for these 19 points and gives the number of points which are not reachable

Table 5
Summary of points not reachable in subnetwork

Betweenness ranking [a]	Point	Points not reachable	Betweenness ranking	Point	Points not reachable
1	16	17	11	38	2
2	26	11	12	19	0
3	22	3	13	2	1
4	11	10	14	9	1
5	31	6	15	14	1
6	5	5	16	23	1
7	8	3	17	29	1
8	32	3	18	34	1
9	28	2	19	36	1
10	20	1			

[a] Betweenness rankings refer to ranks in Table 3.

if that point is removed from the network. Note that points 22, 20, and 19 are ranked inconsistently.

While the use of degree is relatively simply to calculate, it is limited in the ability to distinguish differential centralities. By definition, it only counts the end of the path without regard to the other point on the path. It is presented here for completeness.

4.2. Baboon colony

In this section we illustrate the use of our proposed centrality measure to study the social dynamics of the Gelada baboons. We will discuss this example in depth to illustrate how centrality can be used to better understand the process of structural change in networks. Dunbar and Dunbar (1975), from which our data is drawn, provide a detailed description of the Gelada baboon's social structure and dynamics based on their extensive research in Africa. They studied the types of social groups that constituted a baboon population and how individual members of these groups related to each other. In terms of social organization and group ontogeny, one-male reproductive groups and all-male groups generally remain stable over time. Within the one-male groups, individual females are known to acquire positions of considerable power and often act as co-leaders with the male. These groups can be highly cohesive as a result of the strong bonds among individual adult members.

Table 6
Sex and developmental stages of baboons [a]

Baboon	Age	Sex	Role
1	Adult	F	Dominant
2	2 years	M	
3	Adult	F	
4	Adult	F	
5	Adult	M	Leader
6	3 years	F	
7	3 years	M	Follower
8	Adult	F	
9	1 year	M	
10	3 years	F	
11	2 years	F	
12	1 year	M	
13	1 year	M	
14	Adult	F	Newcomer

[a] From Dunbar and Dunbar (1975).

The individual baboons were chosen by the Dunbars because they could be readily identified in the colony, and were thought likely to be representative of the different stages of group ontogeny, particularly one-male reproductive groups. Our example comprises observations on the colony at two different times. The first set of observations (H22a) was made on 12 baboons. The second set of observations (H22b) was made on the same colony after a mature female (14) and a yearling male (13) were introduced and accepted by the colony. Table 6 gives the sex and developmental status of all baboons.

In collecting the data, only those interactions involving adults and "subadults" of either sex and 3-year-old females were recorded; interactions involving only juveniles and/or infants were considered generally unstable and not recorded. The resulting observations are depicted in the networks shown in Figure 4. The lines connecting two points (baboons) represent nonagonistic interactions (generally grooming behavior) and the frequency of such interactions is recorded by the number beside the line. H22a (Figure 4a) depicts the structure of the group prior to the introduction of a new member. H22b (Figure 4b) is the group after the acceptance and integration of the new adult female (14) and an accompanying juvenile male (13).

We conducted three analyses with our proposed centrality measure. All centralities were calculated using the frequency of contact (shown

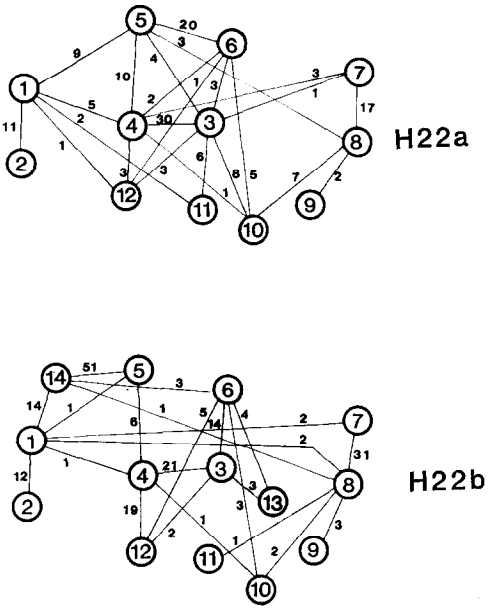

Fig. 4. Network of Baboon colony Study A (H22a) and Study B (H22b) (from Dunbar and Dunbar 1975).

in Figure 4) as weights. The results are tabulated in Table 7. Study A (H22a) lists the ranking and centralities of the individual baboon members. Study B (H22b) lists the rankings and centralities after the introduction of a new adult female and a juvenile male. Study C (H22C) is an investigation where we "removed" the new member baboons (13 and 14) and their linkages to observe what would happen to the centrality rankings of the remaining group.

Our calculations indicate fundamental changes in the structural pattern of baboon relationships over time. Two characteristics of overall group structure are noted. In Study A, the distribution of centralities can be grouped into three distinct classes corresponding to high, middle, and low. The classes are separated by lines. In Table 7. Baboons (3), (4), and (5) are highly central and clustered together. The four baboons on the bottom (2, 9, 11, and 12) comprise another cluster. Note that the "dominant" female (1) is ranked sixth in the group. It would appear that in Study B (H22b), the introduction of the new baboons (most notably 14) nullified centrality distinctions at the top

Table 7
Summary of information centralities: Baboon studies

Study A		Study B		Study C	
Baboon	Information	Baboon	Information	Baboon	Information
4	7.75	14	4.21	4	2.81
3	7.73	5	4.12	8	2.76
5	7.52	1	4.06	3	2.71
		4	4.06	7	2.66
6	6.89	6	3.94	6	2.65
10	6.57	3	3.93	12	2.63
1	6.40	12	3.71	1	2.61
8	6.23	8	3.61	10	2.44
7	5.62	7	3.43	5	2.30
		2	3.15	2	2.21
12	4.53	10	2.92		
11	4.46	13	2.74	9	1.56
2	4.31			11	0.84
9	1.73	9	1.78	–	–
–	–	11	0.88	–	–
Standard deviation	$s = 1.78$		$s = 0.963$		$s = 0.586$
Average	$\bar{I} = 5.81$		$\bar{I} = 3.32$		$\bar{I} = 2.35$
Coefficient of variation	$c = s/\bar{I} = 0.31$		$c = s/\bar{I} = 0.29$		$c = s/\bar{I} = 0.25$

[a] Study A refers to initial observations on 12 baboons; Study B refers to subsequent observations with the addition of two new members (13 and 14); Study C refers to centralities if members 13 and 14 are removed. Baboons are referred to by their identification number.

creating a more uniform ranking overall. It is interesting to note that the new adult female, (14), is the most central. Baboons (3), (4), and (5) still have relatively high rankings in Study B as they did in Study A. After the simulated removal of the new baboons (13 and 14) in Study C, the pattern of centrality rankings still vary in a relatively uniform way and are clustered around the top end. However the rankings do not return to the same ranks as in Study A. This indicates that the introduction of new members and the subsequent formation of new relationships has fundamentally changed the pattern of preexisting relationships in the network. In all three analyses, the bottom group (having the lowest centralities), are clearly discerned.

Another characteristic of the group structure is that the coefficient of variation is remarkably constant for the three analyses. This implies that the larger the mean centrality, the greater the standard deviation and the more likely class distinctions may occur. A large standard deviation indicates a greater separation between centrality values of members of the group.

Individual structural changes

In Study A, adult female baboons (3) (2nd rank) and (4) (1st rank) are the most central baboons in the network or colony. However, the addition of the female (14) to the colony (Study B), resulted in female adult baboons (3) and (4) dropping in overall centrality. On simulating the removal of the new members (13 and 14), baboons (3) and (4) return to their approximately high ranking in the group. It indicates that their social ties were effectively "dampened" under the influence of intense social activities surrounding the inclusion of the new members (13 and 14). This would suggest that throughout the integration of new members (13 and 14), female adult baboons (3 and 4) maintained their ties with the other adult members of the group.

A different sequence of events marks the rise in relative ranking of female adult baboon (8). She ranked moderately central (rank 7) in Study A and sustained this approximate ranking in Study B (rank 8). In Study C, however, she rose in overall centrality (rank 2). Clearly, with the intense grooming activities surrounding inclusion of new member (14), baboon (8) had to increase her activities to just maintain her standing in the group. Effectively adult female (8) strengthened her ties with the other members of the group. Consequently in the simulated removal of (14), her strengthened bonds acted as a buoy and she rose in relative ranking to 2nd place.

Our final finding is reminiscent of a "love triangle". The principals are the harem leader male (5), the dominant female (1) and the female newcomer (14). The dominant female (1) received the attentions (grooming behavior) of the newcomer (14). This action could be interpreted as baboon (14)'s way for seeking acceptance. Since the dominant female was not very central initially (rank 6), her rise in centrality to third place in Study B was largely due to the increased attention she received from baboon (14). She maintained her relatively weak ties with the other members and simply "returned" to 7th place in Study C. It is of interest to note that the newcomer (14), "instinc-

tively" directed her behavior to the dominant female (1) who was not the most central baboon of the group. The Dunbars note that it is difficult to be sure what was meant by "dominant" in the case of mature female baboon (1). This particular adult female was responsible for leading the movement of the group but she did not interact with the male as often as adult female (4). If leadership of the group's movements is important, then it is of some significance that the newcomer female (14) paid most of her attention to female (1). "It is also significant that the relationship was one-sided, with female (1) always being the recipient" (Dunbar and Dunbar 1975: 102).

On the other hand, the harem male (5) ranked high throughout Studies A and B, and then dropped precipitously to a rank of 9 in Study C. This indicates a structural change in the network architecture. Generally, the introduction of baboon (14) created a less stratified colony. However, on closer inspection, we find that the harem male (5) effectively cemented his ties to the newcomer (14) while simultaneously decreasing the frequency of his contacts with the other harem females. Effectively, new individual relations were forged and these emergent patterns could have significant impact on the entire group structure were newcomer (14) to be "removed" by death or accident.

In this regard, Study C completes our picture of these changing relationships. While an "artificial" point in the time, Study C reveals significantly altered relations between the individual baboons indicating fundamental changes in group structure. By failing to sustain his ties with the other harem females, harem leader (5)'s centrality is in large part due to only one other key player, the newcomer (14). The social cost of this behavior could have significant implications for the colony. For instance, the baboon colony may be more open or vulnerable to attack and takeover by a competing adult male. Further investigation into the group dynamics of this situation is beyond the scope of this paper, but the reader is referred to Cunningham (1985) for a technical discussion of this notion. In any event, the simulation allows us to get a view of emerging structural patterns that facilitate our understanding of network processes over time.

5. Conclusions

We have proposed, developed and illustrated by the use of examples a new measure of centrality called information, which may be applied to

nondirected networks. This measure reflects the information contained in all possible paths in a network. We have demonstrated how several widely used measures of centrality may not be internally consistent. As a result these measures may not effectively capture subtle network infrastructures in complex situations. We do not adopt the previous assumptions of "efficiency" represented by the shortest path or geodesic between a pair of points. The utility of the geodesic makes good practical sense in "imposed" networks as in operations research design (e.g. traveling salesman problem). However, our information measure is grounded in a stochastic concept that reflects all paths in a network. All paths are important because they subtly contribute to dynamic network processes.

Albeit a technical paper, we are theoretically motivated in our development. Arguably, networks of human relations or exchanges provide "structural" conveniences upon which theoretical concepts may be suggested. These theoretical notions generate many questions and stimulate research. In that regard, the motivation in developing our technical measure is twofold. We wish to study structural or patterned changes in relations in a group. More fundamental, however, is our concern with the processes that give rise to emerging patterns or structures well before coalitions are socially or analytically recognized. Thus, "events" in group dynamics are fundamentally preempted by ongoing practical and mundane activities in which the social landscape is continuously maintained or reshaped. In any case, what is perceived as social reality is assumed to be indeterminate prior to the actions which reaffirm or challenge it (Moore 1975; Bourdieu 1977; Partridge 1987). Since these actions do not occur in a vacuum but in a social context interacting with others, we hope to capture collective activity as the outcome of instantaneous action over time.

Our view is that centrality is only a descriptive property of a network. An area for future research should be concerned with innovative uses of centrality to describe how networks may change over time or to determine the consequences of new scenarios when nodes or lines are added or deleted. We have attempted to illustrate the calculation of centralities to these prototypical situations. However we regard our efforts in this direction as only a beginning.

Appendix: Theory of centrality

A1. Background

Our measure for estimating centrality has been motivated by the theories of the statistical design of experiments and estimation. A network can be regarded as an incomplete block design with two treatments per block. The anology is made clear by the use of the adjacency matrix in networks and the incidence matrix in the design of experiments. The books by Kempthorne (1982) and John (1971) contain details of the theory associated with the analyses of block designs.

Consider a network with n points and m lines. Define the $n \times m$ matrix $N = (n_{i\alpha})$ $(i = 1, 2, \ldots, n;\ \alpha = 1, 2, \ldots, m)$ by

$$n_{i\alpha} = \begin{cases} 1 & \text{if point } i \text{ is intersected by the } \alpha\text{th line,} \\ 0 & \text{otherwise.} \end{cases}$$

The matrix N contains all of the topographic features of the network and enables the network to be constructed. This matrix is called the incidence matrix in the statistical design of experiments. Note that for networks

$$\sum_{\alpha=1}^{m} n_{i\alpha} = r_i, \qquad \sum_{i=1}^{n} n_{i\alpha} = 2,$$

where r_i is the degree of point i (number of lines which intersect point (i)). The $n \times n$ matrix $\Lambda = NN'$, with elements λ_{ij}, is such that

$$\lambda_{ij} = \begin{cases} r_i & \text{if } i = j, \\ 1 & \text{if points } i \text{ and } j \text{ are incident, } i \neq j, \\ 0 & \text{otherwise, } i \neq j. \end{cases}$$

The adjacency matrix of a network is NN' with the main diagonal elements set equal to zero.

Consider the two points (i) and (j) which are incident on line α. The two possible paths are $(i) \rightarrow (j)$ or $(j) \rightarrow (i)$. We shall model this phenomenon by postulating that there is a signal denoted by $(Y_{i\alpha} - Y_{j\alpha})$

or $(Y_{j\alpha} - Y_{i\alpha})$ generated by the incident points on line α. This signal is stochastic and has the expected value and variance

$$E(Y_{i\alpha} - Y_{j\alpha}) = t_i - t_j$$
$$\text{Var}(Y_{i\alpha} - Y_{j\alpha}) = 1. \tag{5}$$

We define the expected value of a signal as a path. The reciprocal of the variance is our measure of information. Hence the information in a signal between two incident points is unity. We shall identify information with paths. For example, consider a network with three points and two lines such that points (i) and (j) are incident on line α and (j) and (k) are incident on line β. Then the expected value and variance of the signal $(Y_{i\alpha} - Y_{j\alpha}) + (Y_{j\beta} - Y_{k\beta})$ are $t_i - t_k$ and 2. Therefore the measure of information for this signal would be $1/2$. That is, it is equivalent to half of the information for an incident path between (i) and (k).

Our proposed measure of centrality is to consider all the signals in the network $(Y_{i\alpha} - Y_{j\alpha})$ $(i \neq j = 1, 2, \ldots, n;\ \alpha = 1, 2, \ldots, m)$ and estimate the path $(t_i - t_j)$ so that the estimate has minimum variance. The reciprocal of the variance of the estimate is the measure of information for the path $(t_i - t_j)$ (or $(t_j - t_i)$). Thus if V_{ij} is the variance, then the information associated with the path is $I_{ij} = V_{ij}^{-1}$. If all points are reachable there will be $(n-1)$ paths from any point i and hence one can calculate the information of path i to j (or equivalently j to i). A measure of centrality of point i will be a function of the $\{I_{ij}\}$ $j = 1, 2, \ldots, n$ where we define $I_{ij} = \infty$. For computational convenience, we use the harmonic average of the I_{ij} as the centrality of point i, e.g.:

$$I_i = n / \sum_{j=1}^{n} 1/I_{ij}.$$

A2. Reachability

When a network contains n points, there will be $n(n-1)$ possible paths between all pairs of points. (We are counting $(i) \rightarrow (j)$ and $(j) \rightarrow (i)$ as two paths.) In reality, however, there are only $(n-1)$

linearly independent paths. From $(n-1)$ linearly independent paths, one can construct all possible paths. The rank of N determines if all points are reachable. The necessary and sufficient condition for all points to be reachable is that the rank of N is $(n-1)$. When the rank of N is $(n-k)$ for $k>1$, this implies that not all points are reachable as there exist k sub-networks in the network. Points within each sub-network are reachable, but points in different sub-networks are not reachable. When the rank of N is $n-k$, we can arrange the n points in k groups, such that the rth sub-network can be described by the matrix N_r and N can be written as:

$$N = \begin{pmatrix} N_1 & 0 & \dots & 0 \\ 0 & N_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & & N_k \end{pmatrix}.$$

A3. The calculation of centrality

It will be assumed that the rank of N is $(n-1)$ and hence all points are reachable. The incident signals of (i) to the other points can be described by

$$d_{i\alpha} = n_{i\alpha}\left(Y_{i\alpha} - \sum_{j\neq i}^{n} n_{j\alpha}Y_{j\alpha}\right) = n_{i\alpha}\left(2Y_{i\alpha} - \sum_{j=1}^{n} n_{j\alpha}Y_{j\alpha}\right)$$

for $\alpha = 1, 2, \dots, m$.

There will be nm such quantities $(d_{i\alpha})$ for the entire network. In order to estimate the combined paths, we need only deal with a smaller sub-set of data which contains at least $(n-1)$ linearly independent quantities which are functions of the $\{d_{i\alpha}\}$. For this purpose define

$$Q_i = \sum_{\alpha=1}^{m} n_{i\alpha}d_{i\alpha} = \sum_{\alpha=1}^{m} n_{i\alpha}\left(2Y_{i\alpha} - \sum_{j=1}^{n} n_{j\alpha}Y_{j\alpha}\right), \quad i = 1, 2, \dots, n. \tag{6}$$

The quantity Q_i is simply the sum of all signals which involve (i) with the other points. The $\{Q_i\}$ are not linearly independent as $\sum_{j=1}^{n} Q_i = 0$.

However, this is the only dependent relation amongst the $\{Q_i\}$ when N has rank $(n-1)$.

The equations for solving for the combined paths can be obtained by equating Q_i to $E(Q_i)$. This results in

$$Q_i = \sum_{\alpha=1}^{n} n_{i\alpha}\left(2\hat{t}_i - \sum_{j=1}^{n} n_{j\alpha}\hat{t}_j\right) = 2r_i\hat{t}_i - \sum_{j=1}^{n}\lambda_{ij}\hat{t}_j \quad \text{for} \quad i = 1, 2, \ldots, n, \tag{7}$$

where $r_i = \sum_{\alpha=1}^{m} n_{i\alpha}$, $\lambda_{ij} = \sum_{\alpha=1}^{m} n_{i\alpha}n_{j\alpha}$ and t_i has been replaced by $\hat{t}_i$ as a reminder that the solution yields the combined paths. These n equations can be written in matrix notation, e.g.

$$B\hat{t} = Q, \tag{8}$$

where $\hat{t}' = (\hat{t}_2, \hat{t}_2, \ldots, \hat{t}_n)$, $Q' = (Q_1, Q_2, \ldots, Q_n)$ and $B = 2\text{Diag}(r_1, r_2, \ldots, r_n) - NN'$. The $n \times n$ symmetric matrix B has rank $(n-1)$ as the sum of the elements in any row or column in zero. The simultaneous linear equations (8) are exactly the same equations which arise in the statistical design of experiments for obtaining the best linear unbiased estimates of the "treatment" effect vector $\hat{t}$.

An alternate way to arrive at equation (7) is to consider the quantity

$$S = \sum_{i=1}^{n}\sum_{\alpha=1}^{m} n_{i\alpha}(d_{i\alpha} - E(d_{i\alpha}))^2,$$

and minimize S with respect to t_i. For example, set $\partial Q/\partial t_i = 0$ for $i = 1, 2, \ldots, n$. This will result in the same set of equations as (7).

Let $\mathbf{1}$ denote a column vector having n elements each equal to one. The matrix B has rank $(n-1)$ as $\mathbf{1}'B = 0$. Therefore the only linear functions of $\hat{t}$ which can be solved by (8) are those linear functions $l't$ where $\mathbf{1}'l = 0$. In particular, any combined path $(\hat{t}_i - \hat{t}_j)$ satisfies this relationship. One way to solve (8) is to form the matrix $B + J$ where $J = \mathbf{1}\mathbf{1}'$ and solve

$$(B + J)\hat{t} = Q. \tag{9}$$

The solution will be $\hat{t} = CQ$ where $C = A^{-1}$. The variance of any linear function of $\hat{t}$ (say) $l'\hat{t}$ is $V(l'\hat{t}) = l'Cl$ provided $\mathbf{1}'l = 0$.

Let (c_{ij}) be the elements of C. Then the value of $\hat{t}_i$ is given explicitly by $\hat{t}_i = \sum_{j=1}^{n} c_{ij} Q_j$ and any path will be

$$(\hat{t}_i - \hat{t}_i) = \sum_{k=1}^{n} (c_{ik} - c_{jk}) Q_k. \tag{10}$$

The expression given by (10) is the combined path of going from point (i) to (j). It automatically weights the individual paths by the inverse of the variance of the individual paths, or equivalently by the information of the path. Since $V(l'\hat{t}) = l'Cl$, the variance of the combined path $(i) \to (j)$ is $c_{ii} + c_{jj} - 2c_{ij}$ and hence the information associated with this path is

$$I_{ij} = (c_{ii} + c_{jj} - 2c_{ij})^{-1}. \tag{11}$$

The sum of the reciprocals of I_{ij} can be written

$$\sum_{j=1}^{n} 1/I_{ij} = \sum_{j=1}^{n} (c_{ii} + c_{jj} - 2c_{ij}) = nc_{ii} + T - 2R.$$

where $T = \sum_{j=1}^{n} c_{jj}$ and $R = \sum_{j=1}^{n} c_{ij}$. (The sum $R = \sum_{j=1}^{n} c_{ij}$ is the same for every row (or column) and hence is not written with a subscript.)

Therefore the measure of centrality for point I_i can be written

$$I_i = \left[c_{ii} + \frac{T - 2R}{n} \right]^{-1}, \tag{12}$$

where $T = \text{trace } C = \sum_{j=1}^{n} c_{jj}$ and $R = \sum_{j=1}^{n} c_{ij}$.

A4. Centrality defined with a centroid point

Thus the centralities for a network can be easily calculated by inverting an $n \times n$ positive definite matrix and making a few additional simple calculations. The development in Section 3 of this paper was heuristic to show by example how one can obtain the centrality measures from the individual paths.

The definition of centrality defined by equation (2) is the harmonic average of the information associated with the $(n-1)$ paths with regard to a specific point where the path $(t_i - t_i)$ is defined to have

"infinite" information. However an alternate definition of centrality can be derived based on the idea of a path from a specified point to a "centroid" point. Specifically the centroid point can be defined by

$$\bar{t} = \sum_{j=1}^{n} t_j/n$$

and the path of point (i) to the centroid point is $t_i - \bar{t}$. Note that

$$t_i - \bar{t} = t_i - \sum_j t_j/n = \frac{1}{n}\sum_{j=1}^{n}(t_i - t_j),$$

and thus the path of point (i) to the centroid point is (except for a factor of $1/n$), the sum of the paths of point (i) to the other $(n-1)$ points.

Making use of the formula that the variance of $l'\hat{t}$ is $l'Cl$, the $\text{var}(\hat{t}_i - \bar{t}) = c_{ii} - R/n$ where R is the sum of any row of the C matrix. Therefore the centrality as measured by the path of point i to the centroid is

$$I_i' = 1/(c_{ii} - R/n).$$

(We have used a prime ($'$) to identify that this measure of centrality is different than the harmonic average of the paths from i to the other points. When the matrix C is obtained by bordering the matrix B with column and row vectors of unity elements, then $R = 0$ and $I_i' = 1/c_{ii}$, i.e.

$$\begin{bmatrix} B & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix}^{-1} = \begin{bmatrix} C & \mathbf{1}/n \\ \mathbf{1}'/n & 0 \end{bmatrix}.$$

In general the numerical differences between I_i and I_i' are slight. We have chosen to carry out our development of centrality using I_i as it is a more heuristic measure. We have redone our calculations for all of the examples in this paper and all rankings are the same regardless of whether I_i or I_i' were used.

A5. Weighted centralities

The development in the preceding sections can be readily modified when the line α connecting two points carries a weight f_α. Often this

weight is the frequency of communication between the two incident points. To obtain the centralities, consider the sum of the squares of the weighted deviation of the quantities:

$$S = \sum_{i=1}^{n} \sum_{\alpha=1}^{m} n_{i\alpha} f_\alpha \left[d_{i\alpha} - E(d_{i\alpha}) \right]^2 .$$

Minimizing S with respect to t_i results in the simultaneous equations $\partial S / \partial t_i = 0$ for $i = 1, 2, \ldots, n$. These equations are given explicitly by

$$r_i^* \hat{t}_i - \sum_{j \neq i} \lambda_{ij}^* \hat{t}_j = Q_j^* \qquad i = 1, 2, \ldots, n, \tag{13}$$

where

$$\left.\begin{aligned} r_i^* &= \sum_{\alpha=1}^{m} n_{i\alpha} f_\alpha \\ \lambda_{ij}^* &= \sum_{\alpha=1}^{m} n_{i\alpha} n_{j\alpha} f_\alpha \\ Q_i^* &= \sum_{\alpha=1}^{m} n_{i\alpha} f_\alpha d_{i\alpha} \end{aligned}\right\} . \tag{14}$$

Note that they are the same equations as (7). Hence the calculations for the centralities carry over exactly using the * quantities. When $f_\alpha = 1$ for all α, equations (13) reduce to (7).

References

Alba, R.D.
1973 "A Graph-Theoretic Definition of a Sociometric Clique." *Journal of Mathematical Sociology 3*: 113–126.

Allen, M.
1982 "The Identification of Interlock Groups in Large Corporate Networks: Convergent Validation using Divergent Techniques." *Social Networks 4*: 349–366.

Allen, T.
1977 *Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information Within the R&D Organization*. Cambridge, MA: The MIT Press.

Auerbach, D.M., W.W. Darrow, H.W. Jaffee and J.W. Curran
1984 "Cluster of Cases of the Acquired Immune Deficiency Syndrome." *The American Journal of Medicine 76*: 487–492.

Barnes, J.A.
1969 "Network and Political Process." In J.C. Mitchell (ed.), *Social Networks in Urban Situations*. Manchester: Manchester University Press.
1972 *Social Networks* (Addison-Wesley Modular Publishing) *26*: 1–29.
Barnett, G.A. and R.E. Rice
1985 "Longitudinal Non-Euclidean Networks: Applying Galileo." *Social Networks 7*: 263–285.
Bavelas, A.
1948 "A Mathematical Model for Group Structures." *Human Organization 7*: 16–30.
Boissevain, J.
1974 *Friends of Friends: Networks, Manipulators and Coalitions*. New York: St. Martin's Press.
Boissevain, J. and J.C. Mitchell (eds.)
1973 *Network Analysis: Studies in Human Interaction*. Paris: Mouton.
Bonacich, P.
1972 a "A Technique for Analyzing Overlapping Memberships." In H. Costner (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.
1972 b "Factoring and Weighting Approaches to Status Scores and Clique Identification." *Journal of Mathematical Sociology 2*: 113–120.
1987 "Power and Centrality: A Family of Measures." *American Journal of Sociology 92*(5): 1170–1182.
Boorman, S.A. and H.C. White
1976 "Social Structure from Multiple Networks. II. Role Structures." *American Journal of Sociology 81*: 1384–1444.
Bott, E.
1957 *Family and Social Networks*. New York: Free Press.
Bourdieu, P.
1977 *Outline of a Theory of Practice*. New York: Cambridge University Press.
Burt, R.S.
1978 "Cohesion Versus Structural Equivalence as a Basis for Network Subgroups." *Sociological Methods and Research 7*(2): 189–212.
1980 "Models of Network Structure." *Annual Review of Sociology 6*: 79–141.
Cook, K.S., R.M. Emerson, M.R. Gilmore and T. Yamagishi
1983 "The Distribution of Power in Exchange Networks: Theory and Experimental Results." *American Journal of Sociology 89*(2): 275–305.
Cunningham, W.H.
1985 "Optimal Attack and Reinforcement of a Network." *Journal of the Association for Computing Machinery 32*: 549–561.
de Sola Pool, I. and M. Kochen
1979 "Contacts and Influence." *Social Networks 1*: 5–51.
Donninger, C.
1986 "The Distribution of Centrality in Social Networks." *Social Networks 8*: 191–203.
Doreian, P.
1980 "On the Evolution of Group and Network Structure." *Social Networks 2*: 235–252.
Dunbar, R. and P. Dunbar
1975 "Social Dynamics of Gelada Baboons." *Contributions to Primatology 6*: 1–157.
Frombrun, C.
1986 "Structural Dynamics Within and Between Organizations." *Administrative Science Quarterly 31*: 403–421.
Frank, O.
1981 "A Survey of Statistical Methods in Graph Theory." In *Sociological Methodology*. New York: Josey-Bass.

Freeman, L.C.
1979 a "Centrality in Social Networks: Conceptual Clarification." *Social Networks 1*: 215–239.
1980 "The Gatekeeper, Pair-Dependency and Structural Centrality." *Quality and Quantity 14*: 585–592.
Granovetter, M.S.
1973 "The Strength of Weak Ties." *American Journal of Sociology 78*: 1360–1380.
Hage, P.
1979 "Graph Theory as a Structural Model in Cultural Anthropology." *Annual Review of Anthropology 8*: 115–136.
Hage, P. and F. Harary
1983 *Structural Models in Anthropology*. New York: Cambridge University Press.
Harary, F.
1959 "Graph Theoretic Methods in the Management Sciences." *Management Science 5*: 387–403.
1970 "Graph Theory as a Structural Model in the Social Sciences." In *Graph Theory and Its Applications*. New York: Academic Press.
Johannison, B.
1987 "Beyond Process and Structure: Social Exchange Networks." *International Studies of Management and Organization 17*(1): 3–21.
John, P.W.M.
1971 *Statistical Design and Analysis of Experiments*. New York: The Macmillan Company.
Kapferer, B.
1969 "Norms and Manipulation of Relationships in a Work Context in Barnes." In J.C. Mitchell (ed.), *Social Networks in Urban Situations*: 181–244. Manchester: Manchester University Press.
1973 "Social Network and Conjugal Role in Urban Zambia: Towards a Reformulation of the Bott Hypothesis." In J. Boissevain and J.C. Mitchell (eds.) *Network Analysis: Studies in Human Interaction*: 83–110. Paris: Mouton.
Kempthorne, O.
1952 *The Design and Analysis of Experiments*. New York: Wiley.
Klovdahl, A.
1985 "Social Networks and the Spread of Infectious Diseases: The AIDS Example." *Social Science of Medicine 21*: 1203–1216.
Mariolis, P.
1982 "'Region' and 'Subgroup': Organizing Concepts in Social Network Analysis." *Social Networks 4*: 305–328.
Mitchell, C.
1969 *Social Networks in Urban Situations*. Manchester: Manchester University Press, 1969.
1974 "Social Networks." *Annual Review of Anthropology 3*: 279–299.
Mizruchi, M.S.
1984 "Interlock Groups, Cliques, or Interest Groups? Comment on Allen." *Social Networks 6*: 193–199.
Mizruchi, M.S. and D. Bunting
1981 "Influence in Corporate Networks: An Examination of Four Measures." *Administrative Science Quarterly 26*: 475–489.
Moore, S.F.
1975 "Epilogue: Uncertainties in Situations, Indeterminates in Culture." In S.F. Moore and B.G. Myerhoff (eds.), *Symbol and Politics in Communal Ideology*. New York: Cornell University Press.
Nieminen, J.
1974 "On Centrality in a Graph." *Scandinavian Journal of Psychology 15*: 322–336.

Partridge, W.L.
1987 "Toward a Theory of Practice." In E.M. Eddy and W.L. Partridge (eds.), *Applied Anthropology in America*. New York: Columbia University Press.
Peterman, T.A., R.L. Stoneburner, J.R. Allen, H.W. Jaffee and J.W. Curran
1988 "Risk of Human Immunodeficiency Virus Transmission from Heterosexual Adults with Transfusion-Associated Infections". *Journal of the American Medical Association 259*: 55–58.
Sabidussi, G.
1966 "The Centrality Index of a Graph." *Psychometrika 31*: 581–603.
Salzinger, L.L.
1982 "The Ties That Bind: The Effect of Clustering on Dyadic Relationships." *Social Networks 4*: 117–145.
Tichy, N. and C. Fombrun
1979 "Network Analysis in Organizational Settings." *Human Relations 32*(11): 923–965.
Tutzauer, F.
1985 "Toward a Theory of Disintegration in Communication Networks." *Social Networks 7*: 263–285.
White, H.C., S.A. Boorman and R.L. Breiger
1976 "Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions." *American Journal of Sociology 81*: 730–780.