May 29, 2001

Extending Centrality

Martin Everett
School of Computing and Mathematical Sciences
University of Greenwich
Wellington Street
London SE 18 6PF

Stephen P. Borgatti
Dept. of Organization Studies
Carroll Graduate School of Management
Boston College
Chestnut Hill, MA 02467

Extending Centrality

Abstract

In this chapter, we discuss three ways to extend the basic concept of centrality. The first extends centrality to apply to groups in addition to individual actors. This extension makes it possible to evaluate the relative centrality of different teams or departments within an organization, or to assess whether a particular ethnic minority in a society is more integrated than another. The second extends the concept of centrality to apply to 2-mode data in which the data consist of a correspondence between two kinds of nodes, such as individuals and the events they participate in. In the past, researchers have dealt with such data by converting them to standard network data (with considerable loss of information). The extension to 2-mode data means that we can apply the tools and concepts of centrality directly to original 2-mode dataset. The third broadens the centrality concept into a model of the core/periphery structure of a network. With this technique we can evaluate the extent to which a network revolves around a core group of nodes.

Extending Centrality

INTRODUCTION

Centrality is one of the most important and widely used conceptual tools for analysing social networks. Nearly all empirical studies try to identify the most important actors within the network. In this chapter, we discuss three extensions of the basic concept of centrality. The first extension generalizes the concept from that of a property of a single actor to that of a group of actors within the network. This extension makes it possible to evaluate the relative centrality of different teams or departments within an organization, or to assess whether a particular ethnic minority in a society is more integrated than another. The second extension applies the concept of centrality to two-mode data in which the data consist of a correspondence between two kinds of nodes, such as individuals and the events they participate in. In the past, researchers have dealt with such data by converting them to standard network data (with considerable loss of information); the objective of the extension discussed here is to apply the concept of centrality directly to the two-mode data. The third extension uses the centrality concept to examine the core/periphery structure of a network.

It is well known that a wide variety of specific measures have been proposed in the literature dating back at least to the 1950s with the work of Katz (1953). Freeman (1979) imposed order on some of this work in a seminal paper which categorized centrality measures into three basic categories (degree, closeness and betweenness) and presented canonical measures for each category. As a result, these three measures have come to dominate empirical usage, along with the eigenvector-based measure proposed by Bonacich (1972). While many other measures of centrality have been proposed since, these four continue to dominate empirical usage, and so this chapter concentrates on just these. In addition, for the sake of clarity and simplicity, we shall discuss only connected undirected binary networks. However it should be noted that much of the work can be extended without difficulty to directed graphs, valued graphs and graphs with more than one component.

GROUP CENTRALITY

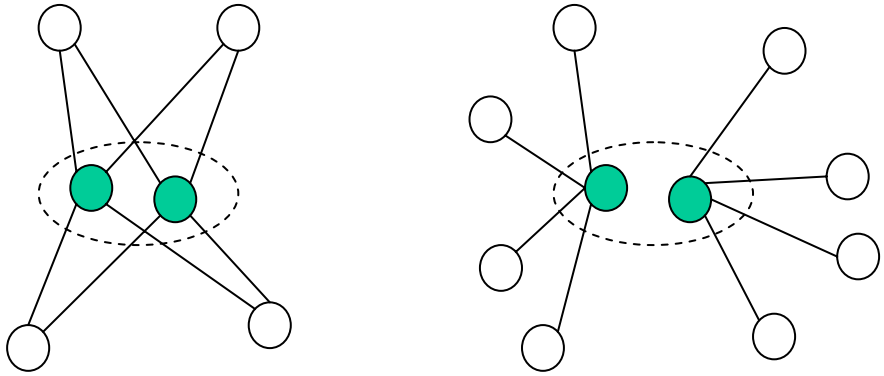
Traditionally centrality measures have been applied to individual actors. However there are many situations when it would be advantageous to have some measure of the centrality of a set of actors. These sets may be defined by attributes of the actors, such as ethnicity, age, club membership, or occupation. Alternatively the sets could be emergent groups identified by a network method such as cliques or structural equivalence. Thus, we can examine informal groups within an organization and ask which ones are most central, and use that in an attempt to account for their relative influence.

In addition, the notion of group centrality can be used solve the inverse problem: how to construct groups that have maximal centrality. A manager may wish to assemble a team with a specific set of skills; if the team were charged with some innovative project it would be an additional benefit if they could draw on the wider expertise available within the organisation. The more central the group, the better positioned they would be to do this.

The notion of group centrality also opens up the possibility of examining the membership of a group in terms of contribution to the group's centrality. If an individual's ties are redundant with those of others, they can be removed from the group without reducing the group's centrality, creating more efficient groups in this respect.

Everett and Borgatti (1999) proposed a general framework for generalizing in this way the three centrality measures discussed in Freeman's paper. They note that for any group centrality measure to be a true generalization of an individual measure, when applied to a group consisting of a single actor it should obtain the same result as the standard individual measure. This immediately implies that a group centrality measure is a measure of the centrality of the whole group with respect to the individuals in the rest of the network, rather than to other groups.

One simple approach that satisfies this condition would be to sum or average the centrality scores in the group. Summing is clearly problematic: larger groups will tend to have higher scores, and when trying to construct a group of maximum centrality we must then restrict the size or the method would always group the entire network together. Averaging solves this problem, however, it takes no account of redundancy or, to put it differently, the fact that actors within the group may be central with respect to or due to the same or different actors. For example consider two groups each of just two actors, as shown in Figure 1. In each group, both actors have degree four. In one group the pair are structurally equivalent (i.e., adjacent to exactly the same four actors), while in the second group the pair are adjacent to four different actors. Simple aggregation methods would result in both these groups having the same centrality score. Clearly the second group, with its larger span of contacts, should have a better score. Thus, the problem is more complicated than simply choosing the *k* individuals with greatest individual centrality, since much of their centrality could be due to ties with the same third parties, or with each other.



Group 1 Group 2

Figure 1

Degree

We define *group degree centrality* as the number of actors outside the group that are connected to members of the group. Since it is a count of the number of actors as opposed to the number of edges then multiple ties to the same actors by different group members are only counted once. This measure needs to be normalized so that we can compare different groups on the same set of actors. Clearly the maximum possible is when every actor outside the group is connected to an actor in the group (in graph theory such a set is said to be dominating). We can therefore normalize by dividing the degree of the group by the number of actors outside the group. As an example we shall look at the EIES data collected by Freeman and Freeman (1979). These data arose from an early experiment on computer-mediated communication. Fifty academics interested in interdisciplinary research were allowed to contact each other via an Electronic Information Exchange System (EIES). The data collected consisted of all messages sent plus acquaintance relationships at two time periods (collected via a questionnaire).The data includes the 32 actors who completed the study. In addition attribute data on primary discipline and number of citations was recorded. The data are available in UCINET 5 (Borgatti, Everett and Freeman, 1999). We shall look at the acquaintance relationship at the start of the study. Two actors are adjacent if they both reported that they have met. The actors are divided into four primary disciplines namely, sociology, anthropology, psychology and statistics. We use these disciplines to form the groups. The results are given in Table 1.

Table 1. Group Degree Centrality for the EIES data

Discipline	Number of Actors	Group Degree	Normalized Group Degree
Anthropology	6	21	81%
Psychology	6	25	96%
Sociology	17	15	100%
Statistics	3	23	80%

Note that although sociology has the lowest (un-normalized) group degree centrality it is a dominating set and so has a normalized group degree centrality of 1.0. Normalization is of greater significance in group centrality than in individual centrality. In individual centrality, the primary purpose of normalization is to enable comparison of centrality scores for individuals in different networks. Within the same network, normalizing centrality makes little difference since normalization is (except in the case of closeness) a

linear transformation affecting all the nodes equally. But in group centrality, different groups in the same network will have different sizes, and so normalization is necessary to compare scores.

Note also that smaller groups need more connections to obtain the same normalized score as larger groups. We can see that the extra connections the statisticians have over the anthropologists does not quite compensate for their smaller size. For small groups to be central they need to work harder than large groups, this has to be taken into consideration when analyzing real data. The converse of this is that it is easier for large groups to have higher centrality scores. There are two reasons for this. First, large groups contain more actors so that each actor requires fewer contacts outside the group in order for the group as a whole to reach more of the outsiders. Second, the more actors there are in the group the fewer there are outside and so the whole group needs to connect to fewer actors to be a dominating set. This effect is particularly strong in small networks.

It is interesting to note that an analysis of individual centrality in the EIES dataset shows that one particular sociologist has direct contact with all non-sociologists. In a sense, then, the connections of all the other sixteen sociologists are redundant in terms of contributing to the degree group centrality. Similarly, two of the anthropologists, two of the psychologists and one of the statisticians do not directly contribute to the group centrality measures of their respective groups. The presence of actors who do not contribute to the group centrality score can be measured in terms of the efficiency of the group. Efficient groups do not have redundancy in terms of supporting actors who do not contribute. We now give a general formulation of this concept.

Let *gpc* be any un-normalized group centrality score, such as group degree centrality. The contribution of a subset K of a group C to *gpc*(C) in a network G is the group centrality score of K with respect to the nodes in G-C. With a slight abuse of notation we shall denote this by *gpc*(K). A group centrality score is *monotone* if for every group C and subset K $gpc(K) \leq gpc(C)$. In essence monotone group centrality means that adding actors to the group increases the group centrality score. (Provided, that is, that we are using measures in which larger values indicate more centrality; if the reverse were true the inequality would need to be reversed). A subset K of C in which $gpc(K)=gpc(C)$ is said to be making a *full contribution*. Let *k* be the size of the smallest subset of C that makes a full contribution. The *efficiency e* of a group C with respect to a monotone group centrality measure can be defined as:

$$e = \frac{k}{|C|}$$

We can see from the EIES data and the observations above that the sociologists have an efficiency of 1/17 (0.06) whereas the efficiencies for the three other groups are 2/3 (0.67). The efficiency is a normalized measure of the maximum number of actors that can be deleted before affecting the group centrality score. A low efficiency means that quite a few actors can be deleted without changing the group centrality value (if they are chosen with care).

Closeness

We can extend the measure of closeness to the group context in a similar way. We define group closeness as the sum of distances from the group to all node outside the group. As is well known in the hierarchical clustering literature (Johnson, 1967), there are many ways to measure the distance from a group to a node outside the group. Let D be the set of all distances (defined in the graph theoretic sense as the length of the shortest path) from a node *x* to a set of node C. Then we can define the distance from *x* to C as the maximum of D, the minimum of D, the mean of D, the median of D or any of a number of other variants. Each of these gives rise to a different group centrality measure, and each is a proper generalization of individual closeness centrality since if the group were a single actor all of these would be identical to each other and to ordinary individual closeness. We can then normalize the group closeness by dividing the summed distance score into the number of non-group members. (This value represents the theoretical minimum for all of the measures mentioned here, if a more esoteric distance is used then this should be replaced by the corresponding optimum value).

The question arises as to which of these should be used in a particular application. This, of course, is dependent on the nature of the data. It is worth noting that the minimum and maximum methods share the property that the distance to a group is defined as the distance to an individual actor within the group. If the data are such that the group can be thought of as an individual unit then the minimum method would be the most appropriate. As an example, consider the group of police informers embedded in a criminal network. Assume that as soon as any one informer knows a bit of information, the information is passed on instantaneously to the police. In this case, it is reasonable to use the minimum distance formulation of group closeness, since the effectiveness of the group is a function of the shortest distance than any informer is from the origin of any bit of information.

Now let us consider the maximum method. Using the maximum method means that everyone within the group are a distance equal to or less than the distance to a given actor. Consider a communication network within an organization, and suppose that everyone who manages a budget needs to know about a regulatory change. If any one department head is unaware of the change, his or her department is not in compliance and may make the organization as a whole liable for penalties. In this case, the maximum method would be more appropriate, as the performance of a group is a function of the time that the last person hears the news. Alternatively it may happen that rumors travel through a network by each actor passing on the rumor to a randomly selected neighbor. The expected time-until-arrival of the rumor to the group will be a function of the all the distances from the group to all other actors. In this case the average method makes sense. The different methods also have some mathematical properties that in different situations may make one more attractive than the others. For example, the minimum method is not very sensitive and it is relatively easy for groups to obtain the maximum value. However, of the closeness methods discussed here, it is the only one that is monotone and so is the only one that can be used to define efficiency.

Betweenness

The extension to betweenness is in the same vein as the extensions discussed above. Group betweenness measures the proportion of geodesics connecting pairs of non-group members that pass through the group. Let C be a subset of node of a graph with node set of V, let $g_{u,v}$ be the number of geodesics connecting u to v, and let $g_{u,v}(C)$ be the number of these geodesics which pass through C. Then the group betweenness of C is given by

$$\sum_{u<v} \frac{g_{u,v}(C)}{g_{u,v}} \qquad u, v \notin C$$

This value can then be normalized by dividing by ½ (|V| - |C|) (|V| - |C|-1), which is the maximum possible.

Social Capital

The notion of group centrality provides a measure of the social capital of an embedded group. Most discussions of social capital distinguish between individual capital and group capital. Individual social capital is easily thought of in terms of centrality. Group social capital is typically thought of in terms of the pattern of ties within the group (e.g., cohesion). This is perhaps because theorists concerned with group social capital typically regard the group as the social universe. However, in organizational theory, the groups we are interested in (e.g., teams, task forces, departments, divisions, whole organizations) are typically embedded in a larger social network (e.g., the organization as a whole, the industry, the economy). This means that the social capital of the group could refer as much to the ties of the group to the network it is embedded in as it does to the ties within the group. The new measures of group centrality provide an effective way to measure this external form of group social capital.

TWO MODE CENTRALITY

We now shift our attention to the application of centrality to a different kind of data, namely 2-mode data. In 2-mode data, there are two kinds of entities, which we shall call actors and events, and a binary relation, such as membership or participation, that connects the actors to the events. The data may be represented by a 2-way, 2-mode affiliation matrix, in which the rows represent actors and the columns represent events, and a 1 in row *i* column *j* indicates that actor *i* attended event *j*. Two-mode data can also be represented as a bipartite graph – a graph in which the nodes can be divided into two classes and the only ties in the network are between nodes of different classes. This type of data is of interest to network analysts when it can reasonably be supposed that two actors participating in the same event indicates the existence or potential for some form of social bond between them.

Bonacich (1991) looked at 2-mode centrality but his methods were not direct extensions of the traditional measures. Since the bipartite graph is simply a graph, we can apply the traditional centrality measures directly to this graph. This approach has been taken by a number of authors particularly with respect to degree centrality and Faust (1997) discusses this conceptualization and suggests some alternatives using Galois lattices. Here we concentrate on the work of Borgatti and Everett (1997) and their approach to normalizing these measures and developing indices of graph centralisation. We shall assume that the bipartite graph representation is of the form G(A+E,R) where A and E are the sets of actors and events, and R is the set of ties connecting them. Let *n* be the size of the node set A and *m* be the size of node set E.

Degree

In the 2-mode context, the degree centrality for an actor is simply the number of events they attend and for an event it is the number of actors attending that event. Clearly the maximum degree for an actor is the total number of events and the maximum degree for an event is the total number of actors. We can use this information to normalize the degree centrality scores. Davis et al. (1941) collected data on a series of social events attended by society women. The data consisted of 18 women and 14 events so that n=18 and m=14.

Table 2. 2-mode degree centrality for the Davis data

ID	Name	Degree	Normalized degree	2-Mode Normalized degree
1	EVELYN	8	25.81	57.14
2	LAURA	7	22.58	50.00
3	THERESA	8	25.81	57.14
4	BRENDA	7	22.58	50.00
5	CHARLOTTE	4	12.90	28.57
6	FRANCES	4	12.90	28.57
7	ELEANOR	4	12.90	28.57
8	PEARL	3	9.68	21.43
9	RUTH	4	12.90	28.57
10	VERNE	4	12.90	28.57
11	MYRNA	4	12.90	28.57
12	KATHERINE	6	19.36	42.86
13	SYLVIA	7	22.58	50.00
14	NORA	8	25.81	57.14
15	HELEN	5	16.13	35.71
16	DOROTHY	2	6.45	14.29
17	OLIVIA	2	6.45	14.29
18	FLORA	2	6.45	14.29
19	E1	3	9.68	16.67
20	E2	3	9.68	16.67
21	E3	6	19.36	33.33
22	E4	4	12.90	22.22

23	E5	8	25.81	44.44
24	E6	8	25.81	44.44
25	E7	10	32.26	55.56
26	E8	14	45.16	77.78
27	E9	12	38.71	66.67
28	E10	5	16.13	27.78
29	E11	4	12.90	22.22
30	E12	6	19.36	33.33
31	E13	3	9.68	16.67
32	E14	3	9.68	16.67

The first column is the raw degrees of the nodes, the second column gives the standard normalization proposed by Freeman for ordinary single mode data. The third column is the two mode normalization. This is calculated by taking the women's degree and dividing by the number of events (14 in this case) and the event's degree and dividing by the number of women (18 in this case) and expressing the answers as a percentage. The two-mode normalization takes account of the special nature of the data and allows the centrality scores to take on the full range of values from zero to one hundred. It should be noted that the 2-mode normalization is non-linear, for example actor Theresa and event E5 both have degree 8 but Theresa has a higher normalized score reflecting the fact that there are fewer events than women.

Closeness and Betweenness

We can take exactly the same approach for closeness and betweenness as we have taken for degree. That is apply the original measures as before but change the way they are normalized to reflect the fact that there are restrictions on which node can be connected. For ordinary closeness we take the raw score and divide this value into the size of the network minus one. In the bipartite case we have a theoretical minimum value of m+2n-2 for the actors and n+2m-2 for the events. We therefore take an event node, calculate its raw closeness centrality score, and divide this value into n+2m-2. For an actor node we do the same thing, but divide the raw score into m+2n-1. Clearly as in the degree case this gives normalization procedure is non-linear.

Betweenness is treated in the same way but the formulas are more complicated. We normalize the events by dividing by $\frac{1}{2}[n^2(p+1)^2+n(p+1)(2r-p-1)-r(2p-r+3)]$ where p is the integer portion of the result of dividing (m-1) by n, and r is the remainder. We normalize the actors by dividing by $\frac{1}{2}[m^2(s+1)^2+m(s+1)(2t-s-1)-t(2s-t+3)]$ where s is the integer portion of the result of dividing (n-1) by m, and t is the remainder.

Centralization

Freeman in his original paper proposed a general measure of centralization to try and capture the extent to which a network consisted of a highly central actor surrounded by peripheral actors. This measure is simply the sum of the differences in centrality of the

most central actor to all the others, normalized by the maximum possible over all connected graphs. This can be expressed as

$$\frac{\sum[c_* - c_i]}{\max \sum[c_* - c_i]}$$

where c_i is the centrality of node *i* and c_* is the centrality of the most central node.

We can apply this formula directly to our 2-mode centrality measures. Note that we should only apply this to the normalized centrality measures since the formula takes the difference between the centrality of one node and that of all other nodes, so we need the scores to be comparable across modes. We also need to determine the denominator in this formula as this is now the maximum over all connected bipartite graphs and not over all connected graphs. In the 1-mode case the graph that achieved the maximum was the star graph. For 2-mode data it is a little more complicated but in general the graphs on which the maximum centralities were achieved to obtain the normalization can be used to construct this denominator. The following formulas give expressions for the maximum and assume the centralities are on the scale of 0 to 1. If percentages are used then the formulas need to be multiplied by 100. The node which achieves the highest centrality score could be either an actor or an event. We denote by n_o the size of the node set which contains the actor with the highest centrality score (this value could be either *n* or *m*) and n_i is the size of the other mode.

Degree

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

Closeness

$$((p+1)n_i + r) - [(1+2p)n_i + 2r]\left(\frac{p(n_i - r)}{2p(2n_i - 1) + 4r + 3n_i - 2} + \frac{r(p+1)}{2p(2n_i - 1) + 4(r-1) + 3n_i}\right)$$

$$- \quad [n_i(p+2) + r - 1]\left(\frac{n_i - r}{n_i(3p+2) + 3r - 2p - 1} + \frac{r}{n_i(3p+2) + 3(r-1) - 2p}\right)$$

$$p = (n_o - 1)\ div\ n_i\ ,\ r = (n_o - 1) \bmod n_i$$

Betweenness

$$(n_o + n_i - 1) \ - \ \frac{p(n_i - r)(2n_o + 2n_i - p - 3) + r(p+1)(2n_o + 2n_i - p - 4)}{n_o^2(s+1)^2 + n_o(s+1)(2t - s - 1) - t(2s - t + 3)}$$

$$p = (n_o - 1)\ div\ n_i,\ r = (n_o - 1) \bmod n_i,\ s = (n_i - 1)\ div\ n_o,\ t = (n_i - 1) \bmod n_o$$

In the equations for closeness and betweenness, parameters *p*, *r*, *s* and *t* are used to simplify the formulas. The parameter *p* is the integer result of dividing n_o*-1* by n_i, and parameter *r* is the remainder. The parameters *s* and *t* are defined analogously.

As an example, consider the betweenness formula for the Davis data. The highest normalized score (24.38%) is achieved by event E8. Summing the difference between 0.2438 and the centrality of every other node gives us the numerator of the centralization formula, and equals 6.3686. The denominator, as given by the equation above is 30.1236, yielding a graph centralization score of 21.14%.

It is interesting to note that it is possible for an event and an actor to have the same centrality score and for this to be the highest score. In this case there are two possible centralizations, one of an actor and one for an event, and these could be quite different for the closeness and betweenness centralizations (they would agree for the degree case). This fact suggests a fundamental problem with this approach, namely that the centralization measures the extent to which actors and events are peripheral to the most central actor or event. It could happen that all the events have similar centrality scores but there is a high degree of centralization amongst the actors taken on their own. Borgatti and Everett propose an extension called single mode centralization. For each mode the difference between the most central node and the centralities of all other node in that mode is calculated. This is exactly the same formula as for all the centralizations except we now restrict the calculation to each mode. Again we need to calculate the formula for the denominator, and these are given below. Note that since we restrict ourselves to a single-mode it is not necessary to use the normalized centrality scores for degree and betweenness but it is necessary for closeness since the normalization is always non-linear. Since the formula for the un-normalized cases are much simpler we present those here. We use the same notation as for the complete centralization case.

Degree (un-normalized)

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

Closeness (normalized)

$$n_o - 1 - [(1 + 2p)n_i + 2r]\left[\frac{p(n_i - r)}{2p(2n_i - 1) + 4r + 3n_i - 2} + \frac{r(p + 1)}{2p(2n_i - 1) + 4(r - 1) + 3n_i}\right]$$

Betweenness (un-normalized)

$\frac{1}{2}(n_o - 1)\left[(p+1)^2 + n_i(p+1)(2r-p-1) - r(2p-r+3)\right]$

We now apply the formula for single mode degree centrality. There are 89 edges in the dataset so that the sums of the degrees of the women and the events will both be 89. The woman with the highest degree has degree 8 and, since there are 18 women, the numerator will be 8x18-89=55. The denominator is given by the formula above and is therefore 17x13=221. This gives a single mode centralization of 25%, a similar calculation for the events results in a figure of 47%. We can see that the women in this case are far less centralized than the events.

The reason for preferring single mode centralization over the traditional method of converting to 1-mode data is that this technique does not destroy information on patterns of overlap. In addition, we are also able to apply methods which are only valid on binary data to the network. In converting to 1-mode it is necessary to dichotomize the data before applying some of the centrality calculations, and this induces further information loss.

Core/Periphery Measures

The notion of core/periphery structures draws on elements from both of the previous sections of this paper. From the discussion of group centrality we draw the basic notion of extending centrality to apply to a group. From the discussion of 2-mode data, we draw on the notion of graph centralization, which we will extend to the group case. The synthesis of these concepts is the notion of a core/periphery structure, which is simultaneously a model of a graph and a generalized measure of centrality. A graph has a core/periphery structure to the extent that it lacks subgroups. Another way of putting it is that all nodes can be regarded as belonging (to a greater or less extent) to a single group, either as core members or peripheral members. The extent to which a node belongs to the core can be thought of as the coreness of the node, and is an individual measure similar to centrality.

Our starting model will be a simple partition of nodes into core and periphery classes, in which the core is a complete subgraph and the periphery is a collection of actors that do not interact with each other. This leaves a number of options for the relations between core and periphery nodes and each of these can give rise to different models. One option is to assume that everyone in the periphery is connected to every member of the core. Table 3 gives an adjacency matrix of this structure. The matrix has been blocked to emphasize the pattern.

Table 3 Idealized core/periphery structure

	1	2	3	4	5	6	7	8	9	10
1		1	1	1	1	1	1	1	1	1
2	1		1	1	1	1	1	1	1	1
3	1	1		1	1	1	1	1	1	1

4	1	1	1		1	1	1	1	1	1
5	1	1	1	1		0	0	0	0	0
6	1	1	1	1	0		0	0	0	0
7	1	1	1	1	0	0		0	0	0
8	1	1	1	1	0	0	0		0	0
9	1	1	1	1	0	0	0	0		0
10	1	1	1	1	0	0	0	0	0	

The pattern can be seen as a generalization of Freeman's (1979) maximally centralized graph, the simple star (see Figure 2). In the star, a single node (the center) is connected to all other nodes, which are not connected to each other. To move to the core/periphery image, we simply add duplicates of the center to the graph, and connect them to each other (see Figure 3). The core of a core/periphery structure can also be seen as a group with maximum group centrality; in this case, the core is in fact a dominating set.

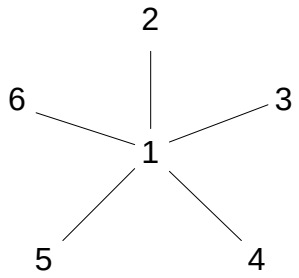


Figure 2. Freeman's Star.

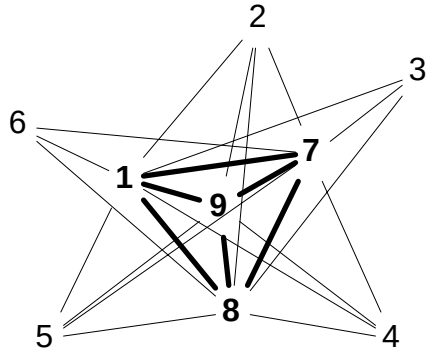


Figure 3. Core-Periphery structure.

The patterns in Table 3 and Figures 2 and 3 are idealized patterns that are unlikely to be actually observed in empirical data. We can readily appreciate that real structures will only approximate this pattern, in that they will have 1-blocks with less than perfect density, and 0-blocks that contain a few ties. A simple measure of how well the real structure approximates the ideal is given by Equation 1 together with Equation 2.

$$\rho = \sum_{i,j} a_{ij}\delta_{ij} \qquad \text{Equation 1}$$

$$\delta_{ij} = \begin{cases} 1 \ \textit{if} \ c_i = CORE \ \ \textit{or} \ \ c_j = CORE \\ \\ 0 \ \textit{otherwise} \end{cases} \qquad \text{Equation 2}$$

In the equations, a_{ij} indicates the presence or absence of a tie in the observed data, c_i refers to the class (core or periphery) that actor i is assigned to, and δ_{ij} (subsequently

called the *pattern matrix*) indicates the presence or absence of a tie in the ideal image. For a fixed distribution of values, the measure achieves its maximum value when and only when A (the matrix of a_{ij}) and Δ (the matrix of δ_{ij}) are identical, which occurs when A has a perfect core/periphery structure. Thus, a structure is a core/periphery structure to the extent that ρ is large. This formulation can be used as the basis for a procedure for detecting core/periphery structures in data. The procedure, which has been implemented in UCINET using a genetic algorithm, begins with a random partition of nodes into two classes (core and periphery), then iteratively reassigns the nodes to maximize a variant of Equation 1.

We can think of the c_i as a discrete coreness measure and assign a value of 1 to the core actors and a value of zero to the peripheral actors. In this case Equation 2 can be written as $\delta_{ij}=c_ic_j$. This can be extended further by allowing the c's to take on values from a continuous range between 0 and 1. Thus, the pattern matrix Δ has large values for pairs of node which are both high in coreness, medium sized values for pairs of node in which one is high in coreness and the other is not, and low values for pairs of node that are both peripheral.

Now that we have a continuous model the simple matching count of Equation 1 would not be appropriate. (Nor, in fact, do we need to restrict the data matrix to contain only binary values). Two possible solutions have been extensively used although many others are possible. The first is to simply correlate matrix A with matrix Δ. We can then optimize the correlation of A and Δ over all values of the vector **c** where the elements of **c** are constrained between zero and one. This is a continuous optimization problem and has been implemented in UCINET using the well-known Nelder-Mead simplex optimization procedure. An alternative is to use the sum of squared differences and optimize this. It is well known that, if the diagonal is not ignored, this equates to finding the principal eigenvector of A and is therefore is simply an eigenvector centrality measure. This gives us a new insight into eigenvector centrality and helps us understand why the smaller separate components have an eigenvector centrality of zero; they simply cannot be part of the core.

In a series of studies, Bernard, Killworth and Sailer collected five sets of data on human interactions in bounded groups and on the actors' ability to recall those interactions (Killworth and Bernard, 1976, 1979; Bernard and Killworth, 1977; Bernard, Killworth and Sailer, 1980, 1982). In each study they obtained measures of social interaction among all actors, and ranking data based on the subjects' memory of those interactions. These data concern interactions in a small business office, again recorded by an "unobtrusive" observer. Observations were made as the observer patrolled a fixed route through the office every fifteen minutes during two four-day periods. The coreness, using the correlation criterion, was calculated using UCINET (which contains the data as a standard dataset) and these have been placed in descending order of coreness in Table 3.

Table 3. Coreness of Office workers

Id	Coreness

22	0.41
3	0.41
7	0.35
35	0.34
2	0.21
14	0.19
32	0.19
13	0.18
30	0.18
15	0.17
40	0.16
4	0.15
23	0.15
9	0.13
19	0.13
34	0.13
18	0.12
31	0.11
16	0.09
29	0.08
17	0.07
24	0.06
25	0.06
5	0.06
27	0.06
36	0.06
28	0.06
10	0.05
26	0.05
33	0.04
39	0.04
37	0.04
6	0.04
11	0.03
12	0.03
8	0.02
20	0.02
38	0.02
21	0.01
1	0.01

We can also use the discrete form of the core/periphery model to compare these results. The results are given in Table 4. This table gives a blockmodel image of the core/periphery structure. We note that the seven actors identified in the core are precisely the top seven actors in term of continuous coreness. It is also interesting to note there is a

significant drop in coreness between actor 10 and the next actor (numbered 22), this gives added support to the discrete model.[1]

Table 4 Discrete core/periphery model of the office data

	22	2	3	16	27	23	10	1	9	4	7	12	13	14	15	8	17	18	11	20	21	5	6	24	25	26	19	28	29	30	31	32	33	34
22				7	6	4	5			1						5						1						1		1		1		
2			14	4		8	1	5	1		1	1	1	1		1			1					1				1			4	1		1
3		14		1		5	1	7	1			1	3	1	1									1				2		3	4	1		1
16	7	4	1		7	6	8				1					4						4				1	1				2			
27	6			7		7	6		1		3	5			1	4						1				1	7			1				
23	4	8	5	6	7		6	4			1	6	4	1		3						2					2		10	6	1		3
10	5	1	1	8	6	6					2				1	4													1		1		3
1		5	7			4				1		1				1												1		1	3			
9		1	1		1							1	3		4			6				1		2		4	5							
4	1							1				2	1		2								14				3		1					
7		1		1	3	1	2							1	1							6	2					1	2				1	
12		1	1		5	6		1	1	2			1							1			1	3		1			3	3	1			1
13		1	3			4			3	1		1			2	1					2	1			1	4				1	1			2
14		1	1			1					1													2				3	1		1	5	3	2
15			1		1		1		4	2	1		2									1	1			1	10					1		
8	5	1		4	4	3	4	1					1					1																
17																											1	1	1			1		
18									6							1																		
11		1																										2		2				
20												1										1							1					
21													2													2								
5	1			4	1	2			1		6		1		1					1						1	2	1	1			1		
6										14	2	1			1												2		2					
24		1	1						2			3		2												3		1		3			1	1
25													1																					
26									4			1	4		1						2	1		3			3						1	1
19				1	1				5	3					10		1					2	2			3								
28	1	1	2	1	7	2		1			1			3			1		2			1		1								2	5	3
29										1	2	3		1			1			1		1	2											1
30	1		3			10	1	1				3	1						2					3							1	1	1	1
31		4	4		1	6		3				1	1	1															1				1
32	1	1	1	2		1	1							5	1		1					1					2		1			3	7
33											1			3										1		1	5		1		3		2
34		1	1			3	3					1	2	2										1		1	3	1	1	1	7	2		

As noted in our discussion of 2-mode data, a way to summarize the pattern of centrality scores in a graph is the notion of centralization. Since the intuitive basis for centralization is the graph in Figure 2, it would seem natural to extend this concept to deal with the core/periphery structure given in Figure 3. We refer to this extension as *concentration*. Since centralization looks at the difference in centrality of the most central actor to all other actors, in order to extend this to the core/periphery case, we need to compare the coreness of the actors in the core with the coreness of those in the periphery. If there is little difference in coreness, then the graph is not highly concentrated.

Suppose that C is a coreness centrality measure on a collection of n actors, and that the actors have been arranged in descending order based on C and the network relabelled so that $c_1 \leq c_2 \leq \ldots.. \leq c_n$. Let the first *j* actors comprise the membership of the core. Then we define *concentration* as follows:

$$\sum_{i=1}^{j}\left(c_i - Max(c_{j+1}, c_{j+2}, .., c_n)\right) + \sum_{k=j+1}^{n}\left(Min(c_1, c_2, .., c_j) - c_k\right)$$

[1] However, as it is a combinatorial algorithm, other runs can produce slightly different results. In such cases it is wise to identify the core as the intersection of all the core members over a number of runs and move the rest into the periphery (or define a category of "semiperiphery").

Note that the coreness of the core and periphery nodes are characterized via the maximum and minimum respectively. Alternatively, we can use the mean for both; both work well in practice. If we assume that the underlying coreness measure can have a maximum value of 1 for every core member and a value of zero for every peripheral member then the numerator has a maximum value of *n* and this can be used as a basis for normalizing concentration.[2]

The concentration measure can be used to compare different networks, just as we typically compare networks with respect to density or centralization. Borgatti and Everett (2000) speculate that groups with high concentration may perform better in certain contexts, due to the short graph theoretic distances among actors and the lack of subgroups that may develop antagonisms or alternative ways of thinking. Similarly, Schenkel, Tiegland and Borgatti (2001) argue that communities of practice will have high concentrations.

The measure can also be used to find the best place to partition a continuous coreness measure into a discrete core and a periphery. We do this by sorting actors in descending order according to coreness, and then repeatedly calculating concentration, taking the core initially to consist of just the top actor, then the top two actors, and so on, and choosing the partition that maximizes concentration.

CONCLUSION

We have discussed three extensions of the original centrality concept: one extends centrality to groups, another extends centrality to 2-mode data, and the third broadens to concept to formulate a model of a core/periphery structure. Each of these has useful application in empirical settings. As noted earlier, group centrality provides a natural way to measure the external aspect of the social capital of groups, thus providing an independent variable in a study predicting group performance. In addition, the technique can be turned around to provide a criterion for forming groups that have maximal centrality. This could be used by organizations to staff teams or taskforces with maximum clout.

The extension to 2-mode data serves a number of important functions. First, we can compare the centrality of members of different modes using a comparable metric. Second, it allows us to directly analyze 2-mode data, using the tools and concepts of network analysis, without resorting to structure-destroying transformations such as multiplying the data matrix by its transpose and dichotomizing. The result is that we can measure the extent to which, for example, an event serves as a unique bridge between different groups of actors. Two-mode tools allow us to mine a wealth of data that can be obtained unobtrusively, such as participation in projects, group memberships, event

[2] It should be noted, however, that core/periphery measures such as the principal eigenvector are usually normalized in such a way that values as extreme as 0 and 1 are not attainable. Hence, a smaller maximum concentration should be used, or alternatively, the coreness measure should be renormalized.

attendance, and so on. This is particularly useful in large networks where data collection by survey is prohibitively expensive and yields unacceptable non-response rates.

The generalization to core/periphery structures represents an advance along several different fronts. First, it extends Freeman's concept of centralization to the case of multiple actors. Centralization measures the extent to which a network revolves around a single highly central actor. But what if there are two or more actors occupying the same central position and playing that same structural role? The centralization measure, by design, gets a lower score in such a case. In contrast, our concentration measure yields the same high score regardless of how many people are in the core. Second, with the core/periphery notion we bring a modeling perspective to the measurement of centrality. We make clear, for example, that the measure of individual coreness is only interpretable when the core/periphery model fits the observed network data (Borgatti and Everett, 1999).

Our objective has been to present the concepts underlying three classes of generalization of centrality. A limitation of our study is that we have only specifically discussed the generalization of a few of the dozens of extant centrality measures. This should not be taken to imply that only the measures we discuss are generalizable. Others can and should be generalized along the lines we have presented here.

REFERENCES

Bernard H and Killworth P. (1977). Informant accuracy in social network data II. Human Communication Research, 4, 3-18.

Bernard H, Killworth P and Sailer L. (1980). Informant accuracy in social network data IV. Social Networks, 2, 191-218.

Bernard H, Killworth P and Sailer L. (1982). Informant accuracy in social network data V. Social Science Research, 11, 30-66.

Bonacich, P. 1972. "Factoring and Weighting Approaches to Status Scores and Clique Identification." *Journal of Mathematical Sociology* 2:113-120.

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

Borgatti, S. P., and Everett, M. G. 1997. Network analysis of 2-mode data. *Social Networks*, 19(3): 243-269.

Borgatti, S.P. & Everett, M.G. Graphs with short paths. Presentation at *Mathematical Sociology Conference, June 2000, Honolulu.*

Borgatti, S. P., Everett, M. G., Freeman, L. C. 1999. *Ucinet 5.0 for Windows.* Natick: Analytic Technologies.

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

Everett, M. G., & Borgatti, S. P. 1999. The centrality of groups and classes. *Journal of Mathematical Sociology.* 23(3): 181-201.

Faust, K. 1997. Centrality in affiliation networks. *Social Networks* 19(2): 157-191.

Freeman, L.C. 1979. "Centrality in Networks: I. Conceptual Clarification." *Social Networks* 1:215-39.

Freeman, S C and L C Freeman (1979). The networkers network: A study of the impact of a new communications medium on sociometric structure. *Social Science Research Reports* No 46. Irvine CA, University of California.

Johnson, S.C. 1976 Hierarchical clustering schemes. *Psychometrika* 32:241-253.
Katz, L. 1953. "A New Index Derived From Sociometric Data Analysis." *Psychometrika* 18:39-43.

Killworth B and Bernard H. (1976). Informant accuracy in social network data. Human Organization, 35, 269-286.

Killworth P and Bernard H. (1979). Informant accuracy in social network data III. Social Networks, 2, 19-46.

Schenkel, A., Tiegland, R., and Borgatti, S.P. (2001). Theorizing structural dimensions of communities of practice: A social network approach. *Presentation at the Academy of Management Conference, Washington, DC August, 2001.*