CONNECTIONS 26(1): 62-70 http://www.insna.org/Connections-Web/Volume26-1/6.Flom.pdf

A New Measure of Linkage Between Two Sub-networks[1]

Peter L. Flom[2], Samuel R. Friedman, Shiela Strauss, Alan Neaigus
National Development and Research Institutes, Inc., New York, NY, USA

Network links between groups of people with and without a certain characteristic are important in several substantive areas. For example, models might study people who "link" those with and without HIV infection; or those who are links between two types of organizations, or between two countries. New measures of this linkage are proposed (for individuals and for entire networks) and reasons why they are superior to some existing measures are detailed through examples.

INTRODUCTION

Most of the literature on social networks involves what Wasserman and Faust (1994) term "one-mode" networks, where all the actors are of the same class, and all may interact (at least in theory) with each other. They also discuss two-mode networks, where the actors are of different classes, and where interest focuses on interactions between the classes, but not within one class. The distinction depends on context and interest; for example, if we were modelling use of e-mail in a university department, we might include both students and professors in a one-mode network. On the other hand, if we were interested only in how e-mail facilitates communication between students and faculty, we could analyse similar data as a two-mode network (since we would then not be interested in faculty-faculty or student-student communication).

There is, however, an additional possibility, where we wish to include all interactions, but focus on those between two groups. We may, for example, wish to determine the extent to which the two groups are linked by some behaviour.

For example, we may wish to model HIV-risk among injection drug users (IDUs) and non-IDUs; a two-mode network would look only at connections between IDUs and non-IDUs, while a traditional one-mode analysis would not differentiate between IDUs and non-IDUs, but would instead treat connections equally. However, since IDUs may have very different infection rates from non-IDUs,

[1] This research was supported by NIDA grants P30 DA121041 and DA13128

[2] National Development and Research Institutes, 71 W. 23rd St., 8th floor, New York, NY 10010; 212.845.4485 (voice); 917.438.0894 (fax); flom@ndri.org

to study HIV spread we may want to look at all connections, but to differentiate those involving both IDUs and non-IDUs from those involving only one group. Here, a two-mode network is inappropriate, because HIV can spread within groups of IDUs and non-IDUs as well as between them, and because the forms of risk linkage can differ between the two groups. Traditional measures (detailed below) do not allow us to focus on the systemic importance of links between groups, but rather treat all links equally.

To better describe such networks, we introduce a new measure of "linkage", used here to describe the concept of a single node being a link between groups in a single connected network. In Figure 1, suppose that A1, A2, and A3 have some trait, while B1, B2, and B3 do not. Then L is a link between the two groups (regardless of whether L has the trait). The idea of linkage is related to that of a bridge between groups, which has been used by other researchers (Elwood, 1995; Morris, Podhista, et al. 1996). Note, however, that Elwood defines a bridge group as one which links a network to people outside the network. In contrast, we define bridge as a link between differentiated groups within a network. Linkage is also related to Everett's concept of "bridge" (cited in Scott, 1991) as a line that connects two cycles, in that linkage involves two different groups (i.e. people with different characteristics), regardless of whether these groups are cyclic. Trotter, Rothenberg and Coyle (1995) identified bridges as one of the key areas for future research in network methods. To the best of our knowledge, however, no currently available measure allows for a numeric estimation of bridgeness or linkage.

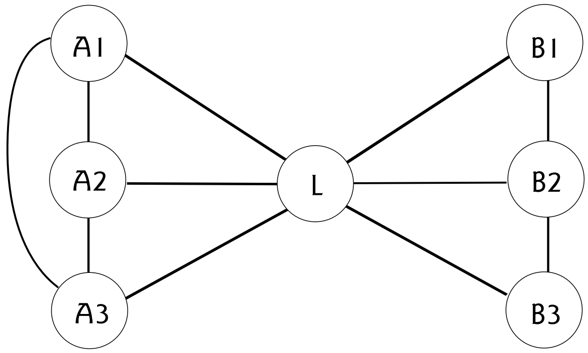

Figure 1

In the next section, we describe existing measures that approximate linkage and propose a new one. Section three examines the performance of existing measures and our new measure in small hypothetical networks. Section four describes extensions of our measure to entire networks (as opposed to single nodes). We close with limitations, suggestions for future research, and conclusions.

EXISTING MEASURES: their inadequacy to the problem, and a proposed solution

The closest existing concept to linkage in this sense is centrality. There are several existing measures of this concept; in this section, we discuss several of these, and discuss why they are not adequate for the task at hand. We then describe our new measure. First, though, we should note that the desired measure is inherently nondirectional; directional measures (which Wasserman and Faust, 1994 refer to as measures of prestige) will not be further considered. Second, the paths considered will be unweighted. Third, we are concerned only with connected networks. The first two of these are possible areas for future work; the last may also be, but the notion of centrality in an unconnected network is somewhat problematic.

Another related concept is segregation. While we are interested in which nodes are key in connecting two groups, segregation is concerned with separation of groups, or limiting interaction between them (Freeman, 1978). There are two key differences between this measure and the one we seek. First it is in the opposite direction – if two groups are totally segregated there is no link between them. In itself, this is a somewhat trivial distinction. It leads, however to a more important one: Segregation, by its nature, applies only to whole networks. While one node might be totally segregated from all others, it does not make sense to speak of the extent to which a node segregates a network, but only of the extent to which it integrates a network.

Centrality

The simplest measure of the centrality of a node, and one proposed by many authors (see Freeman, 1979, for a review) is simply its degree, which is known as the actor degree centrality. Wasserman & Faust (1994) give a standardized form as:

$$C_D(n_i) = \frac{d(n_i)}{g-1} = \frac{\sum_j x_{ij}}{g-1} \qquad (1)$$

where n_i is the node in question, $d(n_i)$ is degree, g is the number of nodes, and x_{ij} indicates the presence or absence of a link from node *i* to node *j*

Closeness centrality focuses on how close a node is to all other nodes in the network. Beauchamp (1965) noted that actors who are central in this sense can communicate information to other actors very effectively. Several authors have developed this notion more formally (see Wasserman & Faust, 1994 and Freeman, 1979 for a review). A standardized form (Sabidussi, 1965; Beauchamp, 1965) is given by

$$C_c(n_i) = \frac{g-1}{\sum_{j=1}^{g} d(n_i n_j)} \qquad (2)$$

where d(a,b) is the distance between a and b, and other terms are as defined above.

Betweenness centrality

One existing measure that does combine aspects of betweenness and centrality, and may be the closest existing measure to the desired one, is actor betweenness centrality (C_B), which was developed by Freeman (1977). In words, betweenness centrality is the proportion of all geodesics on which a particular node lies, but which do not involve that node. In standardized form (for actor n_i) by Wasserman and Faust (1994, p. 190) as

$$C_B(n_i) = \sum g_{jk}(n_i) / g_{jk}\left[(g-1)(g-2)/2\right] \qquad (3)$$

where the sum is over all j < k, with i distinct from j and k, g_{jk} is the number of geodesics connecting j and k, $g_{jk}(n_i)$ is the number of those geodesics which include n_i, and g is the number of nodes. (The term in brackets standardizes the measure to vary from 0 to 1). It should be noted that, while there is a single summation, the sum is over all the geodesics g_{jk} with $j < k$ and i distinct from both j and k. Borgatti (1995) describes this as indexing "the extent to which a node's presence facilitates the flow of that-which-diffuses. If a node that is high on betweenness centrality is removed from the network, then the speed and certainty of transmission from one arbitrary point to another are more damaged than if a node low on betweenness centrality is removed." (p. 114).

Segregation

Freeman (1978) developed a measure of segregation between two groups. Intuitively his measure is the "cross-class edges that are missing in the observed graph, as compared with a graph where edges are generated at random" (p. 416); this is scaled by dividing the expected value by the maximum possible value for a graph of a given size. Since this can only be computed for an entire network, we do not consider it further.

Problems with existing measures

The main problem with all these measures for measuring linkage between two groups is that no distinction is made between nodes in different groups, or between geodesics which remain in one group and those which cross to different groups. They do not account for the fact that the nodes are in different groups, and thus cannot measure which nodes connect the two groups, nor the extent to which they do so.

Proposed solution

We can account for the existence of two groups as follows: Suppose we have a network of n nodes, of which i have the condition and $j = n - i$ do not. First, list all the nodes with the trait as A_1 to A_i, and those without the trait as B_{i+1} to B_n, exclude the node (X) for which we are estimating linkage from whichever list it is on (i.e., A if it has the trait, B if it does not). Then, if x does not have the trait, the linkage for node x is

$$Q(x) = \sum_{a-1}^{i} \sum_{b=i+1}^{n-1} g_{ab}(x) / g_{ab}\, i(j-1) \qquad (4)$$

where $g_{ab}(x)$ is, as in formula 3, the number of geodesics connecting a and b and including x, and i(j-1) standardizes the measure to vary between 0 and 1. If x does have the trait, the summations are from a = 1 to i-1, and b = i to n-1, and the standardization term is (i-1)j. In words, Q_x is the proportion of geodesics linking A_i and B_j that contain X, where X is not an endpoint of the geodesic. (see section 5, below for an example of computing Q).

As a simple example, in Figure 1, suppose that L does have the trait, and relabel L as A4. Table 1 lists values for Q(L) and $C_B(L)$. It can be seen that, while L is clearly different from the other nodes in terms of both Q and C_B, the difference is greater for Q than for C_B. Further, Q = 1 for node L and 0 for all other nodes. A value of 1 indicates that, if the node is deleted, the network is split into two connected components, one with and one without the condition. This may be stated as a theorem, which is proven in the appendix:

> Theorem 1: Given a network N with n nodes, of which some have a condition and some do not, deletion of a single node X will result in two components, one of which has the condition and one of which does not, if, and only if, Q(x) = 1.

Table 1. Q(X) and CB(X) for Figure 1*

X	Q(X)	CB(X)
A1	0.00	0.00
A2	0.00	0.00
A3	0.00	0.00
L	1.00	0.63
B1	0.00	0.00
B2	0.00	0.03
B3	0.00	0.00

* In figure 1, Q1 and Q2 are equivalent; these terms are explained in the text.

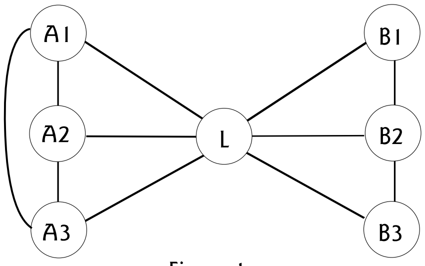

Figure 1

PERFORMANCE OF THE MEASURES IN SMALL, HYPOTHETICAL NETWORKS

Desired characteristics

Here we list some characteristics the measure should have, in order to be able to compare the new measure with others. In Figure 1, we would like the measure to have the following characteristics:

a. Node L1 should be markedly higher on the measure than any other node.

b. All other nodes should be equal.

c. L1 should be 1, or some other distinct value, to indicate that deleting it will split the network into two groups.

Figure 2a is a small network, designed to mimic HIV-risk links (i.e., either sharing needles or having sex) among IDUs and non-IDUs. Group A are IDUs, and group B are not. Here, we would like the measure to have the following characteristics:

a. Node A4 should be markedly higher than any other node.

b. Nodes A3, B4, B5, and B6 should be lowest.

c. Node A4 should be 1, or some other distinctive value.

Figure 2b adds another link between the two groups (the dashed line). Here, we would like:

a. Node A4 to be higher than any other nodes.

b. Nodes A3, B4, B5 and B6 to be lowest.

Comparing Figures 2a and 2b, we would like A4 to decrease and A6 and B3 to increase (although B6 is now more connected to the rest of the network than it was in Figure 2a, removing it from the network would have no effect on the connectedness of the A and B group, as it lies on no geodesics between the two groups).

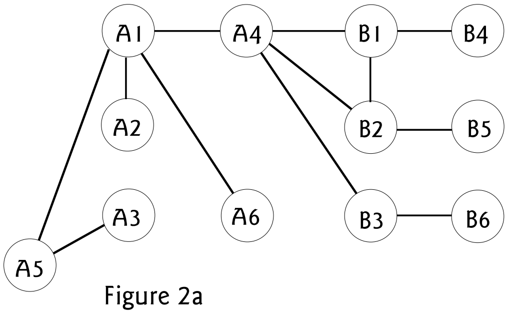

Figure 2a

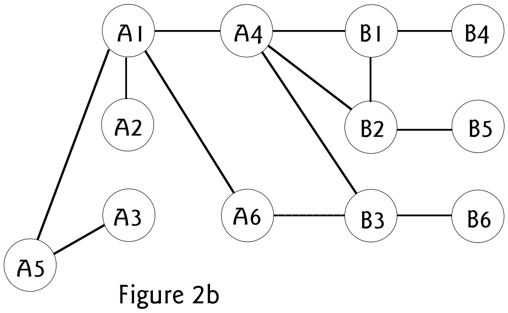

Figure 2b

Table 2. Q(X) and CB(X) for Figures 2a and 2b

	Figure 2a		Figure 2b		
X	Q(X)	CB(X)	Q1(X)	Q2(X)	CB(X)
A1	0.80	0.58	0.72	0.74	0.46
A2	0.00	0.00	0.00	0.00	0.00
A3	0.00	0.00	0.00	0.00	0.00
A4	1.00	0.69	0.79	0.82	0.56
A5	0.20	0.18	0.18	0.16	0.18
A6	0.00	0.00	0.15	0.13	0.09
B1	0.20	0.18	0.19	0.16	0.18
B2	0.20	0.18	0.19	0.16	0.18
B3	0.20	0.18	0.26	0.38	0.23
B4	0.00	0.00	0.00	0.00	0.00
B5	0.00	0.00	0.00	0.00	0.00
B6	0.00	0.00	0.00	0.00	0.00

Multiple crossings

One complication is whether to count as geodesics only those paths which cross exactly once between those with and without the trait, or to count geodesics without regard to how often they cross. For example, in Figure 2b, there are two possible geodesics between B1 and A6: B1-A4-A1-A6 and B1-A4-B3-A6. The former of these crosses once, the latter three times. Below, we calculate Q using each of these definitions of geodesic. Q1 includes only single-cross-geodesics, Q2 includes all geodesics. In some simple networks Q1 and Q2 are identical; where this is so, as in Figure 1 and Figure 2a, we do not show both.

In Figure 1 (see Table 1), it can be seen that Q shows more difference between node L and all other nodes, reflecting the fact that L is the key link between the groups. It can also be seen that the link between A1 and A3 affects C_B, but not Q.

In Figure 2a (see Table 2), Q clearly distinguishes A1 and A4 from the other nodes; while these nodes are also highest in terms of C_B, the differences between A4 and the next highest node are much smaller for C_B than for Q. It is clear from visual inspection that this is "correct" if we are interested in measuring the extent to which a node connects the A group and the B group. Without both A1 and A4, the link would not exist. If we add one link to Figure 2a (the dashed line), both measures change (see Table 2, for Figure 2b). The changes are not very different for the two measures, however. This is the first case in which Q1 is different from Q2, although the differences are small. One fairly important difference is that for B3 all three measures go up (as desired), but C_B goes up more modestly, Q1 goes up slightly more and Q2 goes up substantially. It is difficult to extensively modify such a small network.

EXTENSIONS TO ENTIRE NETWORKS

The measure proposed above is for single nodes in a network with two groups. In this section, we describe extending this measure to entire networks

As Freeman (1977) notes, there are two contrasting views of what centrality for an entire component means: First it could mean the extent to which all points are central; second, it could mean the dominance of a single point. We follow him in using the second definition and use the average difference between the normalized measure of the most central node and that of all other nodes. Thus, Freeman (1977) defined the centrality of a network as

$$CB_{net} = \frac{\sum_{i=1}^{g}\left[C_B\left(n^*\right) - C_B\left(n_i\right)\right]}{g-1} \quad (5)$$

where $C_B\left(n^*\right)_i$ (eq5a) is the largest betweenness centrality value associated with any point in the network (p. 39).

We ca similarly define Q_{net} as

$$Q_{net} = \frac{\sum_{i=1}^{g}\left(Q^* - Q\left(i\right)\right)}{g-1} \quad (6)$$

As examples, for Figure 1, $CB_{net} = .62$ and $Q_{net} = 1$. For Figure 2a, $CB_{net} = .58$ and $Q_{net} = .85$. For Figure 2b, $CB_{net} = .44$, $Q1_{net} = .64$ and $Q2_{net} = .66$. These values seem to indicate that, for whole networks as for nodes, the new measure does a better job of capturing what we want to capture. In Figure 1, the entire connection depends on a single node, and Q_{net} indicates this well. Comparing Figure 2a to Figure 2b,, the addition of the line to Figure 2b makes more of a difference for (either) Q_{net} measure than for CB_{net}, and this makes the importance of that line more obvious.

COMPUTATIONAL CONSIDERATIONS

For this paper, all the numbers were computed by hand by the first author. We have been unable to develop an algorithm for efficiently computing the measure for large groups (i.e. with more than approximately 10 nodes). For small groups, the method used was as follows: First, make a matrix with the columns containing the nodes with the condition, and the rows containing the nodes without it. In each cell, enter the nonterminal nodes in each geodesic between the column and the row. Then, to compute Q for a particular node, count all the geodesics containing that node, and divide by all the geodesics where that node is not a terminal node. If there are two (or more) geodesics between a pair of points, include both. An example may clarify this process; for Figure 2a, we write:

	B1	B2	B3	B4	B5	B6
A1	A4	A4	A4	A4B1	A4B2	A4B3
A2	A1A4	A1A4	A1A4	A1A4B1	A1A4B2	A1A4B3
A3	A5A1A4	A5A1A4	A5A1A5	A5A1A4B1	A5A1A4B2	A5A1A4B3
A4				B1	B2	B3
A5	A1A4	A1A4	A1A4	A1A4B1	A1A4B2	A1A4B3
A6	A1A4	A1A4	A1A4	A1A4B1	A1A4B2	A1A4B3

If we then wish to compute Q for (say) node A1, we count the number of geodesics containing A1, and find there are 24. We divide this by the number of geodesics where A1 <u>might</u> have appeared which is 30, and the result is 0.8.

LIMITATIONS AND FUTURE RESEARCH

One limitation (as noted above) is that the measure deals only with binary data. While a case can be made that this is appropriate in some applications (e.g., people either have sex together, or they do not, and if they do not then there is no risk of sexually transmitting disease), it is also true that, in many applications, people may interact to different degrees. This suggests some measure combining elements of flow-centrality (Freeman, Borgatti, et al. 1991) and the measure proposed in this paper.

A second limitation is that the measure deals with only two groups. The first author is currently working on extending the measure to multiple groups.

DISCUSSION AND CONCLUSIONS

In practice, the differences between Q1 and Q2 may be small. In our example, Q2 seemed preferable. However, preference for one or the other may depend on what the trait and the link are (and their relative computability). For example, if the trait is some disease with multiple strains, each of which protects against the others, then we may wish to consider only links which involve one cross, since a cross depends upon the prior infection (and infectiousness) of the final link in the process. On the other hand, if we are dealing with a disease like hepatitis C with multiple strains with somewhat similar prognoses but in which: (a) infection with one strain does not offer protection against the others; and (b) infection with multiple strains offers different prognoses, we may want to consider all the crossing patterns as possibilities.

The importance of locating a bridge depends largely on context. One example would be that of tracing HIV infection from drug injectors to the general population. Here the link would be having (unprotected) sex, and the two groups would be people who inject and people who don't. If, in a particular sample, it appears that certain characteristics are correlated with linkage, then people with those characteristics might be targeted for interventions, in the hope that these interventions might have particularly great effects.

REFERENCES

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science*, 10: 161-163.

Borgatti, S. P. (1995). Centrality and AIDS. *Connections*, 18: 112-114.

Elwood, W. N. (1995). Lipstick, needle, and company: A case study of the structure of a bridge group in Houston, Texas. *Connections*, 18: 46-57.

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

Freeman, L. C. (1978). Segregation in Social Networks. *Sociological Methods and Research*, 6: 411-429.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40: 35-41.

Freeman, L. C., Borgatti, S. P., & White, D. R. (1991). Centriality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, 13: 141-154.

Morris, M., Podhista, C., Wawer, M. J., & Handcock, M. S. (1996). Bridge populations in the spread of HIV/AIDS in Thailand. *AIDS*, 10: 1265-1271.

Sabidussi, G. (1965). The centrality index of a graph. *Psychometrika*, 31: 581-603.

Scott, J. (1991). *Social network analysis*. Newbury Park: Sage Publications .

Trotter, R. T., Rothenberg, R. B., & Coyle, S. (1995). Drug abuse and HIV prevention research: Expanding paradigms and network contributions to risk reduction. *Connections*, 18: 29-45.

Wasserman, S., & Faust, K. (1994a). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

APPENDIX: PROOF OF THEOREM

Theorem 1: Given a network with n nodes, of which some have a condition and some do not, deletion of a single node X will result in two components, one of which has the condition and one of which does not, if, and only if, Q(x) = 1.

Preliminary: If there is a path connecting two nodes A and B, then there must be a geodesic between them (i.e., the existence of a path implies a shortest path). If A has the condition, and B does not, then there must be at least one pair of adjacent nodes C and D, one of which has the condition and one of which does not. Since adjacency is the shortest possible path, it is, of necessity, a geodesic. Therefore, every path linking A to B includes a geodesic linking some node with the condition to some node without it.

Sufficiency: Suppose the theorem is false. Then, for some network G, with some node X having Q = 1, there is at least one pair of nodes A_i and B_j where X is not on the geodesic, and where X is distinct from A_i and B_j. But if Q = 1 then node X is on all such geodesics

Necessity: If there is no node X with Q = 1, then there is at least one geodesic between A_i and B_j that does not include X. Therefore, removing X leaves that other geodesic. QED.