G Model
SON-725; No. of Pages 8
ARTICLE IN PRESS
Social Networks xxx (2012) xxx–xxx
Contents lists available at SciVerse ScienceDirect
Social Networks
journal homepage: www.elsevier.com/locate/socnet

Categorical attribute based centrality: *E–I* and *G–F* centrality

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

[a] Mitchell Centre for Social Network Analysis, School of Social Sciences, University of Manchester, Arthur Lewis Building, Bridgeford Street, Manchester M13 9PL, UK
[b] Links Centre, Gatton College of Business and Economics, University of Kentucky, Lexington, KY 40506-0034, USA

ARTICLE INFO

Keywords:
Centrality
E–I index
Brokerage roles
Categorical attributes

ABSTRACT

In a paper examining informal networks and organizational crisis, Krackhardt and Stern (1988) proposed a measure assessing the extent to which relations in a network were internal to a group as opposed to external. They called their measure the *E–I* index. The measure is now in wide use and is implemented in standard network packages such as UCINET (Borgatti et al., 2002). The measure is based on a partition-based degree centrality measure and as such can be extended to other centrality measures and group level data. We explore extensions to closeness, betweenness and eigenvector centrality, and show how to apply the technique to sets of subgroups that do not form a partition. In addition, the extension to betweenness suggests a linkage to the Gould and Fernandez brokerage measures, which we explore.

1. Introduction

Many measures of centrality – in particular, all of the measures referred to by Borgatti and Everett (2006) as "radial" measures of centrality – can be written as the row (or column) sums of a cohesion matrix, meaning a matrix whose cells indicate the presence or extent of dyadic cohesion between pairs of nodes. For example, degree centrality can be calculated as the row sums of an adjacency matrix, closeness centrality can be computed as the row sums of a geodesic matrix, and eigenvector centrality can be viewed as weighted row sums of the adjacency matrix where the weights are the centralities of the column nodes. Similarly, the influence measures of Katz and Hubbell can be seen as the row sums of a matrix representing the number of walks of all lengths between every pair of nodes, weighted inversely by length.

As such, the (radial) centrality score of a given node can be decomposed into the separate contributions made by each cell in the row of the cohesion matrix. If the rows and columns of the cohesion matrix were to be sorted according to some categorical attribute of the nodes, such as gender, then for any given row, summing the values for those cells corresponding to, say, males, would give the contribution of male nodes to the centrality of the node corresponding to that row. Thus, using gender as the example, we can partition any node's radial centrality into a portion due to the men in the network, and a portion due to the women in the network. The relative contribution of different groups to the centrality of each node is useful data that can be used to test a variety of hypotheses, such as which individuals derive their central position from one group or multiple groups.

In this paper, we examine decomposing centrality measures into portions that are "due" to different groups of nodes. In doing so, we draw out an obvious correspondence with measures of homophily such as Krackhardt and Stern's (1988) *E–I* index, which in turn allows us to generalize homophily in ways that correspond to the many different kinds of centrality that have been elucidated in the literature. In addition, when we consider betweenness centrality in this light, we are able to draw parallels with Gould and Fernandez brokerage, and indeed extend their measures to apply not just to adjacency but longer paths as well.

2. *E–I* index and degree centrality

Krackhardt and Stern (1988) propose a measure of homophily called the *E–I* index, which is defined as follows:

$$\text{E-I index} = \frac{E - I}{E + I} \quad (1)$$

where E = the number of external (between-group) friendship edges and I = the number of internal (within-group) friendship edges.

The measure varies between +1 and −1, where larger values indicate greater heterophily and smaller values indicate greater homophily. (When used as a measure of homophily it is often helpful to subtract the ratio from 1.) An attractive feature of the measure is that, as a ratio, it is not dependent on the density of the network. In addition, as Krackhardt and Stern point out, the measure allows for variation in peoples definition of friendship and so is not sensitive to certain kinds of measurement errors.

* Corresponding author. Tel.: +44 0 161 275 2515.
E-mail addresses: martin.everett@manchester.ac.uk (M.G. Everett), sborgatti@uky.edu (S.P. Borgatti).

0378-8733/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.socnet.2012.06.002

Krackhardt and Stern apply the measure to the network as whole, obtaining a single value that describes the tendency of the network to be heterophilous. However, the UCINET (Borgatti et al., 2002) implementation of this measure makes a number of extensions. First, since *E–I* cannot always achieve its extremal values, UCINET provides a rescaled value that controls for the density and group sizes. Second, it performs a permutation test to see whether the index is significantly higher or lower than would be expected if the edges were distributed entirely by chance. Third, it calculates the measure not only for the network as a whole, but for groups and individual nodes as well. The latter point means that, for each node in the network, we can compute the extent to which the node has ties with members of its own group (internal ties) versus other groups (external ties).

Of course, at the individual level, $E+I$ is the number of ties a node has, and corresponds to the row sum of a given row in the adjacency matrix. Thus, when we calculate the *E–I* index at the individual level, we are effectively partitioning degree centrality into the portion due to ties with the node's own group and the portion due to ties to all others. The relative importance of outgroup ties is then summarized by the *E–I* statistic.

Table 1 shows the results of partitioning degree centrality and the calculation of *E–I* for an empirical dataset known as the "campnet" data, available in UCINET. The data consist of ties among participants and instructors of a 3-week workshop. The nodes of the network can be partitioned into three groups: female participants, male participants, and instructors. In Table 1 and in most of our analysis as all the instructors are male we use just two groups, namely male and female, but occasionally we use three and we flag when this has been done. The data are directed, but for simplicity we have taken the underlying graph, which consists of a single component. In the table, the number of ties from each node to same-group members and other-group members are shown, along with the *E–I* index for each node.

The relationship between calculating the *E–I* index and partitioning degree centrality can be generalized to other centrality measures as well. The next sections consider closeness centrality, eigenvector centrality and betweenness centrality in the *E–I* context.

2.1. Closeness E–I

Closeness centrality (Freeman, 1979) is defined as the sum of geodesic distances from a node to all others. As such, we can clearly partition this figure into the sum of two quantities: the distances to members of a node's own group (call this I), and the distances to members of other groups (call this E). As a result, we can readily compute the *E–I* index in the usual manner.

Table 1
Degree *E–I*.

Node	Internal	External	Total	*E–I*
HOLLY	2	3	5	0.200
BRAZEY	0	3	3	1.000
CAROL	3	0	3	−1.000
PAM	5	0	5	−1.000
PAT	4	0	4	−1.000
JENNIE	3	0	3	−1.000
PAULINE	4	1	5	−0.600
ANN	3	0	3	−1.000
MICHAEL	4	1	5	−0.600
BILL	3	0	3	−1.000
LEE	2	1	3	−0.333
DON	3	1	4	−0.500
JOHN	2	1	3	−0.333
HARRY	3	1	4	−0.500
GERY	4	0	4	−1.000
STEVE	4	1	5	−0.600
BERT	3	1	4	−0.500
RUSS	4	0	4	−1.000

Table 2
Closeness *E–I*.

Node	Internal	External	Total	*E–I*
HOLLY	14	24	38	0.263
BRAZEY	34	25	59	−0.153
CAROL	14	34	48	0.417
PAM	12	30	42	0.429
PAT	13	30	43	0.395
JENNIE	15	40	55	0.455
PAULINE	12	28	40	0.400
ANN	14	34	48	0.417
MICHAEL	16	20	36	0.111
BILL	22	28	50	0.120
LEE	24	35	59	0.186
DON	22	21	43	−0.023
JOHN	20	18	38	−0.053
HARRY	22	21	43	−0.023
GERY	14	22	36	0.222
STEVE	17	28	45	0.244
BERT	22	29	51	0.137
RUSS	17	23	40	0.150

Table 2 gives the results using the campnet dataset. Note that although the computation of *E–I* does not change when using distances, the interpretation should be reversed: whereas normally a positive *E–I* indicates heterophily (more ties to out-group members than in-group members), now a positive *E–I* indicates that distances to out-group members are greater than distances to in-group members, which is consistent with homophily.

As noted before, it is not just the *E–I* scores that are of interest here. The partial sums that are the within and between scores are useful variables in themselves. The decomposition essentially suggests that there are different ways of getting to a certain level of centrality. Some nodes might do it through a heavy reliance on their fellow in-group members, while others achieve the same overall level using out-group members. We might hypothesize that the degree of reliance on in-group members for centrality might be related to a personality characteristic, such as openness (McCrae and John, 1992). Alternatively, we could use the node profile across these scores to categorize them: for example Pam and Pat have very similar patterns, while Brazey is quite different.

It should be noted that closeness *E–I* has the same problems that closeness centrality does: the measure is technically undefined when the network is disconnected, which is a likely occurrence in the case of directed data. However, just as with closeness centrality, we can get around this using any number of variations on the distance matrix, such as taking the reciprocals of the distances and setting undefined distances to zero (Borgatti, 2005), or subtracting the distances from a constant (Valente and Foreman, 1998; Everett and Borgatti, 2010).

2.2. Eigenvector E–I and related methods

The *E–I* centrality can be applied in the same way to eigenvector centrality (Bonacich, 1972). The internal centrality scores of a given node are proportional to the actors within the group to which that node is connected and the external centrality scores are proportional to the external nodes to which the node is connected. This can easily be expressed in matrix form. Let $A=(a_{ij})$ be the adjacency matrix corresponding to a network G and let X be a subgroup of nodes in G. We then define the matrices A^I and A^E corresponding to X as follows

$$\begin{aligned} A^I_{ij} &= a_{ij} \quad \text{if} \quad i,j \in X \quad \text{and 0 otherwise} \\ A^E_{ij} &= a_{ij} \quad \text{if} \quad i \in X \quad \text{and} \quad j \notin X \quad \text{and 0 otherwise} \end{aligned} \tag{2}$$

Table 3
Eigenvector *E–I*.

	Internal	External	Total	*E–I*
HOLLY	0.132	0.243	0.375	0.296
BRAZEY	0.000	0.097	0.097	1.000
CAROL	0.196	0.000	0.196	−1.000
PAM	0.291	0.000	0.291	−1.000
PAT	0.247	0.000	0.247	−1.000
JENNIE	0.176	0.000	0.176	−1.000
PAULINE	0.224	0.038	0.262	−0.711
ANN	0.179	0.000	0.179	−1.000
MICHAEL	0.266	0.092	0.357	−0.486
BILL	0.243	0.000	0.243	−1.000
LEE	0.073	0.024	0.097	−0.510
DON	0.225	0.092	0.317	−0.420
JOHN	0.090	0.064	0.154	−0.169
HARRY	0.225	0.092	0.317	−0.420
GERY	0.206	0.000	0.206	−1.000
STEVE	0.145	0.024	0.169	−0.720
BERT	0.105	0.024	0.128	−0.631
RUSS	0.161	0.000	0.161	−1.000

Table 4
Weighted *E–I*.

	Internal	External	Total	*E–I*
HOLLY	0.198	0.122	0.320	−0.239
BRAZEY	0.000	0.049	0.049	1.000
CAROL	0.294	0.000	0.294	−1.000
PAM	0.437	0.000	0.437	−1.000
PAT	0.377	0.000	0.377	−1.000
JENNIE	0.264	0.000	0.264	−1.000
PAULINE	0.336	0.019	0.356	−0.893
ANN	0.269	0.000	0.269	−1.000
MICHAEL	0.399	0.046	0.445	−0.793
BILL	0.365	0.000	0.365	−1.000
LEE	0.110	0.012	0.122	−0.802
DON	0.338	0.046	0.384	−0.760
JOHN	0.135	0.032	0.167	−0.617
HARRY	0.338	0.046	0.384	−0.760
GERY	0.309	0.000	0.309	−1.000
STEVE	0.218	0.012	0.230	−0.895
BERT	0.158	0.012	0.170	−0.858
RUSS	0.244	0.000	0.244	−1.000

We note that $\Sigma_X(A^I + A^E) = A$. If $\boldsymbol{x}$ is an eigenvector with corresponding eigenvalue λ then $A\boldsymbol{x} = \lambda\boldsymbol{x}$ and so we can define the internal eigenvector centrality within X as $\boldsymbol{x}^I = 1/\lambda A^I \boldsymbol{x}$ and the external centrality as $\boldsymbol{x}^E = 1/\lambda A^E \boldsymbol{x}$. We note that $\boldsymbol{x}^I + \boldsymbol{x}^E = \boldsymbol{x}$ when we sum over all X and so this is a decomposition of the eigenvector centrality into internal and external scores. These can then be used to find the eigenvector scores *E–I* and hence the group and whole network *E–I* scores. Again using the campnet data we obtain the results as shown in Table 3.

Another way of thinking about the Internal and External scores is as weighted averages. For example Holly has internal connections to Pam and Pat and external connections to Michael, Harry and Don. The sum of eigenvector centrality scores of her internal connections is 0.538 (= 0.291 + 0.247) and her external connections are 0.991; adding these together means she is connected to a total of 1.529. Her eigenvector centrality score is 0.375 and we now weight the internal and external sums so that they add up to this value. Hence the Internal score is $0.375 \times 0.538/1.529 = 0.132$ and the External score is $0.375 \times 0.991/1.529 = 0.243$.

Looking at Table 3 unsurprisingly the results in this case are very similar to those for ordinary degree *E–I*, although the rank-orders are not identical. Of course, we can do the same things with beta centrality (Bonacich, 1987), and the methods of Katz (1953) and Hubbell (1965). These have the advantage of having tunable parameters that would allow the results to differ somewhat from degree centrality.

3. Theoretical reweighting

One benefit of this kind of decomposition – for all radial measures, not just eigenvector – is the ability to weight the partial scores differently so as to give more weight to either internal or external ties. For example, we may be looking at leadership within groups and have a theory that suggests that within-group ties matter more for the perception of leadership than external ties. In constructing a structural measure of leadership we might decide to weight internal ties, say, three times as much as external ties. If the Internal score is the same as the External score then we would want the total to remain the same and so we need to make sure our weights sum to 2. We therefore choose 1.5 and 0.5 as our weights in this case and the results of doing this are shown in Table 4.

Here the focus is on the "total column", which constitutes a new, weighted, version of eigenvector centrality. We can see that it makes significant changes in the rank order of actors. For example, Holly had previously held the highest eigenvector centrality score but with more weight being given to internal scores there are now seven actors with a higher score. This kind of weighting is very consistent with the general concept of eigenvector, which from a substantive view can be seen as differing from degree centrality by the fact that it weights ties to certain alters higher than others.

The *E–I* scores resulting from this reweighting would not normally be of interest, as the weighting necessarily results in all actors getting more homophilous scores, the exception being the extreme *E–I* scores of −1 and 1 which remain the same. We note that the rank orders of the other scores change as well.

One way to think of eigenvector centrality is as a form of iterated degree. We start with degree centrality and then in the next iteration a node's centrality is the sum of the degrees of the nodes they are connected to. This score is then normalized and the process repeated until the scores remain constant. In our approach we have taken the result of this and re-apportioned the eigenvector scores but we could have kept up the spirit of the eigenvector centrality concept by using iterative weighting at each stage. Although we did not pursue that approach here but it may be worth considering.

Finally, we note there is nothing special about eigenvector and reweighting and this technique could have been applied to our previous centrality indices of closeness and degree.

4. Group sizes

We have not yet considered the effect of unequal group sizes. Clearly, if we partition a network into two classes, one very large and one very small, the *E–I* statistic is likely to be negative for virtually every member of the large group, and positive for every member of the small group, especially if the average degree is high. In some cases this is not an issue: after all, the *E–I* statistic is correctly reporting that members of the large group tend to have members of the large group as friends, as do members of the small group. Being a member of a majority has certain consequences, as does being the friend of someone who is a majority member, and so we may well need to measure the extent to which individuals are experiencing this. If so, *E–I* will do a good job.

However, if the research interest is not in measuring the actual outcomes of actor choices but rather their underlying preferences or tendencies, then unequal group sizes will cause *E–I* to produce misleading numbers. In this case, we want to capture the extent to which actors are choosing members of the ingroup at rates that are higher or lower than the relative availability of each group.

One way to handle this is normalization. In this approach, we normalize the values of E and I before applying the $(E - I)/(E + I)$ formula. For example, if we consider degree *E–I*, an obvious approach is to divide the observed value of E and I by their respective

Table 5
Normalized *E–I* scores.

Node	Internal	External	Total	*E–I*
HOLLY	0.286	0.300	0.586	0.024
BRAZEY	0.000	0.300	0.300	1.000
CAROL	0.429	0.000	0.429	−1.000
PAM	0.714	0.000	0.714	−1.000
PAT	0.571	0.000	0.571	−1.000
JENNIE	0.429	0.000	0.429	−1.000
PAULINE	0.571	0.100	0.671	−0.702
ANN	0.429	0.000	0.429	−1.000
MICHAEL	0.800	0.083	0.883	−0.811
BILL	0.600	0.000	0.600	−1.000
LEE	0.400	0.083	0.483	−0.655
DON	0.600	0.083	0.683	−0.756
JOHN	0.400	0.083	0.483	−0.655
HARRY	0.600	0.083	0.683	−0.756
GERY	1.333	0.000	1.333	−1.000
STEVE	1.333	0.071	1.405	−0.898
BERT	1.000	0.071	1.071	−0.867
RUSS	1.333	0.000	1.333	−1.000

Table 6
Crosstab of all dyads in network.

		Same group?		
		1	0	
Is tied?	1	a	b	(a+b)
	0	c	d	(c+d)
		(a+c)	(b+d)	N

maximums. In the case of the campnet dataset, there are 8 female participants, 6 male participants, and 4 instructors. Note we have now moved to using the three categories as this means the groups have very different sizes (the gender split was nearly equal). If Holly is a female participant, she can at most have 7 internal ties and 10 external ties. As shown in Table 1, she has 2 internal ties and 3 external ties. Normalized, these become 0.286 and 0.3, and recomputing *E–I* gives us a score of 0.024. This is fairly different from her original *E–I* score of 0.2, indicating that even her fairly modest heterophily score of 0.2 was largely due to there being more outgroup choices than ingroup choices. The new *E–I* score is just about zero, indicating no preference. Table 5 shows the complete set of normalized degree *E–I* scores. The same strategy can be applied to any of the other centralities discussed above.

An alternative approach is to abandon *E–I* in favor of a measure of homophily that is invariant to differences in group sizes.[1] An obvious possibility in the case of non-valued ties is the point-biserial correlation, first used in this context by Ibarra (1992). To use it, we cross-tabulate two dyadic variables: X and Y; X is a 1/0 variable in which $x_{ij} = 1$ if node i has a tie with node j. Y is also 1/0 and $y_{ij} = 1$ if node i is in the same group as node j. A crosstab of X in Y is shown in schematic form in Table 6.

Using the notation of the cross-tab, the ordinary degree *E–I* index can be written as $(b - a)/(a + b)$. This notation makes it obvious that the *E–I* index counts only ties that are present: it effectively asks, of all observed ties, what is the proportion that falls in the same group as ego? Hence the *E–I* index is a rescaling of $a/(a + b)$, which runs between 0 and 1. If a measure of homophily were to consider non-ties as well, it would automatically take into account group sizes. For example, consider the hypothetical contingency table in Table 7. The *E–I* index is a sizeable −.333, indicating

[1] Ultimately, of course, the two approaches are the same since if we rewrite the *E–I* formula in terms of the un-normalized values of *E* and *I*, the result will be a new formula that "bakes" in the invariance property.

Table 7
Hypothetical contingency table.

		Same group?		
		1	0	
Is tied?	1	10	5	15
	0	40	20	60
		50	25	75

considerable homophily. However, the odds of a tie being internal are no greater than the odds of a non-tie being internal, just as the odds of a same-group pair having a tie is no different from the odds of a cross-group pair having a tie. The odds ratio for the table is 1.0 and the variables X (tie/no tie) and Y (same group/different group) are perfectly independent. In this network, ties are being formed at random with respect to group membership.

The reason the *E–I* index gets it wrong, of course, is because of varying group sizes. For example, in this particular case, it may be that the network is partitioned into a large group and a small group. If choosing ties at random, a node in the large group is likely to have most ties within group, but also most non-ties, which is not captured by EI. A solution is to replace the *E–I* index with a measure of association that is invariant with respect to group size. One example is the point-biserial correlation coefficient, which, following the notation in Table 6, is defined by Eq. (3).

$$\frac{ad - bc}{\sqrt{(a + c)(b + d)(a + b)(c + d)}} \tag{3}$$

When this measure is computed on the data in Table 7, the resulting coefficient is zero, indicating no tendency toward homophily or heterophily, once group sizes have been accounted for. Note that the measure achieves this by taking into account non-ties as well as ties. The actual size of each group is not taken into account directly.

The advantage of this approach – i.e., seeking measures of association that already have size invariance as a property – is that we can connect what we are doing with other bodies of work (e.g., in numerical taxonomy) that have dealt with similar issues (Sneath and Sokal, 1973). The advantage of the normalization approach, however, is that it gives us normalized versions of the internal and external tie counts, which can be useful variables in themselves. In addition, as long as the relative sizes of groups are known, the normalization approach can be used when non-ties are not available, as in true personal network or ego-network research designs.

5. Overlapping groups

When calculating an *E–I* statistic, we normally assume a partition of nodes that divides the nodes into a set of exhaustive, mutually exclusive classes. Typically, these are a priori partitions corresponding to categorical variables like gender, race, ethnicity, month of birth, country, etc. Of course, they can also be the results of a cohesive subgroup analysis, such as a partition obtained via a standard clustering algorithm (e.g., Ward, 1963). However, in many subgroup analyses, the output consists of a set of overlapping groups rather than partitions. Examples include cliques (Luce and Perry, 1949), *n*-cliques (Luce, 1950), and *k*-plexes (Seidman and Foster, 1978). This is not a problem for *E–I* measures, including the generalized measures we present here. The reason is that in calculating a score for any individual node, the *E–I* measure classifies all ties as in-group ties or out-group ties, effectively forming a temporary partition. If a group-level index is needed, the internal and external counts can be aggregated for that group, and an overall *E–I* score computed for the group. Only overall statistics for an entire network pose any kind of problem, as a different estimate

Please cite this article in press as: Everett, M.G., Borgatti, S.P., Categorical attribute based centrality: *E–I* and *G–F* centrality. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.06.002

needs to be calculated for each subgroup. However, the distribution of these *E–I* statistics could well of interest in themselves, and possibly diagnostic of some kind of social property.

6. Betweenness

Having extended *E–I* to several well-known measures of centrality, it is reasonable to consider extending *E–I* to betweenness. However, betweenness is fundamentally different from the other measures we have been considering. Whereas the previous measures were all radial centrality measures, betweenness is a medial measure (Borgatti and Everett, 2006). That is, it is a measure that summarizes a property of walks (specifically, shortest paths) that pass through (rather than originate or terminate at) a given node, on their way from each node in the network to every node in the network.

The most natural analog to our decompositions of radial measures would be to decompose betweenness centrality into the portions due to the groups of each *pair* of nodes in the network, rather than each node. For example, if the attribute of interest is gender, then it makes sense to decompose a node's betweenness into the contributions made by male–male pairs, male–female pairs, female–male pairs (if directed data), and female–female pairs. Table 8 shows the decomposition by gender combinations of betweenness scores for each node in the campnet dataset, which in this case has not been symmetrized in order to preserve richness. To construct this table we submitted the data to the UCINET (Borgatti et al., 2002) geodesic cube routine which is in the cohesion section of the menu. This routine calculates a 3-dimensional matrix $b(i,j,k)$ which gives the proportion of geodesics connecting j and k that pass through i. Submitting this matrix to the block image routine in which the image graphs are constructed by summing the entries with the gender partition yields the results displayed.

As an example of how to interpret the table, consider the first node, Holly. Her betweenness score is 78.33. This comes largely from being along the short paths that connect men to women (54), and women to men (18). Much more rarely (6.33), she dominates the shortest paths between men, and she is never along the shortest paths connecting women to women.

To create *E–I* scores, one approach is to consider to what extent a node's betweenness is due to joining two nodes that are members of the node's own group, and how much is due to joining others outside (either others with both others outside or one outside and one within the node's own group). In that case, the *E–I* score for Holly would be $(72 - 6.33)/(72 + 6.33) = 0.838$. The strong positive value is interpreted as saying that Holly tends to be a heterophilous connector, meaning that, whenever she is along the shortest path between two terminal nodes, it tends to be that one of the nodes is a "foreigner". Alternatively, we could consider the *E* in the *E–I* formula to refer to all cases where a node serves as a shortest path connector between pairs of nodes that are both members of the out-group (e.g., in Holly's case, men), and let *I* refer to all other cases. In that case, Holly's *E–I* score would be $(0 - 78.33)/(0 + 78.33) = -1.0$, which indicates that she never lies on a shortest path connecting "foreigners".

Another way to look at partitioning betweenness centrality is along the lines of Gould and Fernandez (1989) brokerage. They consider triads in which node A has a tie to node B, and B has a tie to node C, but A does not have a tie to C. In these triads, B is thought to be playing a structural role called a broker. If nodes in this network can belong to different groups (e.g., departments), then the broker may find themselves in a variety of different situations: brokering between nodes of its own department, brokering between its own department and another department, brokering between nodes that are both members of another department, and so on.

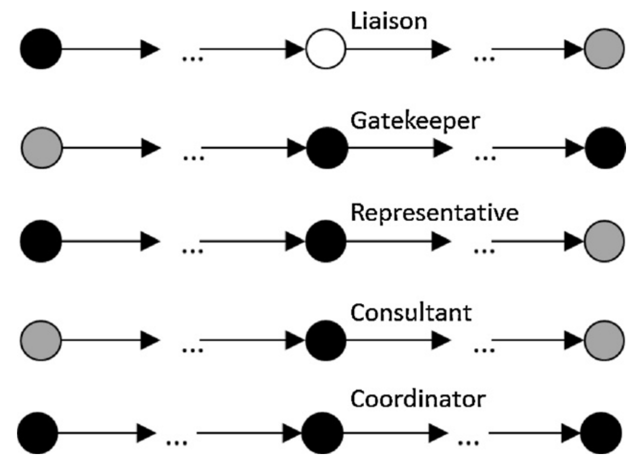

Fig. 1. Gould and Fernandez brokerage roles.

Gould and Fernandez regard these different possibilities as different sub-roles that the broker may play, and give them different names. They are as follows (see also Fig. 1): (1) Coordinator role, where A, B and C all belong to the same group; (2) Gatekeeper role, where A belongs to one group, and B and C belong to another; (3) Representative, where A and B belong to one group, and C belongs to another; (4) Consultant,[2] where A and C belong to one group, and B belongs to another; and (5) Liaison, where A, B and C each belong to a different group.

The brokerage structures of Gould and Fernandez can be extended to the case of betweenness. Instead of requiring B to be adjacent to A and C, we let B be along a shortest path between them and, for consistency with Freeman's betweenness (1979), assign a weight to the $A \rightarrow C$ connection equal to the number of shortest paths from A to B divided into the number of such paths that involve B.

Given a partition of the network we can therefore decompose the betweenness centrality into the various Gould and Fernandez roles. If b is the betweenness score of a given node we obtain

$$b = b^{\text{Coord}} + b^{\text{Gate}} + b^{\text{Rep}} + b^{\text{Con}} + b^{\text{Liaison}} \quad (4)$$

If we only have two groups then the last term is always zero and can be dropped.

As an example, we examine betweenness in a dataset collected from the Human Resources department of a health care company. The company had grown over the years via a series of mergers, and we can group the employees by which company they came from originally. Fig. 2 shows a network in which ties indicate "to whom do you go to for help in reformulating problems". Actors from the original firm are in white, actors from an older old acquisition are in gray and actors from the most recent acquisition are in black. The decomposition of betweenness into the various roles is given in Table 9. The column labeled total corresponds to the raw Freeman betweenness score. The other column labels are Coo = coordinator, Gat = gatekeeper, Rep = representative, Con = consultant and Lia = liaison.

It is clear from the overall betweenness scores that actors TD, KG, BD and PW have high betweenness scores and are therefore playing crucial roles within the network. For most actors, the coordinator and consultation roles are very low scoring and do not contribute much to the overall betweenness scores. The liaison roles are nearly exclusively taken by the original firm and TD and KG are playing a major role in liaising between the old acquisition and the new

[2] Gould and Fernandez referred to this as the Itinerant Broker role, but we find the UCINET term more evocative.

Table 8
Decomposition of betweenness by gender combinations for the directed campnet dataset. The sum of all values in each 2-by-2 matrix is equal to the Freeman betweenness of the node.

```
Matrix: HOLLY

                     1      2
                 Women    Men
                ------ ------

    1 Women      6.333 18.000
    2   Men     54.000  0.000

Matrix: BRAZEY

                     1      2
                 Women    Men
                ------ ------

    1 Women      0.000  0.000
    2   Men      0.000  0.000

Matrix: CAROL

                     1      2
                 Women    Men
                ------ ------

    1 Women      1.333  0.000
    2   Men      0.000  0.000

Matrix: PAM

                     1      2
                 Women    Men
                ------ ------

    1 Women      8.500  0.000
    2   Men     24.000  0.000

Matrix: PAT

                     1      2
                 Women    Men
                ------ ------

    1 Women     10.000 15.000
    2   Men     14.500  0.000

Matrix: JENNIE

                     1      2
                 Women    Men
                ------ ------

    1 Women      3.333  3.000
    2   Men      0.000  0.000

Matrix: PAULINE

                     1      2
                 Women    Men
                ------ ------

    1 Women      4.000  3.000
    2   Men      5.500  0.000

Matrix: ANN

                     1      2
                 Women    Men
                ------ ------

    1 Women      0.500  0.000
    2   Men      0.000  0.000

Matrix: MICHAEL

                     1      2
                 Women    Men
                ------ ------

    1 Women      7.000  2.000
    2   Men     37.833 12.000

Matrix: BILL

                     1      2
                 Women    Men
                ------ ------

    1 Women      0.000  0.000
    2   Men      0.000  0.000

Matrix: LEE

                     1      2
                 Women    Men
                ------ ------

    1 Women      0.000  0.000
    2   Men      5.000  0.000

Matrix: DON

                     1      2
                 Women    Men
                ------ ------

    1 Women      0.000 14.000
    2   Men      2.333  0.000

Matrix: JOHN

                     1      2
                 Women    Men
                ------ ------

    1 Women      0.000  0.000
    2   Men      0.000  0.000

Matrix: HARRY

                     1      2
                 Women    Men
                ------ ------

    1 Women      0.000  0.000
    2   Men      2.333  0.000

Matrix: GERY

                     1      2
                 Women    Men
                ------ ------

    1 Women      7.000  3.000
    2   Men     28.833 15.833

Matrix: STEVE

                     1      2
                 Women    Men
                ------ ------

    1 Women      3.500  2.500
    2   Men      5.667  5.167

Matrix: BERT

                     1      2
                 Women    Men
                ------ ------

    1 Women      3.500  2.500
    2   Men      4.333  3.333

Matrix: RUSS

                     1      2
                 Women    Men
                ------ ------

    1 Women      7.000  4.000
    2   Men     21.667 14.667
```

Table 9
G&F betweenness for the network in Fig. 2.

Node	Coo	Gat	Rep	Con	Lia	Tot
JB	1.50	1.58	7.31	0.00	3.85	14.25
TB	0.87	3.20	5.57	6.03	6.38	22.06
MC	0.00	3.67	0.37	0.00	0.00	4.03
CC	0.00	0.00	0.89	0.20	4.17	5.26
BD	0.00	1.44	3.70	0.20	0.20	5.54
TD	0.87	0.20	10.93	3.03	4.93	19.96
PD	0.00	4.65	1.08	2.33	2.99	11.05
JF	0.00	1.31	5.33	1.00	0.33	7.97
KG	0.87	4.22	7.18	2.03	3.74	18.03
SM	0.00	0.81	0.82	0.00	0.00	1.63
BS	0.00	2.02	0.73	0.33	1.84	4.93
AS	0.00	0.53	0.14	0.00	1.52	2.19
JT	0.00	0.00	0.00	0.00	0.00	0.00
PW	0.00	4.86	6.40	2.03	1.12	14.41
CW	0.20	6.57	0.73	0.50	3.43	11.43
TW	0.00	1.17	0.00	0.00	0.10	1.27
Tot	4.31	36.23	51.18	17.68	34.60	144.01

acquisition. We would note that TD's major role in keeping paths from blues to blacks short is probably not obvious to the actors themselves, and lengthening these paths would be an unintended consequence of TD leaving. We can see that TD is the only node that plays all five betweenness roles. BD acts as a representative for the new acquisition as does TD for the original firm. In contrast the dual role of gatekeeper is taken on by PW for the new acquisition and by KG for the original firm. It is interesting to note the corresponding roles of BD and TT and of PW and KG. No such roles are much in evidence from the old acquisition with just JB having a small gatekeeper role. It would appear that most of the members of the old acquisition have been marginalized.

Burt's (1992) theory of structural holes and Granovetter's (1973) theory of weak ties indicate that having connections to similar others has less benefit than having connections to different others. We may well expect in this case that liaison betweenness has more benefit than the other forms of betweenness. In this example the two highest betweenness actors also had the two highest

Please cite this article in press as: Everett, M.G., Borgatti, S.P., Categorical attribute based centrality: *E–I* and *G–F* centrality. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.06.002

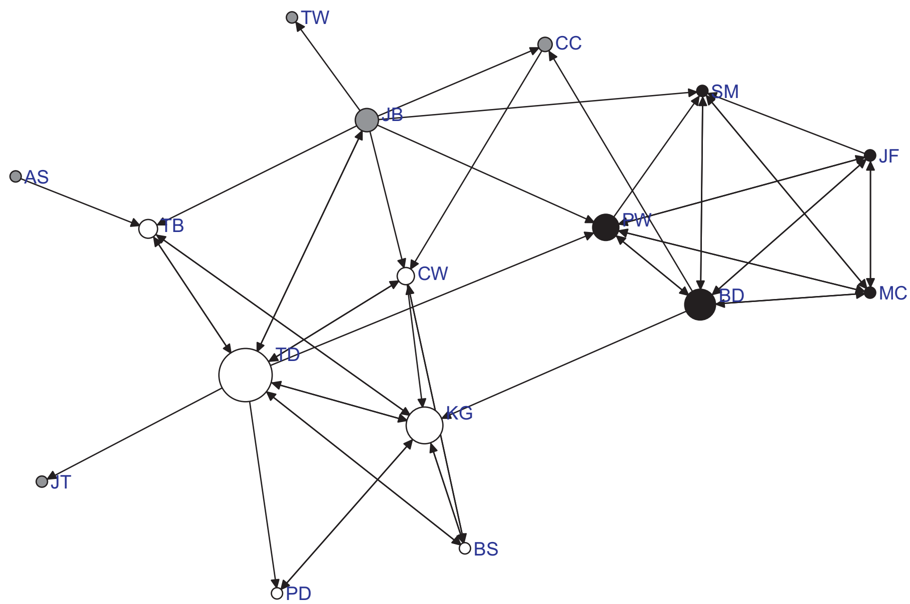

Fig. 2. Members of the HR department of a healthcare organization. Nodes colored by group.

betweenness liaison scores and so we could well expect they would reap the most benefit. But the third highest betweenness actor had a zero liaison betweenness score so may not be in such a good position as the raw overall betweenness suggests.

The above method can be applied to other medial measures such as flow betweenness or random walk betweenness.

The decomposition scores given in Table 9 suffer the same issues with respect to group sizes as the decompositions of radial centrality measures discussed earlier, and indeed we could normalize them in much the same way. In this case, however, we introduce a different approach, which is to compare the observed values with the expected values of a baseline model. For illustration, the baseline model we choose is one that takes the social structure and group sizes as givens and otherwise assumes that the assignment of nodes to groups is independent of network location. In this model, the generic formula for the expected value for any betweenness role can be written as

$$e_h^{[r]} = b_h p^{[r]} = b_h \frac{\sum_{i,k,j}^{G} n_i(n_j-1)(n_k-2)}{N(N-1)(N-2)} \quad (5)$$

where b_h is the betweenness score of node h, n_i is the size of group I, N is the number of nodes in the network as a whole, and the summation is over all G groups. Using the convention that i is the source group, j is the sink group and k is the broker, we can express the probabilities of each role by restricting the values that i, k and j take on in the summation. For the coordinator role we require $i=k=j$. For the gatekeeper role, $i \neq k=j$. For the representative role, $i=k \neq j$. for the consultant role, $i \neq k \neq j$ and $i=j$. For the liaison role, $i \neq k \neq j$ and $i \neq j$.

Table 10 gives the expected values for the healthcare dataset. If we compare the first row to the observed values in Table 9, we can see that actor JB plays the representative role twice as often as we would expect by chance, and plays the gatekeeper role just half as often as we would expect.

Table 10
Expected values for the betweenness decomposition.

Node	Coo	Gat	Rep	Con	Lia	Tot
JB	1.01	3.13	3.14	3.14	3.82	14.25
TB	1.57	4.85	4.86	4.86	5.91	22.06
MC	0.29	0.89	0.88	0.89	1.08	4.03
CC	0.37	1.16	1.15	1.16	1.42	5.26
BD	0.40	1.22	1.22	1.22	1.49	5.54
TD	1.42	4.39	4.40	4.40	5.35	19.96
PD	0.79	2.44	2.43	2.43	2.96	11.05
JF	0.57	1.76	1.76	1.75	2.14	7.97
KG	1.29	3.96	3.97	3.97	4.84	18.03
SM	0.12	0.36	0.36	0.36	0.44	1.63
BS	0.35	1.08	1.09	1.09	1.32	4.93
AS	0.16	0.48	0.48	0.48	0.59	2.19
JT	0.00	0.00	0.00	0.00	0.00	0.00
PW	1.03	3.18	3.17	3.18	3.85	14.41
CW	0.81	2.52	2.51	2.52	3.06	11.43
TW	0.09	0.28	0.28	0.28	0.34	1.27
	10.27	31.70	31.70	31.73	38.61	144.01

7. Conclusion

In this paper we have considered the decomposition of centrality scores based on the contributions of different groups of nodes. In doing this, we are able to show connections with two well-known metrics: the *E–I* homophily index, and the Gould and Fernandez brokerage metrics. Specifically, the *E–I* index can be seen as a measure based on partitioning degree centrality. We show that *E–I* can be generalized to other radial centrality measures such as closeness centrality and eigenvector centrality, creating a closeness *E–I* and an eigenvector EI, respectively. We also show that the Gould and Fernandez brokerage measures, which partition ego network brokerage into five types based on the group memberships of the nodes involved, can be generalized to the whole network case such that what is partitioned is a node's betweenness centrality score. This enables us to characterize the different ways that a node might lie along the shortest paths between two others.

References

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2, 113–120.
Bonacich, P., 1987. Power and centrality: a family of measures. American Journal of Sociology 92, 1170–1182.
Borgatti, S.P., 2005. Centrality and network flow. Social Networks 27 (1), 55–71.
Borgatti, S.P., Everett, M.G., 2006. A graph–theoretic perspective on centrality. Social Networks 28 (4), 466–484.
Borgatti, S.P., Everett, M.G., Freeman, L.C., 2002. Ucinet for Windows: Software for Social Network Analysis. Analytic Technologies, Harvard, MA.

Please cite this article in press as: Everett, M.G., Borgatti, S.P., Categorical attribute based centrality: *E–I* and *G–F* centrality. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.06.002

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge.

Everett, M.G., Borgatti, S.P., 2010. Induced endogenous and exogenous centrality. Social Networks 32 (4), 339–344.

Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. Social Networks 1, 215–239.

Gould, J., Fernandez, J., 1989. Structures of mediation: a formal approach to brokerage in transaction networks. Sociological Methodology, 89–126.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Hubbell, C.H., 1965. An input–output approach to clique identification. Sociometry 28, 377–399.

Ibarra, H., 1992. Homophily and differential returns: sex differences in network structure and access in an advertising firm. Administrative Science Quarterly 37 (3), 422–447.

Katz, L., 1953. A new status index derived from sociometric data analysis. Psychome 18, 34–43.

Krackhardt, D., Stern, R.N., 1988. Informal networks and organizational crises: an experimental simulation. Social Psychology Quarterly 51 (2), 123–140.

Luce, R., 1950. Connectivity and generalized n-cliques in sociometric group structure. Psychometrika 15, 169–190.

Luce, R., Perry, A., 1949. A method of matrix analysis of group structure. Psychometrika 14, 95–116.

McCrae, R.R., John, O.P., 1992. An introduction to the Five-Factor Model and its applications. Journal of Personality 60, 175–215.

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

Sneath, P.H.A., Sokal, R.R., 1973. Numerical Taxonomy. Freeman, San Francisco.

Valente, T.W., Foreman, R.K., 1998. Integration and radiality: measuring the extent of an individual's connectedness and reachability in a network. Social Networks 20, 89–109.

Ward Jr., J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 48, 236–244.

Please cite this article in press as: Everett, M.G., Borgatti, S.P., Categorical attribute based centrality: *E–I* and *G–F* centrality. Soc. Netw. (2012), http://dx.doi.org/10.1016/j.socnet.2012.06.002