Ego network betweenness

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

[a] *University of Westminster, 35 Marylebone Road, London NW1 5LS, UK*
[b] *Department of Organization Studies, Carroll Graduate School of Management,
Boston College, Chestnut Hill, MA 02467, USA*

Abstract

In this paper, we look at the betweenness centrality of ego in an ego network. We discuss the issue of normalization and develop an efficient and simple algorithm for calculating the betweenness score. We then examine the relationship between the ego betweenness and the betweenness of the actor in the whole network. Whereas, we can show that there is no theoretical link between the two we undertake a simulation study, which indicates that the local ego betweenness is highly correlated with the betweenness of the actor in the complete network.
© 2004 Published by Elsevier B.V.

1. Introduction

Many network studies have been undertaken using ego networks. In this paper, we define ego networks to be networks consisting of a single actor (ego) together with the actors they are connected to (alters) and all the links among those alters. These networks are also known as the neighbourhood networks or first order neighbourhoods of ego. The attraction of ego networks is the ease of collection of data compared with collecting data on whole networks. Information on the alters, including how they are connected, is usually obtained entirely from ego. Such structures can be sampled from large populations and can be used to make (usually non-network) statistically significant conclusions about the whole population. There are many areas in which such networks have been studied empirically, for example social support.

---

* Corresponding author.
  E-mail addresses: M.Everett@wmin.ac.uk (M. Everett), borgatts@bc.edu (S.P. Borgatti).

Ego networks have a constrained and simple structure which delivers the benefit of simplicity in data collection but results in the researcher having fewer powerful tools with which to analyze the data. Techniques that have been used to analyze ego networks are usually based on density, connectivity (e.g. *Structural Holes*; Burt, 1992) or the attributes of the alters (Homophily) or combinations of all three. There has also been a great deal of attention to relating the information derived from the ego networks to properties of the whole network. This seems to be a difficult problem although some progress has been made (Frank and Snijders, 1994).

Until relatively recently, empirical work on networks was confined to relatively small networks. The majority of network studies still consist of less than 500 actors, however, this is changing. The advent of the internet and electronic communication monitoring tools has opened the door to collecting data on significantly larger networks. However, it is apparent that many of the tools developed for analyzing networks are not scalable. Closeness centrality on say 10,000 actors is simply not meaningful. As networks become larger computational issues also begin to dominate. Modern computers and algorithmic techniques have greatly improved the size of networks that can be analyzed, but it is still not computationally feasible to do apply a variety of network methods to larger networks. In such cases it would seem logical to look at local networks rather than the whole network.

2. Centrality in ego networks

We shall be concerned with the centrality of ego within an ego network. This concept was first introduced by Freeman (1982) where he looked at betweenness of ego. His paper looked at some of the mathematical properties of the measure in relation to the density of the ego network and showed how the measure could be used in a practical case study. We take a rather different approach here. One of the main objectives of measuring centrality is the need to identify the most important actors within a network. To this end, we shall assume that the ego networks are drawn from the same basic network and the objective is to give a network measure of importance to enable us to compare the centralities of the egos in our sample.

Clearly, since degree centrality is a local property and the ego degree centrality of ego is the same as the degree of the actor in the whole network there is no issue. At the other extreme closeness is about the connections from an actor to all other actors in the network and so is simply not applicable to ego networks. Eigenvector centrality is similarly problematic. The power of eigenvector centrality is that it takes account of the connectivity of the alters, but these are unknown outside of the ego network and so this undermines the fundamental principles behind the measure. Betweenness (Freeman, 1979) examines the extent to which an actor is between all other actors within the network. If an actor is between two other actors then it follows that there is not a connection between the alters on the path connecting them (otherwise this would form a shorter path). There is, therefore, a connection between the betweenness of the actor in the whole network and the betweenness of the actor in the ego network, although it may not be possible to quantify this association. It follows that it may well be useful to calculate the betweenness centrality of ego in ego networks.

It should be noted that the usefulness of a betweenness type measure for ego networks is not new. Burt (1992), in his book *Structural Holes*, provides ample evidence that having

high betweenness centrality, which is highly correlated with having many structural holes, can bring benefits to ego.

In whole network studies, it is common practice to normalize centrality scores. The main reason for this is to allow the researcher to compare centrality scores across different networks. Whilst this may be desirable for ego networks it does pose a few additional problems. Firstly, should the normalization be done using the size of the ego network or the whole network. If the former, then it does not allow comparisons with different egos drawn from the same network since they would have been normalized differently. If it is the latter then we have a number of difficulties. Firstly, the size of the population maybe unknown, secondly, if it is known it is highly likely that ego is drawn from a large population. Normalization assumes that it is possible to obtain the maximum score. Whilst it is conceivable (but highly unlikely) that an actor has say 10,000 other actors as acquaintances, it would be unreasonable to expect in a sexual contact network that an individual could have so many partners! It would, therefore, seem that the best course of action is not too normalize. This could still allow comparisons of ego networks drawn from different populations. Provided they had the opportunity to connect to similar numbers of alters within their networks. There is an additional reason why we recommend against normalization with respect to ego network size. The primary purpose of measuring centrality is to decide which actors are the most important. The larger the number of alters the higher the probability that they are between actors outside of their ego network. Normalization would counter this effect, one of the crucial pieces of data we have on ego is the size of the network, in this case size matters and so we should retain this information.

3. Computing ego network betweenness

The structure of ego networks mean it is very easy to compute the ego betweenness. First, of course, we only have to compute the betweenness of a single actor. In addition, geodesics in the network are either of length 1 or 2. Every pair of non-adjacent alters must have a geodesic of length 2 which passes through ego. We need only consider these geodesics as geodesics of length 1 do not contribute towards betweenness. If $\mathbf{A}$ is the adjacency matrix for a graph G then $\mathbf{A}_{i,j}^2$ contains the number of walks of length 2 connecting i and j. When $i \neq j$, since the walk is of length 2, it must be a path. We need to count the number of paths of length 2 for non-adjacent pairs of actors since these will be geodesics. It follows that $\mathbf{A}^2[\mathbf{1} - \mathbf{A}]_{i,j}$, where $\mathbf{1}$ is a matrix of all 1's, gives the number of geodesics of length 2 joining i to j. The sum of the reciprocal of the entries gives the ego betweenness (this has to be halved if it is a graph).

This can be implemented efficiently by noting from the outset which entries in $\mathbf{A}^2[\mathbf{1} - \mathbf{A}]_{i,j}$ will be non zero.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Since the matrix is symmetric we need only consider the zero entries above the leading diagonal and calculate $\mathbf{A}^2[\mathbf{1} - \mathbf{A}]_{i,j}$ for those entries. This gives

$$\mathbf{A}^2[\mathbf{1} - \mathbf{A}] = \begin{bmatrix} * & * & * & * & * \\ * & * & * & 2 & 1 \\ * & * & * & * & 1 \\ * & * & * & * & 1 \\ * & * & * & * & * \end{bmatrix}$$

The betweenness of ego is simply the sum of the reciprocals of the entries, that is 3.5.

The calculation of all the ego betweenness scores for a whole network would be one order of magnitude faster than calculating the real betweenness scores.

4. Betweenness and ego betweenness

It would be nice if there were a direct link connecting betweenness and ego betweenness. That is given an ego betweenness we could get some bounds on the possible value of the betweenness scores. Unfortunately, this is not possible in general. It is easy to see that ego betweenness is zero if and only if betweenness is zero but beyond this the relationship is much more complicated. We can construct networks in which the differences in these scores is arbitrarily large. For example if we take a star with $n + 1$ vertices, where ego is the centre of the star. We then take a complete graph with p vertices, K_p and connect every actor except ego in the star to K_p. Regardless of the size of p the ego betweenness will be $\frac{1}{2}n(n - 1)$, the betweenness will depend on p and as p increases the betweenness score will decrease. As a consequence of this we can increase the ego betweenness to be as large as we like by increasing n and decrease the betweenness score by increasing p. Hence, we can make the difference in these to be arbitrarily large.

Alternatively, we can also make the betweenness score arbitrarily larger than the ego betweenness score. To do this, take a star with $p + 1$ vertices and make one of the pendant vertices a member of a complete graph K_n. Make the pendant vertex ego. The ego betweenness will now be n and the betweenness score will be np.

These two examples show that we cannot get simple bounds on the values of the betweenness. However, these examples are constructed deliberately to show that there is no formal connection between the two concepts. It may happen that such cases are rare, particularly in real data. We, therefore, wish to investigate whether there is an empirical connection between ego betweenness and actual betweenness. Since we are unlikely to be able to do this mathematically or statistically, then the only avenue open is to examine this by underrating a simulation.

The general approach is simple. Generate some random networks, calculate the betweenness centrality of every actor. From the same network construct the ego network of every actor and from this calculate the ego betweenness of every actor. Compare the two results over all the generated networks.

There are a couple of important issues which need to be addressed before any firm conclusions can be drawn from this process. First we need to decide what we mean by a random network. Clearly, the idea would be that from all the possible social networks that can occur we randomly selected a sample. This is of course simply not possible since we do not have any idea of the distribution of all possible networks. As a first step, we will simply examine Bernoulli networks. That is networks in which the presence or absence of an edge between two actors is decided by a fixed prespecified probability. Such networks do not have the clumpiness we associate with many real networks. That is such networks do not tend to have contrasting regions of highly cohesive subsets and regions of low density within the same network. However, if using simple networks yields negative results then it would be even more likely that more complicated structures would be additionally problematic. We, therefore, see this simple simulation as a first step in understanding the problem.

Second, we have to compare the results. As a first step, we have chosen correlation. This is because we are not trying to predict the actual centrality score of each actor but to determine which actors are more central than others. So that for each randomly generated network we have correlated the betweenness centrality scores with the ego betweenness scores for each actor in the network. Since we repeat this a number of times we obtain a set of correlations. We then calculate the average correlation, the range and standard deviation.

5. Results and conclusions

We performed the above simulations on networks with 25, 50, 100, 200 and 500 actors. We choose the six different values for the probability of a tie namely 0.1–0.6 increasing in steps of 0.1. For each network and each probability we ran the simulation 500 times. The results were as follows.

	25	50	100	200	500
0.6	0.860	0.892	0.931	0.959	0.981
	(0.086)	(0.038)	(0.014)	(0.005)	(0.001)
	[0.534–0.978]	[0.701–0.965]	[0.864–0.962]	[0.942–0.973]	[0.978–0.985]
0.5	0.868	0.854	0.905	0.943	0.974
	(0.070)	(0.052)	(0.020)	(0.007)	(0.002)
	[0.573–0.977]	[0.561–0.946]	[0.835–0.954]	[0.917–0.962]	[0.967–0.979]
0.4	0.921	0.860	0.859	0.914	0.960
	(0.037)	(0.045)	(0.030)	(0.012)	(0.003)
	[0.791–0.984]	[0.663–0.962]	[0.737–0.932]	[0.868–0.942]	[0.949–0.970]
0.3	0.954	0.934	0.880	0.869	0.931
	(0.031)	(0.020)	(0.028)	(0.019)	(0.006)
	[0.739–0.991]	[0.843–0.981]	[0.777–0.939]	[0.790–0.927]	[0.908–0.947]

(*Continued*)

	25	50	100	200	500
0.2	0.930	0.967	0.964	0.936	0.891
	(0.050)	(0.015)	(0.007)	(0.009)	(0.010)
	[0.639–0.994]	[0.865–0.990]	[0.934–0.982]	[0.898–0.954]	[0.861–0.919]
0.1	0.869	0.937	0.975	0.984	0.985
	(0.078)	(0.031)	(0.009)	(0.003)	(0.001)
	[0.468–0.985]	[0.737–0.984]	[0.935–0.991]	[0.971–0.990]	[0.980–0.988]

The values in each cell in the table contain the average correlation with the standard deviation underneath in parenthesis, the range of the correlations is given in brackets at the bottom. The reported average correlations are all very high. The lowest correlation found over all the simulations was 0.468 but this occurred in the smallest network of just 25 actors at the lowest density and in this category the average was 0.869 and with a standard deviation of 0.078. This means that the 0.468 is over five standard deviations away from the mean and is, therefore, very unusual. It should be noted that for the smaller networks with probabilities of 0.1 and 0.2 a lot of the actual betweenness scores are zero and this means care needs to be taken in interpreting some of these results. However, as can be seen by the values reported that for these graphs there is a very strong relationship between ego betweenness and the whole network betweenness scores. This in itself is a significant result but clearly more work needs to be done. We need to examine structures which more commonly arise in real network studies and also give further consideration to the amount of agreement at the individual actor level.

There is a general pattern for the higher density networks, that is those with a density of 0.3 or higher. The correlations start high and as the size increases the correlation decreases to a minimum and then increases again. It is quite possible that this pattern is also present at the two lower values but that we simply have not examined graphs sufficiently large to see it. A preliminary examination of the betweenness scores of the networks reveals that the small networks, particularly those with the lower densities, have a significant number of actors with zero or very low betweenness scores. As size increases it becomes (relatively) unusual for actors to have zero or very low betweenness scores. As already mentioned, if we have a zero score than both betweenness and ego betweenness must agree and, hence, these will give us high levels of association. This fact explains the general trend that the scores start high because of the large number of zero betweenness scores. As the size increases the number of zeros reduces and as a consequence the correlations start to fall. As size continues to increase the edges become more evenly distributed and there is less variation in betweenness scores. In this case the betweenness scores are well approximated by the ego betweenness scores and we see the correlations start to increase.

The smaller sized graphs at the lower densities have larger ranges and consequently they are more unstable and difficult to interpret. One possible explanation for these results is that the small networks with probabilities 0.1 and 0.2 not only have a lot of actors with zero

betweenness but also have the potential for constructions in which the betweenness and ego-betweenness differ greatly as described in Section 4. The occurrence of these is rare but they do have a significant effect. As the size and density increase these odd structures become rarer and so do not occur in the simulations. However, this also requires further investigation.

As the densities increase for fixed sizes we also see a similar decrease followed by an increase. The explanation here is similar to the one above with one minor modification. It is true at the lower densities there are many actors with zero betweenness and as the density increases these become rarer. However, as the increase continues they are re-introduced. This follows from the fact that in a complete graph every actor has a betweenness score of zero.

We can, therefore, summarise the above by noting that ego-betweenness gives a good approximation to betweenness in two situations. Firstly when all the actors have very similar betweenness scores and secondly, when there are highly differentiated scores. This is a very important conclusion. In looking at real data network analysts have tended to find structures with highly differentiated betweenness scores. This is highlighted by the fact that betweenness has been one of the most successful measures of centrality. Classic structures such as core-periphery, scale free and networks in which there are highly separated cohesive subsets all fall into the category of networks with highly differentiated betweenness scores. It may well be that the ego-betweenness measure will perform better on the majority of real rather clumpy data than on the uniform random networks tested above.

To get a preliminary feel for what happens in real data we compared the betweenness scores and ego betweenness scores on the binary symmetric data in UCINET (Borgatti et al., 2002). There are seven separate datasets which contain symmetric binary data. Some of these data sets have more than one symmetric binary relation. However, when these were analyzed separately they all gave very similar results so we only report one representative data set here. The following table gives the correlations together with the size and density of the network.

Network	Size	Density	Correlation
TARO	22	0.17	0.927
PADGETT (PADGB)	16	0.13	0.959
ZACKAR (ZACHE)	34	0.14	0.954
SZCID	16	0.69	0.979
KAPTAIL (KAPTS1)	39	0.21	0.957
KAPMINE (KAPFMM)	15	0.18	0.884
WIRING (RDGAM)	14	0.31	0.996
THUROFF (THURM)	15	0.31	0.968

These results seem to confirm our suspicions. Clearly, we need to do more extensive testing of this hypothesis but at this stage it seems reasonable to conclude that these is

strong evidence that ego betweenness is a good measure for betweenness on many real datasets.

References

Borgatti, S.P., Everett, M.G., Freeman, L.C., 2002. UCINET for Windows: Software for Social Network Analysis. Analytic Technologies, Harvard.

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

Frank, O., Snijders, T.A.B., 1994. Estimating hidden populations using snowball sampling. Journal of Official Statistics 10, 53–67.

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

Freeman, L.C., 1982. Centered graphs and the construction of ego networks. Mathematical Social Sciences 3, 291–304.