Centrality

Introduction

Who is important in a group? Some people are important because of various personal attributes such as knowledge, drive or looks, or because of official title and responsibilities, such as President or CEO. However, individuals can also be important by virtue of the position they hold in a social network – i.e., their centrality.

David Krackhardt () tells the story of a unionization drive in a Silicon Valley firm. The union came the organization and gave a presentation. There was definitely interest on the part of many of the workers, and one worker in particular was extremely excited about unionizing. The union made him their point man for the subsequent campaign. Eventually, there was a vote and the union lost by small margin. A look at the network of friendships among the workers (Figure xx) reveals something interesting about Hal, the union's point man. Hal is really quite peripheral in the network. In contrast, Chris, whom the union never approached, was the informal leader of the employees – he was highly central in the friendship network. Chris was pro-union, but also very friendly with management, who were against unionization. When the election came, he abstained. But had he been lobbied by the union, who could not only have voted for them, but persuaded many others.

Centrality is a node property.[1] Some people think of it in terms of the structural importance or prominence of the node. Others see it as an indicator of how well-connected a node is – i.e., the extent to which its ties can potentially provide access to resources controlled by others in the network. Alternatively, one can see it in terms of how easily a node might influence the rest of the network. Many measures of centrality have been proposed in the literature, each measuring a different aspect of centrality. In this chapter we consider just a few of the most often used measures.

For simplicity of exposition, the various measures are described in the context of undirected (i.e., symmetric) networks. Then, at the end of the chapter, there is a section on centrality in directed networks. Also, it should be noted that centrality is normally

[1] Actually, Everett and Borgatti (19xx) have extended centrality to apply to groups of nodes in addition to individual nodes. But this topic is beyond the scope of this book.

computed with respect to a single relation – if a dataset contains multiple relations for the same set of nodes, a separate set of centrality scores is computed for each relation.

Degree Centrality

Perhaps the simplest measure of centrality is degree, which is simply the number of ties that a node has. Degree centrality is an index of the visibility or exposure of a node in the network – i.e., the “risk” of receiving whatever is flowing through the network (whether it is information in a gossip network or an infection in a sexual network). In addition nodes with high degree in a relation such as trust are in a position to wield influence, since they are trusted by many nodes.

Computing degree is straightforward using software packages like UCINET. As an example, we will use the CAMPNET dataset that comes with UCINET. These data are directed, so for this analysis, we symmetrize it first, using the maximum method. This means that we consider a tie to present between nodes A and B if either A→B or B→A. We call the symmetrized dataset SYMCAMPNET.

As shown in Output x.1, the result of calculating degree is a column vector containing the centrality score for each node. In addition, UCINET calculates a normalized version of raw degree which is just a node’s degree divided by n-1 (the maximum possible) and multiplied by 100 (so that the numbers scale between 0 and 100). These values are shown in the column to the right of simple degree in Output x.1.

Having computed degree, one would typically add it to a node-level database that contains other variables measured on the same nodes, such as gender, organizational rank, race, etc. We can then use conventional statistics to relate centrality to these other variables. For example, we might use a t-test to compare the degree centrality of men and women in an organization.

We can run such a test in UCINET using the tools>statistics>vector>t-test procedure. First we run degree centrality on SYMCAMPNET, resulting in Output x.1, and then we run T-Test. The inputs to this procedure include both the centrality of each node and the gender of each node. It then computes the mean centrality for each gender and calculates the significance of the difference, using a permutation test. The result is shown in Output x.2.

As mentioned earlier, degree centrality can be viewed as a kind of visibility. Nodes with high degree in an organizational network will tend to be the same ones that insiders will list as the important people in the group. An advantage of degree centrality is that it is basically interpretable in all kinds of networks, including disconnected networks. A disadvantage of degree centrality is that it is a relatively coarse measure of centrality. For example if a node is connected to 5 others that would be isolates if not for the tie to the focal node, the centrality of this node is no different from the centrality of a node that is connected to 5 others that well-connected themselves and in the center of a network.

Eigenvector Centrality

Eigenvector centrality can be described from a number of different perspectives (Bonacich, 1972). We present here is a kind of degree centrality in which we count the number of nodes adjacent to a given node (just like degree centrality), but weight each adjacent node by its centrality. The result is that each node's centrality is proportional to the sum of centralities of the nodes it is adjacent to – in effect, a node is only as good as its network.

We can interpret eigenvector centrality as a measure of popularity, in the sense that a node with high eigenvector centrality is connected to nodes who are themselves well-connected. This contrasts with a node that might have many ties but they are to people who have no other ties. We can also view eigenvector centrality as a more sophisticated measure or risk. For example, consider the network of sexual ties in Figure x.1. Nodes A and B both have degree 1. But they don't have the same level of risk because the node that B is having sex with is having sex with many others. Eigenvector centrality captures this difference and assigns B a higher score.

Like degree, there is a normalized version of eigenvector centrality which divides the raw eigenvector score by the maximum possible score achievable in a network of the same size, and then multiplies by 100.

If we use UCINET to calculate eigenvector centrality on the SYMCAMPNET dataset, we obtain the result shown in Output x.3. Note that, in this case, there is considerable agreement between degree centrality and eigenvector centrality (r = 0.xxx), which is not unusual. However, some nodes do show differences in relative centrality, as shown in the scatter diagram given in Figure x.1.

An important limitation of eigenvector centrality is that it should not be used in disconnected networks as it will assign zeros to all members of the smaller components.[2] Furthermore, if a network has a bowtie structure such as shown in Figure x.2, the scores for all the nodes in the smaller subgroup will have uniformly lower scores than the nodes in the larger subgroup. This is not precisely a flaw since in fact the nodes in the smaller group are connected to nodes that are less well connected, but is something one might want to take account of, particularly in the case where the groups correspond to, say, organizational subunits and the size of the subunits is determined by a variable extraneous to the processes being researched.

Closeness Centrality

[2] The term "smaller" refers to both the number of nodes and the number of ties. In other words, if one component has fewer nodes, all of its members get zeros. If the components have the same number of nodes, the component with fewer ties gets zeros.

Closeness centrality is defined as the total number of links separating a node from all others along the shortest possible paths. In other words, to calculate closeness, one begins by calculating, for each pair of nodes in the network, the length of the shortest path from one to the other (aka the geodesic distance). Then for each node, one sums up the total distance from the node to all other nodes.

Closeness can be interpreted as an index of time-until-arrival of something flowing through the network. The greater the raw closeness score, the greater the time it takes on average for information originating at random points in the network to arrive at the node. Equally, one can interpret closeness as the potential ability of a node to reach all other nodes as quickly as possible.

It is important to note that raw closeness is an inverse measure of centrality in that it is nodes with smaller scores that the most central (they are the least distant from other nodes). In fact, the UCINET programs labels raw closeness "farness" to remind the user to interpret the values correctly.

Like degree and eigenvector centrality, there is a normalized version of closeness, and the normalized version reverses the values so that a larger number means that a node is more central. Specifically, the normalized version divides a node's closeness score into n-1, and then multiplies by 100. Hence, a node that is adjacent to every other (such as the center of a star) will have a score of 100).

If we calculate closeness centrality on the SYMCAMPNET dataset, we obtain the result shown in Output x.3. Note that the most central nodes (xx, yy) are not the same as obtained when running degree centrality.

An important fact to note about closeness centrality is that it is inappropriate for disconnected networks (in which the distances between some pairs of nodes is undefined), and it is rarely useful in directed networks, but this is discussed more fully at the end of this chapter. It has also been frequently noticed that closeness centrality often exhibits little variance, which means it doesn't strongly distinguish between the most central and least central nodes, and can fail to correlate highly with other variables.

Betweenness Centrality

Betweenness centrality measures how often a given node falls along the shortest path between two other nodes. More specifically, it is calculated for a given focal node by computing, for each pair of nodes other than the focal node, what proportion of all the shortest paths from one to the other pass through the focal node. These proportions are summed across all pairs and the result is a single value for each node in the network.

Betweenness is zero for a given node when it never along the shortest path between any two others. It reaches its maximum value when the node lies along every shortest path between every pair of nodes.

Betweenness is typically interpreted in terms of the potential for controlling flows through the network – i.e., playing a gate-keeping or toll-taking role. In a sense, nodes with high betweenness are in a position to threaten the network with disruption of operations.

Programs like UCINET compute both raw betweenness and a normalized version that divides raw betweenness by the maximum score possible in a network of that size, and then multiplies by 100. The center of a star-shaped network will have a normalized betweenness score of 100. If we calculate betweenness centrality on the SYMCAMPNET dataset, we obtain the result shown in Output x.3. Note that …

It is useful to note that, in general, the variance of betweenness is quite high, providing effective discrimination between nodes and potentially correlating well with other variables.

It is also worth realizing that the ability to exploit a high betweenness position varies with the ease with which nodes can create ties. For example, suppose that a given node has high betweenness, meaning that many nodes need that node to reach other nodes via efficient paths. In principle, this node has power because it can threaten to stop transmitting, making nodes use less efficient paths to reach one another. But this threat only works if nodes cannot easily create new ties to go around the recalcitrant node.

An excellent example is provided by the study of medieval Russian trade networks studied by Forrest Pitts (1979). He notes that in the 12th century, Moscow was just another principality identical in all respects to hundreds of others. Soon, however, it began to grow, outstripping the other principalities in the region. The question is why – was it good leadership or something more structural? Pitts notes that every principality was located on a river, which was used for trade. The rivers constitute highly durable and difficult-to-create ties in a network of principalities. In this network, Moscow turned out to have the highest betweenness centrality. It was therefore in an excellent position to make demands (e.g., exact tolls) on the traders. Since the traders could not easily create new ties (e.g., redirect rivers), Moscow could effectively enforce its demands.

Directed Networks

To optimally use the centrality measures discussed above with directed data, we modify most of them slightly.

For degree centrality, we split the concept of degree into two separate measures: in-degree and out-degree. In-degree counts the number of incoming ties (arcs) whereas out-degree counts the number of outgoing ties. Another way to look at it is that in-degree consists of the column sums of the adjacency matrix, and out-degree is the row sums of the matrix. As a check on the computation, it is useful to know that whereas any given node can have more (or less) in-degree than out-degree, the average in-degree for all nodes must, by mathematical necessity, equal the average out-degree. Depending on the social relation in question, we often interpret out-degree as the “gregariousness” of the node and the in-degree as the “prestige” of the node. It is not unusual to regard node gregariousness with some suspicion as it may reflect differences in interpretation by different respondents (e.g., one respondent views everyone as “friend” while another respondent reserves the term for only the closest ties).

For eigenvector centrality, there are a couple of useful approaches for handling directed data. First of all, the mathematical notion of eigenvectors does apply to non-symmetric adjacency matrices. However, for many directed networks, the solutions require complex numbers (the kind that contain both a real and an imaginary component). One class of networks which are guaranteed to have real-valued solutions is the set of strongly connected networks (these are directed networks in which every node can reach every other node by some directed path). So for these networks we can compute two eigenvectors: a right eigenvector (similar to out-degree) and a left eigenvector (similar to in-degree). If the relation being measured is who gives advice to whom, the right eigenvector can be seen as a measure of potential to influence others via both direct and indirect ties. A node has a high score on the right eigenvector if it has many ties to nodes that themselves have many ties to others (who have many ties … etc). The left eigenvector can be seen as a measure of risk of being influenced (or infected) by others. A node has a high score if it receives ties from many nodes who themselves receive many ties from nodes who receive … and so on.

As an example, consider the data presented by Casciaro () on investing relations among Italian banks. The data consist of a matrix X in which xij indicates the percentage of bank j’s shares that are owned by bank i. The network is shown in figure xx. Table

A closely related approach is the method of Katz (), which Burt () called prominence. In this method, we start by dividing each value in the adjacency matrix by the row sum, so that after normalization each row adds to 1. The purpose of this is strictly technical: it provides a mathematical guarantee that the iterative process we describe next will converge on the desired solution. After some additional transformations of the matrix, the following iterative process is begun: First, we sum the columns, just as if we were computing in-degree. This produces our first order estimate of the prominence of each node. Then we recalculate the sum, but this time we weight each value by the first-order prominence scores. This gives a second-order set of estimates in which a high score indicates that a node is chosen by (many) people who are themselves often chosen. This process is repeated until the relative proportions of scores stop changing. The result is a more nuanced version of in-degree centrality in which a node is prominent if it is chosen by many prominent nodes.

The one problem with this method is that the normalization of rows does change the data in important ways. The ties of gregarious nodes will be weighted downward, while the ties of nodes with little out-degree will be weighted upward. This is reasonable only when having a lot of ties implies giving each one less attention.

A different approach that does not require normalization of the rows is that of hubs and authorities. Hubs are nodes that send ties to nodes that have lots of incoming ties. Authorities are nodes that receive ties from nodes that have lots of outgoing ties. In this approach, each node receives two scores: one that reflects how much of a hub it is, and one that indicates how much of an authority it is. The scores are related in that the a node's hub score is proportional to the sum of the authority scores of the nodes it sends ties to, and a node's authority score is proportional to the sum of the hub scores of the nodes that point to it.

For closeness centrality, we adopt the same strategy and split the concept into in-closeness and out-closeness. The in-closeness variable measures the extent to which a node is easily reached by others – i.e., there are short directed paths from others to the node. The out-closeness variable measures the extent to which a node can easily reach others – i.e., the paths going from the node to all others are relatively short.