Cohesion

The graph-theoretic terms discussed in the previous chapter have very specific and concrete meanings which are highly shared across the field of graph theory and other fields like social network analysis that use graph theory. For those terms, the conceptual idea and its measurement are one and the same thing. For example we defined degree as a certain quantity that is calculated in a certain way. In contrast, in this section we consider the more abstract social science concept of cohesion, which can be measured in a variety of different ways, none of which can be said to be "the" way. In general, measures of cohesion are built upon the more fundamental concepts described in the previous section.

The idea of cohesion is connectedness or "knittedness". Most people think about cohesion as a group or network level property. Some groups are more cohesive than others. There is a Spanish word – "enrededado" – that expresses it nicely. It means mixed up together, like a big clump of electrical wires. It is particularly appropriate because the word is based on the word for network, which is "red". A measure of group cohesion, then, is a single value that characterizes the stuck-togetherness of the group.

But we can also talk about dyadic or relational cohesion. This refers to the attraction or closeness of pairs of nodes, some of whom will be very cohesive while others are quite distant or uninvolved. Actually, the idea can be applied to any level of analysis. For example, at the node level, we can refer to the extent that a node is connected with all other nodes in the network as the node's cohesion with the rest of the group. In fact, it is key thesis of this book that one of the main types of centrality is nothing more than node-level cohesion.

Because group-level cohesion is more familiar and easier to understand, it would make sense to discuss that first. However, since many of the group-level measures are built on dyadic level measures, it is more logical and convenient to discuss the dyadic level measures first.

Relational Cohesion

The simplest, most fundamental measure of relational cohesion is adjacency itself. If you and I have a tie (let's say a trust tie) then we are more cohesive than if we didn't have a tie. Of course, we have to be careful to think about what kind of relation is being measured. But even a conflict or dislike tie is a measure of dyadic cohesion – it is just that it is an inverse measure. If the data consist of valued ties (e.g., strengths or frequencies), so much the better, because then we have degrees of cohesion instead of simple presence or absence.

It's useful to note that some nodes that are not adjacent may still be indirectly related. All nodes that belong to the same component are far more cohesive than a pair of nodes that are on separate components. If a virus is spreading in one component, it will eventually reach every node in the component – but it cannot jump to another component. This

suggests another measure of dyadic cohesion, namely reachability. Two nodes are reachable if there exists a path – no matter how long – from one to the other. We typically represent this as a matrix R in which $r_{ij} = 1$ if i can reach j by some path, and $r_{ij} = 0$ if no such path exists (i.e., the nodes are on separate components).

Of course, if we are using the existence of a path from one node to another as a measure of cohesion, it is only a small stretch to consider counting the number of links in the shortest path between two nodes as an inverse measure of dyadic cohesion. The geodesic distance matrix is in fact extremely similar to the adjacency matrix of a graph: where there are 1s in the adjacency matrix, there are 1s in the distance matrix. But where there are 0s in the adjacency matrix, there are a range of values in the distance matrix, providing a more nuanced account of lack of adjacency.

One problem with geodesic distance is that the distance between nodes in separate components is technically undefined (or, popularly, infinite). A solution is to use the reciprocal of geodesic distance ($1/d_{ij}$) with the convention that if the distance is undefined, then the reciprocal is zero. This also has the advantage of making it so that larger values indicate more cohesion.

An entirely different approach is based on the notion of a tie being buttressed by the ties that the two nodes have in common with third parties. The basic idea is that if a tie is embedded in a locally dense region of the network, it will be harder for the tie to break apart, and for one party to treat to the other badly (because anything they do will be observed by others). Thus, a tie between nodes A and B is said to be embedded to the extent there exist third parties C such that both A and B are adjacent to C (Feld, 19xx). Such ties have also been referred to as Simmelian ties (Krackhardt, 19xx).

Group-Level Cohesion

The group-level counterpart of the simplest type of dyadic cohesion is the simple sum of each of the dyadic cohesions. This gives the total amount of cohesion in the network. If the data consist of frequencies of interactions between people, then the sum gives the total amount of interaction going on in the system. If the adjacency matrix consists of 1s and 0s, indicating the presence or absence of ties, then the sum gives the total number of ties in the group. We can then compare that total with the total for other groups of similar size, to get a sense of the relative amount of cohesion in each.

Alternatively, we could normalize this total by dividing by the maximum possible, facilitating comparisons across graphs of different sizes. In an undirected graph without reflexive loops, the maximum is given by n*(n-1)/2, where n is the number of nodes in the graph. Dividing by this number gives the proportion of all dyads that are actually tied, a measure known as graph density. Equation 1 gives the full formula, where T is the number of ties in the network and n is the number of nodes.

$$Density = \frac{2T}{n(n-1)}$$ Equation 1

Another way to think of density is as the average of all values in the adjacency matrix of the graph (not counting the diagonal values, which represent self-loops). Taking this perspective makes the generalization of density to valued ties quite natural, as we simply take the average tie strength.

A closely related measure of cohesion is the average degree of the network. If we compute the degree (number of ties) for each node, and then average these degrees, we obtain the average degree of the network. Equation 2 gives the formula and shows how average degree ($\bar{d}$) is related to density. The average degree is easier to interpret than density because it is literally the average number of ties that each node has. It is also the average of the row sums of the adjacency matrix.

$$\bar{d} = \frac{2T}{n} = Density * (n-1)$$ Equation 2

One problem with both density and average degree is that they don't take into account the broader structure of the network. For example, a network that is divided into two very dense components (see Figure xx) will have high density and average degree scores, but will only be cohesive within local regions. Globally, the network will be very non-cohesive because none of the nodes in one component can reach any of the nodes in the other.

This suggests a number of measures of cohesion (or non-cohesion) based on the number and size of components in the network. The first, based only on the number of components, is typically normalized as shown in Equation 3, where c is the number of components and n is the number of nodes in the graph. This normalized measure – called the component ratio – achieves its maximum value of 1.0 when every node is an isolate, and its minimum value of 0 when there is just one component. Obviously, this is an inverse measure of cohesion as larger values indicate less cohesion.

$$CR = \frac{c-1}{n-1}$$ Equation 3

A more sensitive measure along these same lines is called fragmentation (Borgatti, 19xx) or connectedness (Krackhardt, XXXX). Fragmentation is defined as the proportion of pairs of nodes that cannot reach each other by any path. In other words, the proportion of pairs of nodes that are not located in the same component. The formula for fragmentation is given in Equation 4, where r_{ij} is 1 if nodes i and j are in the same component and 0 otherwise.

$$F = 1 - \frac{\sum_{i,j} r_{ij}}{n(n-1)} = 1 - Connectedness$$ Equation 4

Another approach to operationalizing cohesion is based on the lengths of paths connecting pairs of nodes. Perhaps the most obvious measure of this type is the average geodesic distance, also known as the characteristic path length. This is literally the average of all the values in the distance matrix, not including the main diagonal. It can be seen as a measure of how long things typically take to flow from one randomly chosen node to another, at least for something that flows along shortest paths. Clearly, if things flowing through the network can reach nodes quickly, the network is in this sense cohesive.

A variation on this form of cohesion is the diameter of the graph, which is simply the largest value in the distance matrix (i.e., the length of largest shortest path).It can be seen as the maximum amount of time that it would take to diffuse something to every single person in a network when traveling exclusively via shortest paths. All variations on the mean, the mode, the maximum and so on can be seen as measures of network cohesion when applied to the geodesic distance matrix.

A difficulty with the average geodesic distance and its variants is that it cannot be applied to disconnected graphs, i.e., ones with multiple components, since some distances are not defined. One strategy is to replace the missing values with an arbitrary large value.[1] Another strategy is to take reciprocals of the valid distances and assign zeroes to the missing cells. This is called breadth, and is defined according to Equation 5, where d_{ij} is the geodesic distance from i to j and $1/d_{ij}$ is defined to be zero when d_{ij} is undefined. Like fragmentation, breadth is an inverse measure of cohesion.

$$B = 1 - \frac{\sum_{i,j} \frac{1}{d_{ij}}}{n(n-1)} \qquad \text{Equation 4}$$

A completely different approach to cohesion is robustness. How difficult is it to disconnect the network by removing nodes or lines? If you need to remove quite a few nodes or lines to increase the number of components in the graph, then the network is highly robust and in this sense cohesive. Equivalently, if you have to remove many lines or nodes, then there must be many fully independent paths between the nodes, again suggesting cohesion. This suggests that the graph-theoretic concepts of cutpoint and vertex cutset, as well as bridge and edge cutset, might be useful. Indeed, both the vertex connectivity and the edge connectivity of a graph can be seen as measures of cohesion. The greater the value, the more independent paths there are between all pairs of nodes, and the more cohesive the network.

[1] Valente suggests subtracting the valid distances from an arbitrary constant and assigning 0 for the cells corresponding to missing distances. The correlation of this with the original strategy of simply assigning a constant to missing values is a perfect -1.0.

- discuss transitivity as local cohesion? (Or as normalized aggregate of embeddedn/simmelian ties)
- Discuss centrality and cp structures?

Dyad Level	Group Level
Adjacency	Density
Strength of tie	ATS (Average Tie Strength) (D-Bar?)
Geodesic Distance	CPL (Characteristic Path Length) Breadth
Reachability	Connectedness; Fragmentation; Component Ratio
Simmelian/Embedded ties	Transitivity

As we shall see in later chapters, if can regard some network level cohesion measures as aggregations of dyadic cohesion across all pairs of nodes to obtain a summary for the entire network, then shall also be able to view centrality as an aggregation of dyadic cohesion up to the node level. In short, sum measures just sum the cohesion of the ties associated with a given node.