1 Homophily and assortative mixing

Networks, and particularly social networks, often exhibit a property called *homophily* or *assortative mixing*, which simply means that the attributes of vertices correlate across edges. That is, when I observe some edge (i, j) in a network and then examine the attributes of i and j, I see that those attributes are similar to each other.

In social networks, people have a very strong tendency to associate with others who are similar to them, e.g., in age, nationality, language, socioeconomic status, educational level, political beliefs, and many others. This property, however, is only a statement of pattern, and does not say much about the underlying mechanism. For instance, if we observe a pattern of homophily in a social network, e.g., on political beliefs or obesity, we generally cannot distinguish between (i) the edge forming as a result of the attributes being similar, or (ii) the attributes becoming more similar as a result of the edge.

In some networks, the opposite pattern is also seen, which we call *disassortative mixing* and which represents a pattern of association between vertices with dissimilar attributes. For instance, dating or sexual contact networks are largely disassortative, with the large majority of edges running between men and women, with a smaller number of edges running among men or among women. Similarly, in food webs, predators tend to be connected to their prey, rather than to each other.

There are two basic ways a network can be assortative, distinguished by where the attribute is a label or enumerative value (i.e., an unordered type like vertex color or shape) or a scalar value. We will cover them both, and then consider one type of scalar mixing that is of particular interest.

1.1 Enumerative attributes

Enumerative attributes are those that lack any particular ordering, and are sometimes also called categorical variables. They represent things like vertex color, shape, ethnicity, gender, etc.

A good measure of this form of assortativity, i.e., the degree to which like things are connected, is called the *modularity*. This measure is defined as the sum of the differences between the observed and expected fractions of edges for each pair of types. There are two cases where this measure yields zero. The first is the trivial case where all vertices of are the same type. Here, there is only one type of edge, and thus the observed fraction of edges of this form is exactly the same as its expected value. The second is when there is no assortativity, i.e., when edges among a given pair of types occur no more or less frequently than we would expect at random. (Under what circumstances would this measure yield its maximum value of 1? Note that its minimum value is not zero, but is in fact -1. Under what circumstances would this value be achieved?)

In order to calculate the modularity, we must first have a complete labeling of the vertices. Given such a labeling, the modularity for an undirected network is given by

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \ , \tag{1}$$

where A is the adjacency matrix, k_i is the degree of vertex i, m is total the number of edges in the network, c_i is the label of vertex i, and $\delta(c_i, c_j)$ is the Kronecker delta function, which equals 1 when its arguments are the same and 0 otherwise.

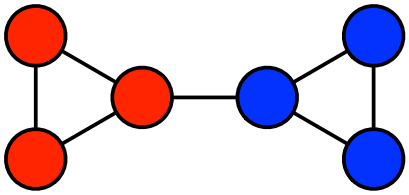
To briefly dissect this equation, let us begin with the summation. The sum is over all pairs i, j and thus considers every adjacency in A, including those that do not represent edges in the network. The delta function serves as a kind of filter, selecting only those pairs i, j whose labels are the same $c_i = c_j$. Thus, the summation is effectively only over edges whose endpoints are of the same type. The inner term is the difference between the observed fraction of all edges between i and j, which is $A_{ij}/2m$, and its expected value under a random graph with the same degree distribution. If the degrees of vertices i and j are k_i and k_j, then the probability that i and j are connected, if edges are distributed at random conditioned on respecting these degrees, is the number of chances they have to connect $k_i k_j$ divided by the total number of such pairs, which is $(2m)^2$. Because both terms have a common denominator of $1/2m$, we may factor this out, which gives the leading constant.

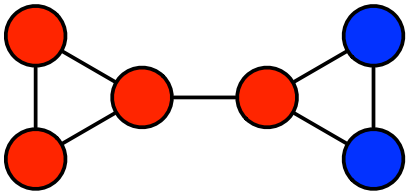
Notice that the summation is only over edges among vertices of the same type. This allows us to rewrite and simplify the equation by dropping many of the zeros in the total summation. Let e_{rr} be the fraction of edges among vertices of type r and let a_r be the fraction of the network's total degree associated with vertices of type r. (Note that the normalization is different for e_{rr} and a_r; in the former case, we count edges once and divide by m, while in the latter case, each edge contributes to the degree of two vertices and thus we divide by $2m$.) The modularity is then simply

$$Q = \sum_r \left(e_{rr} - a_r^2 \right) \ , \tag{2}$$

where a_r is the observed density of edges with type r and a_r^2 is the expected density. This form of Eq. (1) is more compact, and may be used even when we do not have the full adjacency matrix, but instead have only a matrix containing the number of connections between vertex types. Note also that both equations are defined only for undirected graphs. Directed or weighted versions may be defined, but these are not commonly used.

To illustrate the modularity calculation, consider two distinct labelings of the same small modular network with $m = 7$ edges shown below. To begin, we construct a 2×2 matrix based on each labeling, which counts the fraction of edges of a given type:





labeling 1	red	blue
red	3/7	1/14
blue	1/14	3/7

labeling 2	red	blue
red	4/7	2/14
blue	2/14	1/7

From these matrices, the modularity is straightforward to calculate, yielding $Q_1 = 5/14 = 0.357$ for the first, and $Q = 6/49 = 0.122$ for the second. In this case, labeling 1 yields a higher modularity than labeling 2 because it has a larger fraction of within-type edges, i.e., it displays more assortative mixing.

1.2 Scalar attributes

When vertex attributes are scalar values, like age, weight, or income, we may say not only when two vertices have the same value, as in the previous case, but also when two vertices are close in value.

A good measure for this form of assortativity is a kind of network-based generalization of the Pearson correlation coefficient. We begin by adapting the usual definition of covariance to our network context, letting x_i denote the scalar value associated with vertex i. The mean value observed at either end of an edge is simply $\mu = (1/2m)\sum_i k_i x_i$, which weights each observed scalar value by the number of edges in which it participates.

The covariance of x across edges is then

$$\mathrm{cov}(x_i, x_j) = \frac{\sum_{ij} A_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{ij} A_{ij}}$$

$$= \frac{1}{2m}\sum_{ij} A_{ij} x_i x_j - \mu^2$$

$$= \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right) x_i x_j \ , \qquad (3)$$

where we have reused the definition of μ, and simplified considerably, to get to the last line. Note that this form of the covariance is remarkably similar to the definition of the modularity given in Eq. (1), except that instead of a delta function selecting only edges for which the attribute is exactly the same, we now weight every edge by the product $x_i x_j$. When the values of the last term tend to agree, i.e., small values are multiplied by small values and large values multiplied by large values, the covariance will be positive, indicating assortative mixing, while when the opposite is true, the covariance is negative, indicating disassortative mixing.

Finally, just as in the case of the modularity, and in the case of the traditional definition of the Pearson correlation coefficient, it is useful to normalize the covariance so that it ranges from -1 to 1. This version has the form

$$r = \frac{\sum_{ij} \left(A_{ij} - k_i k_j \,/2m\right) x_i x_j}{\sum_{ij} \left(k_i \delta_{ij} - k_i k_j \,/2m\right) x_i x_j} \ , \tag{4}$$

which is called the *assortativity coefficient*, and is exactly the covariance divided by the variance.

1.3 Vertex degrees

Vertex degree is a particular form of scalar attribute that is of particular interest, as it reveals important insights about the large-scale structure of the network. In this case, we have $x_i = k_i$, which slightly simplifies Eq. (4).

Assortative mixing by degree produces a network in which the high-degree vertices tend to connect to each other in dense, high-degree core, while the low-degree vertices also connect to each other, producing a sparse, low-degree periphery. In these networks, degree correlates with centrality. By contrast, disassortative mixing by degree produces a network in which the high-degree vertices tend to connect to low-degree vertices, producing star-like structures. In these networks, centrality correlates less strongly with degree. (See Figure 7.12 in *Networks* for a good visualization of this distinction.)

1.4 A few comments

Within social networks, homophily is a nearly overpowering phenomenon. It is sufficiently strong that merely knowing the attribute labels or values for a subset of a vertex's neighbors can allow a researcher (or marketer) to make a fairly good guess about the label or value at the vertex. That is, knowing your friends' values tells me a great deal about your value, even if you have not disclosed it to me. That being said, homophily is merely a correlation, and thus the values of your friends' friends is only moderately predictive of your value.

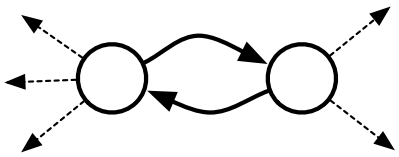
2 Transitivity

Reciprocity and *transitivity* are two terms that refer to the same idea, at different scales. Transitivity is the more general term, referring to a clique of size ℓ. The idea here is that a transitive property is one in which if a and b have a relationship, and b and c have one, then a and c also have one. For a set of ℓ items, transitivity implies a fully connected component or "clique" of size ℓ. The most commonly studied forms of transitivity are for $\ell = 2$ (reciprocity or bidirectional link density) and $\ell = 3$ ("clustering coefficient" or triangle density). Higher-ordered versions of transitivity are sometimes studied in small social networks, e.g., examining tight-knit groups of friends, but are rarely studied in other types of networks.

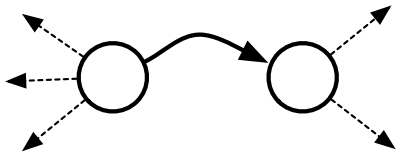
The reason is that perfect transitivity among many vertices is a fairly unusual situation. A more typical situation is approximate transitivity, where instead of a fully connected component, we see a largish subgraph that has relatively more edges among its members than to the rest of the network. (How might this idea of approximate transitivity bump into our notion of sparse networks as the network gets larger?) Approximate transitivity, however, is a slippery concept as there is only one way to be perfectly transitivity, but many ways to be approximately so. We will return to this idea later in the semester, under the name "community structure" or modular networks.

2.1 Reciprocity

In directed networks, not all edges are bidirectional, and the fraction of those edges that are bidirectional can tell us interesting things about the network, depending on what kind of network it is. For instance, these figures show one reciprocated and one unreciprocated edge.



reciprocated



unreciprocated

In social networks, bidirectional or reciprocated edges can indicate whether a friendship is perceived as being equal or whether one party views it as stronger than the other. That is, reciprocated edges can tell us something about social status. In transportation networks, they represent reachability, while in communication networks, they may represent influence.[1] Reciprocity can be calculated in

[1] But, you should be wary of anyone claiming to either measure or test for social influence. In general, most studies of "influence" are actually studies of correlation, and as with homophily, correlation does not imply causation. Keep your wits and skepticism about you.

any directed network, e.g., in food webs, predation and parasitism are directed relationships, as is genetic regulation in a gene-regulatory network. However, the meaning of a reciprocated link depends on the context and the underlying processes in that domain. For example, a reciprocated link in a food web means something different than a reciprocated link in a social network.

The reciprocity of a network is given by the fraction of reciprocated links. To compute this value, we simply count the relative frequency of cliques of size 2 in a directed network. When edges have unit weight, a network's reciprocity is defined as

$$r = \frac{1}{m} \sum_{ij} A_{ij} A_{ji} \ . \tag{5}$$

In this way, if both the forward and backward direction of a particular edge appear in the network, the reciprocity score is incremented twice. The normalization factor counts all edges that could be reciprocated.

Reciprocity may also be defined as a vertex-level measure, in which we count only the fraction of 2-cliques attached to some vertex i:

$$r_i = \frac{1}{k_i} \sum_{j} A_{ij} A_{ji} \ , \tag{6}$$

which may be a useful way of estimating a vertex-level covariate for additional analysis (e.g., to compare with the vertex degree, or centrality score). This measure is sometimes called the *local reciprocity*.

In empirical social networks, a value of $r \simeq 0.25$ is not unusual. While this may like a surprisingly small value, this value is observed even in very large networks where the probability of two directed edges forming a bidirectional loop is roughly $O(1/n)$. In some cases, a small value is found in part because the survey technique limits the number of responses, or because different people interpret the word "friend" to mean different things. In other cases, a small value may reflect the presence of status-driven relationships. In undirected networks, reciprocity is always 1.

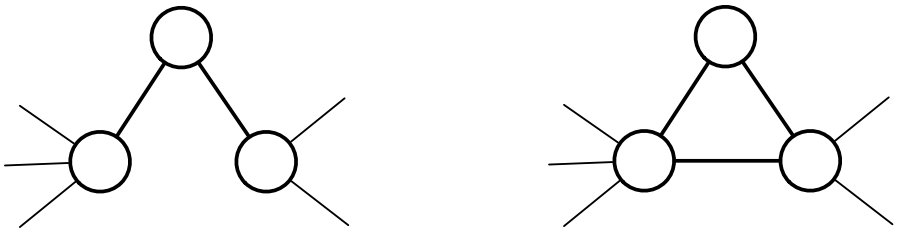
2.2 Clustering coefficient

Structurally, reciprocity measures the fraction of 2-cliques that appear in a directed network. Another commonly used measure counts the fraction of 3-cliques or triangle density in an undirected network (directed versions also exist, but their calculation is slightly more tricky, as there are many more ways three vertices could be connected in a directed network), and is called the *clustering*

coefficient.[2] This measure is defined mathematically as

$$C = \frac{\text{(number of closed paths of length 2)}}{\text{(number of paths of length 2)}} \ . \tag{7}$$

A path of length two has the natural definition, which is a sequence of vertices i, j, k for which the edges (i, j) and (j, k) are in the network. A "closed path" of length two is simply a path of length two plus the edge (k, i).



Thus, C is a number between 0 and 1, and measures the density of triangles in the network. It takes the value $C = 1$ only for a fully connected component, i.e., a clique of size n. The opposite extreme, a value of $C = 0$ can occur on a number of different networks, the most obvious being any bipartite network (including trees).

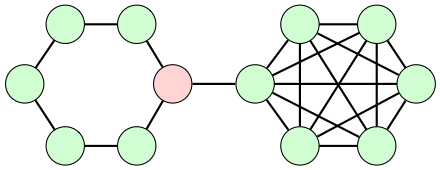
In the above formulation, as in our discussion of paths for betweenness, we count distinct orderings of the vertices as being different. If we collapse these counts, we may simplify Eq. (7) in the case of undirected graphs by counting only closed and "open" triangles. Let us define a "connected triple" or "open triad" as any trio of nodes i, j, k in which at least two pairs are connected. A "closed triad" is then any such trio in which we have added the final, missing connection (hence the word "closed"). The clustering coefficient is then

$$C = \frac{\text{(number of triangles)} \times 3}{\text{(number of connected triples)}} \ , \tag{8}$$

where the factor of 3 comes from the symmetry of the triangle.

To illustrate this measure, consider again our simple cycle-and-clique network. The clique has 6 vertices and thus both 20 triangles and 60 connected triples. The cycle contains 6 connected triples and no triangles. Finally, there are 2 connected triples starting from the cycle and 5 connected

[2]This name is not a particularly good one, as the term "clustering" is used in several other contexts in the study of networks and vector data sets. Readers should be wary of these alternative usages when reading the literature.



triples starting from the clique that use the edge joining the cycle to the clique. Thus, the clustering coefficient is $C = 60/73 = 0.82$. This value is relatively large for social networks, which generally have clustering coefficients of $C \simeq 0.20$ or so. A non-trivial clustering coefficient is generally a distinguishing feature of social networks, with most biological and technological networks containing far fewer triangles, and which exhibit clustering coefficients close to 0.

The clustering coefficient can also be defined as a vertex-level measure:[3]

$$C_i = \frac{\text{(number of pairs of neighbors of } i \text{ that are connected)}}{\text{(number of pairs of neighbors of } i\text{)}} \ , \tag{9}$$

which is the natural definition of triangle density where we require that vertex i be the middle vertex of every connected triple. In our cycle-plus-clique example above, the highlighted vertex has a local clustering coefficient of $C_i = 0$ because it participates in no triangles. Its immediate neighbor in the clique is a more interesting case, with $C_i = 10/15 = 0.67$.

3 Other neighborhood measures

There are many measures of vertex similarity that depend on the network structure, and there is a steady supply of new measures being developed. Among these others, there are a few that have achieved widespread usage. One class considers the similarity of the neighbor sets of two vertices i and j. If these sets are identical, we call the vertices *structurally equivalent*. Approximate equivalence is a generally more useful notion, and there are a number of ways to quantify this idea.[4] A common measure of approximate structural equivalence is the *Jaccard coefficient*, defined mathematically as

$$J_{ij} = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} \ , \tag{10}$$

where $N(i)$ is the set of vertices with connections to vertex i, i.e., i's neighbor set. When the neighbor sets are identical, $J_{ij} = 1$, and when they are disjoint, $J_{ij} = 0$. One nice property of

[3] Two ideas from the sociological literature that are closely related to the local clustering coefficient are *structural holes* and *redundancy*, which we won't cover here.

[4] You've already encountered one, cosine similarity, in the problem set.

the Jaccard coefficient is that it naturally controls for the overall size of the neighbor sets, while a simple count of the number of shared neighbors would not.

4 At home

1. Read Chapter 7.9–7.13 (pages 198–231) in *Networks*
2. Next time: degree distributions