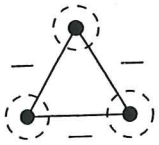


In this network all three vertices dislike each other, so there is an odd number of minus signs around the loop, but there is no problem dividing the network into three clusters of one vertex each such that everyone dislikes the members of the other clusters. This network is clusterable but not balanced.

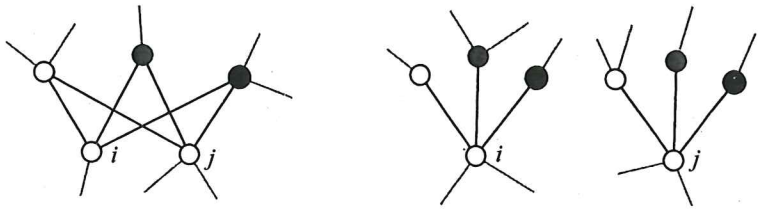
7.12 SIMILARITY

Another central concept in social network analysis is that of similarity between vertices. In what ways can vertices in a network be similar, and how can we quantify that similarity? Which vertices in a given network are most similar to one another? Which vertex v is most similar to a given vertex u? Answers to questions like these can help us tease apart the types and relationships of vertices in social networks, information networks, and others. For instance, one could imagine that it might be useful to have a list of web pages that are similar—in some appropriate sense—to another page that we specify. In fact, several web search engines already provide a feature like this: "Click here for pages similar to this one."

Similarity can be determined in many different ways and most of them have nothing to do with networks. For example, commercial dating and matchmaking services try to match people with others to whom they are similar by using descriptions of people's interests, background, likes, and dislikes. In effect, these services are computing similarity measures between people based on personal characteristics. Our focus in this book, however, is on networks, so we will concentrate on the more limited problem of determining similarity between the vertices of a network using the information contained in the network structure.

There are two fundamental approaches to constructing measures of network similarity, called *structural equivalence* and *regular equivalence*. The names are rather opaque, but the ideas they represent are simple enough. Two vertices in a network are structurally equivalent if they share many of the same network neighbors. In Fig. 7.9a we show a sketch depicting structural equivalence between two vertices i and j—the two share, in this case, three of the same neighbors, although both also have other neighbors that are not shared.

Regular equivalence is more subtle. Two regularly equivalent vertices do not necessarily share the same neighbors, but they have neighbors *who are*



(a) Structural equivalence (b) Regular equivalence

Figure 7.9: Structural equivalence and regular equivalence. (a) Vertices i and j are structurally equivalent if they share many of the same neighbors. (b) Vertices i and j are regularly equivalent if their neighbors are themselves equivalent (indicated here by the different shades of vertices).

themselves similar. Two history students at different universities, for example, may not have any friends in common, but they can still be similar in the sense that they both know a lot of other history students, history instructors, and so forth. Similarly, two CEOs at two different companies may have no colleagues in common, but they are similar in the sense that they have professional ties to their respective CFO, CIO, members of the board, company president, and so forth. Regular equivalence is illustrated in Fig. 7.9b.

In the next few sections we describe some mathematical measures that quantify these ideas of similarity. As we will see, measures for structural equivalence are considerably better developed than those for regular equivalence.

7.12.1 COSINE SIMILARITY

We start by looking at measures of structural equivalence and we will concentrate on undirected networks. Perhaps the simplest and most obvious measure of structural equivalence would be just a count of the number of common neighbors two vertices have. In an undirected network the number n_{ij} of common neighbors of vertices i and j is given by

$$n_{ij} = \sum_k A_{ik} A_{kj}, \tag{7.46}$$

which is the ijth element of $\mathbf{A}^2$. This quantity is closely related to the "cocitation" measure introduced in Section 6.4.1. Cocitation is defined for directed

networks whereas we are here considering undirected ones, but otherwise it is essentially the same thing.

However, a simple count of common neighbors for two vertices is not on its own a very good measure of similarity. If two vertices have three common neighbors is that a lot or a little? It's hard to tell unless we know, for instance, what the degrees of the vertices are, or how many common neighbors other pairs of vertices share. What we need is some sort of normalization that places the similarity value on some easily understood scale. One strategy might be simply to divide by the total number of vertices in the network n, since this is the maximum number of common neighbors two vertices can have in a simple graph. (Technically the maximum is actually $n-2$, but the difference is small when n is large.) However, this unduly penalizes vertices with low degree: if a vertex has degree three, then it can have at most three neighbors in common with another vertex, but the two vertices would still receive a small similarity value if the divisor n were very large. A better measure would allow for the varying degrees of vertices. Such a measure is the *cosine similarity*, sometimes also called *Salton's cosine*.

In geometry, the inner or dot product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is given by $\mathbf{x}\cdot\mathbf{y} = |\mathbf{x}|\,|\mathbf{y}|\cos\theta$, where $|\mathbf{x}|$ is the magnitude of $\mathbf{x}$ and θ is the angle between the two vectors. Rearranging, we can write the cosine of the angle as

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}|\,|\mathbf{y}|}. \tag{7.47}$$

Salton [290] proposed that we regard the ith and jth rows (or columns) of the adjacency matrix as two vectors and use the cosine of the angle between them as our similarity measure. Noting that the dot product of two rows is simply $\sum_k A_{ik}A_{kj}$ for an undirected network, this gives us a similarity

$$\sigma_{ij} = \cos\theta = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}}. \tag{7.48}$$

Assuming our network is an unweighted simple graph, the elements of the adjacency matrix take only the values 0 and 1, so that $A_{ij}^2 = A_{ij}$ for all i, j. Then $\sum_k A_{ik}^2 = \sum_k A_{ik} = k_i$, where k_i is the degree of vertex i (see Eq. (6.19)). Thus

$$\sigma_{ij} = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{k_i k_j}} = \frac{n_{ij}}{\sqrt{k_i k_j}}. \tag{7.49}$$

The cosine similarity of i and j is therefore the number of common neighbors of the two vertices divided by the geometric mean of their degrees. For the

vertices i and j depicted in Fig. 7.9a, for instance, the cosine similarity would be

$$\sigma_{ij} = \frac{3}{\sqrt{4 \times 5}} = 0.671\ldots \tag{7.50}$$

Notice that the cosine similarity is technically undefined if one or both of the vertices has degree zero, but by convention we normally say in that case that $\sigma_{ij} = 0$.

The cosine similarity provides a natural scale for our similarity measure. Its value always lies in the range from 0 to 1. A cosine similarity of 1 indicates that two vertices have exactly the same neighbors. A cosine similarity of zero indicates that they have none of the same neighbors. Notice that the cosine similarity can never be negative, being a sum of positive terms, even though cosines in general can of course be negative.

7.12.2 PEARSON COEFFICIENTS

An alternative way to normalize the count of common neighbors is to compare it with the expected value that count would take on a network in which vertices choose their neighbors at random. This line of argument leads us to the *Pearson correlation coefficient*.

Suppose vertices i and j have degrees k_i and k_j respectively. How many common neighbors should we expect them to have? This is straightforward to calculate if they choose their neighbors purely at random. Imagine that vertex i chooses k_i neighbors uniformly at random from the n possibilities open to it (or $n - 1$ on a network without self-loops, but the distinction is slight for a large network), and vertex j similarly chooses k_j neighbors at random. For the first neighbor that j chooses there is a probability of k_i/n that it will choose one of the ones i chose, and similarly for each succeeding choice. (We neglect the possibility of choosing the same neighbor twice, since it is small for a large network.) Then in total the expected number of common neighbors between the two vertices will be k_j times this, or $k_i k_j/n$.

A reasonable measure of similarity between two vertices is the actual number of common neighbors they have minus the expected number that they

would have if they chose their neighbors at random:

$$\begin{aligned}\sum_k A_{ik}A_{jk} - \frac{k_i k_j}{n} &= \sum_k A_{ik}A_{jk} - \frac{1}{n}\sum_k A_{ik}\sum_l A_{jl} \\ &= \sum_k A_{ik}A_{jk} - n\langle A_i\rangle\langle A_j\rangle \\ &= \sum_k [A_{ik}A_{jk} - \langle A_i\rangle\langle A_j\rangle] \\ &= \sum_k (A_{ik} - \langle A_i\rangle)(A_{jk} - \langle A_j\rangle), \end{aligned} \tag{7.51}$$

where $\langle A_i\rangle$ denotes the mean $n^{-1}\sum_k A_{ik}$ of the elements of the ith row of the adjacency matrix. Equation (7.51) will be zero if the number of common neighbors of i and j is exactly what we would expect on the basis of random chance. If it is positive, then i and j have more neighbors than we would expect by chance, which we take as an indication of similarity between the two. Equation (7.51) can also be negative, indicating that i and j have fewer neighbors than we would expect, a possible sign of dissimilarity.

Equation (7.51) is simply n times the covariance $\mathrm{cov}(A_i, A_j)$ of the two rows of the adjacency matrix. It is common to normalize the covariance, as we did with the cosine similarity, so that its maximum value is 1. The maximum value of the covariance of any two sets of quantities occurs when the sets are exactly the same, in which case their covariance is just equal to the variance of either set, which we could write as σ_i^2 or σ_j^2, or in symmetric form as $\sigma_i\sigma_j$. Normalizing by this quantity then gives us the standard Pearson correlation coefficient:

$$r_{ij} = \frac{\mathrm{cov}(A_i, A_j)}{\sigma_i\sigma_j} = \frac{\sum_k (A_{ik} - \langle A_i\rangle)(A_{jk} - \langle A_j\rangle)}{\sqrt{\sum_k (A_{ik} - \langle A_i\rangle)^2}\sqrt{\sum_k (A_{jk} - \langle A_j\rangle)^2}}. \tag{7.52}$$

This quantity lies strictly in the range $-1 \le r_{ij} \le 1$.

The Pearson coefficient is a widely used measure of similarity. It allows us to say when vertices are both similar or dissimilar compared with what we would expect if connections in the network were formed at random.

7.12.3 OTHER MEASURES OF STRUCTURAL EQUIVALENCE

There are many other possible measures of structural equivalence. For instance, one could also normalize the number n_{ij} of common neighbors by dividing by (rather than subtracting) the expected value of $k_i k_j/n$. That would give us a similarity of

$$\frac{n_{ij}}{k_i k_j/n} = n\frac{\sum_k A_{ik}A_{jk}}{\sum_k A_{ik}\sum_k A_{jk}}. \tag{7.53}$$

This quantity will be 1 if the number of common neighbors is exactly as expected on the basis of chance, greater than one if there are more common neighbors than that, and less than one for dissimilar vertices with fewer common neighbors than we would expect by chance. It is never negative and has the nice property that it is zero when the vertices in question have no common neighbors. This measure could be looked upon as an alternative to the cosine similarity: the two differ in that one has the product of the degrees $k_i k_j$ in the denominator while the other has the square root of the product $\sqrt{k_i k_j}$. It has been suggested that Eq. (7.53) may in some cases be a superior measure to the cosine similarity because, by normalizing with respect to the expected number of common neighbors rather than the maximum number, it allows us to easily identify statistically surprising coincidences between the neighborhoods of vertices, which cosine similarity does not [195].

Another measure of structural equivalence is the so-called *Euclidean distance*,[32] which is equal to the number of neighbors that differ between two vertices. That is, it is the number of vertices that are neighbors of i but not of j, or vice versa. Euclidean distance is really a dissimilarity measure, since it is larger for vertices that differ more.

In terms of the adjacency matrix the Euclidean distance d_{ij} between two vertices can be written

$$d_{ij} = \sum_k (A_{ik} - A_{jk})^2. \tag{7.54}$$

As with our other measures it is sometimes convenient to normalize the Euclidean distance by dividing by its possible maximum value. The maximum value of d_{ij} occurs when two vertices have no neighbors in common, in which case the distance is equal to the sum of the degrees of the vertices: $d_{ij} = k_i + k_j$. Dividing by this maximum value the normalized distance is

$$\frac{\sum_k (A_{ik} - A_{jk})^2}{k_i + k_j} = \frac{\sum_k (A_{ik} + A_{jk} - 2A_{ik}A_{jk})}{k_i + k_j} = 1 - 2\frac{n_{ij}}{k_i + k_j}, \tag{7.55}$$

where we have made use of the fact that $A_{ij}^2 = A_{ij}$ because A_{ij} is always zero or one, and n_{ij} is again the number of neighbors that i and j have in common. To within additive and multiplicative constants, this normalized Euclidean distance can thus be regarded as just another alternative normalization of the number of common neighbors.

[32] This is actually a bad name for it—it should be called *Hamming distance*, since it is essentially the same as the Hamming distance of computer science and has nothing to do with Euclid.

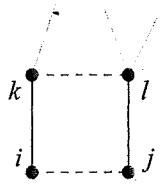
7.12.4 REGULAR EQUIVALENCE

The similarity measures discussed in the preceding sections are all measures of structural equivalence, i.e., they are measures of the extent to which two vertices share the same neighbors. The other main type of similarity considered in social network analysis is regular equivalence. As described above, regularly equivalent vertices are vertices that, while they do not necessarily share neighbors, have neighbors who are themselves similar—see Fig. 7.9b again.

Quantitative measures of regular equivalence are less well developed than measures of structural equivalence. In the 1970s social network analysts came up with some rather complicated computer algorithms, such as the "REGE" algorithm of White and Reitz [320,327], that were intended to discover regular equivalence in networks, but the operation of these algorithms is involved and not easy to interpret. More recently, however, some simpler algebraic measures have been developed that appear to work reasonably well. The basic idea [45, 162,195] is to define a similarity score σ_{ij} such that i and j have high similarity if they have neighbors k and l that themselves have high similarity. For an undirected network we can write this as

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl}, \tag{7.56}$$

or in matrix terms $\boldsymbol{\sigma} = \alpha \mathbf{A} \boldsymbol{\sigma} \mathbf{A}$. Although it may not be immediately clear, this expression is a type of eigenvector equation, where the entire matrix $\boldsymbol{\sigma}$ of similarities is the eigenvector. The parameter α is the eigenvalue (or more correctly, its inverse) and, as with the eigenvector centrality of Section 7.2, we are normally interested in the leading eigenvalue, which can be found by standard methods.



Vertices i and j are considered similar (dashed line) if they have respective neighbors k and l that are themselves similar.

See Section 11.1 for a discussion of computer algorithms for finding eigenvectors.

This formula however has some problems. First, it doesn't necessarily give a high value for the "self-similarity" σ_{ii} of a vertex to itself, which is counterintuitive. Presumably, all vertices are highly similar to themselves! As a consequence of this, Eq. (7.56) also doesn't necessarily give a high similarity score to vertex pairs that have a lot of common neighbors, which in the light of our examination of structural equivalence in the preceding few sections we perhaps feel it should. If we had high self-similarity scores for all vertices, on the other hand, then Eq. (7.56) would automatically give high similarity also to vertices with many common neighbors.

We can fix these problems by introducing an extra diagonal term in the similarity thus:

$$\sigma_{ij} = \alpha \sum_{kl} A_{ik} A_{jl} \sigma_{kl} + \delta_{ij}, \tag{7.57}$$

or in matrix notation

$$\boldsymbol{\sigma} = \alpha \mathbf{A} \boldsymbol{\sigma} \mathbf{A} + \mathbf{I}. \tag{7.58}$$

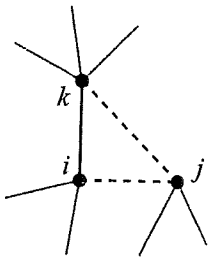
However, while expressions like this have been proposed as similarity measures, they still suffer from some problems. Suppose we evaluate Eq. (7.58) by repeated iteration, taking a starting value, for example, of $\boldsymbol{\sigma}^{(0)} = 0$ and using it to compute $\boldsymbol{\sigma}^{(1)} = \alpha \mathbf{A} \boldsymbol{\sigma} \mathbf{A} + \mathbf{I}$, and then repeating the process many times until σ converges. On the first few iterations we will get the following results:

$$\boldsymbol{\sigma}^{(1)} = \mathbf{I}, \tag{7.59a}$$

$$\boldsymbol{\sigma}^{(2)} = \alpha \mathbf{A}^2 + \mathbf{I}, \tag{7.59b}$$

$$\boldsymbol{\sigma}^{(3)} = \alpha^2 \mathbf{A}^4 + \alpha \mathbf{A}^2 + \mathbf{I}. \tag{7.59c}$$

The pattern is clear: in the limit of many iterations, we will get a sum over even powers of the adjacency matrix. However, as discussed in Section 6.10, the elements of the rth power of the adjacency matrix count paths of length r between vertices, and hence this measure of similarity is a weighted sum over the numbers of paths of even length between pairs of vertices.



In the modified definition of regular equivalence vertex i is considered similar to vertex j (dashed line) if it has a neighbor k that is itself similar to j.

But why should we consider only paths of even length? Why not consider paths of all lengths? These questions lead us to a better definition of regular equivalence as follows: vertices i and j are similar if i has a neighbor k that is itself similar to j.[33] Again we assume that vertices are similar to themselves, which we can represent with a diagonal δ_{ij} term in the similarity, and our similarity measure then looks like

$$\sigma_{ij} = \alpha \sum_k A_{ik} \sigma_{kj} + \delta_{ij}, \tag{7.60}$$

or

$$\boldsymbol{\sigma} = \alpha \mathbf{A} \boldsymbol{\sigma} + \mathbf{I}, \tag{7.61}$$

in matrix notation. Evaluating this expression by iterating again starting from $\boldsymbol{\sigma}^{(0)} = 0$, we get

$$\boldsymbol{\sigma}^{(1)} = \mathbf{I}, \tag{7.62a}$$

$$\boldsymbol{\sigma}^{(2)} = \alpha \mathbf{A} + \mathbf{I}, \tag{7.62b}$$

$$\boldsymbol{\sigma}^{(3)} = \alpha^2 \mathbf{A}^2 + \alpha \mathbf{A} + \mathbf{I}. \tag{7.62c}$$

[33] This definition is not obviously symmetric with respect to i and j but, as we see, does in fact give rise to an expression for the similarity that is symmetric.

In the limit of a large number of iterations this gives

$$\boldsymbol{\sigma} = \sum_{m=0}^{\infty} (\alpha \mathbf{A})^m = (\mathbf{I} - \alpha \mathbf{A})^{-1}, \tag{7.63}$$

which we could also have deduced directly by rearranging Eq. (7.61). Now our similarity measure includes counts of paths at all lengths, not just even paths. In fact, we can see now that this similarity measure could be defined a completely different way, as a weighted count of all the paths between the vertices i and j with paths of length r getting weight α^r. So long as $\alpha < 1$, longer paths will get less weight than shorter ones, which seems sensible: in effect we are saying that vertices are similar if they are connected either by a few short paths or by very many long ones.

Equation (7.63) is reminiscent of the formula for the Katz centrality, Eq. (7.10). We could call Eq. (7.63) the "Katz similarity" perhaps, although Katz himself never discussed it. The Katz centrality of a vertex would then be simply the sum of the Katz similarities of that vertex to all others. Vertices that are similar to many others would get high centrality, a concept that certainly makes intuitive sense. As with the Katz centrality, the value of the parameter α is undetermined—we are free to choose it as we see fit—but it must satisfy $\alpha < 1/\kappa_1$ if the sum in Eq. (7.63) is to converge, where κ_1 is the largest eigenvalue of the adjacency matrix.

In a sense, this regular equivalence measure can be seen as a generalization of our structural equivalence measures in earlier sections. With those measures we were counting the common neighbors of a pair of vertices, but the number of common neighbors is also of course the number of paths of length two between the vertices. Our "Katz similarity" measure merely extends this concept to counting paths of all lengths.

Some variations of this similarity measure are possible. As defined it tends to give high similarity to vertices that have high degree, because if a vertex has many neighbors it tends to increase the number of those neighbors that are similar to any other given vertex and hence increases the total similarity to that vertex. In some cases this might be desirable: maybe the person with many friends *should* be considered more similar to others than the person with few. However, in other cases it gives an unwanted bias in favor of high-degree nodes. Who is to say that two hermits are not "similar" in an interesting sense? If we wish, we can remove the bias in favor of high degree by dividing by vertex degree thus:

$$\sigma_{ij} = \frac{\alpha}{k_i} \sum_k A_{ik} \sigma_{kj} + \delta_{ij}, \tag{7.64}$$

or in matrix notation $\sigma = \alpha \mathbf{D}^{-1}\mathbf{A}\sigma + \mathbf{I}$, where, as previously, $\mathbf{D}$ is the diagonal matrix with elements $D_{ii} = k_i$. This expression can be rearranged to read:[34]

$$\sigma = (\mathbf{I} - \alpha \mathbf{D}^{-1}\mathbf{A})^{-1} = (\mathbf{D} - \alpha \mathbf{A})^{-1}\mathbf{D}. \tag{7.65}$$

Another useful variant is to consider cases where the last term in Eqs. (7.60) or (7.64) is not simply diagonal, but includes off-diagonal terms too. Such a generalization would allow us to specify explicitly that particular pairs of vertices are similar, based on some other (probably non-network) information that we have at our disposal. Going back to the example of CEOs at companies that we gave at the beginning of Section 7.12, we might, for example, want to state explicitly that the CFOs and CIOs and so forth at different companies are similar, and then our similarity measure would, we hope, correctly deduce from the network structure that the CEOs are similar also. This kind of approach is particularly useful in the case of networks that consist of more than one component, so that some pairs of vertices are not connected at all. If, for instance, we have two separate components representing people in two different companies, then there will be no paths of any length between individuals in different companies, and hence a measure like (7.60) or (7.64) will never assign a non-zero similarity to such individuals. If however, we explicitly insert some similarities between members of the different companies, our measure will then be able to generalize and extend those inputs to deduce similarities between other members.

This idea of generalizing from a few given similarities arises in other contexts too. For example, in the fields of machine learning and information retrieval there is a considerable literature on how to generalize known similarities between a subset of the objects in a collection of, say, text documents to the rest of the collection, based on network data or other information.

7.13 HOMOPHILY AND ASSORTATIVE MIXING

Consider Fig. 7.10, which shows a friendship network of children at an American school, determined from a questionnaire of the type discussed in Section 3.2.[35] One very clear feature that emerges from the figure is the division of

[34] It is interesting to note that when we expand this measure in powers of the adjacency matrix, as we did in Eq. (7.63), the second-order (i.e., path-length two) term is the same as the structural equivalence measure of Eq. (7.53), which perhaps lends further credence to both expressions as natural measures of similarity.

[35] The study used a "name generator"—students were asked to list the names of others they considered to be their friends. This results in a directed network, but we have neglected the edge