after, or simply eager to collaborate, these nodes seem to be important in some sense. Therefore, the degree is a very natural measure of centrality in social networks.

The *average degree* of a network indicates how connected the nodes are on average. As we shall see later (Section 3.2), the average degree may not be representative of the actual distribution of degree values. This is the case when the nodes have heterogeneous degrees, as in many real-world networks.

3.1.2 Closeness

Another way to measure the centrality of a node is by determining how "close" it is to the other nodes. This can be done by summing the distances from the node to all others. If the distances are short on average, their sum is a small number and we say that the node has high centrality. This leads to the definition of *closeness centrality*, which is simply the inverse of the sum of distances of a node from all others.

The closeness centrality of a node i is defined as

$$g_i = \frac{1}{\sum_{j \neq i} \ell_{ij}}, \tag{3.1}$$

where ℓ_{ij} is the distance from i to j and the sum runs over all the nodes of the network, except i itself. An alternative formulation is obtained by multiplying g_i by the constant $N-1$, which is just the number of terms in the sum at the denominator:

$$\tilde{g}_i = (N-1)g_i = \frac{N-1}{\sum_{j \neq i} \ell_{ij}} = \frac{1}{\sum_{j \neq i} \ell_{ij}/(N-1)}. \tag{3.2}$$

This way we discount the graph size and make the measure comparable across different networks. Since what matters is not the actual value of g_i but its ranking compared to the closeness centrality of the other nodes, the relative centrality of the nodes remains the same as by using Eq. (3.1), because the ranking is not altered if the values are multiplied by a constant. The expression $\sum_{j \neq i} \ell_{ij}/(N-1)$ is the *average distance* from the focal node i to the rest of the network. So we find that closeness can be expressed equivalently as the inverse of the average distance.

NetworkX has a function to compute the closeness centrality:

```
nx.closeness_centrality(G, node) # closeness centrality
                                 # of node
```

3.1.3 Betweenness

Many phenomena taking place in networks are based on diffusion processes (Chapter 7). Examples include the transmission of information across a social network, the traffic of goods through a port, and the spread of epidemics in the network of physical contacts between the individuals of a population. This has suggested a third notion of centrality, called *betweenness*: a node is the more central, the more often it is involved in these processes.

Naturally, betweenness centrality has a different implementation for each distinct type of diffusion. The simplest and most popular implementation considers a simple process where signals are transmitted from each node to every other node, by following shortest paths. This approach is often used in transportation networks to provide an estimate of the traffic handled by the nodes, assuming that the number of shortest paths that traverse a node is a good approximation for the frequency of use of the node. The centrality is then estimated by counting how many times a node is crossed by those paths. The higher the count, the more traffic is controlled by the node, which is therefore more influential in the network.

Given two nodes, there may be more than one shortest path between them in the network, all having the same length. For instance, if nodes X and Y are not connected to each other but have two common neighbors S and T, there are two distinct shortest paths of length two running from X to Y: $X-S-Y$ and $X-T-Y$. Let σ_{hj} be the total number of shortest paths from h to j and $\sigma_{hj}(i)$ the number of these shortest paths that pass through node i. The betweenness of i is defined as

$$b_i = \sum_{h \neq j \neq i} \frac{\sigma_{hj}(i)}{\sigma_{hj}}. \tag{3.3}$$

In Eq. (3.3) the sum runs over all pairs of vertices h and j, distinct from i and from each other. If no shortest path between h and j crosses i [$\sigma_{hj}(i) = 0$], the contribution of the pair (h, j) to the betweenness of i is 0. If all shortest paths between h and j cross i [$\sigma_{hj}(i) = \sigma_{hj}$], the contribution is 1. If a node is a *leaf* (i.e. it has only one neighbor), it cannot be crossed by any path. Therefore its betweenness is zero. Since the potential contributions come from all pairs of nodes, the betweenness grows with the network size.

Let us work through the example in Figure 3.1(a). For node **1**, the only pair of nodes that has a shortest path going through this node is (**2**, **4**). However, there are two shortest paths of equal length between **2** and **4**: the other path goes through node **3** and not **1**. Therefore the betweenness of node **1** is 1/2. Next, consider node **3**. The shortest paths between the three node pairs (**1**, **5**), (**2**, **5**), and (**4**, **5**) go through **3**. As we observed earlier, there are two equivalent shortest paths between nodes **2** and **4**, only one of which goes through **3**, contributing 1/2 to the sum. The total gives a betweenness centrality of 3.5 for node **3**. The remaining nodes **2**, **4**, and **5** have no shortest paths going through them, therefore their betweenness is zero.

A node has high betweenness if it occupies a special position in the network, such that it is an important station for the communication patterns running through the network. For that to happen, it is not necessary to have many neighbors. Generally we observe a correlation between the degree of a node and its betweenness, so that well-connected nodes have high betweenness and vice versa [Figure 3.1(a)]. However, there are many exceptions. Nodes bridging different regions of a network typically have high betweenness, even if their degree is low, as illustrated in Figure 3.1(b).

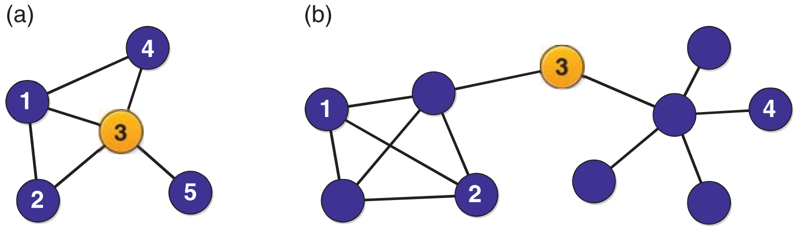

Fig. 3.1 Illustrations of node betweenness centrality. (a) The orange node has high degree ($k_3 = 4$) as well as high betweenness ($b_3 = 3.5$). (b) The orange node has low degree ($k_3 = 2$) but keeps the network connected, acting as the only bridge between nodes in the two subnetworks. For example, the shortest path between nodes **1** and **2** does not go through the orange node, but the path between **1** and **4** does. In fact, all the shortest paths between the four nodes in one subnetwork and the five nodes in the other subnetwork go through the orange node. Therefore its betweenness is $b_3 = 4 \times 5 = 20$.

The concept has a straightforward extension to links. The betweenness centrality of a link is the fraction of shortest paths among all possible node couples that pass through that link. Links with very high betweenness centrality often join cohesive regions of the network, called *communities*. Therefore betweenness can be used to locate and remove those links, allowing for the separation and consequent identification of communities (Chapter 6).

The betweenness centrality depends on the size of the network. If we wish to compare the centrality of nodes or links in different networks, the betweenness values should be normalized.

For node betweenness, the maximum number of paths that could go through node i is the number of pairs of nodes excluding i itself. This is expressed by $\binom{N-1}{2} = \frac{(N-1)(N-2)}{2}$. The normalized betweenness of node i is therefore obtained by dividing b_i in Eq. (3.3) by this factor.

NetworkX has functions to compute the normalized betweenness centrality of nodes and links:

```
nx.betweenness_centrality(G)      # dict nodes ->
                                  # betweenness centrality
nx.edge_betweenness_centrality(G) # dict links ->
                                  # betweenness centrality
```

3.2 Centrality Distributions

Before the advent of online social media, the social networks that one could study were typically built through personal interviews and surveys, which could not involve very