CHAPTER 2

# Representing and Measuring Networks

This chapter presents some of the fundamentals of how networks are represented, measured, and characterized. It provides basic concepts and definitions that are the basis for the language of network research. Sprinkled throughout are observations from case studies that illustrate some of the concepts. More discussion about observed social and economic networks appears in Chapter 3.

## 2.1 ▪ Representing Networks

As networks of relationships come in many shapes and sizes, there is no single way of representing networks that encompasses all applications. Nevertheless, there are some representations that serve as a useful basis for capturing many applications. Here I focus on a few standard ways of denoting networks that are broad and flexible enough to capture a multitude of applications and yet sufficiently simple to be compact, intuitive, and tractable. As we proceed, I try to make clear what these concepts capture and what they omit.

### 2.1.1 Nodes and Players

The set $N = \{1, \ldots, n\}$ is the set of *nodes* that are involved in a network of relationships. Nodes are also referred to as "vertices," "individuals," "agents," or "players," depending on the setting. It is important to emphasize that nodes can be individual people, firms, countries, or other organizations; a node can even be something like a web page belonging to a person or organization.

### 2.1.2 Graphs and Networks

The canonical network form is an undirected graph, in which two nodes are either connected or they are not. For such graphs it cannot be that one node is related to a

second without the second being related to the first. This behavior is generally true of many social and/or economic relationships, such as partnerships, friendships, alliances, and acquaintances. Such networks are central to most of the chapters that follow. However, there are other situations that are better modeled as directed networks, in which one node may be connected to a second without the second being connected to the first. For instance, a network that keeps track of which authors cite which other authors, or which web pages have links to which others would naturally take the form of a directed graph.

The distinction between directed and undirected networks is not a mere technicality. It is fundamental to the analysis, as the applications and modeling of the two types are quite different. In particular, when links are necessarily reciprocal, then it is generally the case that joint consent is needed to establish and maintain the relationship. For instance, to form a trading partnership, both partners need to agree to it. To maintain a friendship the same is generally true, as it is for a business relationship or an alliance. In the case of directed networks, one individual may direct a link at another without the other's consent, which is generally true in citation networks or in links between web pages. These distinctions result in some basic differences in the modeling of network formation, as well as different conclusions about which networks will arise, which are optimal, and so on.

In what follows the default is that the network is undirected, and I mention explicitly when directed networks are considered. Let us begin with the formal definitions of graphs that represent networks.

A *graph* $(N, g)$ consists of a set of nodes $N = \{1, \ldots, n\}$ and a real-valued $n \times n$ matrix $g$, where $g_{ij}$ represents the (possibly weighted and/or directed) relation between $i$ and $j$. This matrix is often referred to as the *adjacency matrix,* as it lists which nodes are linked to each other, or in other words which nodes are adjacent to one another.[1] In the case in which the entries of $g$ take on more than two values and can track the intensity level of relationships, the graph is referred to as a *weighted* graph. Otherwise, it is standard to use the values of either 0 or 1, and the graph is *unweighted.* In much of what follows, $N$ will be fixed or given. Thus, I often refer to $g$ as being a network or graph.

A network is *directed* if it is possible that $g_{ij} \neq g_{ji}$, and a network is *undirected* $g_{ij} = g_{ji}$ for all nodes $i$ and $j$. Parts of the literature refer to directed graphs as *digraphs.*

For instance, if $N = \{1, 2, 3\}$, then

$$g = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{2.1}$$

is the (undirected and unweighted) network with a *link* between nodes 1 and 2, a link between nodes 2 and 3, but no link between nodes 1 and 3 (Figure 2.1). Nodes

1. There are more general graph structures that can represent the possibility of multiple relationships between different nodes; for instance, having different links for being friends, relatives, coworkers, and the like. These are sometimes referred to as a *multiplex networks.* One can also allow for relationships that involve more than two nodes at a time. For example, see Diestel [200] and Page and Wooders [518] for some more general representations.


FIGURE 2.1 A network with two links.

are also often referred to as *vertices* and links as *edges* or *ties*; links are sometimes called *arcs* in the case of directed graphs.

Self-links or *loops* often do not have any real meaning or consequence, and so whether we set $g_{ii} = 1$ or $g_{ii} = 0$ as a default is usually (but not always!) irrelevant. Unless otherwise indicated, assume that $g_{ii} = 0$ for all $i$.[2]

There are equivalent ways of representing a graph. Instead of viewing $g$ as an $n \times n$ matrix, it is sometimes easier to describe a graph by listing all links or edges in the graph. That is, we can view a graph as a pair $(N, g)$, where $g$ is the collection of links that are listed as a subsets of $N$ of size 2. For instance, the network $g$ in Figure 2.1 can be written as $g = \{\{1, 2\}, \{2, 3\}\}$, or simplifying notation a bit, $g = \{12, 23\}$. Thus $ij$ represents the link connecting nodes $i$ and $j$. Then we can write $ij \in g$ to indicate that $i$ and $j$ are linked under the network $g$; that is, writing $ij \in g$ is equivalent to writing $g_{ij} = 1$. I alternate between the different representations as is convenient. It will also be useful to write $g' \subset g$, to indicate that

$$\{ij : ij \in g'\} \subset \{ij : ij \in g\}.$$

Let the shorthand notation of $g + ij$ represent the network obtained by adding the link $ij$ to an existing network $g$, and, let $g - ij$ represent the network obtained by deleting the link $ij$ from the network $g$. We can represent directed networks in an analogous manner, viewing $ij$ as a directed link and distinguishing between $ij$ and $ji$. Let $G(N)$ be the set of all undirected and unweighted networks on $N$.

In some cases the specific identity of the node in a position in the network is of interest, and in other situations only the structure of the network is important. The idea that two networks or graphs have the same structure is captured through the concept of an isomorphism. The networks $(N, g)$ and $(N', g')$ are *isomorphic* if there exists a one-to-one and onto function (a bijection) $f : N \to N'$, such that $ij \in g$ if and only if $f(i)f(j) \in g'$. Thus, $f$ just relabels the nodes, and the networks are the same up to that relabeling.

Given a subset of nodes $S \subset N$ and a network $g$, let $g|_S$ denote the network $g$ restricted to the set of nodes $S$, so that

$$[g|_S]_{ij} = \begin{cases} 1 & \text{if } i \in S,\ j \in S,\ g_{ij} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus $g|_S$ is the network obtained by deleting all links except those that are between nodes in $S$. An example is pictured in Figure 2.2.

---

2. Sometimes graphs without any self-links (and without multiple links) are referred to as *simple graphs.* Unless otherwise stated, the term graph refers to a simple graph in this book. If self-links and multiple links between nodes are permitted, the resulting structure is termed a *multigraph.*

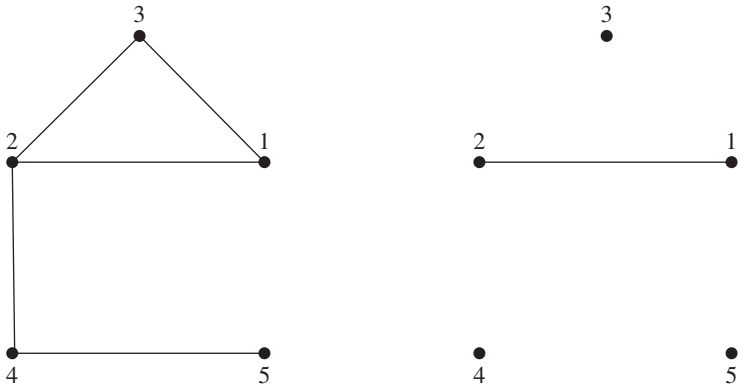


**FIGURE 2.2** A network and the network restricted to $S = \{1, 2, 5\}$.

For any network $g$, let $N(g)$ be the set of nodes that have at least one link in the network $g$. That is, $N(g) = \{i | \exists j \text{s.t. } ij \in g, \text{ or } ji \in g\}$.[3]

## 2.1.3 Paths and Cycles

Much of the interest in networked relationships comes from the fact that individual nodes benefit (or suffer) from indirect relationships. Friends might provide access to favors from their friends, and information might spread through the links of a network. To capture the indirect interactions in a network, it is important to model paths through the network. In the case of an undirected network, a path is an obvious object. As there are multiple definitions in the case of a directed network, I return to those after providing definitions for an undirected network.

A *path* in a network $g \in G(N)$ between nodes $i$ and $j$ is a sequence of links $i_1i_2, i_2i_3, \ldots, i_{K-1}i_K$ such that $i_ki_{k+1} \in g$ for each $k \in \{1, \ldots, K-1\}$, with $i_1 = i$ and $i_K = j$, and such that each node in the sequence $i_1, \ldots, i_K$ is distinct.[4] A *walk* in a network $g \in G(N)$ between nodes $i$ and $j$ is a sequence of links $i_1i_2, \ldots, i_{K-1}i_K$ such that $i_ki_{k+1} \in g$ for each $k \in \{1, \ldots, K-1\}$, with $i_1 = i$ and $i_K = j$. The distinction between a path and a walk is whether all involved nodes are distinct. A walk may come back to a given node more than once, whereas a path is a walk that never hits the same node twice.[5]

---

3. In this case it matters whether $g_{ii} = 1$, in which case $i \in N(g)$, or whether $g_{ii} = 0$, in which case $i \notin N(g)$.

4. A path may also be defined to be a subnetwork that consists of the set of involved nodes and the set of links between these nodes.

5. The definition of a path given here is the standard one from the graph theory literature. In some of the network literature, the term *path* is used more loosely and can refer to a walk, so that nodes can be visited more than once. This ambiguity can cause some confusion, which should be borne in mind when reading the literature.

A *cycle* is a walk $i_1 i_2, \ldots, i_{K-1} i_K$ that starts and ends at the same node (so $i_1 = i_K$) and such that all other nodes are distinct ($i_k \neq i_{k'}$ when $k < k'$ unless $k = 1$ and $k' = K$). Thus a cycle is a walk such that the only node that appears more than once is the starting/ending node. A cycle can be constructed from any path by adding a link from the end to the starting node; and conversely, deleting the first or last link of a cycle results in a path.

A *geodesic* between nodes $i$ and $j$ is a shortest path between these nodes; that is, a path with no more links than any other path between these nodes.

To summarize:

- A walk is a sequence of links connecting a sequence of nodes.
- A cycle is a walk that starts and ends at the same node, with all nodes appearing once except the starting node, which also appears as the ending node.
- A path is a walk in which a node appears at most once in the sequence.
- A geodesic between two nodes is a shortest path between them.

Note that for the convention of setting $g_{ii} = 0$, then $g^2 = g \times g$ tells us how many walks there are of length 2 between any two nodes.

For instance for the network

$$g = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix},$$

$g^2$ is

$$g^2 = \begin{pmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{pmatrix}.$$

So, for instance there are two walks between 1 and 4 of length 2 (passing between 2 and 3, respectively). There are two walks from 1 back to 1 (passing through 2 and 3, respectively). For this network $g^3$ is

$$g^3 = \begin{pmatrix} 0 & 4 & 4 & 0 \\ 4 & 0 & 0 & 4 \\ 4 & 0 & 0 & 4 \\ 0 & 4 & 4 & 0 \end{pmatrix}.$$

There are four walks of length 3 between 1 and 2 (namely, (12,24,42), (13,34,42), (12,21,12), and (13,31,12)). Note that some walks have cycles in them (and hence the use of the term *walk* rather than *path*). The $k$th power of the network, $g^k$, keeps track of all possible walks of length $k$ between any two nodes, including walks with many cycles within them.

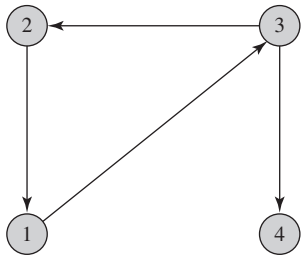


**FIGURE 2.3** Directed path from 2 to 4 (via 1 and 3), directed cycle from 1 to 3 to 2 to 1, and directed walk from 3 to 2 to 1 to 3 to 4.

### 2.1.4 Directed Paths, Walks, and Cycles

In the case of directed networks, there are different possible definitions of paths and cycles. The definitions depend on whether we want to keep track of the direction of the links, and in various applications depend on whether communication is restricted to following the direction of the links or can move in both directions along a directed link, as for example in a network of links between web pages.

In the case in which direction is important, the definitions are just as stated above for undirected networks, but the ordering of the nodes in each link now takes on an important role. For instance, we might be interested in knowing whether one web page can be found from another by following (directed) links starting from one page and leading to the other. This case deals with directed paths, directed walks, and directed cycles.

A *directed walk* in a network $g \in G(N)$ is a sequence of links $i_1i_2, \ldots, i_{K-1}i_K$ such that $i_ki_{k+1} \in g$ (that is, $g_{i_ki_{k+1}} = 1$) for each $k \in \{1, \ldots, K-1\}$.

A *directed path* in a directed network $g \in G(N)$ from node $i$ to node $j$ is a sequence of links $i_1i_2, \ldots, i_{K-1}i_K$ such that $i_ki_{k+1} \in g$ (that is, $g_{i_ki_{k+1}} = 1$) for each $k \in \{1, \ldots, K-1\}$, with $i_1 = i$ and $i_K = j$, such that each node in the sequence $i_1, \ldots, i_K$ is distinct.

A *directed cycle* in a network $g \in G(N)$ is a sequence of links $i_1i_2, \ldots, i_{K-1}i_K$ such that $i_ki_{k+1} \in g$ (that is, $g_{i_ki_{k+1}} = 1$) for each $k \in \{1, \ldots, K-1\}$, with $i_1 = i_K$.

These definitions are illustrated in Figure 2.3.

In cases where the direction of the link just indicates who initiated the link, but where links can conduct in both directions, we can keep track of undirected paths. There we think of $i$ and $j$ being linked if either $g_{ij} = 1$ or $g_{ji} = 1$. In that case, we can simply define the undirected network that comes from considering $i$ and $j$ to be linked if there is a directed link in either direction. In general, I refer to such paths, walks, and cycles as undirected.

To be more specific, given a directed network $g$ let $\widehat{g}$ denote the undirected network obtained by allowing an undirected link for each directed one present in $g$. That is, let $\widehat{g}_{ij} = \max(g_{ij}, g_{ji})$. Then we say that there is an *undirected path* between nodes $i$ and $j$ in $g$ if there is a path between them in $\widehat{g}$. An undirected cycle or walk is defined in a similar fashion. In Figure 2.3 there is no directed path from node 4 to any other node, but there is an undirected path from node 4 to each of the other nodes.

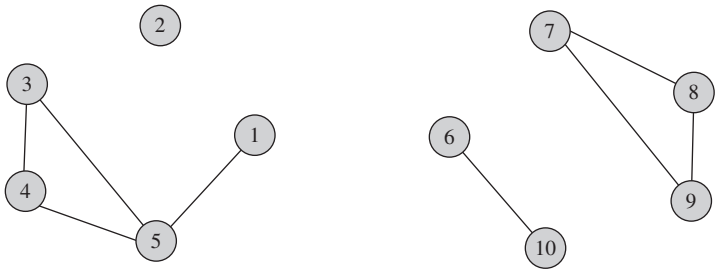


**FIGURE 2.4** A network with four components.

### 2.1.5 Components and Connected Subgraphs

In many applications it is important to track which nodes can reach which other nodes through paths in the network. This tracking ability plays a critical role in phenomena like contagion, learning, and the diffusion of various behaviors through a social network. Looking at the path relationships in a network naturally partitions a network into different connected subgraphs that are commonly referred to as *components*. Again, definitions are first provided for undirected networks and later for directed ones.

A network $(N, g)$ is *connected* (or path-connected) if every two nodes in the network are connected by some path in the network. That is, $(N, g)$ is connected if for each $i \in N$ and $j \in N$ there exists a path in $(N, g)$ between $i$ and $j$.

A *component* of a network $(N, g)$ is a nonempty subnetwork $(N', g')$ such that $\emptyset \neq N' \subset N$, $g' \subset g$,

- $(N', g')$ is connected, and
- if $i \in N'$ and $ij \in g$, then $j \in N'$ and $ij \in g'$.

Thus the components of a network are the distinct maximal connected subgraphs of a network. In the network shown in Figure 2.4 there are four components: the node 2 together with an empty set of links, the nodes {1, 3, 4, 5} together with links {15, 35, 34, 45}, the nodes 6 and 10 together with the link {6–10}, and the nodes {7, 8, 9} together with the links {78, 79, 89}. Note that under this definition of component, a completely isolated node that has no links is considered a component.[6]

The set of components of a network $(N, g)$ is denoted $C(N, g)$. In cases for which $N$ is fixed or obvious, I simply denote the components by $C(g)$. The component containing a specific node $i$ is denoted $C_i(g)$.

Components of a network partition the nodes into groups within which nodes are path-connected. Let $\Pi(N, g)$ denote the partition of $N$ induced by the network $(N, g)$. That is, $S \in \Pi(N, g)$, if and only if $(S, h) \in C(N, g)$ for some $h \subset g$. For example, the network in Figure 2.4 induces the partition $\Pi(N, g) =$

6. This inclusion is a matter of convention, and one can also find definitions of components that only allow for subnetworks with links.

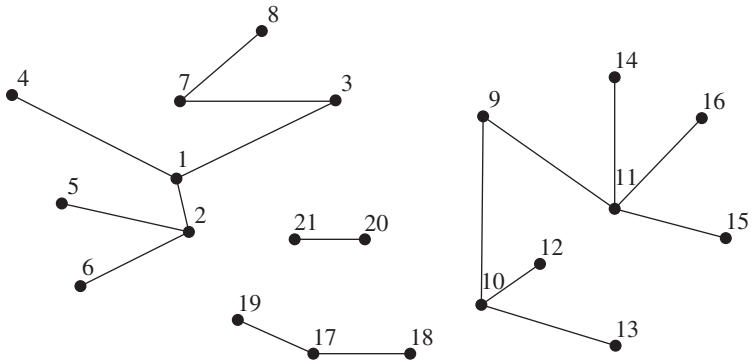


**FIGURE 2.5** Four trees in a forest.

{{1, 3, 4, 5}, {2}, {6, 10}, {7, 8, 9}} over the set of nodes. Thus, a network is connected if and only if it consists of a single component (so that $\Pi(N, g) = \{N\}$).

A link $ij$ is a *bridge* in the network $g$ if $g - ij$ has more components than $g$.[7]

In the case of a directed network, there are again several different approaches to defining those terms. One way is to again ignore the directed nature of links and to consider the undirected network that has a link present if one is present in either direction. This method defines one notion of connection and components. In some applications in which direction is important, for instance in the transmission of information, we want to keep track of the directed nature of the network. In such cases, I refer to *strongly connected* graphs or subgraphs, so that each node can reach every other one by a directed path. Further definitions are specified as needed.

### 2.1.6 Trees, Stars, Circles, and Complete Networks

There are a few particular network structures that are commonly referred to.

A *tree* is a connected network that has no cycles.

A *forest* is a network such that each component is a tree. Thus any network that has no cycles is a forest, as in the example pictured in Figure 2.5.

A particularly prominent forest network is the star. A *star* is a network in which there exists some node $i$ such that every link in the network involves node $i$. In this case, $i$ is referred to as the *center* of the star.

Here are a few facts about trees that are easy to derive (see Exercise 2.2) and are worth mentioning.

- A connected network is a tree if and only if it has $n - 1$ links.
- A tree has at least two leaves, where leaves are nodes that have exactly one link.
- In a tree, there is a unique path between any two nodes.

7. There are variations on this definition, with some requiring that the components connected by the bridge both involve more than one node.

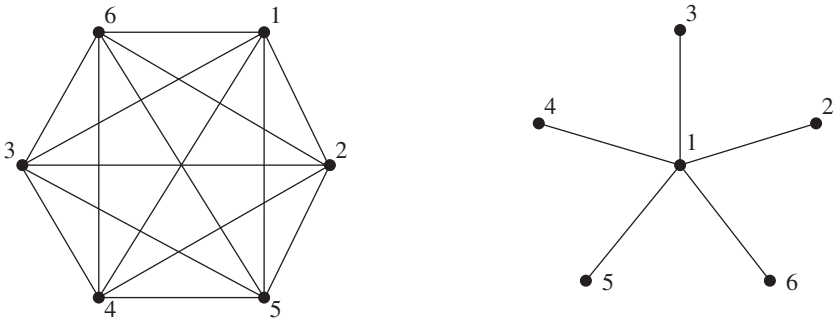


**FIGURE 2.6** A complete network on six nodes and a star network on six nodes.

A *complete network* is one in which all possible links are present, so that $g_{ij} = 1$ for all $i \neq j$ (see Figure 2.6).

A *circle* (also known as a *cycle-graph*) is a network that has a single cycle and is such that each node in the network has exactly two neighbors.

In the case of directed networks, there can be many different stars involving the same set of nodes and having the same center, depending on which directed links are present between any two linked nodes. On occasion it is useful to distinguish between these stars, for instance, indicating whether links go in or out from the center node.

### 2.1.7 Neighborhood

The *neighborhood of a node* $i$ is the set of nodes that $i$ is linked to.[8]

$$N_i(g) = \{j : g_{ij} = 1\}.$$

Given some set of nodes $S$, the *neighborhood of* $S$ is the union of the neighborhoods of its members. That is

$$N_S(g) = \bigcup_{i \in S} N_i(g) = \{j : \exists i \in S,\, g_{ij} = 1\}.$$

We can also talk about extended neighborhoods of a node, for instance of all the nodes that can be reached by walks of length no more than 2, and so on. The *two-neighborhood of a node* $i$ is

$$N_i^2(g) = N_i(g) \cup \left( \bigcup_{j \in N_i(g)} N_j(g) \right).$$

8. Note that whether $i$ is in $i$'s neighborhood depends on whether $g_{ii} = 1$ is allowed. As I am following a default convention of $g_{ii} = 0$, $i$ is generally not considered to be in $i$'s neighborhood. This definition ensures that $i$'s degree is the number of other nodes that $i$ is linked to, which is then the cardinality of $i$'s neighborhood.

Inductively, all nodes that can be reached from $i$ by walks of length no more than $k$ make up the *k-neighborhood* of $i$, which can be defined by

$$N_i^k(g) = N_i(g) \cup \left( \bigcup_{j \in N_i(g)} N_j^{k-1}(g) \right).$$

Similar definitions of $k$-neighborhoods hold for any set of nodes $S$, so that $N_S^k(g) = \cup_{i \in S} N_i^k(g)$ is the set of nodes that can be reached from some node in $S$ by a walk of length no more than $k$. Generally the *extended neighborhood* of a node $i$ is all of the nodes it is walk-connected to, or $N_i^n(g)$.

The above definitions also work for directed networks, in which case the nodes in $N_i^k(g)$ are those nodes that can be reached from $i$ by a directed walk.

### 2.1.8 Degree and Network Density

The *degree* of a node is the number of links that involve that node, which is the cardinality of the node's neighborhood. Thus node $i$'s degree in a network $g$, denoted $d_i(g)$, is

$$d_i(g) = \#\{j : g_{ji} = 1\} = \#N_i(g).$$

In the case of a directed network, the above calculation is the node's *in-degree.* The *out-degree* of node $i$ is the corresponding calculation $\#\{j : g_{ij} = 1\}$. These definitions coincide for an undirected network (see Figure 2.7). The *density* of a network keeps track of the relative fraction of links that are present, and is simply the average degree divided by $n - 1$.

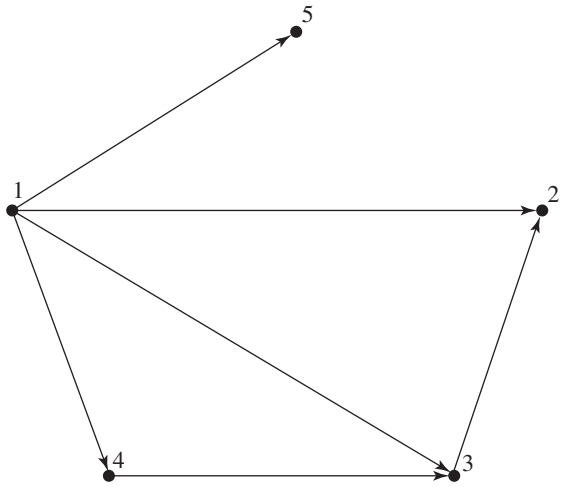


**FIGURE 2.7** A directed network on five nodes. Node 1 has in-degree 2 and out-degree 4.

## 2.2 ▪ Some Summary Statistics and Characteristics of Networks

While a small network can be usefully described directly by its graph $g$ and is easily illustrated in a figure, larger networks can be more difficult to envision and describe. Moreover, it is important to be able to compare networks and classify them according to their properties and thus to have a stable of summary statistics that provide meaningful insight into their structures.

### 2.2.1 Degree Distributions

A fundamental characteristic of a network is its degree distribution. The *degree distribution* of a network is a description of the relative frequencies of nodes that have different degrees. That is, $P(d)$ is the fraction of nodes that have degree $d$ under a degree distribution $P$.[9]

For instance, a *regular* network is one in which all nodes have the same degree. A network is *regular of degree* $k$ if $P(k) = 1$ and $P(d) = 0$ for all $d \neq k$. Such a network is quite different from the random network described in Section 1.2.3, in which there is a great deal of heterogeneity in the degrees of nodes, and the distribution is a Poisson distribution.

Beyond the degenerate degree distribution associated with a regular network, and the Poisson degree distribution associated with Poisson random networks discussed in Section 1.2.3, another prominent distribution is what is referred to as a *scale-free* degree distribution. These distributions date to Pareto [526], and they appear in a wide variety of settings including networks describing incomes, word usage, city populations, and degrees in networks (as is discussed in more detail in Chapter 3).[10]

A *scale-free distribution* (or power distribution) $P(d)$ satisfies[11]

$$P(d) = cd^{-\gamma}, \tag{2.2}$$

where $c > 0$ is a scalar (which normalizes the support of the distribution to sum to 1).[12] Thus if we increase the degree by a factor $k$, then the frequency is reduced by a factor of $k^{-\gamma}$. As this is true regardless of the starting degree $d$, the relative probabilities of degrees of a fixed relative ratio are the same independent of the scale of those degrees. That is, $P(2)/P(1)$ is the same as $P(20)/P(10)$. Hence the term *scale-free*. Scale-free distributions are often said to exhibit a *power law*, with reference to the power function $d^{-\gamma}$.

---

9. $P$ can be a frequency distribution if we are describing data, or it can be a probability distribution if we are working with random networks.

10. For an informative overview, see Mitzenmacher [470].

11. One has to be careful about defining the value at $d = 0$, as it might not be well defined; so let us keep track of nodes with degree at least 1.

12. When the support is $\{1, 2, \ldots\}$, then the scalar is the inverse of what is known as the Riemann zeta function, $z(\gamma) = \sum_{d=1}^{\infty} \frac{1}{d^{\gamma}}$.

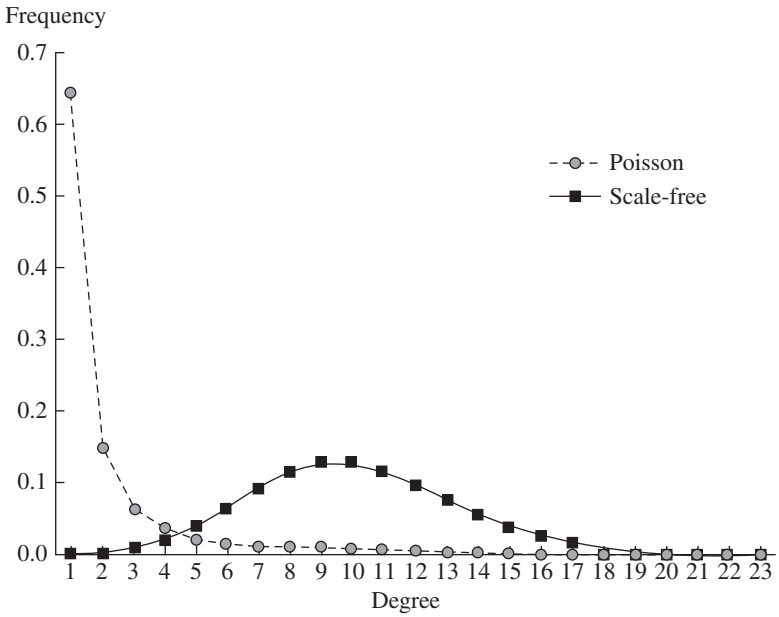


**FIGURE 2.8** Comparing a scale-free distribution to a Poisson distribution.

Generally, given a degree distribution $P$, let $\langle d \rangle_P$ denote the expected value of $d$, and $\langle d^2 \rangle_P$ denote the expectation of the square of the degree, and so on. I often omit the $_P$ notation when $P$ is fixed.

Scale-free distributions have "fat tails." That is, there tend to be many more nodes with very small and very large degrees than one would see if the links were formed completely independently so that degree followed a Poisson distribution. We can see this comparison in Figure 2.8, which shows plots of these degree distributions when the average degree is 10. The figure compares the Poisson degree distribution from (1.4) with the scale-free distribution from (2.2).

The fatter tail of the scale-free distribution is obvious in the lower tail (for lower degrees), while for higher degrees it is harder to see the differences. If we convert the plot to a log-log plot (i.e., log(frequency) versus log(degree) instead of the raw numbers), then the differences in the upper tail (for higher degrees) become more evident (Figure 2.9).

Figure 2.9 points out another interesting aspect of scale-free distributions: they are linear when plotted on a log-log plot. That is, we can rewrite (2.2) by taking logs of both sides to obtain:

$$\log(f(d)) = \log(c) - \gamma \log(d).$$

This form is useful when trying to estimate $\gamma$ from data, as then a linear regression can be used.

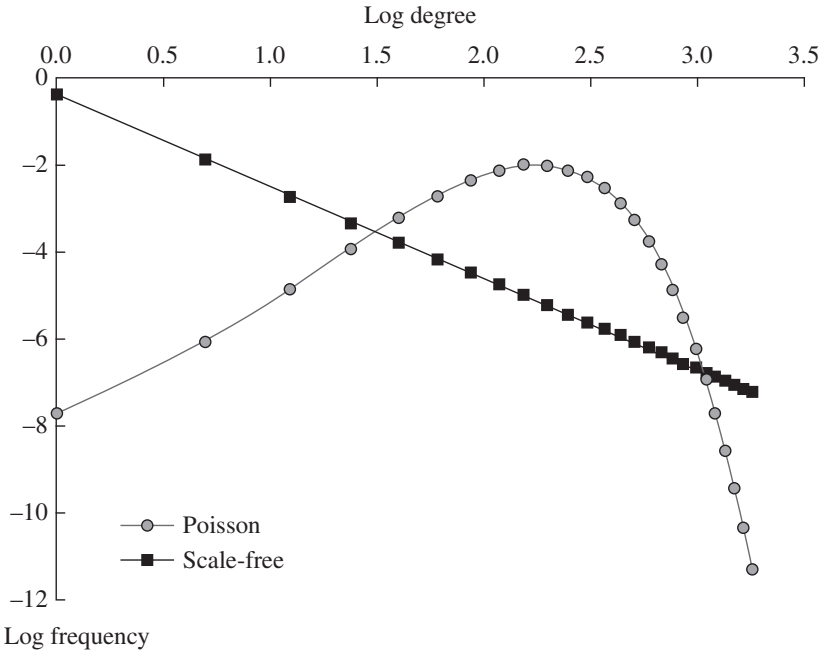


**FIGURE 2.9** Comparing a scale-free distribution to a Poisson distribution: log-log plot.

## 2.2.2 Diameter and Average Path Length

The *distance* between two nodes is the length of (number of links in) the shortest path or *geodesic* between them. If there is no path between the nodes, then the distance between them is infinite. This concept leads us to another important characteristic of a network: its diameter. The *diameter* of a network is the largest distance between any two nodes in the network.[13]

To see how diameter can vary across networks with the same number of nodes and links, consider two different networks in which each node has on average two links, as in Figure 2.10. The first network is a circle, and the second is a tree. Even though both networks have approximately an average degree of 2, they are clearly very different in structure. The degree distribution reflects some aspect of the difference in that the circle is regular, so that every node has exactly two links, while in the binary tree almost half of the nodes have degree 3 and nearly half have degree 1 (the exception is the root node, which has degree 2). However, we need other measures to clearly distinguish these networks. For instance, the diameter of

13. Related measures, working with cycles rather than paths, are the girth and circumference of a network. The *girth* is the length of the smallest cycle in a network (set to infinity if there are no cycles), and the *circumference* is the length of the largest cycle (set to 0 if there are no cycles).

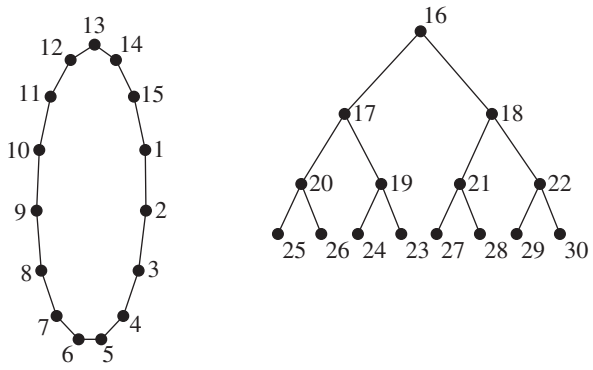


**FIGURE 2.10** Circle and tree.

a circle of $n$ nodes is either $n/2$ or $(n-1)/2$, while the diameter of a binary tree of $n$ nodes is roughly $2\log_2(n+1) - 2$.[14]

The diameter is one measure of path length, but it only provides an upper bound. *Average path length* (also referred to as *characteristic path length*) between nodes is another measure that captures related properties. The average is taken over geodesics, or shortest paths. Clearly, the average path length is bounded above by the diameter; in some cases it can be much shorter than the diameter. Thus, it is often useful to see whether the diameter is being determined by a few outliers (literally), or whether it is of the same order as the average geodesic.

Many networks are not fully connected and may consist of a number of separate components. In such cases, one often reports the diameter and average path length in the largest component, being careful to specify whether that component is a giant component (containing a nontrivial fraction of the networks nodes).[15]

Recalling that raising the adjacency matrix $g$ to a power $k$ provides as its $ij$th entry the number of walks of length $k$ between nodes $i$ and $j$, we can easily calculate shortest path lengths. That is, the shortest path length between nodes $i$ and $j$ can be found by finding the smallest $\ell$ such that the $ij$th entry $g^\ell$ is positive: that entry is the number of shortest paths between those nodes. The same calculation provides shortest directed paths in the case of directed networks.

---

14. This measurement holds precisely if there is an integer $K$ such that $n = 2^K - 1$ in the case of a binary tree.

15. There is a way to circumvent these problems. As Newman [503] suggests, the measure

$$\frac{n(n+1)}{2\sum_{ij}\frac{1}{\ell(i,j,g)}},$$

where $\ell(i, j, g)$ is the length of the shortest path between $i$ and $j$ in $g$ and is set to infinity if the nodes are not connected. This measure can be calculated regardless of component structure. So rather than averaging path lengths, one looks at the reciprocal of the average of the reciprocal path lengths. Taking the reciprocal twice leads to something similar to averaging path lengths directly, but working with the reciprocals eliminates the influence of infinite path lengths.

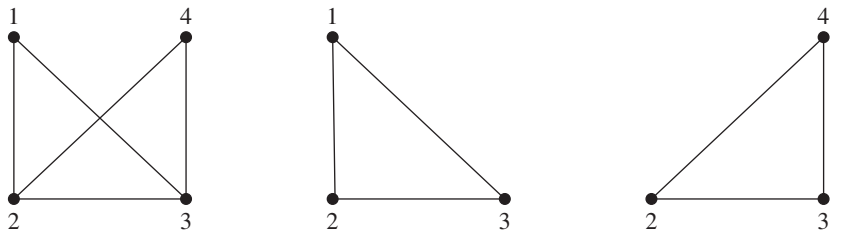


**FIGURE 2.11** A network on four nodes and its two cliques.

Calculating shortest path lengths for all pairs of nodes, through successive powers of the adjacency matrix $g$, then provides a basic method of calculating diameter. There are more computationally efficient algorithms for calculating or estimating diameter,[16] and most network software programs include such calculations as built-in features.

### 2.2.3 Cliquishness, Cohesiveness, and Clustering

One fascinating and important aspect of social networks is how tightly clustered they are. For example, the extent to which my friends are friends with one another captures one facet of this clustering. There are a variety of concepts that measure how cohesive or closely knit a network is.

An early concept related to this is the idea of a clique. A *clique* is a maximal completely connected subnetwork of a given network.[17] That is, if some set of nodes $S \subset N$ are such that $g|_S$ is the complete network on the nodes $S$, and for any $i \in N \setminus S$ $g|_{S\cup\{i\}}$ is not complete, then the nodes $S$ are said to form a clique.[18] Cliques are generally required to contain at least three nodes; otherwise each link could potentially define a clique of two nodes. Note that a given node can be part of several cliques at once. For example, in Figure 2.11 both nodes 2 and 3 are in two different cliques.

One measure of cliquishness is to count the number and size of the cliques in a network. One difficulty with this measure is that removing one link from a large clique can change the clique structure completely. For instance, removing one link from a complete network among four nodes changes the clique structure from having one clique involving four nodes to two cliques of three nodes. More generally, the clique structure is very sensitive to slight changes in a network.

---

16. For instance, there are more efficient ways of calculating powers of $g$ when it is diagonalizable (see Section 2.4.1). Computational efficiency can be important when $n$ is large.

17. Note the distinction between a clique and a component. A clique must be completely connected and not be a strict subset of any subnetwork that is completely connected, while a component must be path-connected and not be a strict subset of any subnetwork that is path-connected. Neither implies the other.

18. An early definition of this is from Luce and Perry [443].

The most common way of measuring some aspect of cliquishness is based on transitive triples or clustering.[19] Examining undirected networks, the most basic clustering measure is simply to perform the following exercise. Look at all situations in which two links emanate from the same node (e.g., $ij$ and $ik$ both involve node $i$) and ask how often $jk$ is then also in the network. So if $i$ has relationships with both $j$ and $k$, how likely on average is it that $j$ and $k$ are related in the network? This clustering measure is represented by

$$Cl(g) = \frac{\sum_i \#\{jk \in g | k \neq j,\ j \in N_i(g),\ k \in N_i(g)\}}{\sum_i \#\{jk | k \neq j,\ j \in N_i(g),\ k \in N_i(g)\}} = \frac{\sum_{i;j\neq i;k\neq j;k\neq i} g_{ij}g_{ik}g_{jk}}{\sum_{i;j\neq i;k\neq j;k\neq i} g_{ij}g_{ik}}.$$

I will often refer to this as *overall* clustering to distinguish it from the other measures of clustering that follow.

Another measure that has also been used in the literature is similar to the clustering coefficient $Cl(g)$, except that instead of considering the fraction of fully connected triples out of the potential triples in which at least two links are present, the measure is computed on a node-by-node basis and then averaged across nodes. This measure is based on the following definition of *individual clustering for a node i*:

$$Cl_i(g) = \frac{\#\{jk \in g | k \neq j,\ j \in N_i(g),\ k \in N_i(g)\}}{\#\{jk | k \neq j,\ j \in N_i(g),\ k \in N_i(g)\}} = \frac{\sum_{j\neq i;k\neq j;k\neq i} g_{ij}g_{ik}g_{jk}}{\sum_{j\neq i;k\neq j;k\neq i} g_{ij}g_{ik}}.$$

Thus, $Cl_i(g)$ looks at all pairs of nodes that are linked to $i$ and then considers how many of them are linked to one another.[20] Another way to write the individual clustering coefficient is then

$$Cl_i(g) = \frac{\#\{jk \in g | k \neq j,\ j \in N_i(g),\ k \in N_i(g)\}}{d_i(g)(d_i(g) - 1)/2}.$$

The *average clustering* coefficient is then

$$Cl^{Avg}(g) = \sum_i Cl_i(g)/n.$$

Note that this calculation is different from that for the overall clustering coefficient $Cl(g)$, where the average is taken over all triples. Under average clustering, one computes a clustering for each node and then averages across nodes. This method gives more weight to low-degree nodes than does the clustering coefficient method.

As an illustration of these two measures, let us compute them relative to the Florentine marriage network pictured in Figure 1.1. To compute the overall

---

19. Clustering in the sense used here comes from the recent random network literature (e.g., see Newman [503]). *Clustering* has an interesting history as a term, growing out of the earlier sociology literature and based on partitioning signed graphs into subsets in which nodes within elements of the partition have only positive relationships between them, and only negative relationships exist across elements of the partition (e.g., see Chapter 6 in Wasserman and Faust [650]).

20. A convention is to set $Cl_i(g) = 0$ if $i$ has no more than one link.

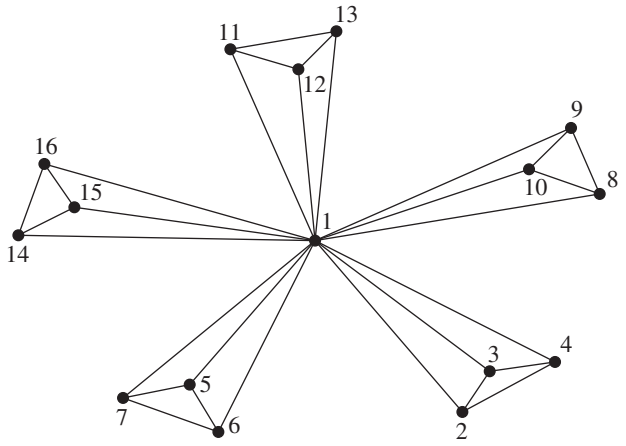


**FIGURE 2.12** Differences in clustering measures.

clustering coefficient, first count how many configurations of the form $ij$, $jk$ there are in the network. For instance, there is one with Medici-Barbadori and Barbadori-Castellan. There are three with the Ridolfi at the center (Strozzi-Ridolfi-Tornabuon, Strozzi-Ridolfi-Medici, and Tornabuon-Ridolfi-Medici). There are fifteen with the Medici at the center, and so forth. Totaling across all such configurations, we find that there are 47 such configurations in the network. Out of those 47 configurations, 9 of them are completed. Thus, $Cl(g) = 9/47$. In terms of the average clustering, compute the clustering for each family separately. For instance, the Barbadori have one possible pair and they are not connected, so the Barbadori have $Cl_{\text{Barbadori}}(g) = 0/1 = 0$. The Pazzi, Acciaiuol, Ginori, Lambertes, and Pucci are all 0 by convention (they each have one or no links). The Bischeri, Castellan, Ridolfi, and Tornabuon each have clustering 1/3. The Strozzi are also 1/3 (2 out of 6). The Peruzzi are 2/3, the Medici 1/15, and the Guadagni and Albizzi have at least two neighbors each, but still have clusterings of 0.[21] When we average across all of these, we get $Cl^{\text{Avg}}(g) = .15$ or 3/20. This value is a bit less than the overall clustering $Cl(g)$, as a number of 0s are included in the average clustering.

The above calculations show that it is possible for these two common measures of clustering to differ. While in that example the average clustering is less than the overall clustering, it can also go the other way. Moreover, it is not uncommon to generate networks in which the two measures produce very different values. For instance, consider the following variation of a star network. Begin with a large number of triads (complete networks among three nodes), and then add a center node, to which every other node is connected (Figure 2.12).

As the number of nodes involved gets large, average clustering goes to 1, while overall clustering goes to 0! To see this note that all of the nodes other than the center node have individual clustering measures of 1. Thus when averaged the average clustering coefficient converges to 1. However, for the overall clustering

21. This relates back to the important role of the Medici, as many of their neighbors were not directly connected but were connected only indirectly through the Medici.

coefficient, each time that a new triad is added the number of possible pairs of links goes up by 3 times the number of links the center node already has (plus 12), while the number of those pairs that are completed only increases by 12. Thus overall clustering goes to 0. Clearly the two measures are capturing different aspects of clustering and so there is no "right" or "wrong" measure. This example shows that such simple coefficients cannot give a full picture of the interrelatedness of a network but only an impression of some aspect of it.

In the case of directed networks one has further choices for measuring clustering. One option is simply to ignore the direction of a link and consider two nodes to be linked if there is a directed link in either direction between them. Based on this derived undirected network, one can then apply the above measures of overall and average clustering. A different approach is to keep track of the percentage of *transitive triples.* This approach considers situations in which node $i$ has a directed link to $j$, and $j$ has a directed link to $k$, and then asks whether $i$ has a directed link to $k$ (i.e., the usual notion of transitivity of relationships).[22] The fraction of times in a network that the answer is "yes" is the *fraction of transitive triples:*

$$Cl^{TT}(g) = \frac{\sum_{i;j\neq i;k\neq j} g_{ij}g_{jk}g_{ik}}{\sum_{i;j\neq i;k\neq j} g_{ij}g_{jk}}.$$

The above fraction of transitive triples is a standard measure, but much of the empirical literature has instead simply ignored the directed nature of relationships.[23]

### 2.2.4 Centrality

Most of the measures discussed to this point are predominately macro in nature; that is, they describe broad characteristics of a network. In many cases, we might also be interested in micro measures that allow us to compare nodes and to say something about how a given node relates to the overall network. For instance, as we saw in the Florentine marriage example in Section 1.2.1, the idea of how central a node is can be very important. In particular, notions that somehow capture a node's position in a network are useful. As such, many different measures of centrality have been developed, and they each tend to capture different aspects of the concept, which can be useful when working with information flows, bargaining power, infection transmission, influence, and other sorts of important behaviors on a network.

Measures of centrality can be categorized into four main groups depending on the types of statistics on which they are based:[24]

1. degree—how connected a node is;
2. closeness—how easily a node can reach other nodes;

---

22. Alternatively, one could examine the percentage of times that $k$ has a directed link to $i$ so that a directed cycle emerges. This calculation can yield very different results, depending on the context.

23. There are also hybrid measures (mixing ideas of directed and undirected links) in which one counts the percentage of possible directed links among a node's direct neighbors that are present on average, as in, for example, Adamic's [2] study of the world wide web.

24. See Borgatti [94] for more discussion on categorizing measures of centrality.

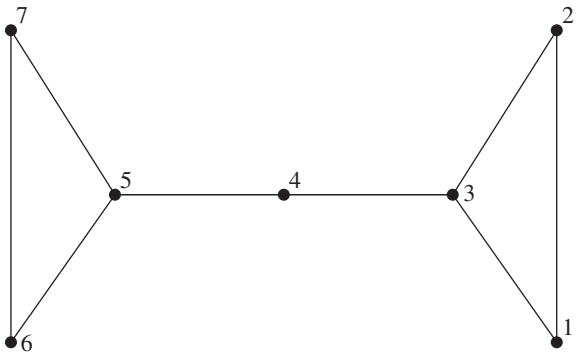


**FIGURE 2.13** A central node with low degree centrality.

3. betweenness—how important a node is in terms of connecting other nodes; and
4. neighbors' characteristics—how important, central, or influential a node's neighbors are.

Given how different these notions are, even without looking at formal definitions it is easy to see that they capture complementary aspects of a node's position, and any particular measure will be better suited for some applications and less appropriate for others. Let me discuss some of the more standard definitions of each type.

***Degree Centrality*** Perhaps the simplest measure of the position of a given node in a network is simply to keep track of its degree. A node with degree $n - 1$ would be directly connected to all other nodes, and hence quite central to the network. A node connected to only two other nodes (for large $n$) would be, at least in one sense, less central. The *degree centrality* of a node is simply $d_i(g)/(n - 1)$, so that it ranges from 0 to 1 and indicates how well a node is connected in terms of direct connections.

Of course, degree centrality clearly misses many of the interesting aspects of a network. In particular, it does not measure how well located a node is in a network. It might be that a node has relatively few links, but lies in a critical location in the network. For many applications a centrality measure that is sensitive to a node's influence or marginal contribution to the network is important. For example, consider the network in Figure 2.13.

In this network the degree of nodes 3 and 5 are three, and the degree of node 4 is only two. Arguably, node 4 is at least as central as nodes 3 and 5, and far more central than the other nodes that each have two links (nodes 1, 2, 6, and 7). There are several senses in which we see a powerful or central role for node 4. If one deletes node 4, the component structure of the network changes. This change might be very important for applications involving information transmission, where node 4 is critical to path-connecting nodes 1 and 7. This aspect would be picked up by a measure such as betweenness. We also see that node 4 is relatively close to all other nodes in that it is at most two links away from any other node, whereas

every other node has at least one node linked to it a distance of three or more. This aspect would be important in applications in which something is being conveyed or transmitted through the network (e.g., an opinion or favor), and there is a decay of the connection strength with distance. In that case, being closer can either help a node make use of other nodes (e.g., having access to favors) or can enhance its influence (e.g., conveying opinions). Thus we are brought to the next category of centrality measures.

***Closeness Centrality*** This second class of measures tracks how close a given node is to any other node. One obvious closeness-based measure is just the inverse of the average distance between $i$ and any other node $j$: $(n-1)/\sum_{j\neq i}\ell(i,j)$, where $\ell(i,j)$ is the number of links in the shortest path between $i$ and $j$. There are various conventions for handling networks that are not connected, as well as other possible measures of distance, which leads to a family of closeness measures.

A richer way of measuring centrality based on closeness is to consider a decay parameter $\delta$, where $1>\delta>0$ and then consider the proximity between a given node and every other node weighted by the decay. In particular, let the *decay centrality* of a node be defined as

$$\sum_{j\neq i}\delta^{\ell(i,j)},$$

where $\ell(i,j)$ is set to infinity if $i$ and $j$ are not path-connected. This centrality measure is related to the symmetric connections model of Jackson and Wolinsky [361], as it is just the benefit that a node receives in that model of a network. As $\delta$ approaches 1, it is easy to see that decay centrality measures how large a component a node lies in. As $\delta$ approaches 0, then decay centrality gives infinitely more weight to closer nodes than farther ones, so it becomes proportional to degree centrality. For intermediate values of $\delta$, a node is rewarded for how close it is to other nodes, but in a way that very distant nodes are weighted less than closer nodes.

***Betweenness Centrality*** A measure of centrality that is based on how well situated a node is in terms of the paths that it lies on was first proposed by Freeman [255]. We first discussed it in Section 1.2.1 in the context of the Florentine marriages (see Figure 1.1).

Let $P_i(kj)$ denote the number of geodesics (shortest paths) between $k$ and $j$ that $i$ lies on, and let $P(kj)$ be the total number of geodesics between $k$ and $j$. We can estimate how important $i$ is in terms of connecting $k$ and $j$ by looking at the ratio $P_i(kj)/P(kj)$. If this ratio is close to 1, then $i$ lies on most of the shortest paths connecting $k$ to $j$, while if it is close to 0, then $i$ is less critical to $k$ and $j$. Averaging across all pairs of nodes, the *betweenness centrality* of a node $i$ is

$$Ce_i^B(g)=\sum_{k\neq j:i\notin\{k,j\}}\frac{P_i(kj)/P(kj)}{(n-1)(n-2)/2}.$$

For the network in Figure 2.13, we find that $Ce_4^B(g)=9/15$, $Ce_3^B(g)=8/15$, and $Ce_1^B(g)=0$. These values make it clear that nodes 3, 4, and 5 are much more central than the other nodes, and that 4 is the most central node in terms of connecting the other pairs of nodes.

### *Prestige-, Power-, and Eigenvector-Related Centrality Measures*

Beyond these fairly direct measures of centrality, there are more intricate ones. One of the more elegant, both mathematically and in terms of the ideas that it captures, is a notion developed by Bonacich [91]. Bonacich's measure is based on ideas that trace back to Seeley [585], Katz [381], and earlier work of Bonacich [90], and it is useful to start by discussing those ideas. These measures are based on the premise that a node's importance is determined by how important its neighbors are. That is, we might like to account not only for the connectivity or closeness of a node to many other nodes, but also for its proximity to many other "important" nodes. This notion is central to such phenomena as citation rankings and Google page rankings. The difficulty is that such a measure becomes self-referential. The centrality of a node depends on how central its neighbors are, which depends on the centrality of their neighbors, and so forth. There are various approaches to dealing with this issue. The following is a nice application of some basic ideas from matrix algebra and fixed-point theory.

Define the *Katz prestige* of a node $i$, denoted $P_i^K(g)$, to be a sum of the prestige of $i$'s neighbors divided by their respective degrees. Here I use the term *prestige* as in Katz [381], but it is also a measure of centrality. So $i$ gains prestige from having a neighbor $j$ who has high prestige. However, this measure is corrected by how many neighbors $j$ has, so that if $j$ has more relationships then $i$ obtains less prestige from being connected to $j$, all else being equal. This correction for the number of relationships that $j$ has might be thought of as correcting for the relative access or time that $i$ spends with $j$. That is, the Katz prestige of a node $i$ is

$$P_i^K(g) = \sum_{j \neq i} g_{ij} \frac{P_j^k(g)}{d_j(g)}. \tag{2.3}$$

This definition is self-referential, so it is not immediately obvious that it is uniquely (or always) defined. It does provide a series of equations and unknowns, so in principle it is solvable. We can see this as follows. Let $\widehat{g}_{ij} = g_{ij}/d_j(g)$ be the normalized adjacency matrix $g$ so that the sum across any (nonzero) *column* is normalized to 1.[25] The relationship (2.3) can then be rewritten as

$$P^K(g) = \widehat{g} P^K(g), \tag{2.4}$$

or

$$(\mathit{I\!I} - \widehat{g}) P^K(g) = 0, \tag{2.5}$$

where $P^K$ is written as a $n \times 1$ vector, and $\mathit{I\!I}$ is the identity matrix.

So, calculating the Katz prestige associated with the nodes of a given network reduces to finding the unit eigenvector of $\widehat{g}$, which is a standard calculation (see Section 2.4 for background on eigenvectors). Note that the Katz prestige is only determined up to a scale factor, so that if $P^K(g)$ solves (2.4) and (2.5), then so does $cP^K$ for any scalar $c$.

---

25. Let $0/0 = 0$, so that if $d_j(g) = 0$, then set $\widehat{g}_{ij} = 0$.

Katz prestige turns out to be more novel in directed networks than in undirected ones. If in-degree is the same as out-degree for every node, then it is easy to check that the solution to (2.4) is the list of nodes' degrees (or any rescaling of them), so that $[P^K(g)]_i = d_i(g)$. This equality provides a justification for degree centrality but not for a new measure. In the case of a directed network, the normalization in $\widehat{g}_{ij} = g_{ij}/d_j(g)$ is generally by in-degree, so that columns still sum to 1, with the interpretation that directed links to a given node have equal access to that node. In that case, the measure of Katz prestige differs for in-degree and out-degree.[26]

When applied to the network in Figure 2.13, the Katz prestige measures are the $P_4^K(g) = 2$, $P_3^K(g) = 3$, and $P_1^K(g) = 2$. Thus more "prestige" is given to nodes 3 and 5 than to the middle node 4, which has the same prestige as nodes 1, 2, 6, and 7. Here we see the importance of the weighting in the Katz prestige calculation. The middle node 4 is linked to two prestigious nodes, but only receives 1/3 of their time each. So its prestige is $(3)/3 + (3)/3 = 2$. Nodes 3 and 5 are linked to three nodes each. Although each of these three nodes is less prestigious, 3 and 5 receive 1/2 of each of their weight: $2/2 + 2/2 + 2/2 = 3$.

In a variation on this idea that avoids reduction to degree centrality, one does not normalize the network of relations $g$. The measure is known as *eigenvector centrality*[27] and was proposed by Bonacich [90]. Let $C^e(g)$ denote the eigenvector centrality associated with a network $g$. The centrality of a node is proportional to the sum of the centrality of its neighbors: $\lambda C_i^e(g) = \sum_j g_{ij} C_j^e(g)$. In matrix notation:

$$\lambda C^e(g) = g C^e(g), \tag{2.6}$$

where $\lambda$ is a proportionality factor. Thus from (2.6) $C^e(g)$ is an eigenvector of $g$, and $\lambda$ is its corresponding eigenvalue. Given that it generally makes sense to look for a measure with nonnegative values, the standard convention is to use the eigenvector associated with the largest eigenvalue, which is nonnegative for the networks considered here (see Section 2.4).

Note that the definition of eigenvector centrality also works for weighted and/or directed networks, without any changes to the expressions. Thus the Katz prestige is a form of eigenvector centrality when we have adjusted the network adjacency matrix to be weighted.

Katz [381] introduced another way of tracking the power or prestige of a node. The idea presumes that the power or prestige of a node is simply a weighted sum of the walks that emanate from it. A walk of length 1 is worth $a$, a walk of length 2 is worth $a^2$, and so forth, for some parameter $0 < a < 1$. This scheme gives higher weights to walks of shorter distance, as in the connections model. So it is a method of looking at all walks from a given node and weighting them by distance.

Note that $g\mathbb{1}$ (where $\mathbb{1}$ is the $n \times 1$ vector of 1s) is the vector of degrees of nodes, which tells us how many walks of length 1 emanate from each node. Based on what we saw in Section 2.1.3, $g^k\mathbb{1}$ is the vector whose $i$th entry is the total

26. However, if one normalizes by out-degree, then the measure will be out-degree.
27. See Section 2.4 for background on eigenvectors.

number of walks of length $k$ that emanate from each node. Thus the vector of the power of nodes, or prestige of nodes, can be written as

$$P^{K2}(g, a) = ag\mathbb{1} + a^2g^2\mathbb{1} + a^3g^3\mathbb{1}\cdots. \tag{2.7}$$

We can rewrite (2.7) as

$$P^{K2}(g, a) = \left(1 + ag + a^2g^2\cdots\right) ag\mathbb{1}. \tag{2.8}$$

For small enough $a > 0$, this is finite and can be expressed as[28]

$$P^{K2}(g, a) = (\mathit{I\!I} - ag)^{-1} ag\mathbb{1}. \tag{2.9}$$

Another way to interpret (2.8) is to note that we can start by assigning some base value of $ad_i(g)$ to node $i$. This value is expressed as the vector $ag\mathbb{1}$. Then a given node receives its base value, plus $a$ times the base value of each node it has a direct link to, plus $a^2$ times the base value of each node that it has a walk of length 2 to and weighted by the number of walks to the given node, plus $a^3$ times the base value of each node it has a walk of length 3 to, and so forth.

The measure introduced by Bonacich can be thought of as a direct extension of the above measure of power or prestige. It is often called *Bonacich centrality* and can be expressed as

$$Ce^B(g, a, b) = (\mathit{I\!I} - bg)^{-1}ag\mathbb{1}, \tag{2.10}$$

where $a > 0$ and $b > 0$ are scalars, and $b$ is sufficiently small so that (2.10) is well defined.[29]

Bonacich centrality can be thought of as a variation on the second prestige measure of Katz, where again we start with base values of $ad_i(g)$ for each node, but then we evaluate walks of length $k$ to other nodes by a factor of $b^k$ times the base value of the end node, allowing $b$ to differ from $a$. So $b$ is a factor that captures how the value of being connected to another node decays with distance, while $a$ captures the base value on each node. When $b = a$, the two measures coincide.

Normalizing $a = 1$, we can calculate the Bonacich centrality of the network in Figure 2.13 for a couple of values of $b$, which are listed in Table 2.1 along with other centrality measures for the same network. Degree centrality favors nodes 3 and 5, but treats nodes 1, 2, 6, 7, and 4 similarly, and so misses some aspects of the structure of the network. Closeness differentiates the three types of nodes, favoring node 4, which is similar to betweenness centrality, but with less spread. Decay centrality treats nodes 3, 4, and 5 as being more central than nodes 1, 2, 6,

---

28. From (2.8), if $P^{K2}(g, a)$ is finite, then it follows that $P^{K2}(g, a) - agP^{K2}(g, a) = ag\mathbb{1}$ or $(\mathit{I\!I} - ag)P^{K2}(g, a) = ag\mathbb{1}$. A sufficient condition for $P^{K2}(g, a)$ to be finite is that $a$ be smaller than 1 divided by the norm of the largest eigenvalue of $g$; and for the latter to be true it is sufficient that $a$ be smaller than 1 divided by the maximum degree of any agent.

29. Note that the scalar $a$ is no longer relevant, as it simply multiplies all of the terms. It is only useful in comparing to the corresponding Katz measure. This is not to say that the Bonacich measure is the same as that of Katz, as being able to change $b$ without forcing $a$ to adjust in the same manner can lead to important differences.

TABLE 2.1
Centrality comparisons for Figure 2.13

| Measure of centrality | Nodes 1, 2, 6, and 7 | Nodes 3 and 5 | Node 4 |
|---|---|---|---|
| Degree (and Katz prestige $P^K$) | .33 | .50 | .33 |
| Closeness | .40 | .55 | .60 |
| Decay centrality ($\delta = .5$) | 1.5 | 2.0 | 2.0 |
| Decay centrality ($\delta = .75$) | 3.1 | 3.7 | 3.8 |
| Decay centrality ($\delta = .25$) | .59 | .84 | .75 |
| Betweenness | .0 | .53 | .60 |
| Eigenvector centrality | .47 | .63 | .54 |
| Katz prestige-2 $P^{K2}$, $a = 1/3$ | 3.1 | 4.3 | 3.5 |
| Bonacich centrality $b = 1/3$, $a = 1$ | 9.4 | 13.0 | 11.0 |
| Bonacich centrality $b = 1/4$, $a = 1$ | 4.9 | 6.8 | 5.4 |

and 7 for any $\delta$, but the relative rankings of 3 and 5 relative to 4 depend on $\delta$. With a lower $\delta$ the results resemble those for like-degree centrality and favor nodes 3 and 5, while for higher $\delta$ they resemble those for closeness or betweenness and favor node 4. The eigenvector centralities and self-referential definitions of Bonacich and Katz prestige-2 all favor nodes 3 and 5, to varying extents. As $b$ decreases the Bonacich favors closer connections and higher-degree nodes, while for higher $b$, longer paths become more important.

These measures are certainly not the only measures of centrality, and it is clear from the above that the measures capture different aspects of the positioning of the nodes. Given how complex networks can be, it is not surprising that there are many different ways of viewing position, centrality, or power in a network.

## 2.3 ▪ Appendix: Basic Graph Theory

Here I present some basic results in graph theory that will be useful in subsequent chapters.[30]

### 2.3.1 Hall's Theorem and Bipartite Graphs

A *bipartite* network $(N, g)$ is one for which $N$ can be partitioned into two sets $A$ and $B$ such that if a link $ij$ is in $g$, then one of the nodes comes from $A$ and the other comes from $B$. A bipartite network is pictured in Figure 2.14. Settings with two classes of nodes are often referred to as *matching* settings (and in some cases *marriage markets*), where one group is referred to as "women" and the other as

30. Excellent texts on graph theory are Bollobás [85] and Diestel [200].

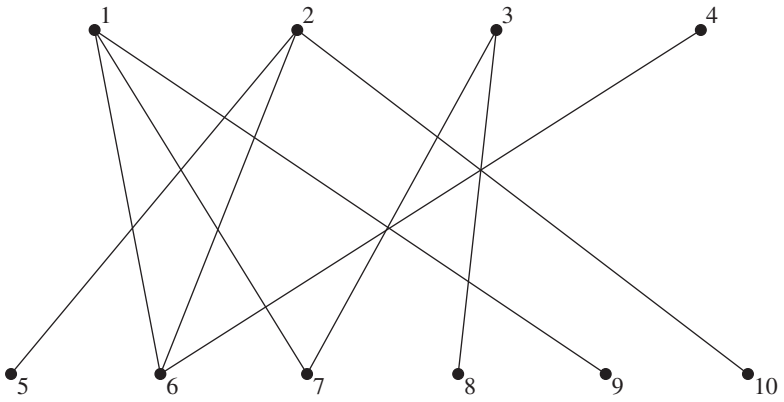


**FIGURE 2.14** A bipartite network.

"men." It has applications to markets, where for instance one of the sets consists of buyers and the other of sellers, as well as to such processes as the assignment of students to schools, researchers to labs, and so forth. (See Roth and Sotomayor [568] for an overview of the matching literature.)

One interpretation of a bipartite graph in a matching setting is that it represents the potential relationships that might occur. The object is often to determine a matching for some set of nodes, say $S \subset A$, which is a pairing of the nodes in $S$ with nodes in $B$ such that each node in $S$ is assigned to a distinct node of $B$ and the pairings are feasible as defined by $g$. That is, a matching for $S \subset A$ relative to $g$ is a mapping $\mu : S \to B$ such that $i\mu(i) \in g$ for each $i \in S$ and $j \neq i$ implies $\mu(j) \neq \mu(i)$.

It is clear that if we wish to assign each element of $S \subset A$ to a distinct element of $B$, then the number of neighbors of $S$ in $B$ must be at least as large as the size of $S$. Moreover, this condition must be true for any subset of $S$, since we wish to match each element to a different element from $B$. Hall's theorem states that this condition is not only necessary but also sufficient for such a matching to exist.

**Theorem 2.1 (Hall)** *Consider a bipartite graph* $(N, g)$ *with an associated bipartition of nodes* $\{A, B\}$. *There exists a matching of a set* $S \subset A$ *if and only if* $|N_{S'}(g)| \geq |S'|$ *for all* $S' \subset S$.

As we shall see in Chapter 10, this theorem is useful for working with networked models of markets, which are often bipartite in structure.

### 2.3.2 Set Coverings and Independent Sets

Given a network $(N, g)$, an *independent* set of nodes $A \subset N$ is a set such that if $i \in A$ and $j \in A$ and $i \neq j$, then $ij \notin g$. An independent set of nodes $A$ is maximal if it is not a proper subset of any other independent set of nodes as illustrated in Figure 2.15.

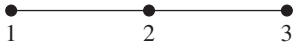


FIGURE 2.15 A three-node network. Independent sets: {1}, {2}, {3}, {1,3}; maximal independent sets: {1,3}, {2}.

The following observation (e.g., see Galeotti et al. [274]) is straightforward but useful when characterizing the equilibria of games played on networks.

**Observation 2.1** *Consider a network* $(N, g)$ *and a network* $(N, g')$ *such that* $g \subset g'$. *Any independent set A of* $g'$ *is an independent set of g, but if* $g' \neq g$, *then there exist (maximal) independent sets of g that are not (maximal) independent sets of* $g'$.

The proof of this observation is Exercise 2.8.

Independent sets are closely related to the equilibria of some games played on networks, as first pointed out by Bramoullé and Kranton [103]. To see how independent sets relate to equilibria, consider the following game played on a network.[31] Each player chooses whether to buy a product (e.g., a book). If a player does not buy the book, then he or she can freely borrow the book from any of his or her neighbors who bought it.[32] Indirect borrowing is not permitted, so a player cannot borrow the book from a neighbor of a neighbor. If none of a player's neighbors has bought the book, then the player would prefer to pay the cost of buying the book himself or herself rather than not having any access to the book. This problem is what is known as a classic *free-rider* problem, but defined on a network. A (pure strategy) equilibrium in this game is simply a specification of which players buy the book such that (1) no player who buys the book regrets it, and (2) no player who did not buy the book would rather buy the book. It is easy to see that the (pure strategy) equilibria of this game are precisely the situations in which the players who buy the book form a maximal independent set. This follows because (1) implies that if some player buys a book, then it must be that none of his or her neighbors buy the book, and (2) implies that any player who does not buy a book must have at least one neighbor who bought the book. Thus the first part implies that the set of people who buy the book must be independent, and the second part implies that the set must be maximal.

### 2.3.3 Colorings

Related to the concept of independent sets is that of colorings. One of the basic applications is to scheduling problems. For example, consider a network in which the nodes represent researchers who will attend a conference.[33] A link indicates

31. For more detailed definitions of game-theoretic concepts and a discussion of games played on networks see Chapter 9.

32. Assume that if some player buys the book, and several neighbors wish to borrow it, then they can coordinate on when they borrow it so that they can each borrow it without rivalry.

33. This example is from Bollobás [85].

that the two researchers wish to attend each other's presentations. The conference organizer wishes to know how many different time slots are needed (running parallel sessions within time slots) to ensure that each researcher can attend all of the presentations he or she would like to, and also present his or her own work. This problem is equivalent to coloring the associated graph. Suppose we have a different color to code each time slot of the conference. We want to color the nodes so that no two neighboring nodes have the same color. What is the minimum number of colors needed? That number is called the *chromatic number* of the graph.[34]

If we color the nodes of a network in $k$ colors, then we have produced $k$ independent sets. The coloring problem can then be thought of as finding the minimum number of independent sets needed to partition the set of nodes.[35] This challenging problem has resulted in some celebrated results. The most famous is probably the four-color theorem. That theorem concerns *planar* graphs. Without providing a formal topological definition, a planar graph is one that can be drawn on a piece of paper without having any two links cross each other (so that links can only intersect at one of their involved nodes). The four-color theorem states that every planar graph has a chromatic number of no more than four. This theorem was conjectured in the mid-1800s, and some false proofs were provided before it was proven in 1977 by Appel, Haken, and Koch [19].[36] An overview of coloring problems would take us beyond the scope of this text, but the problems are so central to graph theory and important in their applications that they at least deserve mention.

### 2.3.4 Eulerian Tours and Hamilton Cycles

The mathematician Leonhard Euler asked (and answered) a question that concerns paths in a graph. The puzzle traces back to a question concerning the old Prussian city of Königsberg, which lay on the Pregel River. The city was cut into four pieces by the river and had seven bridges. The question was whether it was possible to design a walk that started at some point in the city, crossed each bridge exactly once, and returned to the starting point. The four parts of the city can be thought of as the vertices or nodes of a graph, and the seven bridges as edges or links of the graph (Figure 2.17). The question then amounts to asking whether there exists

34. This problem is known as the vertex coloring problem. There are also edge coloring problems, and a recent generalization called list coloring problems. The edge coloring problem is to color the edges so that no two adjacent edges have the same color. The minimal number of colors needed has application, for instance, to having enough time slots for scheduling bilateral meetings of neighboring nodes, so that no node needs to be in more than one meeting at once. For an introduction to these problems, see Bollobás [85] or Diestel [200].

35. But note that the sets need not be maximal independent sets. For instance, node 1 is in its own element of the partition in Figure 2.16, but it is not a maximal independent set as it is not connected to 6. If we change the partition and color 6 to be the same color as 1, then we have another four-coloring. But then 2 is in its own element of the partition and does not form a maximal independent set.

36. That proof involved a computer verification that a series of 1,482 cases each reduces to being four-colorable. Shorter proofs have since been provided.

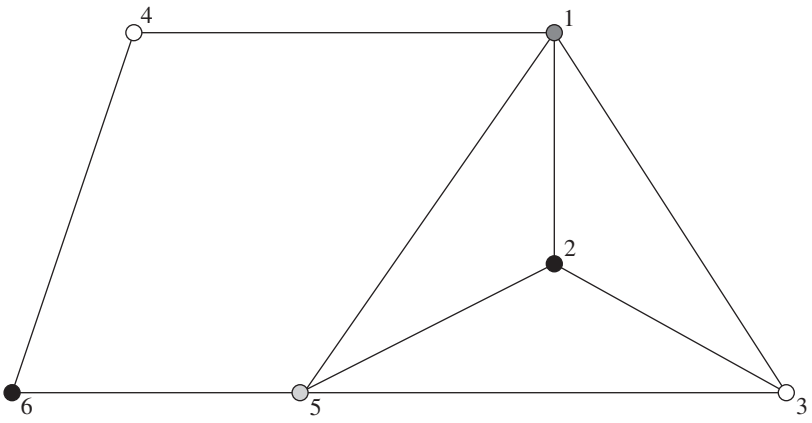


**FIGURE 2.16** A planar network on six nodes with chromatic number 4.

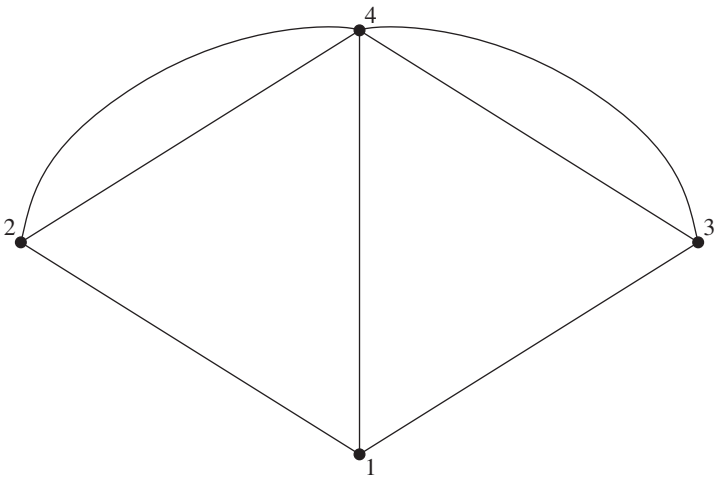


**FIGURE 2.17** Multigraph for the Königsberg bridge problem.

a walk in the graph that contains each link in the graph exactly once and starts and ends at the same node.[37] Such a closed walk is said to be an *Eulerian tour* or circuit.

A walk is said to be *closed* if it starts and ends at the same node. It is clear that to have a closed walk that involves every link of a network exactly once, each node in

37. The graph here is actually a *multigraph,* as there is more than one link between some pairs of nodes. The general problem of finding Eulerian tours can be stated in either context.

the network must have an even degree.[38] This result holds because each time a node is "entered" by one link on the walk it must be "exited" by a different link, and each time the node is visited, it must be by a link that has not appeared previously on the walk. Euler's simple but remarkable theorem is that this condition is necessary *and sufficient* for such a closed walk to exist.

**Theorem 2.2 (Euler)** *A connected network $g$ has a closed walk that involves each link exactly once if and only if the degree of each node is even.*

The proof is straightforward and appears as Exercise 2.9.

One can ask a related question for nodes rather than links: when is it possible to find a closed walk that involves each node in the network exactly once? Such a closed walk must be a cycle, and is referred to as a *Hamilton cycle* or a *Hamiltonian.* A further question is whether there exists a *Hamilton path* that includes each node exactly once. Clearly a network that has a Hamilton cycle also has a Hamilton path, while the converse is not true (consider a line).

Determining whether a network has a Hamilton cycle is a much more challenging question than whether it has an Euler tour; this has been an active area of research in graph theory for some time. It has direct applications to the traveling salesman problem, in which a salesman must visit each city on a trip exactly once, cities are nodes on a network, and the path must follow the links.

The seminal theorem on Hamilton cycles is due to Dirac [201]. Stronger theorems have since been developed, as we shall shortly see, but it is worth stating on its own, as it has an intuitive proof that illuminates the proofs of some of the later results.

**Theorem 2.3 (Dirac)** *If a network has $n \geq 3$ nodes and each node has degree of at least $n/2$, then the network has a Hamilton cycle.*

***Proof of Theorem 2.3.*** First, the network must be connected, because if the minimum degree is $n/2$, then the smallest component has more than half the nodes, and so the network cannot consist of more than one component. Next, consider a longest path in this network, and if there is more than one longest path then pick any one of them. Let $i$ be the starting node on the path and $j$ the ending node. It must be that each of $i$'s neighbors lies on the path, and so at least $n/2$ of the nodes in the path are neighbors of $i$. To see this, note that if it were not the case, then by starting with an omitted neighbor of $i$ and then moving to $i$, we could find a longer path. Similarly, $j$ has at least $n/2$ neighbors on the path. It is then easy to check that since $i$ and $j$ each have at least $n/2$ neighbors on the path, at least one of the nodes on the path that is a neighbor of $i$, say $k$, must have the previous node on the path be a neighbor of $j$. Thus consider the cycle formed as follows: $ik, k+1 \ldots j$, $jk-1, k-2 \ldots i$ (where the dots correspond to the original path). The claim is that this cycle is a Hamilton cycle. If this cycle does not include all nodes, then since the network is connected, there is some node outside the cycle connected to some node in the cycle. Then it is possible to make a path including that node and

38. Note that a closed walk is not necessarily a cycle (as it may visit some of the intermediate nodes more than once), but a cycle is a closed walk.

all nodes in the cycle, which contradicts the assumption that the original path was of maximal length. ■

An example of a strengthening of the Dirac theorem is the following theorem.

**Theorem 2.4 (Chvátal [157])** *Order the nodes of a network of $n \geq 3$ nodes in increasing order of their degrees, so that node 1 has the lowest degree and node n has the highest degree. If the degrees are such that $d_i \leq i$ for some $i < n/2$ implies $d_{n-i} \geq n - i$, then the network has a Hamilton cycle.*

This theorem also has a converse. If a degree sequence does not have this property, then one can find a network that has a degree sequence with at least as high a degree in each entry that does not have a Hamilton cycle. While it is clear that there are networks that have low average degree and have Hamilton cycles (e.g., simply arrange nodes in a circle), this converse shows that guaranteeing the existence of Hamilton cycles either requires strong conditions on basic characteristics like degree sequences or requires much more information about the structure of the network.

## 2.4 ▪ Appendix: Eigenvectors and Eigenvalues

Given an $n \times n$ matrix $T$, an *eigenvector* $v$ is a nonzero vector such that

$$Tv = \lambda v, \tag{2.11}$$

for some scalar $\lambda$, which is called the *eigenvalue* of $v$. Generally, we are interested in nonzero solutions to this equation (noting that a vector of 0s always solves (2.11)).

Eigenvectors come in two flavors: *left-hand* and *right-hand eigenvectors*, which are also known as *row* and *column eigenvectors,* respectively. These terms refer to whether the eigenvector multiplies the matrix $T$ from the left-hand or right-hand side, and correspondingly whether it is a row or column vector. So a left-hand (row) eigenvector is a $1 \times n$ vector $v$ such that

$$vT = \lambda v, \tag{2.12}$$

whereas a right-hand (column) eigenvector is an $n \times 1$ vector $v$ that satisfies (2.11) for some eigenvalue $\lambda$. As the definition at the start of this section suggests, *eigenvector* without a modifier usually refers to a right-hand eigenvector.

Basically, eigenvectors are vectors that, when acted upon by the matrix $T$, give back some rescaling of themselves, rather than being distorted to some new vector or new direction. So they serve as a sort of fixed point of the transformation $T$, and for many matrices (but not all), there will be as many eigenvector-eigenvalue pairs as there are dimensions: $n$.[39]

---

39. We have to be careful here to restrict attention to some normalization of each eigenvector, so it has norm 1, for instance. Otherwise, note that if $v$ is an eigenvector of $T$, then so is $kv$ for any scalar $k$, as (2.11), as well as (2.12), are satisfied if $v$ is rescaled.

The usefulness of eigenvectors can be seen in some of their applications. We have already seen important applications—calculating centrality or power, and in particular calculating Katz prestige (and also the eigenvector centrality). The idea is that a given agent's prestige is a weighted average of his or her neighbors' prestige, where the weights correspond to weights from a social network. This measure then presents a self-referential problem, as the prestige has to be derived from the prestige. In this case, we look for an eigenvector with an eigenvalue of 1 since the prestige returns the prestige without rescaling. The existence of an eigenvector with an eigenvalue of 1 in this context is implied by the Perron-Frobenius theorem (see Meyer [466]).

The Perron-Frobenius theorem implies that if $T$ is a nonnegative *(column) stochastic* matrix, so that the entries of each of its columns sum to 1, then there exists a nonnegative right-hand eigenvector $v$ that solves (2.11) and has a corresponding eigenvalue $\lambda = 1$. The same is true of row stochastic matrices and left-hand eigenvectors. If in addition $T^t$ has all positive entries for some $t$, then all other eigenvalues have a magnitude less than 1.[40]

Eigenvectors are also quite useful for examining the steady state or limit point of some system. Here we might think of $T$ as a transition matrix. Starting with some column vector $v$, the system transitions to a new vector $Tv$. A steady state of such a system, or a convergence point, is often a point such that $v = Tv$, so that once the system reaches $v$, it stays there. Again, $v$ is an eigenvector of $T$ that has a unit eigenvalue. These systems play a central role in Markov chains (see Section 4.5.8), where the $v$s represent probabilities of being in different states of a system, and the entries of $T$ represent probabilities of transferring from one state to another. This matrix is again stochastic (as probabilities sum to 1) and has a unit eigenvector.

Calculating the eigenvalues and corresponding eigenvectors of a matrix can be done using different methods, as the eigenvector calculation is basically a set of linear equations. If one knows $\lambda$, then (2.11) and (2.12) are systems of $n$ equations in $n$ unknowns. A useful way to solve for the eigenvalues associated with $T$ is to rewrite (2.11) as

$$(T - \lambda I\!I)v = 0,$$

where $I\!I$ is the identity matrix (with 1 for each diagonal entry and 0 elsewhere). For this equation to have a nonzero solution $v$, $T - \lambda I\!I$ must be a singular (noninvertible) matrix.[41] Thus the *characteristic equation* of $T$ is

$$\det(T - \lambda I\!I) = 0,$$

where $\det(\cdot)$ indicates determinant. The solutions to this equation are the eigenvalues of $T$.

---

40. The Perron-Frobenius theorem implies that the largest eigenvalue of any nonnegative matrix is real valued, and its corresponding eigenvector is nonnegative. Other eigenvalues can be complex valued.

41. This is a matrix in which some rows are linear combinations of other rows, or similarly for columns, which corresponds to having a determinant of 0.

### 2.4.1 Diagonal Decompositions

There are some particularly useful ways to rewrite a matrix $T$. To begin, let $V$ be the matrix of left-hand eigenvectors—so that each row is one of the eigenvectors of $T$. Then we can write

$$VT = \Lambda V, \tag{2.13}$$

where $\Lambda$ is the matrix with the eigenvalues corresponding to each row of $V$ on its diagonal:

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}.$$

From (2.13) it follows that if $V$ is invertible, then

$$T = V^{-1}\Lambda V. \tag{2.14}$$

Equation (2.14) is the *diagonal decomposition* of $T$. If it exists, then $T$ is said to be *diagonalizable.*

It is sometimes useful to note that $V^{-1}$, if well-defined, is the matrix of right-hand (column) eigenvectors of $T$, and that they have the same matrix of eigenvalues as $V$. To see this, note that from (2.13) $VTV^{-1} = \Lambda VV^{-1} = \Lambda$. Thus $V^{-1}VTV^{-1} = V^{-1}\Lambda$, and so $TV^{-1} = V^{-1}\Lambda = \Lambda V^{-1}$, and $V^{-1}$ is the vector of right-hand eigenvectors.

The decomposition (2.14) is useful for calculating higher powers of $T$ (which, for instance recalling Section 2.1.3, is useful in calculating the walks of $T$ if $T$ has entries consisting of 0 or 1). From (2.14) it follows that

$$T^2 = V^{-1}\Lambda VV^{-1}\Lambda V = V^{-1}\Lambda^2 V,$$

and more generally that

$$T^t = V^{-1}\Lambda^t V,$$

which is useful for calculating speeds of convergence, as in Section 8.3.6.[42]

## 2.5 ▪ Exercises

**2.1** ***Paths and Connectedness*** Given a network $(N, g)$, define its complement to be the network $(N, g')$ such that $ij \in g'$ if and only if $ij \notin g$. Show that if a network is

42. This formulation can help substantially from a computational perspective as well. Raising $T$ to a power directly, for a large matrix, can be computationally intensive. Instead, raising $\Lambda$ to a power is much easier since it involves raising only the diagonal entries to a power.

not connected, then its complement is. Provide an example of a four-node network that is connected and is such that its complement is also connected.

**2.2** ***Facts about Trees*** Show the following:

(a) A connected network is a tree if and only if it has $n - 1$ links.
(b) A tree has at least two leaves, where leaves are nodes that have exactly one link.
(c) In a tree, there is a unique path between any two nodes.

**2.3** ***Diameter and Degree*** Consider a sequence of networks such that each network in the sequence is connected and involves more nodes than the previous network. Show that if the diameter of the networks is bounded, then the maximal degree of the networks is unbounded. That is, show that if there exists a finite number $M$ such that the diameter of every network in the sequence is less than $M$, then for any integer $K$ there exists a network in the sequence and a node in that network that has more than $K$ neighbors.

**2.4** ***Centrality Measures*** Consider a two-link network among three nodes. That is, let the network consist of links 12 and 23.

(a) Calculate the Katz prestige (based on (2.5)) of each node, and compare it to the degree centrality and betweenness centrality for this network.
(b) Calculate the second measure due to Katz (based on (2.9)) for each node, when $a = 1/2$, which is the Bonacich centrality of each node when $b = 1/2$ and $a = 1/2$. How does this compare to Bonacich centrality when $b = 1/4$ and $a = 1/2$? Which nodes are relatively favored when $b$ increases and why? What happens as we continue to increase $b$ to $b = 3/4$?

**2.5** ***Average versus Overall Clustering*** Consider a network $(g, N)$ such that each node has at least two neighbors ($n_i(g) \geq 2$ for each $i \in N$). Compare the average clustering measure of a network to the overall clustering measure in the following two cases:

(a) $Cl_i(g) \geq Cl_j(g)$ when $d_i(g) \geq d_j(g)$, and
(b) $Cl_i(g) \leq Cl_j(g)$ when $d_i(g) \geq d_j(g)$.

Hint: Write the average clustering as $\sum_i Cl_i(g) \left(\frac{1}{n}\right)$ and argue that overall clustering can be written as $\sum_i Cl_i(g) \left(\frac{d_i(g)(d_i(g)-1)/2}{\sum_j d_j(g)(d_j(g)-1)/2}\right)$. Then compare these different weighted sums.

**2.6** ***Cohesiveness and Close-Knittedness*** There are various measures of how introspective or cohesive a given set of nodes is. Consider a set of nodes $S \subset N$. Given $1 \geq r \geq 0$ Morris [487] defines the set of nodes $S$ to be *$r$-cohesive* with respect to a network $g$ if each node in $S$ has at least a fraction $r$ of its neighbors in $S$. That is, $S$ is $r$-cohesive relative to $g$ if

$$\min_{i \in S} \frac{|N_i(g) \cap S|}{d_i(g)} \geq r, \tag{2.15}$$

where $0/0$ is set to 1.

Young [668] defines the set of nodes $S$ to be *r-close-knit* with respect to a network $g$ if each subset of $S$ has at least a fraction $r$ of its links remaining in $S$. Given $S'$ and $S$, let $d(S', S, g) = |\{ij | i \in S', j \in S\}|$ be the number of links between members of $S'$ and members of $S$. Then $S$ is $r$-close-knit relative to $g$ if

$$\min_{S' \subset S} \frac{d(S', S, g)}{\sum_{i \in S'} d_i(g)} \geq r,$$

where $0/0$ is set to 1.

Show that if a set of nodes $S$ is $r$-close-knit relative to $g$ then it is $r$-cohesive. Provide an example showing that the converse is false.

**2.7** ***Independent Sets*** Show that there is a unique network on $n$ nodes that is connected and is such that a maximal independent set of that network involves all nodes except node $i$. Show that there are two maximal independent sets of that network.

**2.8** ***Independent Sets and Equilibria*** Prove Observation 2.1.

**2.9** ***Euler Tours*** Prove Theorem 2.2. (Hint: First argue that any longest walk that does not involve any link more than once must be closed.)