# Chapter 1

# Introduction

This chapter provides an introduction to the analysis of networks through the presentation of several examples of research. This provides not only some idea of why the subject is interesting, but also of the range of networks studied, approaches taken and methods used.

## 1.1 Why Model Networks?

Social networks permeate our social and economic lives. They play a central role in the transmission of information about job opportunities, and are critical to the trade of many goods and services. They are the basis of the provision of mutual insurance in developing countries. Social networks are also important in determining how diseases spread, which products we buy, which languages we speak, how we vote, as well as whether or not we decide to become criminals, how much education we obtain, and our likelihood of succeeding professionally. The countless ways in which network structures affect our well-being make it critical to understand: (i) how social network structures impact behavior, and (ii) which network structures are likely to emerge in a society. The purpose of this monograph is to provide a framework for an analysis of social networks, with an eye on these two questions.

As the modeling of networks comes from varied fields and employs a variety of different techniques, before jumping into formal definitions and models, it is useful to start with a few examples that help give some impression of what social networks are and how they have been modeled. The following examples illustrate widely different perspectives, issues, and approaches; previewing some of the breadth of the range of

topics to follow.

## 1.2 A Set of Examples:

The first example is a detailed look at the role of social networks in the rise of the Medici.

### 1.2.1 Florentine Marriages

The Medici have been called the "godfathers of the Renaissance." Their accumulation of power in the early fifteenth century in Florence, was orchestrated by Cosimo de' Medici despite the fact that his family started with less wealth and political clout than other families in the oligarchy that ruled Florence at the time. Cosimo consolidated political and economic power by leveraging the central position of the Medici in networks of family inter-marriages, economic relationships, and political patronage. His understanding of and fortuitous position in these social networks enabled him to build and control an early forerunner to a political party, while other important families of the time floundered in response.[1]

Padgett and Ansell [491] provide powerful evidence for this by documenting the network of marriages between some key families in Florence in the 1430's. The following figure provides the links between the key families in Florence at that time, where a link represents a marriage between members of the two linked families.[2]

As mentioned above, during this time period the Medici (with Cosimo de' Medici playing the key role) rose in power and largely consolidated control of the business and politics of Florence. Previously Florence had been ruled by an oligarchy of elite families. If one examines wealth and political clout, however, the Medici did not stand out at this point and so one has to look at the structure of social relationships to understand why it was the Medici who rose in power. For instance, the Strozzi had

[1] See Kent [367] and Padgett and Ansell [491] for detailed analyses, as well as more discussion of this example.

[2] The data here were originally collected by Kent [367], but were first coded by Padgett and Ansell [491], who discuss the network relationships in more detail. The analysis provided here is just a teaser that offers a glimpse of the importance of the network structure. The interested reader should consult Padgett and Ansell [491] for a much richer analysis.

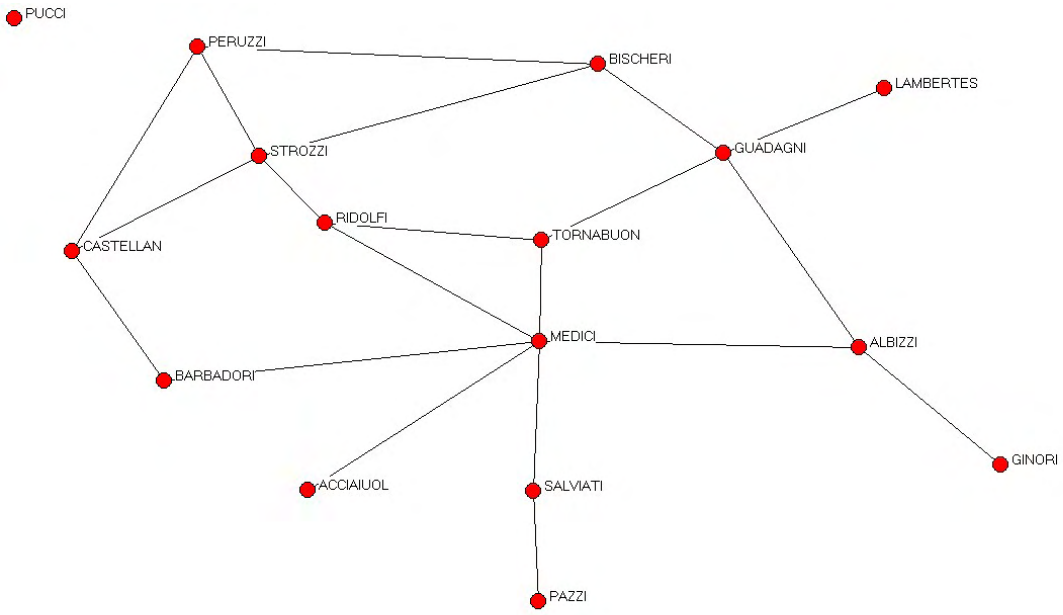



**Figure 1.2.1 15th Century Florentine Marriges Data from Padgett and Ansell [491] (drawn using UCINET)**

both greater wealth and more seats in the local legislature, and yet the Medici rose to eclipse them. The key to understanding this, as Padgett and Ansell [491] detail, can be seen in the network structure.

If we do a rough calculation of importance in the network, simply by counting how many families a given family is linked to through marriages, then the Medici do come out on top. However, they only edge out the next highest families, the Strozzi and the Guadagni, by a ratio of 3 to 2. While this is suggestive, it is not so dramatic as to be telling. We need to look a bit closer at the network structure to get a better handle on a key to the success of the Medici. In particular, the following measure of betweenness is illuminating.

Let $P(ij)$ denote the number of shortest paths connecting family $i$ to family $j$. [3] Let $P_k(ij)$ denote the number of these paths that family $k$ lies on. For instance, the shortest path between the Barbadori and Guadagni has three links in it. There are two such paths: Barbadori - Medici - Albizzi - Guadagni, and Barbadori - Medici -

[3] Formal definitions of path and some other terms used in this chapter appear in Chapter 2. The ideas should generally be clear, but the unsure reader can skip forward if they wish. Paths represent the obvious thing: a series of links connecting one node to another.

Tournabouni - Guadagni. If we set $i$ = Barbadori and $j$ = Guadagni, then $P(ij) = 2$. As the Medici lie on both paths, $P_k(ij) = 2$ when we set $k$ = Medici, and $i$ = Barbadori and $j$ = Guadagni. In contrast this number is 0 if we set $k$ = Strozzi, and is 1 if we set $k$ = Albizzi. Thus, in a sense, the Medici are the key family in connecting the Barbadori to the Guadagni.

In order to get a fuller feel for how central a family is, we can look at an average of this betweenness calculation. We can ask for each pair of other families, what fraction of the total number of shortest paths between the two the given family lies on. This would be 1 if we are looking at the fraction of the shortest paths the Medici lie on between the Barbadori and Guadagni, and 1/2 if we examine the corresponding fraction that the Albizzi lie on. Averaging across all pairs of other families gives us a sort of betweenness or power measure (due to Freeman [237]) for a given family. In particular, we can calculate

$$\sum_{ij:i\neq j,k\notin\{i,j\}} \frac{P_k(ij)/P(ij)}{(n-1)(n-2)/2} \tag{1.1}$$

for each family $k$, where we set $\frac{P_k(ij)}{P(ij)} = 0$ if there are no paths connecting $i$ and $j$, and the denominator captures that a given family could lie on paths between up to $(n-1)(n-2)/2$ pairs of other families. This measure of betweenness for the Medici is .522. That means that if we look at all the shortest paths between various families (other than the Medici) in this network, the Medici lie on over half of them! In contrast, a similar calculation for the Strozzi comes out at .103, or just over ten percent. The second highest family in terms of betweenness after the Medici is the Guadagni with a betweenness of .255. To the extent that marriage relationships were keys to communicating information, brokering business deals, and reaching political decisions, the Medici were much better positioned than other families, at least according to this notion of betweenness.[4] While aided by circumstance (for instance, fiscal problems resulting from wars), it was the Medici and not some other family that ended up consolidating power. As Padgett and Ansell [491] put it, "Medician political control was produced by network disjunctures within the elite, which the Medici alone spanned."

[4]The calculations here are conducted on a subset of key families (a data set from Wasserman and Faust [615]), rather than the entire data set which consists of hundreds of families. As such, the numbers differ slightly from those reported in footnote 31 of Padgett and Ansell [491]. Padgett and Ansell also find similar differences in centrality between the Medici and other families in terms of a network of business ties.

This analysis shows that network structure can provide important insights beyond those found in other political and economic characteristics. The example also illustrates that the network structure is important beyond a simple count of how many social ties each member has, and suggests that different measures of betweenness or centrality will capture different aspects of network structure.

This example also suggests a series of other questions that we will be addressing throughout this book. For instance, was it simply by chance that the Medici came to have such a special position in the network or was it by choice and careful planning? As Padgett and Ansell [491] say (footnote 13), "The modern reader may need reminding that all of the elite marriages recorded here were arranged by patriarchs (or their equivalents) in the two families. Intra-elite marriages were conceived of partially in political alliance terms." With this perspective in mind we then might ask why other families did not form more ties, or try to circumvent the central position of the Medici. We could also ask whether the resulting network was optimal from a variety of perspectives: was it optimal from the Medici's perspective, was it optimal from the oligarchs' perspective, and was it optimal for the functioning of local politics and the economy of 15th century Florence? These types of questions are ones that we can begin to answer through explicit models of the costs and benefits of networks, as well as models of how networks form.

### 1.2.2 Friendships Among High School Students

The next example comes from the The National Longitudinal Adolescent Health Data Set, known as "Add Health."[5] These data provide detailed social network information for over ninety thousand high school students from U.S. high schools interviewed during the mid 1990s; together with various data on the students' socio-economic background, behaviors and opinions. The data provide a number of insights and illustrate some features of networks that are discussed in more detail in the coming chapters.

Figure 1.2.2 shows a network of romantic relationships as found through surveys of

[5] Add Health is a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu). The network data that I present in this example were extracted by James Moody from the Add Health data set.

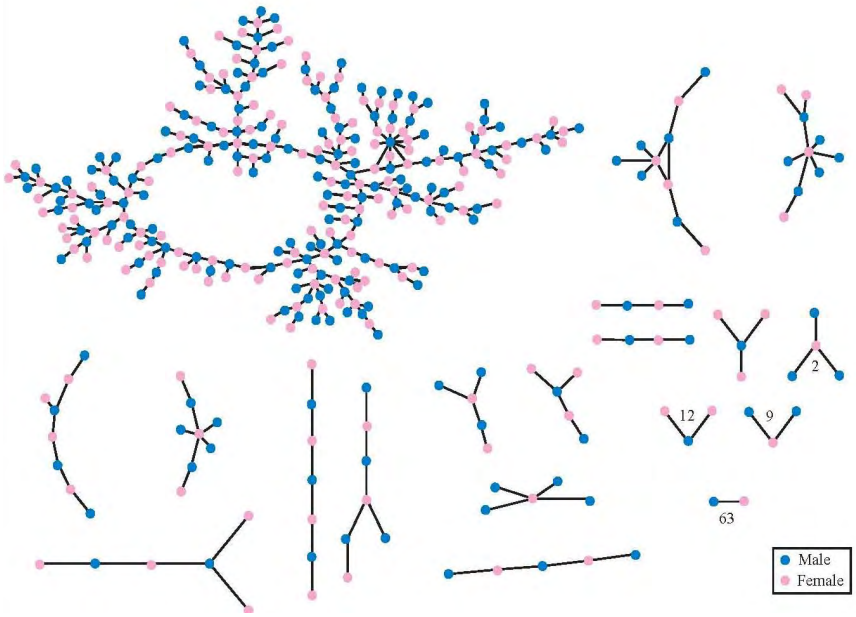



**Figure 1.2.2.** A Figure from Bearman, Moody and Stovel [47] based the Add Health Data Aet. A Link Denotes a Romantic Relationship, and the Numbers by Some Components Indicate How Many Such Componets Appear.

students in one of the high schools in the study. The students were asked to list the romantic liasons that they had during the six months previous to the survey.

There are several things to remark about Figure 1.2.2. The network is nearly a *bipartite* network, meaning that the nodes can be divide into two groups, male and female, so that links only lie between groups (with a few exceptions). Despite its nearly bipartite nature, the distribution of the degrees of the nodes (number of links each node has) turns out to closely match a network where links are formed uniformly at random (for details on this see Section 3.2.3), and we see a number of features of large random networks. For example, we see a "giant component," where over one hundred of the students are connected via sequences of links in the network. The next largest component (maximal set of students who are each linked to one another via sequences of links) only has ten students in it. This component structure has important implications for the diffusion of disease, information, and behaviors, as discussed in detail in Chapters 7, 8, and 9. Next, note that the network is quite "tree-like" in that there are very few loops or cycles in the network. There is a very large cycle visible in the giant component, and then a couple of smaller cycles present, but very few overall.

The absence of many cycles means that as one walks along the links of the network until hitting a dead-end, most of the nodes that are met are new ones that have not been encountered before. This is important in navigation of networks. This feature is found in many random networks in cases where there are enough links so that a giant component is present, but there are also few enough links so that the network is not fully connected. This contrasts with what we see in the denser friendship network pictured in Figure 1.2.2, where there are many cycles, and a shorter distance between nodes.

The network pictured in Figure 1.2.2 is also from the Add Health data set and connects a population of high school students.[6] Here the nodes are coded by their race rather than sex, and the relationships are friendships rather than romantic relationships. This is a much denser network than the romance network, and also exhibits some other features of interest.

A strong feature present in Figure 1.2.2 is what is known as "homophily," a term due to Lazarsfeld and Merton [404]. That is, there is a bias in friendships towards similar individuals; in this case the homophily concerns the race of the individuals. This bias is above what one would expect due to the makeup of the population. In this school, 52 percent of the students are white and yet 86 percent of whites' friendships are with other whites. Similarly, 38 percent of the students are black and yet 85 percent of blacks' friendships are with other blacks. Hispanics are more integrated in this school, comprising 5 percent of the population, but having only 2 percent of their friendships with Hispanics.[7] If friendships were formed without race being a factor, then whites would have roughly 52 percent of their friendships with other whites rather than 85 percent.[8] This bias is referred to as "inbreeding homophily" and has strong consequences. As we can see in the figure, it means that the students end up somewhat segregated by race, and this will impact the spread of information, learning, and the

[6] A link indicates that at least one of the two students named the other as a friend in the survey. Not all friendships were reported by both students. For more detailed discussion of these particular data see Curarrini, Jackson and Pin [171].

[7] The Hispanics in this school are exceptional compared to what is generally observed in the larger data set of 84 high schools. Most racial groups (including Hispanics in many of the other schools) tend to have a greater percentage of own-race friendships than the percentage their race in the population, regardless of their fraction of the population. See Currarini, Jackson and Pin [171] for details.

[8] There are a variety of possible reasons for the patterns observed, as it could be that race is correlated with other factors that affect friendship opportunities. For more discussion of this with respect to these data see Moody [458] and Currarini, Jackson and Pin [171]. The main point here is that the resulting network has clear patterns and those will have consequences.

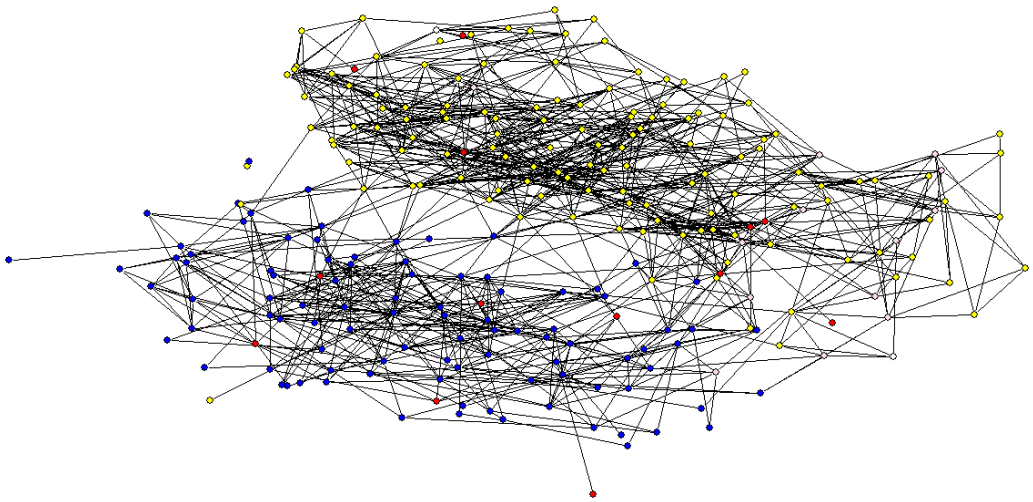


**Figure 1.2.2 "Add Health" Friendships among High School Students Coded by Race: Blue=Black, Yellow=White, Red=Hispanic, Green=Asian, White=Other**

speed with which things propagate through the network; themes that are explored in detail in what follows.

### 1.2.3 Random Graphs and Networks

The examples of Florentine marriages and high school friendships suggest the need for models of how and why networks form as they do. The last two examples in this chapter illustrate two complementary approaches to modeling network formation.

The next example of network analysis comes from the graph-theoretic branch of mathematics, and has recently been extended in various directions by the computer science, statistical physics, and economics literatures (as will be examined in some of the following chapters). This is perhaps the most basic model of network formation that one could imagine: it simply supposes that a completely random process is responsible for the formation of the links in a network. The properties of such random networks provide some insight into the properties that some social and economic networks have. Some of the properties that have been extensively studied are how links are distributed across different nodes, how connected the network is in terms of being able to find paths from one node to another, what the average and maximal path lengths are, how many isolated nodes there are, and so forth. Such random networks will serve as a very useful benchmark against which we can contrast observed networks; as such comparisons help identify which elements of social structure are not the result of mere randomness, but must be traced to other factors.

Erdös and Rényi [211], [212], [213] provided seminal studies of purely random networks.[9] To describe one of the key models, fix a set of $n$ nodes. Each link is formed with a given probability $p$, and the formation is independent across links.[10] Let us examine this model in some detail, as it has an intuitive structure and has been a springboard for many recent models.

Consider a set of nodes $N = \{1, \ldots, n\}$, and let a link between any two nodes, $i$

[9] See also Solomonoff and Rapoport [576] and Rapoport [524], [525], [526], for related predecessors.

[10] Two closely related models that they explored are as follows. In one of the alternative models, a precise number $M$ of links is formed out of the $n(n-1)/2$ possible links. Each different graph with $M$ links has an equal probability of being selected. In the second alternative model, the set of all possible networks on the $n$ nodes is considered and one is randomly picked uniformly at random. This can also be done according to some other probability distribution. While these models are clearly different, they turn out to have many properties in common. Note that the last model nests the other two (and any other random graph model on a fixed set of nodes) if one chooses the right probability distributions over all networks.

and $j$, be formed with probability $p$, where $0 < p < 1$. The formation of links is independent. This is a binomial model of link formation, which gives rise to a manageable set of calculations regarding the resulting network structure.[11] For instance, if $n = 3$, then a complete network forms with probability $p^3$, any given network with two links (there are three such networks) forms with probability $p^2(1-p)$, any given network with one link forms with probability $p(1-p)^2$, and the empty network that has no links forms with probability $(1-p)^3$. More generally, any given network that has $m$ links on $n$ nodes has a probability of

$$p^m(1-p)^{\frac{n(n-1)}{2}-m} \tag{1.2}$$

of forming under this process.[12]

We can calculate some statistics that describe the network. For instance, we can find the degree distribution fairly easily. The degree of a node is the number of links that the node has. The degree distribution of a random network describes the probability that any given node will have a degree (number of links) of $d$.[13] The probability that any given node $i$ has exactly $d$ links is

$$\left( \begin{array}{c} n-1 \\ d \end{array} \right) p^d(1-p)^{n-1-d}. \tag{1.3}$$

Note that even though links are formed independently, there will be some correlation in the degrees of various nodes, which will affect the distribution of nodes that have a given degree. For instance, if $n = 2$, then it must be that both nodes have the same degree: the network either consists of two nodes of degree 0, or two nodes of degree 1. As $n$ becomes large, however, the correlation of degree between any two nodes vanishes, as the possible link between them is only one out of the $n-1$ that each might have. Thus, as $n$ becomes large, the fraction of nodes that have $d$ links will

[11] See Section 4.5.4 for more background on the binomial distribution.

[12] Note here that there is a distinction between the probability of some specific network forming and some network architecture forming. With four nodes the chance that a network forms with a link between nodes 1 and 2 and a link between nodes 2 and 3 is $p^2(1-p)^4$. However, the chance that a network forms which contains two links involving three nodes is $12\ p^2(1-p)^4$, as there are 12 different networks we could draw that have this same shape. The difference between these counts is whether we pay attention to the labels of the nodes in various positions.

[13] The degree distribution of a network is often given for an observed network, and thus is a frequency distribution. Here, when dealing with a random network, one can talk about the degree distribution before the network has actually formed, and so we refer to probabilities of nodes having given degrees, rather than observed frequencies of nodes with given degrees.

approach the expression in (1.3). For large $n$ and small $p$, this binomial expression is approximated by a Poisson distribution, so that the fraction of nodes that have $d$ links is approximately[14]

$$\frac{e^{-(n-1)p}((n-1)p)^d}{d!}. \tag{1.4}$$

Given the approximation of the degree distribution by a Poisson distribution, the class of random graphs where each link is formed independently with an identical probability is often referred to as the class of *Poisson random networks*, and I will use this terminology in what follows.

To provide a better feeling for the structure of such networks, I generated a couple of Poisson random networks for different $p$'s. I chose $n = 50$ nodes as this produces a network that is easy to visualize. Let us start with an expected degree of 1 for each node. This is equivalent to setting $p$ at roughly .02. Figure 1.2.3 pictures a network generated with these parameters.[15] This network exhibits a number of features that are common to this range of $p$ and $n$. First, we should expect some isolated nodes. Based on the approximation of a Poisson distribution (1.4) with $n = 50$ and $p = .02$, we should expect about 37.5 percent of the nodes to be isolated (i.e., have $d = 0$), which is roughly 18 or 19 nodes. There are 19 isolated nodes in the network, by chance. Figure 1.2.3 compares the realized frequency distribution of degrees with the Poisson approximation.

The distributions match fairly closely. The network also has some other features that are common to random networks with $p$'s and $n$'s in this relative range. In graph theoretical terms, the network is a "forest," or a collection of trees. That is, there are no cycles in the network (where a cycle is a sequence of links that lead from one node back to itself, as described in more detail in Section 2.1.3). The chance of there being a cycle is relatively low with such a small link probability. In addition, there are six components (maximal subnetworks such that every pair of nodes in the subnetwork is connected by a path or sequence of links) that involve more than one node. And one

[14] To see this, note that for large $n$ and small $p$, $(1-p)^{n-1-d}$ is roughly $(1-p)^{n-1}$. Then, we write $(1-p)^{n-1} = (1 - \frac{(n-1)p}{n-1})^{n-1}$ which, if $(n-1)p$ is either constant or shrinking (if we allow $p$ to vary with $n$), is approximately $e^{-(n-1)p}$. Then for fixed $d$, large $n$, and small $p$, $\begin{pmatrix} n-1 \\ d \end{pmatrix}$ is roughly $\frac{(n-1)^d}{d!}$.

[15] The networks in Figures 1.2.3 and 1.2.3 were generated and drawn using the random network generator in UCINET [89]. The nodes are arranged to make the links as easy as possible to distinguish.

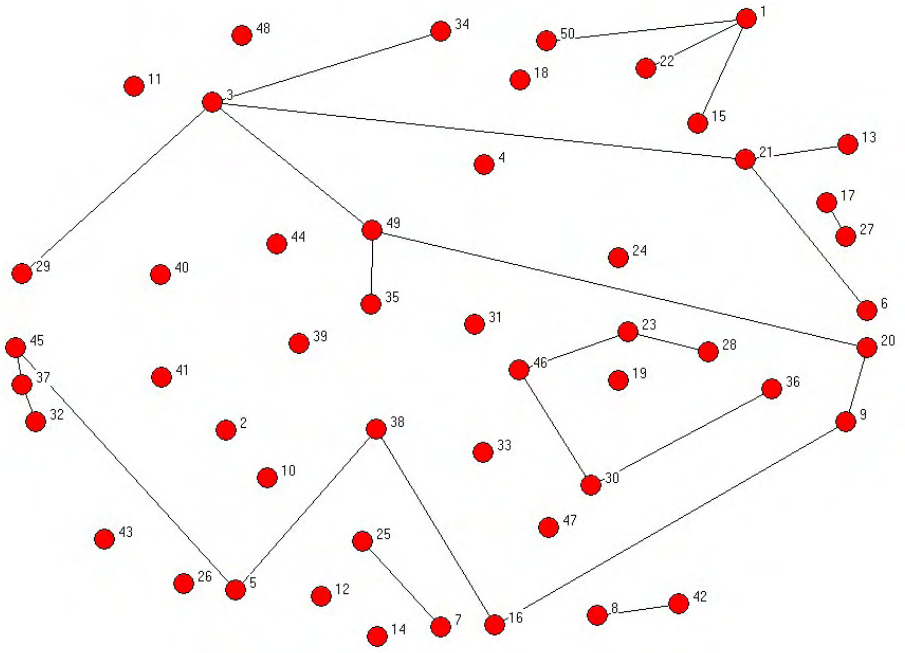



**Figure 1.2.3. A Randomly Generated Network with Probability .02 on each Link**

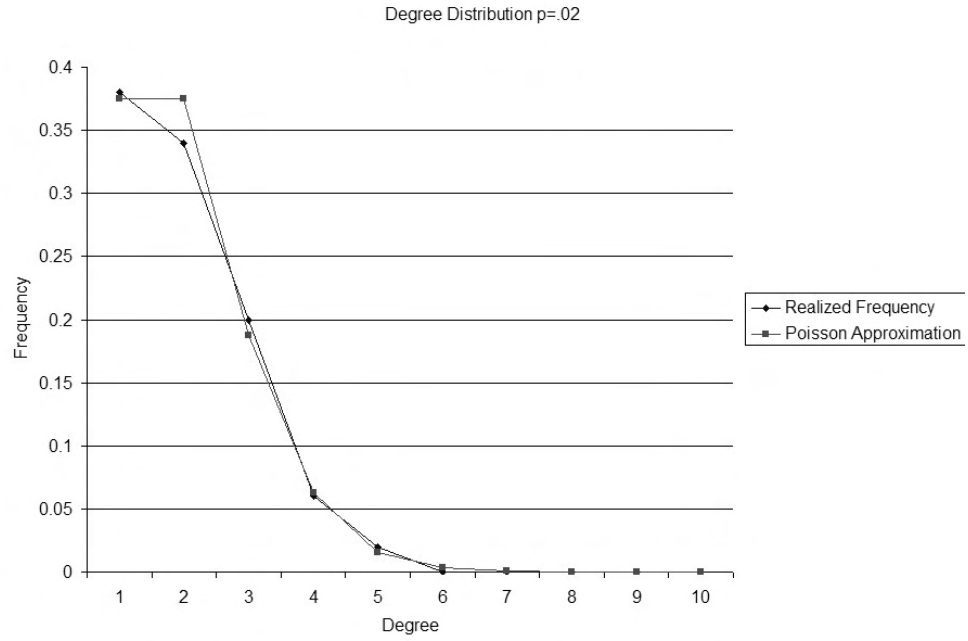



**Figure 1.2.3 Frequency Distribution of a Randomly Generated Network and the Poisson Approximation for a Probability of .02 on each Link**

of the components is much larger than the others: involving 16 nodes, while the next largest component only has 5 nodes in it. As we shall discuss shortly, this is to be expected.

Next, let us start with the same number of nodes, but increase the probability of a link forming to $p = \log(50)/50 = .078$, which is roughly the threshold where isolated nodes should start to disappear. (This threshold is discussed in more detail in Chapter 4.) Indeed, based on the approximation of a Poisson distribution (1.4) with $n = 50$ and $p = .08$, we should expect about 2 percent of the nodes to be isolated (with degree 0), or roughly 1 node out of 50. This is exactly what we see in the realized network in Figure 1.2.3 (again, by chance). With the exception of the single isolated node, the rest of the network is connected into one component.

As shown in Figure 1.2.3, the realized frequency distribution of degrees is again similar to the Poisson approximation, although, as one should expect at this level of randomness, not a perfect match.

The degree distribution tells us a great deal about a network's structure. Let us

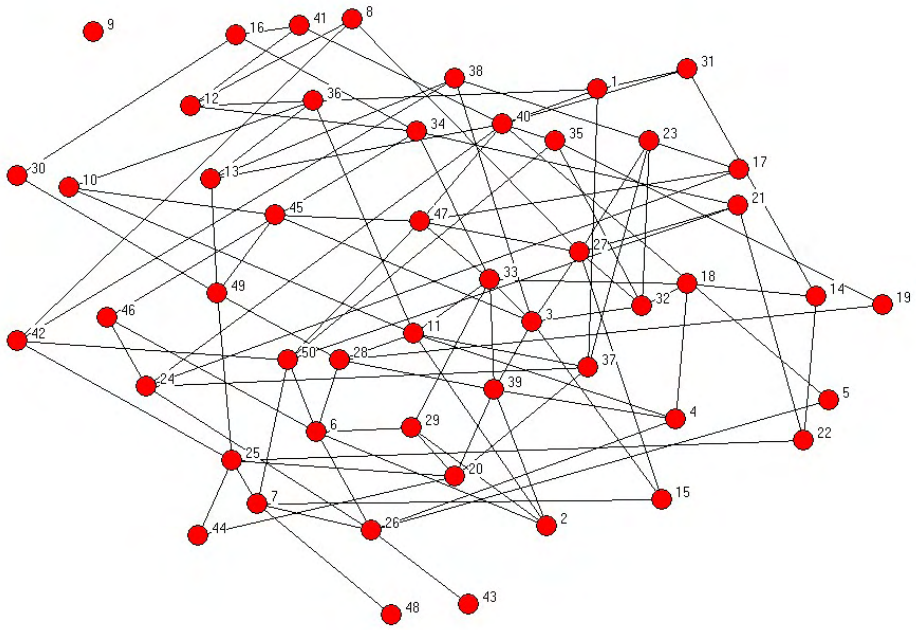



**Figure 1.2.3 A Randomly Generated Network with Probability .08 of each Link**

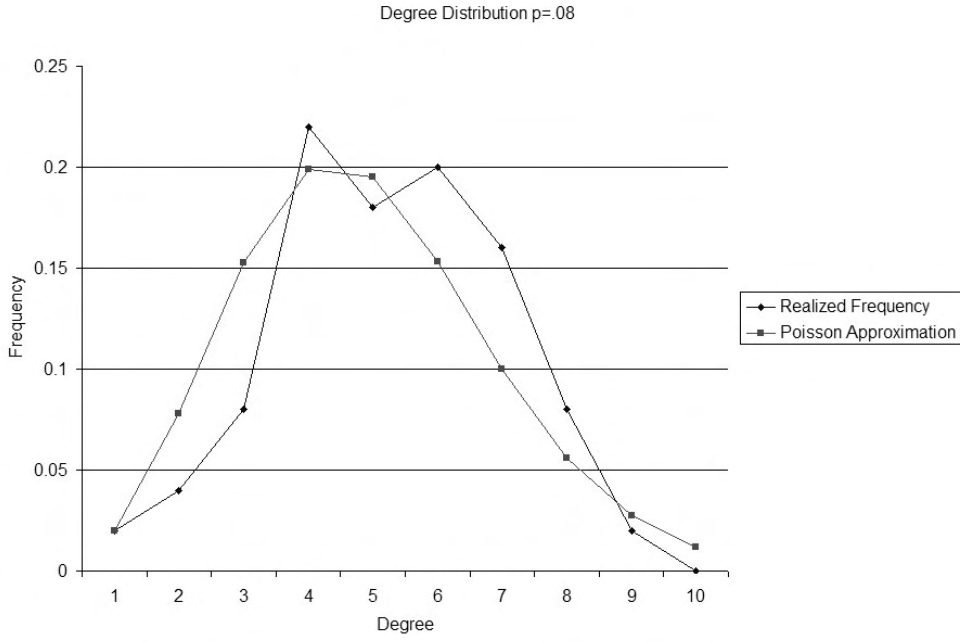



**Figure 1.2.3 Frequency Distribution of a Randomly Generated Network and the Poisson Approximation for a Probability of .08 on each Link**

examine this in more detail, as it provides a first illustration of the concept of a *phase transition*, where the structure of a random network changes as we change the formation process.

Consider what fraction of nodes are completely isolated; i.e., what fraction of nodes have degree $d = 0$? From (1.4) it follows that this is approximated by $e^{-(n-1)p}$ for large networks, provided the average degree $(n-1)p$ is not too large. To get a more precise expression, let us examine the threshold where this fraction is just such that we expect to have one isolated node on average. That is where $e^{-(n-1)p} = \frac{1}{n}$. Solving this yields $p(n-1) = log(n)$, or right at the point where average degree $(n-1)p$ is $log(n)$. Indeed, this is a threshold for a "phase transition," as we shall see in Section 4.2.2. If the average degree is substantially above $log(n)$, then probability of having any isolated nodes goes to 0, while if the average degree is substantially below $log(n)$, then the probability of having at least some isolated nodes goes to 1. In fact, as we shall see in Theorem 4.2.1, this is the threshold such that if the average degree is significantly above this level then the network is path-connected with a probability converging to 1 as $n$ grows (so that any node can be reached from any other via a path in the network), while below this level the network will consist of multiple components with a probability going to 1.

Other properties of random networks are examined in much more detail in Chapter 4. While it is clear that completely random networks are not always a good approximation for real social and economic networks, the analysis above (and in Chapter 4) shows us that much can be deduced in such models; and that there are some basic patterns and structures that we will see emerging more generally. As we build more realistic models, similar analyses can be conducted.

### 1.2.4 The Symmetric Connections Model

Although random network formation models give us some insight into the sorts of characteristics that networks might have, and exhibit some of the features that we see in the Add Health social network data, it does not provide as much insight into the Florentine marriage network. There, marriages were carefully arranged. The last example comes from the game-theoretic, economics literature and provides a basis for the analysis of networks that are formed when links are chosen by the agents in the network. Through this example, we can begin to look at the questions about which networks might be best for a society and which networks might arise if the players have discretion in choosing their links.


**Figure 1.2.4** ***The utilities to the players in a three-link four-player network in the symmetric connections model.***

It is a simple model of social connections that was developed by Jackson and Wolinsky [343]. In this model, links represent social relationships, for instance friendships, between players. These relationships offer benefits in terms of favors, information, etc., and also involve some costs. Moreover, players also benefit from indirect relationships. A "friend of a friend" also results in some indirect benefits, although of a lesser value than the direct benefits that come from a "friend." The same is true of "friends of a friend of a friend," and so forth. The benefit deteriorates with the "distance" of the relationship. This is represented by a factor $\delta$ that lies between 0 and 1, which indicates the benefit from a direct relationship and is raised to higher powers for more distant relationships. For instance, in the network where player 1 is linked to 2, 2 is linked to 3, and 3 is linked to 4: player 1 gets a benefit of $\delta$ from the direct connection with player 2, an indirect benefit of $\delta^2$ from the indirect connection with player 3, and an indirect benefit of $\delta^3$ from the indirect connection with player 4. The payoffs to this four players in a three-link network is pictured in Figure 1.2.4.

For $\delta < 1$ this leads to a lower benefit from an indirect connection than a direct one. Players only pay costs, however, for maintaining their direct relationships.[16]

Given a network $g$,[17] the net utility or payoff $u_i(g)$ that player $i$ receives from a network $g$ is teh sum of benefits that the player gets for his or her direct and indirect connections to other players less the cost of maintaining his or her links. In particular, it is

$$u_i(g) = \sum_{j \neq i:\ i \text{ and } j \text{ are path-connected in } g} \delta^{\ell_{ij}(g)} - d_i(g)c,$$

where $\ell_{ij}(g)$ is the number of links in the shortest path between $i$ and $j$, $d_i(g)$ is the

[16] In the most general version of the connections model the benefits and costs may be relation specific, and so are indexed by $ij$. One interesting variation is where the cost structure is specific to some geography, so that linking with a given player depends on their physical proximity. That variation has been studied by Johnson and Gilles [349] and is discussed in Exercise 6.13.

[17] For complete definitions, see Chapter 2. For now, all that is important is that this tells us which pairs of players are linked.

number of links that $i$ has ($i$'s degree), and $c > 0$ is the cost for a player of maintaining a link.

The highly stylized nature of the connections model allows us to begin to answer questions regarding which networks are "best" (most "efficient") from society's point of view, as well as which networks are likely to form when self-interested players choose their own links.

Let us define a network to be *efficient* if it maximizes the total utility to all players in the society. That is, $g$ is efficient if it maximizes $\sum_i u_i(g)$.[18]

It is clear that if costs are very low, it will be efficient to include all links in the network. In particular, if $c < \delta - \delta^2$, then adding a link between any two agents $i$ and $j$ will always increase total welfare. This follows because they are each getting at most $\delta^2$ of value from having any sort of indirect connection between them, and since $\delta^2 < \delta - c$, the extra value of a direct connection between them increases their utilities (and might also increase, and cannot decrease, the utilities of other agents).

When the cost rises above this level, so that $c > \delta - \delta^2$ but $c$ is not too high (see Exercise 1.3), it turns out that the unique efficient network structure is to have all players arranged in a "star" network. That is, there should be some central player who is connected to each other player, so that one player has $n-1$ links and each of the other players has 1 link. The idea behind why a star among all players is the unique efficient structure in this middle cost range, is as follows. A star involves the minimum number of links needed to ensure that all pairs of players are path connected, and it has each player within two links of every other player. The intuition behind this dominating other structures is then easy to see. Suppose for instance we have a network with links between 1 and 2, 2 and 3, and 3 and 4. If we change the link between 3 and 4 to be one between 2 and 4, we end up with a star network. The star network has the same number of links as our starting network, and thus the same cost and payoffs from direct connections. However, now all agents are within two links of each other whereas before some of the indirect connections involved paths of length three. This is pictured in Figure 1.2.4.

As we shall see, this is the key to the set of efficient networks having a remarkably simple characterization: either costs are so low that it makes sense to add links, and then it makes sense to add all links, or costs are so high that no links make sense, or

[18] This is just one of many possible measures of efficiency and societal welfare, which is a well-studied subject in philosophy and economics. How we measure efficiency has important consequences in network analysis and is discussed in more detail in Chapter 6.

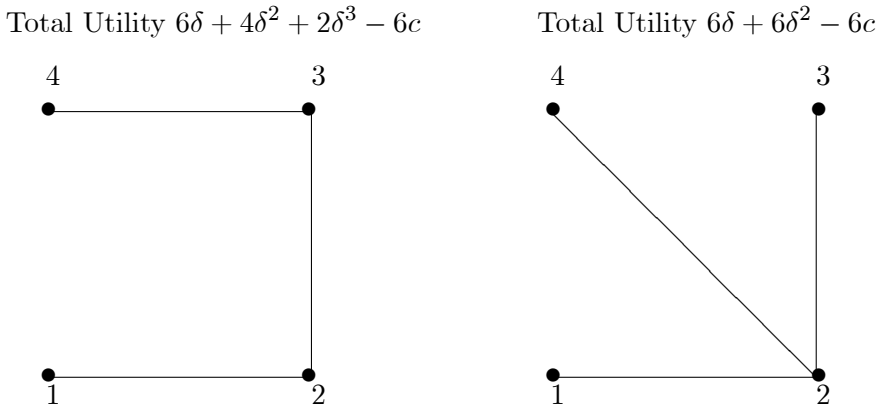



**Figure 1.2.4** ***The Gain in Total Utility from Changing a "Line" into a "Star".***

costs are in a middle range and the unique efficient architecture is a star network. This characterization of efficient networks being either stars, empty or complete, actually holds for a fairly general class of models where utilities depend on path length and decay with distance, as is shown in detail in Section 6.3.

We can now compare the efficient networks with those that arise if agents form links in a self-interested manner. To capture how agents will act, let us consider a simple equilibrium concept introduced in Jackson and Wolinsky [343]. This concept is called "pairwise stability" and involves checking two things about a network: first, no agent would raise his or her payoff by deleting some link that he or she are directly involved in; and second, no two agents would both benefit by adding a link between themselves. This stability notion captures the idea that links are bilateral relationships and require the consent of both individuals. If some individual would benefit by terminating some relationship that he or she is involved in, then that link would be deleted; while if two individuals would each benefit by forming a new relationship, then that link would be added.

In the case where costs are very low $c < \delta - \delta^2$, as we have already argued, the direct benefit to the agents from adding or maintaining a link is positive, even if they are already indirectly connected. Thus, in that case the unique pairwise stable network will be the efficient one which is the complete network. The more interesting case comes when $c > \delta - \delta^2$, but $c$ is not too high, so that the star is the efficient network.

If $\delta > c > \delta - \delta^2$, then a star network (that involves all agents) will be both pairwise stable and efficient. To see this we need only check that no player wants to delete a link, and no two agents both want to add a link. The marginal benefit to the center

player from any given link already in the network is $\delta - c > 0$, and the marginal benefit to a peripheral player is $\delta + (n-2)\delta^2 - c > 0$. Thus, neither player wants to delete a link. Adding a link between two peripheral players only shortens the distance between them from two links to one, and does not shorten any other paths - and since $c > \delta - \delta^2$ adding such a link would not benefit either of the players. While the star is pairwise stable, in this cost range so are some other networks. For example if $c < \delta - \delta^3$, then four players connected in a "circle" would also be pairwise stable. In fact, as we shall see in Section 6.3, many other (inefficient) networks can be pairwise stable.

If $c > \delta$, then the efficient (star) network will not be pairwise stable, as the center player gets only a marginal benefit of $\delta - c < 0$ from any of the links. This tells us that in this cost range there cannot exist any pairwise stable networks where there is some player who just has one link, as the other player involved in that link would benefit by severing it. For various values of $c > \delta$ there will exist nonempty pairwise stable networks, but they will not be star networks: as just argued, they must be such that each player has at least two links.

This model makes it clear that there will be situations where individual incentives are not aligned with overall societal benefits. While this connections model is highly stylized, it still captures some basic insights about the payoffs from networked relationships and it shows that we can model the incentives that underlie network formation and see when resulting networks are efficient.

This model also raises some interesting questions that we will examine further in the chapters that follow. How does the network that forms depend on the payoffs to the players for different networks? What are alternative ways of predicting which networks will form? What if players can bargain when they form links, so that the payoffs are endogenous to the network formation process (as is true in many market and partnership applications)? Uow does the relationship between the efficient networks and those which form based on individual incentives depend on the underlying application and payoff structure?

## 1.3 Exercises

EXERCISE **1.1** *A Weighted Betweenness Measure*

Consider the following variation on the betweenness measure in (1.1). Any given shortest path between two families is weighted by inverse of the number of intermediate

**Figure 1.3 Differences in Betweenness measures.**

nodes on that path. For instance, the shortest path between the Ridolfi and Albizzi involves two links and the Medici are the only family that lies between them on that path. In contrast, between the Ridolfi and the Ginori the shortest path is three links and there are two families, the Medici and Albizzi, that lie between the Ridolfi and Ginori on that path.

More specifically, let $\ell_{ij}$ be the length of the shortest path between nodes $i$ and $j$ and let $W_k(ij) = P_k(ij)/(\ell_{ij} - 1)$, (setting $\ell_{ij} = \infty$ and $W_k(ij) = 0$ if $i$ and $j$ are not connected). Then the weighted betweenness measure for a given node $k$ be defined by

$$WB_k = \sum_{ij:i\neq j,k\notin\{i,j\}} \frac{W_k(ij)/P(ij)}{(n-1)(n-2)/2}. \tag{1.5}$$

where we take the convention that $\frac{W_k(ij)}{P(ij)} = 0/0 = 0$ if there are no paths connecting $i$ and $j$.

Show that

- $WB_k > 0$ if and only if $k$ has more than one link in a network and some of $k$'s neighbors are not linked to each other,
- $WB_k = 1$ for the center node in a star network that includes all nodes (with $n \geq 3$), and
- $WB_k < 1$ unless $k$ is the center node in a star network that contains all nodes.

Calculate this measure for the the network pictured in Figure 1.3 for nodes 5 and 6.

Contrast this measure with the betweenness measure in (1.1).

EXERCISE **1.2** *Random networks*

Fix the probability of any given link forming in a Poisson random network to be $p$ where $1 > p > 0$. Fix some arbitrary network $g$ on $k$ nodes. Now, consider a sequence of random networks indexed by the number of nodes $n$, as $n \to \infty$. Show that the probability that a copy of the $k$ node network $g$ is a subnetwork of the random network on the $n$ nodes goes to 1 as $n$ goes to infinity.

[Hint: partition the $n$ nodes into as many separate groups of $k$ nodes as possible (with some leftover nodes) and consider the subnetworks that end up forming on each of these groups. Using the expression in (1.2) and the independence of link formation, show that the probability that the none of these match the desired network goes to 0 as $n$ grows.]

EXERCISE **1.3** *The Upper Bound for a Star to be Efficient*

Find the maximum level of cost in terms of $\delta$ and $n$, for which a star is an efficient network in the symmetric connections model.

EXERCISE **1.4** *The Connections Model with Low Decay*$^*$

Consider the symmetric connections model in a setting where $1 > \delta > c > 0$.

Show that if $\delta$ is close enough to 1, so that there is "low decay" and $\delta^{n-1}$ is nearly $\delta$, then in every pairwise stable network every pair of players have some path between them and that there are at most $n-1$ total links in the network.

In a case where $\delta$ is close enough to 1 so that any network that has $n-1$ links and connects all agents is pairwise stable, what fraction of the pairwise stable networks are also efficient networks?

How does that fraction behave as $n$ grows (adjusting $\delta$ to be high enough as $n$ grows)?

EXERCISE **1.5** *Homophily and Balance Across Groups*

Consider a society of two groups, where the set $N_1$ comprises the members of group 1 and the set $N_2$ comprises the members of group 2, with cardinalities $n_1$ and $n_2$, respectively. Suppose that $n_1 > n_2$. For an individual $i$, let $d_i$ be $i$'s degree (total number of friends) and let $s_i$ denote the number of friends that $i$ has that are within own group. Let $h_k$ denote a simple homophily index for group $k$, defined by $h_k = \frac{\sum_{i \in N_k} s_i}{\sum_{i \in N_k} d_i}$. Show that if $h_1$ and $h_2$ are both above 0 and below 1, and the average degree in group 1 is at least as high as the average degree in group 2, then $h_1 > h_2$. What are $h_1$ and $h_2$ in the case where friendships are formed in percentages that correspond to the relevant populations.

# Chapter 2

# Representing and Measuring Networks

With some feeling for network analysis under our belts, this chapter presents some of the fundamentals of how networks are represented, measured and characterized. This provides basic concepts and definitions that are the basis for the language of network research. Sprinkled throughout are some 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 will encompass 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 simple enough to be compact, intuitive, and tractable. As we proceed, I will try to make clear what is being admitted and what is being ruled out.

### 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 will also be referred to as "vertices," "individuals," "agents," or "players," depending on the setting. It is important to emphasize that nodes might be individual people, firms, countries, or other organizations; or a node might even be something like a web page belonging to some person or organization.

### 2.1.2 Graphs and Networks

The canonical network form is an undirected graph, where two nodes are either connected or not. This applies to situations where two nodes are either in a relationship with each other or not, but it cannot be that one is related to the second without the second being related to the first. This is generally true of many social and/or economic relationships, such as partnerships, friendships, alliances, acquaintances, etc. This sort of network will be central to most of the chapters that follow. However, there are other situations that we will examine that are better modeled as directed networks, where 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 are quite different. In particular, when links are necessarily reciprocal, then it will generally be the case that joint consent is needed to establish and/or maintain the relationship. For instance, in order to form a trading partnership, both partners need to agree. To maintain a friendship the same is generally true, as is maintaining a business relationship, alliance, etc. 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 and which are optimal, etc.

In what follows the default is that the network is undirected and I will be explicit 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

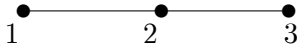



**Figure 2.1.2.** ***A network with two links.***

are linked to each other, or in other words which nodes are adjacent to one another.[1] In the case where the entries of $g$ take on more than two values, and can keep track of the intensity of 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. As such, I will 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* if it is required that $g_{ij} = g_{ji}$ 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 where there is a *link* between nodes 1 and 2, a link between nodes 2 and 3, but no link between nodes 1 and 3.

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

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

[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, co-workers, etc. 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 Page and Wooders [493] for some more general representations.

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

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 of all the links or edges that are in the graph. That is, we can then view a graph as a pair $(N, g)$, where $g$ is the a collection of links which are just listed as a subsets of $N$ of size 2. So, for instance, the network $g$ in (2.1) can equivalently be written as $g = \{\{1, 2\}, \{2, 3\}\}$, or simplifying notation a bit $g = \{12, 23\}$. So, we write $ij$ to represent 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 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 we will be interested in the exact labels of which nodes are in which positions in a network, and in other situations we will just care about the structure of the network. The idea that two networks or graphs are 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

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

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 2.1.2.

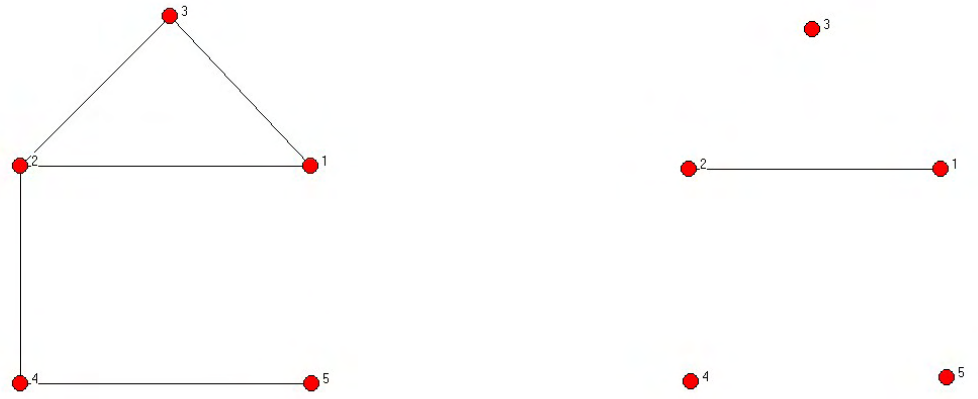



**Figure 2.1.2 A Network and the Network Restricted to** $S = \{1, 2, 5\}$

For any network $g$, let $N(g)$ be the set of nodes that have at least one link in the network $g$. That is, $N(g) = \{i \mid \exists j \text{ s.t. } ij \in g, \text{ or } ji \in g\}$.[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. In order to capture the indirect interactions in a network, it is important to model paths through a 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 $i_1 i_2, \ldots, i_{K-1} i_K$ such that $i_k i_{k+1} \in g$ 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

[3]Here 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, so that it consists of the set of involved nodes

network $g \in G(N)$ between nodes $i$ and $j$ is a sequence of links $i_1 i_2, \ldots, i_{K-1} i_K$ such that $i_k i_{k+1} \in g$ 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 the 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]

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 briefly 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 where a node appears at most once in the sequence.
- A geodesic between two nodes is a shortest path between them.

Note that if we follow 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 if we start with a network

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

and the set of links.

[5] The definition of path here is the standard one from the graph theory literature. In some of the network literature, the term path is used more loosely and is actually that of a "walk," so that nodes can be visited more than once. This can cause some confusion, which the reader should be aware of when reading the broader literature.

then $g^2$ 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). Then $g^3$ 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 we are seeing walks that have some 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.

### 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 or not, and in various applications will depend on whether things like communication are restricted only to follow 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 where direction is important, the definitions are just as stated above for undirected networks, but where the ordering of the nodes in each link now takes on an important role. For instance, we might be interested in knowing whether one can find one web page from another by following (directed) links starting from one page and leading to the other. In that case, I will refer to directed paths, directed walks, and directed cycles.

A *directed walk* in a network $g \in G(N)$ is a sequence of links $i_1 i_2, \ldots, i_{K-1} i_K$ such that $i_k i_{k+1} \in g$ (that is, $g_{i_k i_{k+1}} = 1$) for each $k \in \{1, \ldots, K-1\}$.

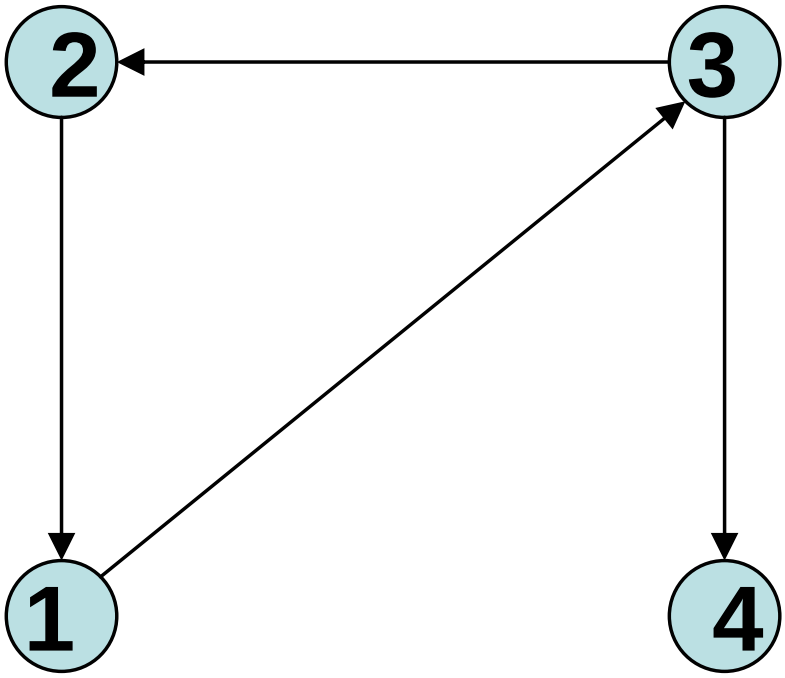



**Figure 2.1.4 A Directed Path from 2 to 4 (via 1 and 3); a Directed Cycle from 1 to 3 to 2 to 1; and a Directed Walk from 3 to 2 to 1 to 3 to 4.**

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 $i_K = j$, and 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 2.1.4.

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 will 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 to be present whenever there is a directed link 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 $\widehat{g}$. Similarly, we

define an undirected cycle or walk.

In Figure 2.1.4 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.

### 2.1.5 Components and Connected subgraphs

An important thing to keep track of in many applications of networks is which nodes can reach which other nodes through paths in the network. This plays a critical role in things 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 networks.

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 $\emptyset \neq N' \subset N$, $g' \subset g$, and

- $(N', g')$ 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 below 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 where $N$ is fixed or obvious, I use a notation where components are simply denoted $C(g)$. 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 $(N, g)$.

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

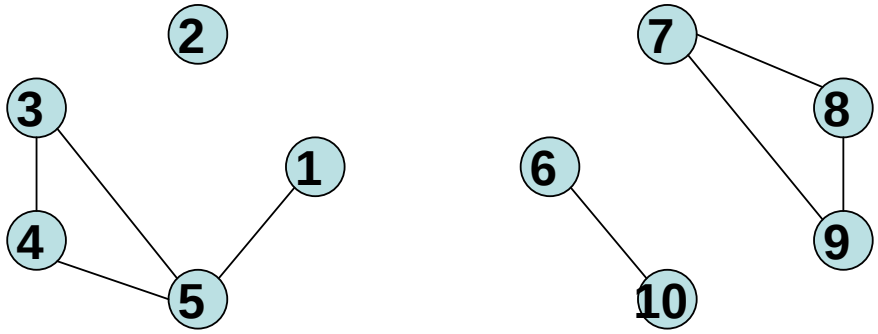



**Figure 2.1.5.** *A network with four components.*

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 2.1.5 induces the partition $\Pi(N, g) = \{\{1, 3, 4, 5\}, \{2\}, \{6, 10\}, \{7, 8, 9\}\}$ over the set of nodes.

So, a network is connected if and only if it consists of a single component (and so $\Pi(N, g) = \{N\}$).

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. One ways is to again work by ignoring the directed nature of links, and to consider the undirected network that has a link present if one is present in either direction. This defines one notion of connection and components. In some applications, where direction is important, for instance in the transmission of information, we will want to keep track of the directed nature of the network. In such cases, I will refer to *strongly connected* graphs or subgraphs, so that each node can reach each other via a directed path. Further definitions are specified as needed in what follows.

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

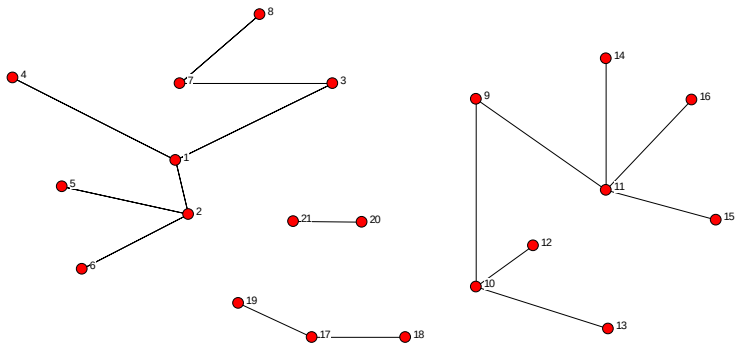



**Figure 2.1.6 Four Trees in a Forest**

### 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 2.1.6.

A particularly prominent forest network is a star. A *star* is a network such that 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.

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

The *complete network* is one where all possible links are present, so one where where $g_{ij} = 1$ for all $i \neq j$.

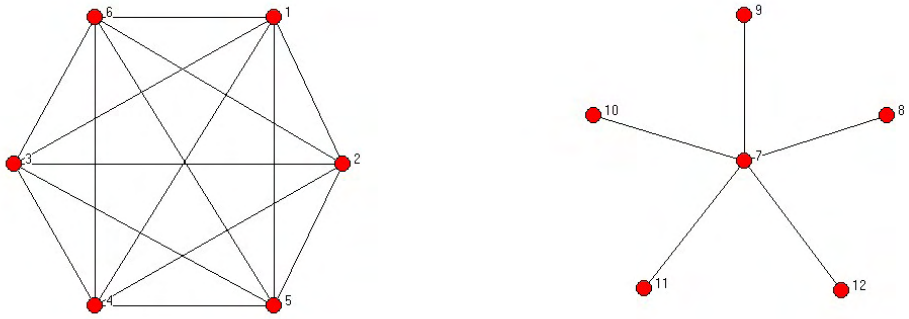



**Figure 2.1.6 A Complete Network on Six Nodes and a Star Network on Six Nodes**

A *circle* (also known as a cycle-graph) is a network that has a single cycle and 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 will be useful to distinguish between these.

### 2.1.7 Neighborhood

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

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

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

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 paths of no more than length 2, etc. The *two-neighborhood of a node* $i$ is

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

Inductively, all the nodes that can be reached from $i$ by paths of length no more than $k$ is 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 path of length no more than $k$.

Generally when referring to the *extended neighborhood* of a node $i$, that is all of the nodes it is path connected to or $N_i^n(g)$.

The above definitions also work for directed networks, where then the interpretation is that the nodes in $N_i^k(g)$ are the nodes that can be reached from $i$ via a directed path.

### 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 $i$'s neighborhood. Thus, a 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 would be the *in-degree*. The *out-degree* of node $i$ is the corresponding calculation $\#\{j : g_{ij} = 1\}$. These coincide in the case of an undirected network.

The *density* of a network keeps track of the relative fraction of links that are present, and is simply the average degree divided by $n - 1$.

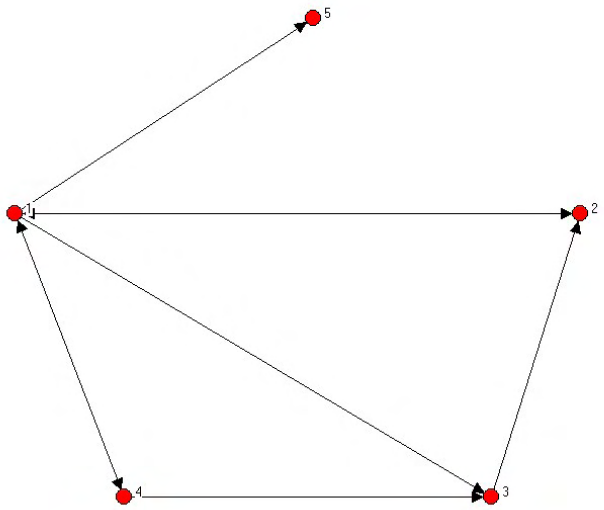



**Figure 2.1.8 A Directed Network on Five Nodes, Node 1 has In-Degree 2 and Out-Degree 4**

## 2.2 Some Summary Statistics and Characteristics of Networks

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

### 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]

[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.

For instance, a *regular* network is one where 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 we described in Section 1.2.3, where 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 that we saw in Section **??**, another prominent degree distribution is what is referred to as a "scale-free" degree distribution. These distributions date to Pareto [501], and appear in a wide variety of settings ranging from incomes , to word usage, to city populations, to 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 depends on the support of the distribution).[12] This means that if we increase the degree by a factor $k$, then we end up with a frequency that goes down by a factor of $k^{-\gamma}$. As this is true regardless of the starting degree $d$, it means that relative probabilities of degrees of a fixed relative ratio are the same regardless 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 $d^{-\gamma}$.

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, etc. I will often omit the $_P$ notation when $P$ is fixed.

Such 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 the following plots of these degree distributions when the average degree is ten (comparing the Poisson degree distribution from (1.4) with the secale free distribution from (2.2)) :

[10] For an informative overview, see Mitzenmacher [446].

[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, $z(\gamma) = \sum_{d=1}^{\infty} \frac{1}{d^\gamma}$.

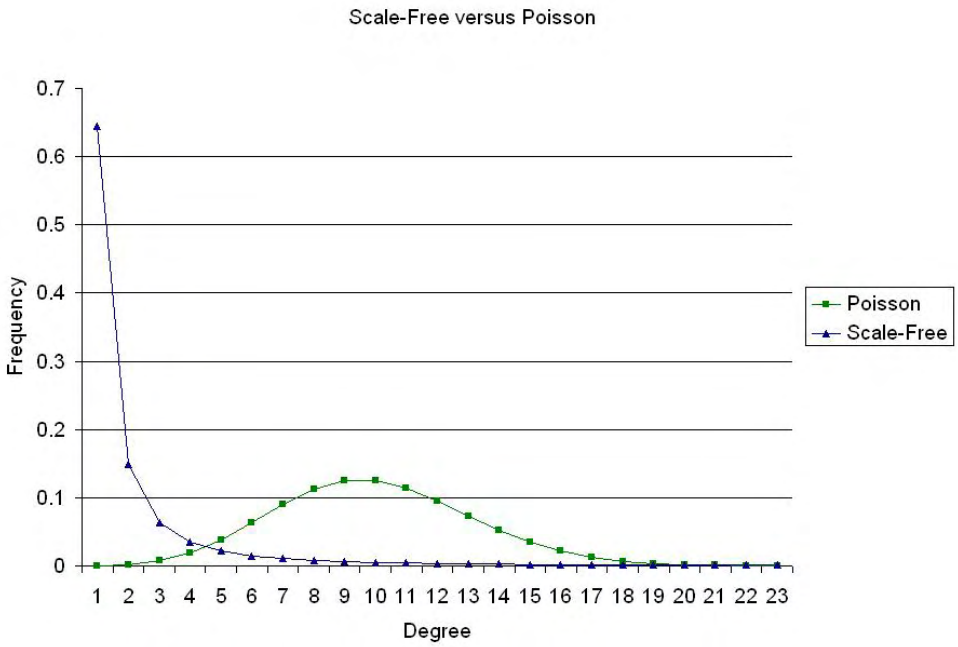



**Figure 2.2.1. Comparing a Scale-Free Distribution to a Poisson Distribution**

We can easily see the fatter tail of the scale-free distribution 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 (that is, we plot the log(frequency) versus the log(degree) instead of the raw numbers), then the differences in the upper tail (for higher degrees) become more evident:

Figure **??** also 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 is useful in trying to estimate $\gamma$ from data, as then linear regression can be used.

### 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.

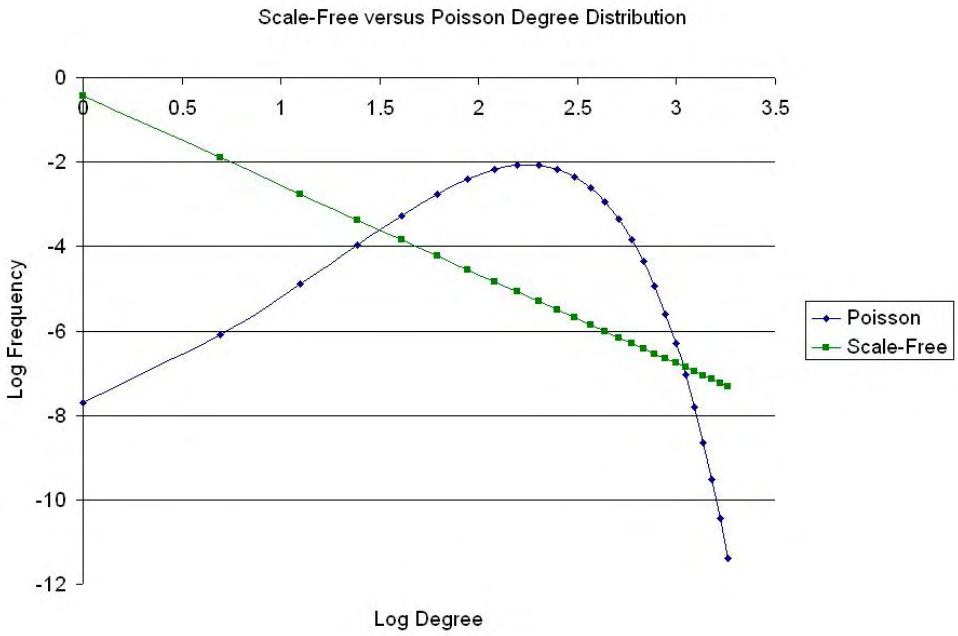



**Figure ??** Comparing a Scale-Free Distribution to a Poisson Distribution: LOG-LOG Plot

This leads us to another important characteristic of a network: namely 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 where each node has on average two links as in Figure 2.2.2 - the first network is a circle and the second is a tree.

Despite the fact that both networks have an average degree of 2, they are very clearly different in structure. The degree distribution picks up 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 (almost) half of the nodes have degree 1 (the exception is the root node that has degree exactly 2). However, we need other measures to pick up clear differences in these networks. For instance, the

[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 zero if there are no cycles).

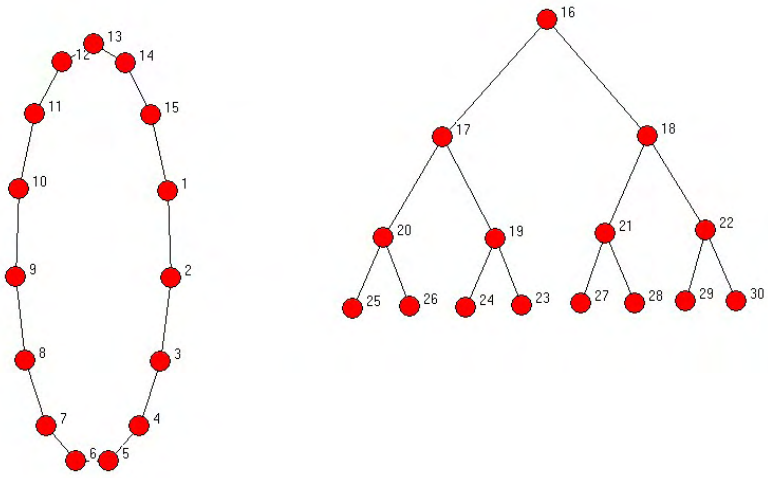



**Figure 2.2.2.** Circle and Tree

diameter of 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 $2log_2(n+1)$.[14]

While the diameter is one measure of path length, it only provides an upper bound on path length. *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 will be bounded above by the diameter; and in some cases 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 within the largest component, being careful to specify whether or not that component is a "giant" component (containing a non-trivial fraction of nodes).[15]

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

[15] There is a way to circumvent these problems. As Newman [480] 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 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

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, and then that entry is the number of shortest paths between those nodes. This same calculation provides shortest directed paths in the case of directed networks. 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 each other captures one facet of this. 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 3 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 2.2.3 both nodes 2 and 3 are in two different cliques.

One measure of cliqueshness is to count the number and size of the cliques of a network. One difficulty with this is that removing one link from a large clique can change the clique structure completely. For instance, removing one link from a complete

reciprocal twice leads to something similar to averaging path lengths directly, but working with the reciprocals eliminates the influence of infinite path lengths.

[16] For instance, there are more efficient ways of calculating powers of $g$ when it is diagonalizable (see Section 2.4.1), and 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 [**?**].

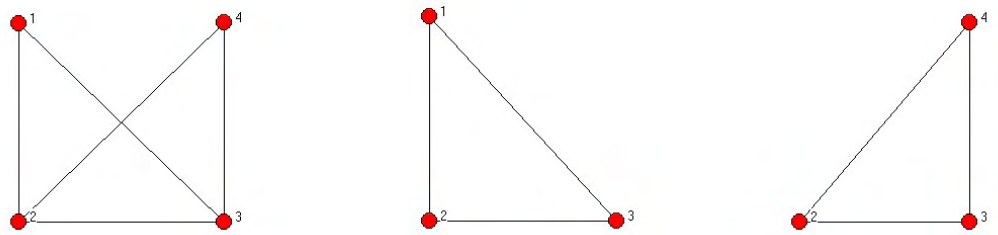



**Figure 2.2.3 A Network on Four Nodes and its Two Cliques**

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.

The most common way of measuring some aspect of cliqueshness is based on "transitive triples" or "clustering."[19] Examining undirected networks, the most basic clustering measure is simply to perform the following exercise. Look across all situations where two links both emanate from the same node, so for instance $ij$ and $ik$ both involve node $i$, and ask what proportion of the time it is that $jk$ 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, in order to distinguish it from the other

[19] Clustering in the way it is used here is a term that has come out of the recent random network literature (e.g., see Newman [480]). "Clustering" has an interesting history as a term, growing out of the earlier sociology literature, based on partitioning signed graphs into subsets where 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 [615]).

measures of clustering that follow.

A different measure that has also been used in the literature is similar to the clustering coefficient $Cl(g)$ above, except that instead of considering the fraction of fully connected triples out of the potential triples where at least two links are present, the measure does this on a node-by-node basis, and then averages across nodes. This 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 across all pairs of nodes that are linked to $i$, and then considers how many of them are linked to each other.[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 is a different calculation than 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 gives more weight to low-degree nodes than the clustering coefficient.

As an illustration of these two measures, let us compute them relative to the Florentine Marriage Network pictured in Figure **??**.

To compute the overall clustering coefficient, we first count how many configurations of the form $ij$, $jk$ there are in the network. For instance, there is one with Barbadori-Medici and Barbadori-Castellan. There are three with the Ridolfi at the center. 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 such configurations, 9 of them are completed. Thus, $Cl(g) = 9/47$. In terms of the average clustering, we compute the clustering for each family separately. For instance, the Barbadori have one possible pair and they are not connected, so the Barbadori have $Cl_{Barbadori}(g) = 0/1 = 0$. The Pazzi, Acciaiuol, Ginori, Lambertes, and Pucci are all 0 by convention (they each have one or no links). The Bischeri, Castellan,

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

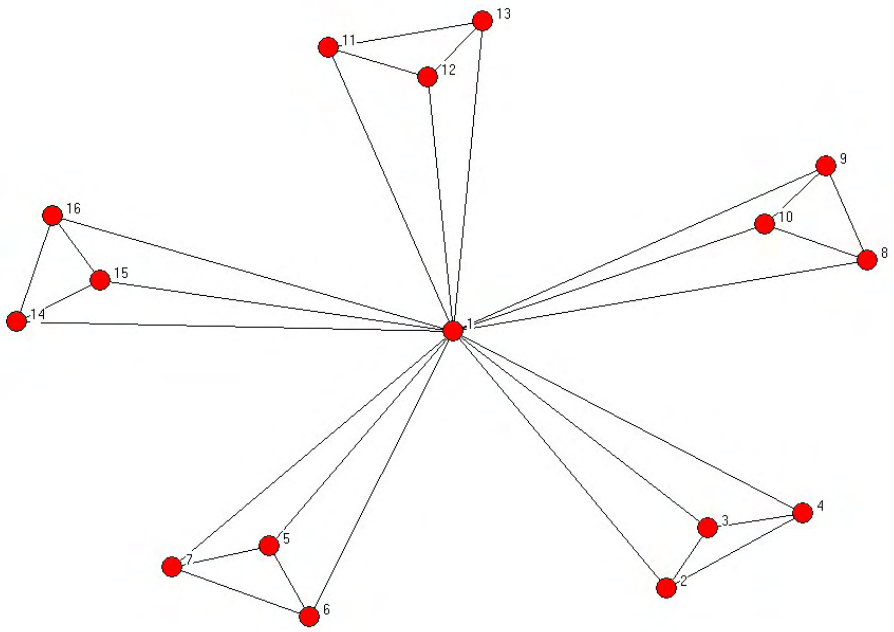



**Figure 2.2.3.** Differences in Clustering Measures

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 $Cl^{Avg}(g) = .15$ or 3/20. This is a bit less than the overall clustering $Cl(g)$, as we are including a number of 0's in the average clustering.

The above calculations show that it is possible to have these two common measures of clustering differ. While in that example the average clustering is lower than the overall clustering, it can also go the other way. Moreover, it is not hard to generate networks where the two measures can produce very different numbers for the same network. 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 we connect every other node. This is pictured in Figure 2.2.3.

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

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

node have individual clustering measures of 1. Thus, when averaging across them we end up with an average clustering coefficient converging to 1. However, with regards to the overall clustering coefficient, each time that a new triad is added the number of possible pairs of links goes up by 2 times how many links the center node already had (plus 3), while the number of those pairs that are completed only goes up by 3. Thus, overall clustering goes to 0.

Clearly, the two different measures of clustering are capturing different things and so there is no "right" or "wrong" measures. This does point out that such simple coefficients cannot give a full picture of the inter-relatedness of a network, but only an impression of some aspect of it.

In the case of directed networks one has further choices with regard to 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 looks at situations where 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 $k$; which follows the usual notion of "transitivity" of relationships.[22] The percentage of times in a network that the answer is "yes" is the *fraction of transitive triples*. This fraction is represented as follows.

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

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

### 2.2.4 Centrality

Most of the measures I have 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

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

[23] There are also hybrid measures (mixing ideas of directed and undirected links) where 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 [1] study of the www.

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 position that a node has, 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. The four main measurements that go into centrality measures are:[24]

1. degree - how connected a node is,
2. closeness - how easily a node can reach other nodes,
3. betweenness - how important a node is in terms of connecting other nodes,
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 foresee that they will capture complementary aspects of a node's position, and any particular measure will be better suited for some applications and less well suited 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 2 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 tells us how well a node is connected, in terms of direct connections.

Of course, degree centrality is clearly missing many of the interesting aspects of a network. In particular, it completely misses any aspect of 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, where the influence or marginal contribution

[24] See Borgatti **??** for more discussion about categorizing measures of centrality.

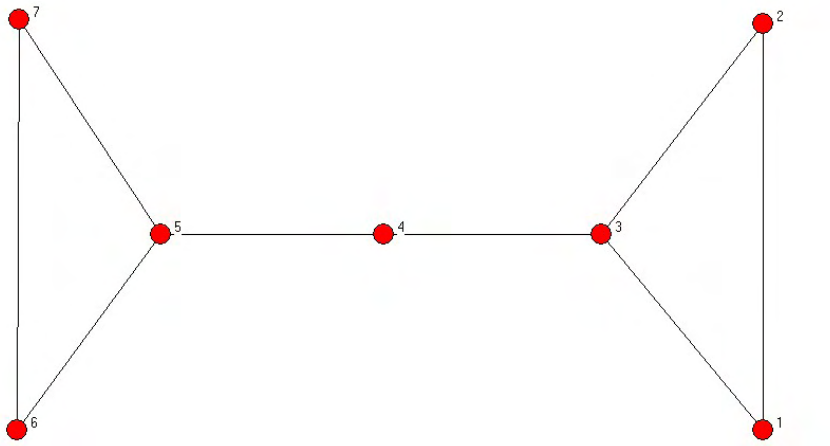



**Figure 2.2.4.** A Central Node with Low Degree Centrality

of a node to a network is important, we would prefer to have a centrality measure that would pick this up. For example, consider the network in Figure 2.2.4.

In the network in Figure 2.2.4 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 of node 4. If one deletes node 4, the component structure of the network changes. This might be very important if we are thinking about something like information transmission, where node 4 is critical in path-connecting nodes 1 and 7. This will be picked up by a measure such as betweenness. We also see that node 4 is relatively close to all of the other nodes, in that it is at most two links away from any other node, whereas each other node has at least one node at a distance of three or more. This would be important in applications where something is being conveyed or transmitted through the network (say an opinion or favor) and there is a decay of the 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 to have influence (e.g., conveying opinions). This brings us to the

next category of centrality measures.

### Closeness Centrality

This second class of measures keeps track of how close a given node is to each other node. One obvious "closeness"-based measure is just the inverse of the average distance between $i$ and any other node: $(n-1)/\sum_{j\neq i}\ell(i,j)$, 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 whole 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 each 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 is a centrality measure that is related to the symmetric connections model of Jackson and Wolinsky [343], as it is just the benefit that a node gets in that model of a network.

As $\delta$ gets close to one, it is easy to see that decay centrality measures how large a component a node lies in. As $\delta$ gets close to 0, then decay centrality gives infinitely more weight to closer nodes than farther nodes, 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 [237], and we first discussed it in Section **??** in the context of the Florentine marriages (see (**??**).

Letting $P_i(kj)$ denote the number of geodesics (shortest paths) between $k$ and $j$ that $i$ lies on, and $P(kj)$ the total number of geodesics between $k$ and $j$, we get an idea of how important $i$ is in terms of connecting $k$ and $j$ by looking at the ratio $P_i(kj)/P(kj)$. If this is close to 1, then $i$ lies on most of the shortest paths connecting $k$ to $j$, while if this 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

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

When we examine the network in Figure 2.2.4 with respect to betweenness centrality, we find that $Ce_4^B(g) = 9/15$, $Ce_3^B(g) = 8/15$, and $Ce_1^B(g) = 0$. This makes 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 [84]. Bonacich's measure is based on ideas that trace back to Seeley [**?**], Katz [361], and earlier work of Bonacich [83], 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 not only account for the fact that a node is connected or close to many other nodes, but that it is close to many other "important" nodes. This notion is central to citation rankings and things like 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 this 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 am using the term "prestige" to keep with the original, but one can also think of this as a way to measure "centrality." So, $i$ gains prestige from having a neighbor $j$ who has high prestige. However, this is corrected by how many neighbors $j$ has, so that as $j$ has more relationships then $i$ gets less prestige from being connected to $j$, all else held equal. This correction for the number of relationships that $j$ has might be thought of as correcting for the relative access or time that $i$ gets to spend 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 is a self-referential definition, so it is not immediately obvious that it is uniquely

(or always) defined. It does provide us with a series of equations and unknowns, so in principle 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 (non-zero) *column* is normalized to 1.[25] The relationship in (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 $c$.

Katz Prestige turns out to be more novel in directed networks than undirected ones. If indegree is the same as outdegree 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 provides a justification of degree centrality but not a new measure. In the case of a directed network, the normalization in $\widehat{g}_{ij} = g_{ij}/d_j(g)$ is generally by indegree, 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 from indegree and outdegree.[26]

When applied to the network in Figure 2.2.4, the Katz Prestige is the degree $P_4^K(g) = 2$, $P_3^K(g) = 3$, and $P_1^K(g) = 2$. This gives more "prestige" to the nodes 3 and 5 than to the middle node 4, which ends up with 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 gets 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 get 1/2 of each of their weight (so $2/2 + 2/2 + 2/2 = 3$).

A variation on this idea that avoids boiling down to degree centrality is where one does not normalize the network of relations $g$, is what is known as *eigenvector*

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

[26] However, if one normalizes by outdegree, then the measure will be outdegree.

*centrality*,[27] and was proposed by Bonacich [83]. Letting $C^e(g)$ denote the eigenvector centrality associated with a network $g$, the idea is that the centrality of a node is proportional to the sum of the centrality of its neighbors. Thus, we write $\lambda C_i^e(g) = \sum_j g_{ij} C_j^e(g)$. In terms of matrix notation,

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

where $\lambda$ is a proportionality factor. Thus, $C^e(g)$ is an eigenvector of $g$, and $\lambda$ is its corresponding eigenvalue. Given that it generally makes sense to look for a measure where these are nonnegative, the standard convention is to look for the eigenvector associated with the largest eigenvalue, which is nonnegative here.[28]

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

Katz [361] also introduced another way of keeping track of the power or prestige of a node. The idea is based on presuming that the power or prestige of a node is simply a weighted sum of the walks that it has emanating 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 gives higher weights to walks of shorter distance, in a similar way as in the connections model. So, it is a way of looking at all of the walks from some given node to other nodes in the network and weighting them by distance.

Note that $g\ \mathbb{1}$ (where $\mathbb{1}$ is the $n \times 1$ vector of 1's) 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 **??**, $g^k\ \mathbb{1}$ is the vector where each entry is the total 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

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

We can rewrite this as

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

For small enough $a > 0$, this is finite and then can be expressed as[29]

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

[27] Again, see Section 2.4 for background on eigenvectors.

[28] See Section 2.4.

[29] 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 $(\mathbb{I} -$

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 is expressed as the vector $ag\ \mathbb{1}$. Then a given node gets 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. This is often called "Bonacich Centrality," and can be expressed as

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

where $a > 0$ and $b > 0$ are scalars, and $b$ is sufficiently small so that (2.10) is well defined. [30]

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 someone decays with distance, while $a$ captures the base value on each node. In the case where $b = a$, the two measures clearly coincide.

Normalizing $a = 1$, we can calculate the Bonacich centrality of the network in Figure 2.2.4 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 seems to miss some aspects of the structure of the network. Closeness provides differences among 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 1,2,6, and 7 for any $\delta$, but the relative rankings of 3 and 5 relative to 4 depends on $\delta$. With a lower $\delta$ it looks more like degree centrality and favors nodes 3 and 5, while for higher $\delta$ it looks more like closeness or betweenness and favors node 4. The eigenvector centalities and self-referential defintions of Bonacich

$ag)P^{K2}(g, a) = ag\ \mathbb{1}$. A sufficient condition for this to be finite is that $a$ be smaller than 1 over the norm of the largest eigenvalue of $g$; and for this it is sufficient that $a$ be smaller than 1 over the maximum degree of any agent.

[30] Note that the scalar $a$ is no longer relevant, as it simply multiplies all of the terms. It is only useful in comparing back 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 2.2.4

| | Nodes 1,2,6,7 | 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 | 0.0 | .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 | 11 |
| Bonacich Centrality $b = 1/4$, $a = 1$ | 4.9 | 6.8 | 5.4 |

and Katz Prestige-2 all favor nodes 3 and 5, to varying extents. As $b$ gets lower it favors closer connections and favors higher degree nodes, while for higher $b$ longer paths become more important.

While these are some measures of centrality, and ones that I sometimes refer to in what follows, they 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 within a network.

## 2.3 Appendix: Some Basic Graph Theory

Here I present some basic results in graph theory that will be useful in other chapters.[31]

### 2.3.1 Hall's theorem and Bipartite Graphs

A *bipartite* network $(N, g)$ is one where one can partition the set of nodes $N$ 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 networ is pictured in Figure 2.3.1.

[31] Excellent texts on graph theory are Bollobás [79] and Diestel [**?**].

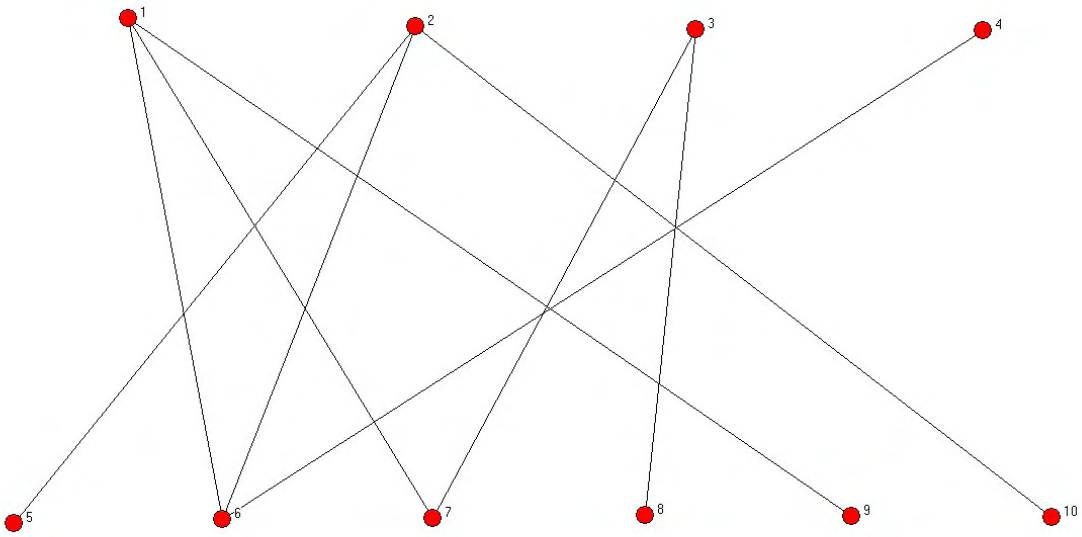



**Figure 2.3.1 A Bipartite Network**

This is often referred to as a "matching" setting (and in some cases a "marriage market"), where one group is referred to as "women" and the other as "men." It has applications to markets, where for instance one of the sets consists of buyers and the other of sellers, as well as things such as the assignment of students to schools, researchers to labs, and so forth. (See Roth and Sotomayor [**?**] 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 then 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 must be true for any subset of $S$, since we wish to match each element

**Figure 2.3.2.** ***Indepdendent Sets: {1}; {2};{3}; {1,3}; Maximal Independent Sets: {1,3}, {2}.***

to a different element from $B$. What Hall's Theorem [300] states is that this is not only a necessary condition, but also a sufficient condition for such a matching to exist.

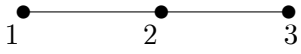
THEOREM **2.3.1** *[Hall's Theorem] Consider a bipartite graph $(N, g)$ with an associated bipartition of nodes $\{A, B\}$. There exists a matching of a set $S \subset A$ if and only if $|N_{S'}(g)| \geq |S'|$ for all $S' \subset S$.*

As we shall see in Chapter 10, this will be a useful theorem when it comes to 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.




The following observation (from Galeotti et al [256]) is straightforward, but useful when characterizing equilibria of games played on networks.

OBSERVATION **2.3.1** *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 in $g'$.*

The proof of this 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 [95]. To see how independent sets relate to equilibria, consider the following game played on a network.[32] Each player

[32] For more detailed definitions of game theoretic concepts and discussion of games played on networks see Chapter **??**.

chooses whether to buy a product (say a book) or not. If a player does not buy the book, then he or she can freely borrow the book from any of its neighbors who bought it.[33] 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 have 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 at all. This is what is known as a classic "free-rider" problem, but on a network. Each player would prefer have some neighbor buy the book and then "free-ride" by borrowing the book, rather than having to buy the book himself or herself. A (pure strategy) equilibrium in this game is simply a specification of which players buy the book such that: (i) no player who buys the book regrets it, and (ii) 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 where the players who buy the book form a maximal independent set. This follows from the observations that (i) implies that if some player buys a book then it must be that none of his or her neighbors buy the book, and (ii) 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 where the nodes are researchers who will attend a conference.[34] A link indicates 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) in order to make sure 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 that we need? That

[33] 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.

[34] This example is from Bollobás [79].

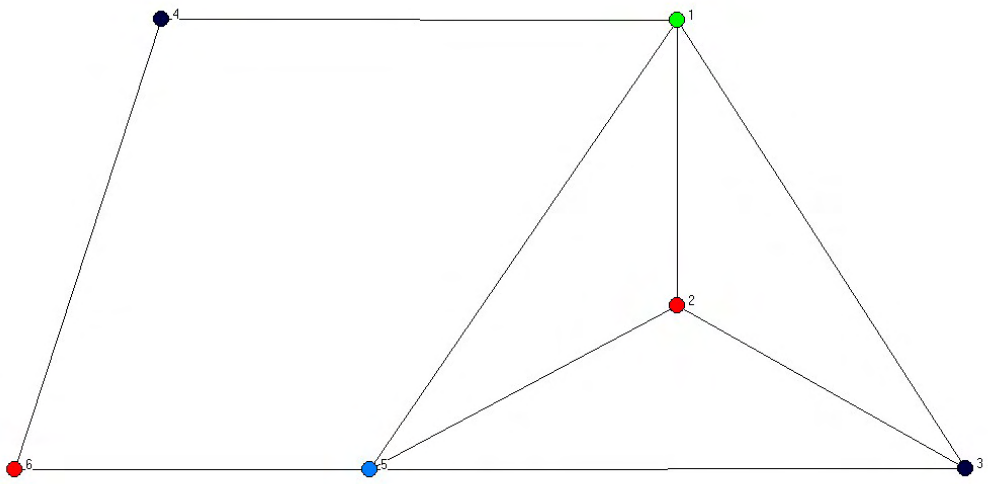



**Figure 2.3.3 A Planar Network on Six Nodes with Chromatic Number 4**

number is called the *chromatic number* of the graph.[35]

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.[36] This is a challenging problem that 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

[35] This is what 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 [79] or Diestel [**?**].

[36] 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 2.3.3, 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 green along with 1, then we have another 4-coloring. But then 2 is in its own element of the partition and does not form a maximal independent set.

without having any two links cross each other (so that links can only intersect at one of their involved nodes). The four color theorem is that every planar graph has a chromatic number of no more than four. This theorem was conjectured in the middle 1800's, and some false proofs were provided before it was proven in 1977 by Appel and Haken [**?**].[37] 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. The question then amounts to asking whether their exists a walk in the graph that contains each link in the graph exactly once and starts and ends at the same node.[38] Such a closed walk is said to be a Eulerian tour or circuit.

A walk is said to be *closed* if it starts and ends at the same node. It is clear that in order to have a closed walk that involves every link of a network exactly once it must be that each node in the network has an even degree.[39] This follows since 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 there to exist such a closed walk.

[37] That proof involved a computer verification that a series of 1482 cases each reduce to being 4-colorable. Shorter proofs have since been provided.

[38] 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.

[39] 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.

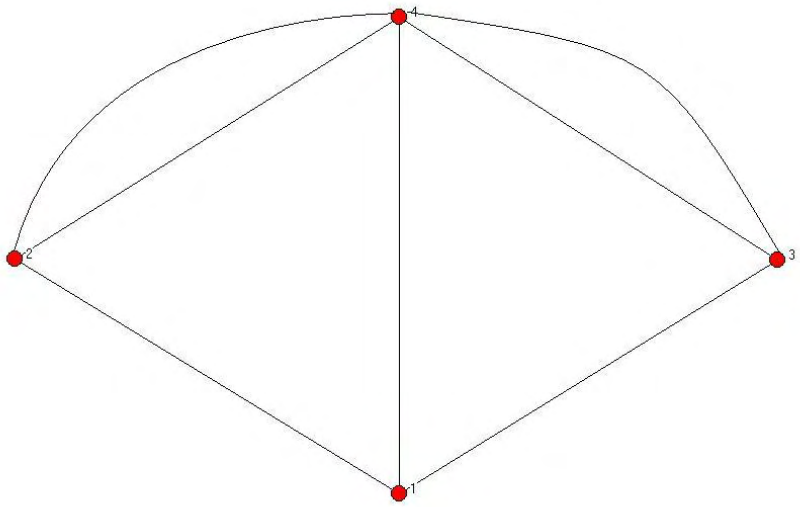



**Figure 2.3.4 The Multi-Graph for the Königsberg Bridge Problem**

THEOREM **2.3.2** *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 related question is whether there exists a "Hamilton path" that hits each node exactly once. Clearly a network that has a Hamilton cycle has a Hamilton path, while the converse is not true (consider a line).

Discovering whether or not a network has a Hamilton cycle is a much more challenging question than whether it has a Euler tour; and this has been an active area of research in graph theory for some time. It has direct applications to the "traveling salesman problem," where 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 [**?**]. 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 helps one see the paths to proving some of the later

results.

THEOREM **2.3.3** *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.*

The proof is as follows. First, let us argue that the network must be connected. This follows since 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. Let $i$ be the node that the path starts at and $j$ be the node it ends at. 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 this 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, we know that $j$ 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: $ik$, $k+1 \ldots j$, $j\ k-1$, $k-2 \ldots i$ (where the $\ldots$ correspond to the original path). The claim is that this is a Hamilton cycle. If this cycle does not include all nodes, then since the network is connected there is some node outside of the cycle connected to some node in the cycle. Then it is possible to make a path including that node and all nodes in the cycle, which contradicts the fact that the original path was of maximal length.

An example of a strengthening of the Dirac Theorem, is the following theorem due to Chvátal [**?**].

THEOREM **2.3.4** *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 else one needs 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 non-zero solutions to this equation (noting that a vector of 0's always solves (2.11)).

Eigenvectors come in two flavors: *left-hand eigenvectors* and *right-hand eigenvectors*, which are also known as *row eigenvectors* and *column eigenvectors*, respectively. This refers 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 an $1 \times n$ 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, and for many matrices, but not all, there will be as many eigenvector-eigenvalue pairs as there are dimensions, $n$.[40]

The usefulness of eigenvectors can be seen in terms of some of their applications. We have already seen one important application, in terms of calculating centrality or power, and in particular in terms of 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 then provides a self-referential problem, as the prestige has to be derived from the prestige. In that case, we look for an eigenvector with an eigenvalue of 1. The existence of an eigenvector with an eigenvalue of 1 in this context is implied by the Perron-Frobenius Theorem.

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 will exist

[40] 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 we rescale $v$.

a nonnegative right-hand eigenvector $v$ that solves (**??**) 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.[41]

Another type of application where eigenvectors are quite useful is in examining the steady state or limit point of some system. Here we might think of $T$ as a transition matrix. So, starting with some column vector $v$, the system transitions to some new vector $Tv$. A steady state of such a system, or a convergence point, will often be a point such that $v = Tv$, so that once one reaches $v$, one stays there. Again, this is an eigenvector of $T$, which has a unit eigenvalue. These 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 $T$ represents probabilities of transferring from one state to another. This is again a stochastic matrix (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 in a set of unkowns. 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

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

where $I$ is the identity matrix (with 1's on the diagonal and 0's elsewhere). In order for this equation to have a nonzero solution $v$, it must be that $T - \lambda I$ is a singular (non-invertible) matrix.[42] Thus, the *characteristic equation* of $T$ is that

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

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

[41] 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.

[42] This is a matrix where some rows are linear combinations of other rows, or similarly for columns, and this corresponds to having a determinant that is 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

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

This is the *diagonal decomposition* of $T$, if it exists, and then $T$ is said to be *diagonalizable*.

It is sometimes useful to note that $V^{-1}$, if it is 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) it follows that $VTV^{-1} = \Lambda VV^{-1} = \Lambda$. 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.

This is useful in calculating higher powers of $T$ (which, for instance recalling Section **??**, is useful in calculating the walks of $T$ if $T$ has 0,1 entries). From (**??**) it follows that

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

and more generally that

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

This helps in calculating speeds of convergence, as in Section **??**.[43]

[43]It is worth noting that this 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 just involves raising the diagonal entries to a power.

## 2.5 Exercises

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

EXERCISE **2.2** *Facts about Trees.*

Show the following:

- 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.

EXERCISE **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.

EXERCISE **2.4** *Centrality Measures*

Consider a two link network among three nodes. That is let the network consist of links 12 and 23.

Calculate the Katz-prestige (based on (2.5)) of each node, and compare this to the degree centrality and betweenness centrality for this network.

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 $b = 3/4$?

EXERCISE **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. $Cl_i(g) \geq Cl_j(g)$ whenever $d_i(g) \geq d_j(g)$, and

b. $Cl_i(g) \leq Cl_j(g)$ whenever $d_i(g) \geq d_j(g)$.

Hint: Write 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.

EXERCISE **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 [464] defines the set of nodes $S$ to be *r-cohesive* with respect to a network $g$ if each node in $S$ has at least $r$ 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 [632] 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 $r$ of its links staying 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

$$\min_{S' \subset S} \frac{d(S', S, g)}{\sum_{i \in S'} d_i(g)} \geq r, \tag{2.16}$$

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.

EXERCISE **2.7** *Independent Sets*

Show that there is a unique network on $n$ nodes that is connected and 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.

EXERCISE **2.8** *Independent Sets and Equilibria*

Prove Observation 2.3.1.

EXERCISE **2.9** *Euler Tours*

Prove Theorem 2.3.2. (Hint: First argue that any longest walk that does not involve any link more than once must be closed.)

# Chapter 3

# Empirical Background on Social and Economic Networks

There are numerous and extensive case studies of a variety of social and economic networks. Through such studies we have learned an immense amount about the structure of networks. In this chapter, I discuss some of the basic stylized facts and hypotheses that have come out of decades of empirical research on social and economic networks. As this literature is much too large to survey here, I focus on the fundamental characteristics of networks, mainly dealing with the structural aspects of networks, and some of the hypotheses that we return to in later chapters.

I begin with two cautions regarding some of the stylized facts from the literature. First, examining the structure of any given social network is a formidable task that faces significant hurdles associated with how to define and measure links or relationships. For instance, a primary tool for estimating social networks is to use various sorts of surveys or interviews of involved parties. Given that individuals have hundreds or even thousands of social relationships, getting them to recall the relevant ones with any desired accuracy is difficult.[1] In addition, it may be impossible to contact or observe all nodes in the network, and when contacted they may have reasons to distort or conceal relationships. Even recent studies of web pages, co-authorship, email, and citation networks, where data are more easily obtained, have other measurement idiosyncrasies. Beyond this, networks change over time and overlap in various ways. Close friends may fail to interact for long periods. Much of the information that we have about

[1] There is a sizeable literature on techniques for measuring social networks, as well as dealing with other measurement issues such missing data, biases in responses, etc. For instance, see Marsden [427], Bernard [52], and Bernard et al [53].

the structure of social networks comes from limited measurements of links that often take a static and discrete view of something that is inherently dynamic and volatile. Second, as there are biases and idiosyncrasies associated with each data set, and data are often collected and encoded in different ways, little has been done to systematically determine the prevalence of characteristics across ranges of social settings.[2] Thus, much of what is discussed below is based on what might be termed anecdotal evidence gleaned from various case studies, and the stylized facts reported below should be interpreted with the appropriate caution, and there is a need for broader systematic studies and comparisons of networks across social settings.

## 3.1 The Prevalence of Social Networks

Social relationships play a critical role not only in day-to-day life and behavior, but also in determining long run welfare. They affect the opinions that we hold and the information that we see, and are also often the key to accessing resources. While this is self-evident and gives sociology its foundation, quantifying the extent to which social relationships play roles in various aspects of life is an illuminating exercise.

One of the most robust and well-studied roles of social networks is in obtaining employment. There have been a number of studies of how social contacts matter in obtaining information about job openings. Such studies began in the late 1940's and there is now a rich base of information on this subject.[3] One of the earliest studies, by Myers and Shultz [472], was based on interviews with textile workers. They found that 62 percent had found out about and applied to their first job through a social contact, in contrast with only 23 percent who applied by direct application, and the remaining 15 percent who found their job through an agency, ads, etc. A study by Rees and Shultz [531] showed that these numbers were not particular to textile workers, but applied very broadly. For instance, the percentage of those interviewed who found their jobs through the use of social contacts as a function of their profession was: typist 37.3 percent, accountant 23.5 percent, material handler 73.8 percent, janitor 65.5 percent, and electrician 57.4 percent. Moreover, the prevalent use of social contacts in finding jobs is robust across race and gender.[4]

[2]There are authors, such as Watts [620] and Newman [480], who have looked across (a few) case studies to suggest some common features.

[3]For a recent overview of research on social networks in labor markets see Ioannides and Loury [**?**].

[4]See Corcoran, Datcher, and Duncan [**?**] for comparisons across race and gender, and Pellizzari [510] for data across countries.

The role of social networks is not unique to labor markets, but has been documented much more extensively. For example, networks and social interactions play a role in crime: Reiss [533], [534] finds that two thirds of criminals commit crimes with others, and Glaeser, Sacerdote and Scheinkman [**?**] find that social interaction is important in determining criminal activity, especially with respect to petty crime, youth activity in crime, and in areas with less intact households. Networks have also been studied with regards to various markets: Uzzi [598] finds that relation specific knowledge is critical in the garment industry and he documents how social networks play a key role in that industry; and Weisbuch, Kirman, Herreiner [626] study repeated interactions in the Marseille fish market and discuss the importance of the network structure. Social networks also serve a vital role in the provision of social insurance. For instance, Fafchamps and Lund [221] show that social networks are critical to the understanding of risk-sharing in rural Philippines, and De Weerdt [176] analyzes risk-sharing in parts of Africa. The set of case studies is much more extensive than this list indicates, and also includes extensive analyses of networks in disease transmission, in the diffusion of language and culture, in the collaboration on scientific research and invention, in the citation of articles, in the formation of opinions, in political activity, in choices of products to buy, and interactions of boards of firms - just to name a few other applications.[5]

I now turn to discuss some of the regularities and stylized facts about social networks that some of these studies have revealed.

## 3.2 Observations about the Structure of Networks

The following are characteristics that have been exhibited by a variety of social networks.

### 3.2.1 Diameter and Small Worlds

The stylized fact that large social networks exhibit features of "small worlds" (see Milgram [444]) is one of the earliest, best-known, and most extensively studied aspects of social networks. The term "small worlds" embodies the idea that that large networks

[5]The analysis also moves beyond social networks per se, to include things like analyses of the co-appearance of literary (comic-book) superheroes.

tend to have small diameter and small average path length.[6]

Stanley Milgram [444] pioneered the study of path length through a clever experiment where people had to route a letter to another person who was not directly known to them. Letters were distributed to subjects in Kansas and Nebraska, who were told the name, profession, and some approximate residential details about a "target" person who lived in Massachusetts. The subjects were asked to pass the letter on to someone whom they knew well and would be likely to know the target or to be able to pass it on to someone else, etc., with the objective of getting the letter to the target. While roughly a quarter of the letters reached their targets, the median number of hops for a letter to reach a target was 5 and the maximum was 12. Given that the letters should not be expected to have taken the shortest path, this is a startlingly small number. In addition, given the chains of interactions needed to get a letter from an initial subject to the target, the fact that a quarter of the letters reached their targets is also an impressive figure, especially in light of the fact that response rates in many voluntary surveys are on the order of twenty to thirty percent.[7]

To get some feeling for why many social networks exhibit small diameters, it is useful to think about neighborhood sizes. Most people have thousands of acquaintances. Depending on whether one keeps track of strong relationships or casual acquaintances, this might vary from the order of tens or hundreds to the order of thousands for a typical adult in a developed country (e.g., see Pool and Kochen [518], which is a key early study of small worlds). If we take a conservative estimate that a given individual has 100 relatives, friends, colleagues, and acquaintances with whom they are in somewhat regular contact, then we end up with a very rough calculation (ignoring clustering and treating the network as if it were a tree) of $100^2$ or 10,000 friends of friends, and 1 million friends of friends of friends. By the time we move out four links, we have covered a nontrivial portion of most countries. While this overestimates the reach of a network since it treats the network like a tree and ignores the clustering exhibited in most networks, it still provides a feeling for orders of magnitude. If we count more casual acquaintances, and use a figure on the order of 1000 acquaintances per person, then a tree network reaches a million nodes within a path distance of two and reaches

[6] See Watts [620] for more discussion. This stylized fact is captured in the famous "six degrees of separation" of John Gaure's play, and actually dates to a 1929 play called "Chains" by Frigyes Karinthy.

[7] This study has been replicated and extended a number of times. A recent example is research by Watts [**?**], who used email messages in a study involving nearly 50000 subjects in 157 countries and found similar sized chains.

a billion within a path distance of three.

Other examples provide similar impressions of path length and diameter measurements of observed networks. Watts and Strogatz [623] report a mean distance of 3.7 in a network among actors where a link indicates that two actors have been in a movie together. Studies of networks of co-authorship in scientific journals also report relatively small path lengths and diameters on larger numbers of nodes. Here a link represents the co-authorship of a paper during some time period covered by the study. The well-known and prolific mathematician Paul Erdös had many co-authors, and as a fun distraction many mathematicians (and economists for that matter) have found the shortest path(s) from themselves to Erdös. For example, an author who co-authored a paper with Erdös has an Erdös number of 1. An author who never directly co-authored a paper with Erdös, but who co-authored with a co-author of Erdös has an Erdös number of 2, and so forth. There are also some interesting patterns that emerge in such networks in terms of how they grow. [8] These networks are of scientific interest themselves, as they tell us something about how research is conducted and also how information and innovation might be disseminated. Similar studies have now been conducted in various fields, including mathematics (Grossman and Ion [**?**], de Castro and Grossman [**?**]), biology and physics (Newman [**?**], [**?**]), and economics (Goyal, van der Leij and Moraga-González [286]). Various statistics from these studies give us some impression of the network structure, as shown in Table 3.1.[9]

Here we see that despite the non-comparabilities of the networks along some dimensions (e.g., average degree, clustering, and size of the largest component), the average path length and diameters of each of the networks are very comparable. Moreover, these are of an order substantially smaller than the number of nodes in the network. This gives us an impression of the "small-world" nature of social networks.

To see how dramatic this effect can be, consider the average number of links it

[8] A web site (www.oakland.edu/enp/) maintained by Jerry Grossman, Patrick Ion, and Rodrigo de Castro provides a part of that graph. The American Mathematical Society website also provides platform that gives a shortest path between two authors. A similar analysis is of the "Kevin Bacon" network (see the web site at the computer science department at the University of Virginia, www.cs.virginia.edu/oracle/), where a link indicates that two actors appeared in the same movie. In 2004, William Tozier auctioned (on eBay) a promise to co-author an article, as that would provide the purchaser with an Erdös number of 5 as Tozier's is 4. This led to a winning bid of over one thousand dollars and a resulting controversy, as well as a number of other such auctions (see *Science News Online*, June 12, 2004, vol. 165, no. 24).

[9] As these networks are not connected (there are many isolated authors), the figures for average path length and diameter are reported for the largest component.

Table 3.1: Co-Authorship Networks

| | Biology | Economics | Math | Physics |
|---|---|---|---|---|
| number of nodes | 1520521 | 81217 | 253339 | 52909 |
| average degree | 15.5 | 1.7 | 3.9 | 9.3 |
| average path length | 4.9 | 9.5 | 7.6 | 6.2 |
| diameter of the largest component | 24 | 29 | 27 | 20 |
| overall clustering | .09 | .16 | .15 | .45 |
| fraction of nodes in the largest component | .92 | .41 | .82 | .85 |

takes to get from one web page to another on the world wide web. Lada Adamic [1] analyzed a sample of 153,127 web sites.[10] She found that there existed a (undirected) path starting at one page and ending at another in 85.4 percent of the possible cases; and that in those cases the average minimum path length was only 3.1. In looking for directed paths, she found that of the 153,127 web sites, there was a strongly connected component of 64,826 sites (so that any web site in that component could be reached via a directed path from any other web site in that component). The average minimum directed path length in that component was 4.2. Again, while not all pairs of sites are path-connected, the fact that it takes so few clicks to get from many of the sites to many others is impressive.[11]

### 3.2.2 Clustering

Another interesting observation about social networks is that they tend to have high clustering coefficients relative to what would emerge if the links were simply determined by an independent random process. Ideas behind clustering have been important in sociology since Simmel [560] who pointed out the interest in triads (triples of mutually connected nodes). A variety of large socially generated networks exhibit clustering

[10]This was based on a data set collected by Jim Pitkow of Xerox PARC. The initial data set contained 259,794 web sites and consisting of over 50 million pages. The network was trimmed of any "leaf nodes".

[11]It is worth noting that the data were collected via an algorithm that followed links in order to locate nodes, and such web-crawling algorithms necessarily introduce some bias in the portion of the overall network that they identify and particularly with respect to path structure.

measures much larger than would arise if the network were generated at random. For instance, let us re-consider the networks of researchers that have been analyzed in various fields of study. For instance, Newman [480] reports overall clustering coefficients of 0.496 for computer science, and 0.45 for physics, while Grossman [298] reports an overall coefficient of 0.15 in mathematics. To get an idea of how this compares with the clustering that would appear in a purely random network, let us consider the physics network which has 52,909 nodes and an average degree of 9.27. A purely random network that had this average degree would have a probability of any given link forming of 9.27/52908, or roughly .00018. For such a purely random network, the chance that link $ik$ is present when $ij$ and $jk$ are present is simply the probability that $ik$ is present, which is then .00018. This tells us that the clustering of .45 is roughly 2500 times greater than the clustering we would see in a random network of the same size and connectivity. We can also examine analogous numbers for a similar network constructed for researchers in economics. The data of Goyal, van der Leij, and Moroga-Gonzalez [286] covering papers published in economics journals in the 1970's has a total of 33770 nodes and an average degree of .894.[12] The clustering they report for that network is .193, whereas the corresponding clustering for a purely random network of the same degree is on the order of .894/33770 or .000026. Here the observed clustering is almost 10000 times larger than in the random network.[13]

Similarly high clustering has been observed in a variety of other contexts. For example, Watts [620] reports a clustering coefficient of 0.79 for the network consisting of movie actors linked by movies in which they have co-starred. Several studies have also analyzed clustering in the world wide web. Adamic [1] gives a clustering measure of 0.1078 on the world wide web data set mentioned in Section **??**. To get a feeling for how large this clustering measure is, note that we expect a purely random graph with the same number of links to have a clustering coefficient of 0.00023, so that the observed network has about 469 times more clustering than if links were formed independently.

[12] The data in Table **??** are from the 1990's rather than 1970's, and have more nodes, higher average degree and slightly lower clustering.

[13] Note in such collaboration networks, as there may be many co-authors on any given paper, clustering in this particular application partly reflects the fact that a multi-co-authored paper provides a complete set of connections between the authors. Given large numbers of co-authors per paper in physics, this partly explains the high clustering number there. The economics data exhibits less of this, as there less than 4 percent of all papers involved more than two co-authors, while roughly 25 percent of all papers had two co-authors.

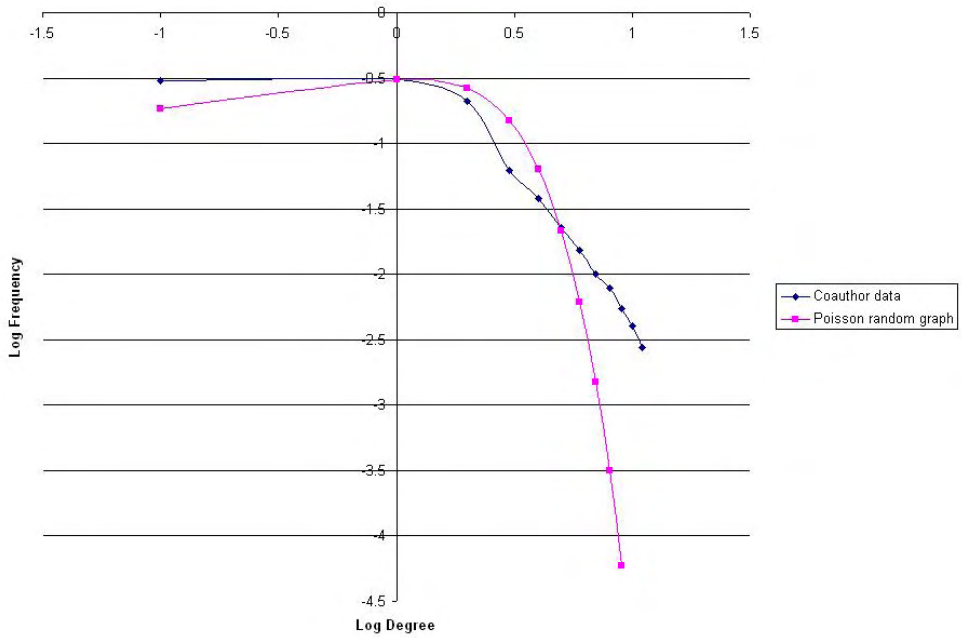



**Figure 3.2.3.** Comparison of the Degree Distributions of a Co-Authorship Network and a Poisson Random Network with the same Average Degree

### 3.2.3 Degree Distributions

As we saw in Table 3.1, networks differ in their average numbers of links. For instance, in Table 3.1, the number of co-authors per paper varies dramatically across fields, and there are other differences in social structure across fields. For instance, in the economics data set, there are on average 1.6 authors per paper (and only 12 percent of papers have more than two authors), while in the biology data there are on average 3.8 authors per paper. Although the average degree of a network provides a rough feeling for connectivity, there is much more information that we would like to know. For instance, how variable is the degree across the nodes of the network? We get a much richer feeling for the structure of a social network by examining the full distribution of node degrees rather than just looking at the average.

To see an example of such a distribution, consider Figure 3.2.3 below, which provides a log-log plot of the frequency distribution of degrees from the economics co-authorship data from Goyal, van der Leij, Moraga-Gonzalez [286].

The degrees of economists in the data set range from 0 to over 50. The distribution

also has an interesting shape. It clearly exhibits some curvature. However it also shows "less" curvature than the distribution of degrees generated from a network with the same number of links, but where the links are chosen independently with identical probability (a Poisson random graph, as discussed in Section **??**). What this indicates is that there are more economists with very high degree and more with very low degree than we would see in a network where links were generated uniformly at random.

This "fat-tailed" property is not unique to this network, as discussed already in Section 2.2.1. There has been a good deal of attention paid to the observation that the degree distributions of many observed large networks tend to exhibit "fat tails." Price [521] was the first to document such distributions in a network setting, observing that citation networks among scientific articles seemed to follow a power law (both in terms of in and out degree). It has been loosely said that these distributions approximate a "scale-free" or "power-law" distribution, at least in the upper tail. This refers to a frequency of a given degree being proportional to the degree raised to a power, so that the probability or frequency of a given degree can be expressed as

$$P(d) = cd^{-\gamma}, \tag{3.1}$$

where $c > 0$ and $\gamma > 1$ are parameters of the distribution, and hence the term "power-law." The "scale-free" aspect refers to the fact that if we consider the probability of a degree $d$ and compare that to a degree $d'$, then the ratio of $P(d)/P(d') = (d/d')^{-\gamma}$. Now suppose that we double the size of each of these degrees. We find that $P(2d)/P(2d') = (d/d')^{-\gamma}$. It is easy to see that rescaling $d$ and $d'$ by any factor will still give us this same ratio of probabilities, and hence the relative probabilities of different degrees just depends on their ratio and not on their absolute size. This explains the term "scale-free."[14]

For example, Figure 3.2.3 shows the degree distribution from the data set of Albert, Jeong and Barabási [8], which is the distribution of in-degrees from the network of links

[14] A discrete distribution, such as above where $d$ can only take on values $\{0, 1, 2, \ldots\}$, is sometimes hard to work with in terms of estimating expected values and conditional expectations, and so in some cases it is useful to use an approximation in the form of a continuous distribution where $d$ can take on non-integer values. The canonical continuous distribution satisfying a power law is a Pareto distribution, named for Vilfredo Pareto [501] who studied the distribution of wealth across individuals, among other things. The cumulative distribution function for a Pareto distribution with support $[1, \infty)$ and where $\gamma > 1$ is $F(d) = 1 - d^{-\gamma+1}$. The corresponding density is then $f(d) = (\gamma - 1)d^{-\gamma}$, which is of a similar form to the probability given in (3.1). If one tries to estimate the cumulative distribution function corresponding to (3.1), one would end up with $Prob[d \leq d'] = \sum_0^{d'} cd^{-\gamma}$, which does not have a nice closed form, but is approximately $1 - c'd^{-\gamma+1}$, where $c' > 0$ is a constant.

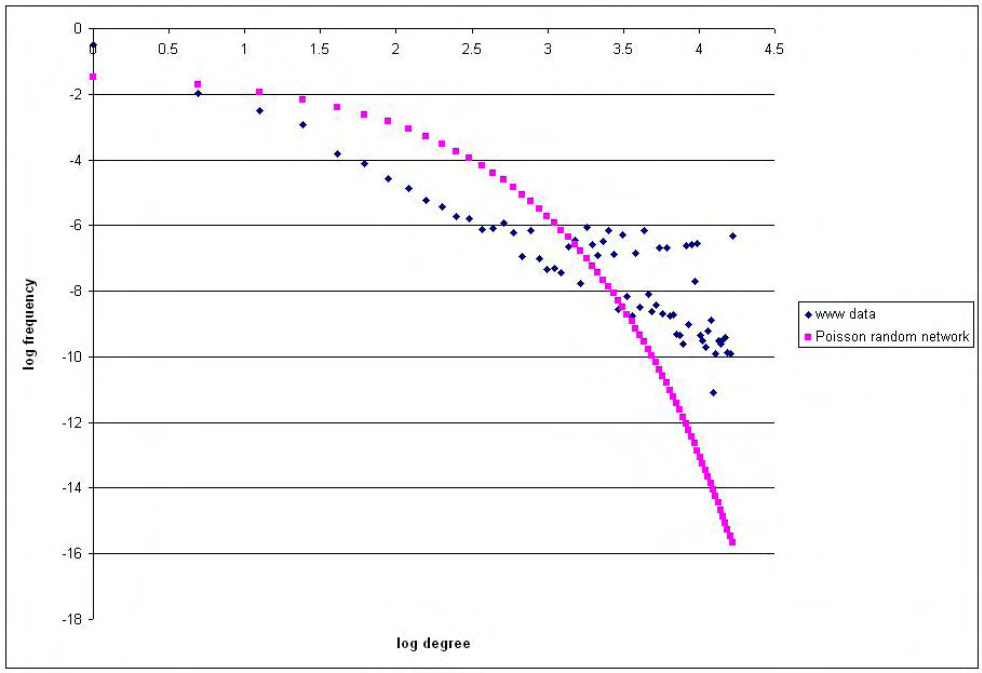



**Figure 3.2.3.** Distribution of In-Degrees of Notre Dame WWW from Albert, Jeong, and Barabasi [8] compared to a Poisson Random Network

among web pages on the Notre Dame world-wide-web in the late 1990s.

If we estimate the $\gamma$ for a scale-free distribution of the type $P(d) = cd^{-\gamma}$ from a log-log regression on these data, we find an estimate of -2.56.[15]

Such scale-free distributions and fat tails appear well beyond network applications, such as word usage (Estoup [**?**] and Zipf [639]), plant classifications (Willis [629] and Yule [636]), city size (Auerbach [**?**] and Zipf [639]), and article citations (Price [521]).[16] There is a natural explanation for them (discussed in detail in Section **??**).

There is a very important caution to be mentioned here. It is clear that many network degree distributions exhibit "fat tails" when compared to a Poisson random graph, and that many of the other mentioned applications also have such fat tails. However, it is not so clear that these distributions are really power distributions. Most of these claims are made simply by examining log-log plots that appear approximately linear, and then fitting a regression and finding a coefficient. On a plot of log of

[15] This is as estimated by Jackson and Rogers [337].

[16] See Mitzenmacher [446] for an overview of some of the literature on power laws.

frequency versus log of degree, most of the data can end up occupying only a small portion of the figure. For example, in Figure 3.2.3, *less than ten percent* of the data fall in the range below -4 on the vertical scale.[17] The few studies that have fit more than one distribution to a network have found that the degree distribution that best fit tended *not* to be a power distribution (e.g., Pennock et al [513] and Jackson and Rogers [337]).

Table 3.2 from Jackson and Rogers [337] provides a look at how close to scale-free versus independently random a network is. They examine a class of degree distributions where the cumulative distribution function $F$ is given by

$$F(d) = 1 - (rm)^{1+r} \left(d + rm\right)^{-(1+r)}, \tag{3.2}$$

where $m > 0$ is the average in-degree and $r$ is a parameter that varies between 0 and $\infty$ and captures how randomly the links are formed. (See Section **??** for details.) In the extreme where $r$ tends to 0 this converges to a scale free distribution, and in the extreme where $r$ tends to $\infty$ this converges to a negative exponential distribution, which is the proper analog of the degree distribution of a purely random network that is growing over time.[18]

The following figures show how the degree distribution changes as $r$ is varied, changing from a scale-free distribution with fat-tails to one with uniformly random attachment.

[17] Here the points on the graph seem misleading, as here are more points below -4 on the scale. However, the frequency on the vertical scale provides the log weights, and so points higher on the scale represent orders of magnitude more data points. The point with log(degree) equal to 0 corresponds to more than 20 percent of the data!

[18] Instead of starting with a fixed number of nodes and randomly putting in links all at once, consider a process where at each time a new node is born. That new node forms some links randomly with the existing nodes. As nodes age, they will gain links as more nodes are added, while newborn nodes will have only their initial number of links. Rather than a Poisson distribution, this leads to a negative exponential distribution of degrees, which fits some networks very well. This is discussed in more detail in Chapter **??**.

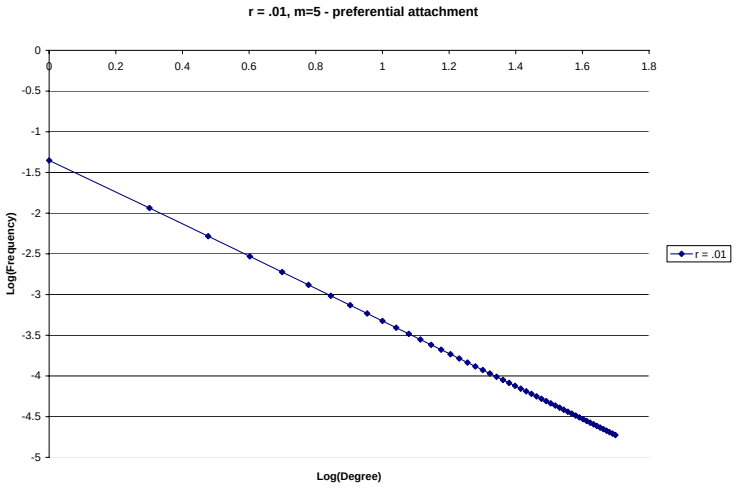



**Figure 3.2.3.** Degree Distribution with low $r$ - Essentially Preferential Attachment

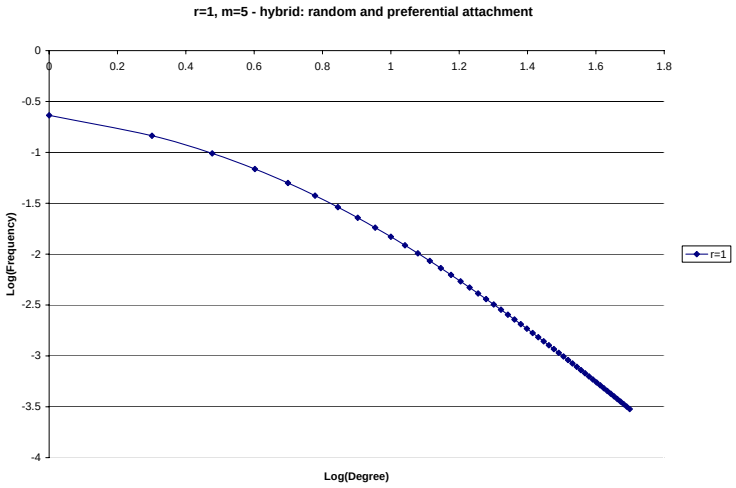



**Figure 3.2.3.** Degree Distribution with medium $r$ - A Mixture of Uniform and Preferential Attachment

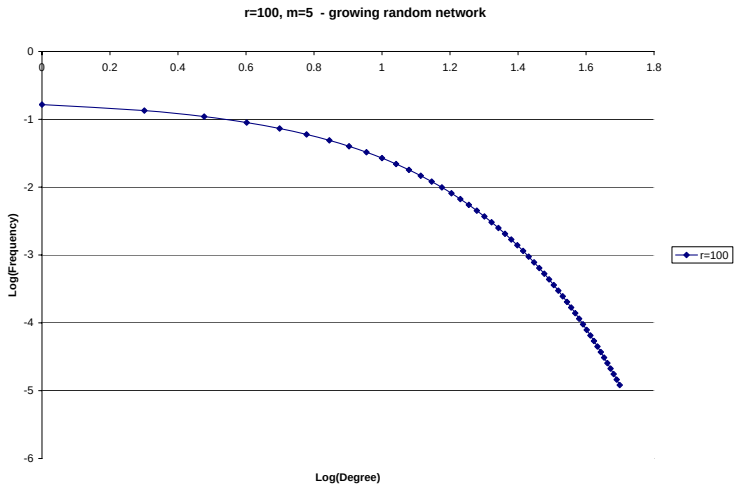



**Figure 3.2.3.** Degree Distribution with high $r$ - A Growing Random Network

Fits to a few networks[19] give us an idea of the variation across applications.[20] In Chapter **??**, the derivation of a variation on this degree distribution as well as techniques for fitting it to data are detailed.

[19]The www data are from an analysis of the links between web pages on the Notre Dame domain of the world wide web from Albert, Jeong, and Barabási [**?**]. The co-authorship data are from the above cited study by Goyal, van der Leij, and Moraga-González [286]. The citation network consists of the network of citations among all papers that have cited Milgram's [444] paper or have the phrase "small worlds" in the title, and is from Garfield [258]. The prison data record friendships among inmates in a study by MacRae [422], the ham radio data record interactions between ham radio operators from Killworth and Bernard [372], and the high school romance data collected romantic relationships between high school students over a period of a year and a half in a US high school and is from Bearman, Moody, and Stovel [47]. The number of nodes, average degree, and clustering numbers are as reported by the studies. The estimates on randomness are from Jackson and Rogers [337]. The fits on these estimated $r$'s are high, with $R^2$'s ranging between 93 and 99 percent.

[20]The clustering figure for the co-author data is actually for overall clustering, as the average number is not available but is likely to be higher given that the clustering is decreasing in degree. The clustering for the high school romance network is special because that network is mainly heterosexual in its relationships, and so completed triads do not appear. Even if one looks for larger cycles, there are only five present in the whole network, which would be characteristic of a large network formed at random between two groups.

Table 3.2: Comparisons Across Applications

| | WWW | Citations | Co-author | Ham Radio | Prison | High School Romance |
|---|---|---|---|---|---|---|
| Number of Nodes | 325729 | 396 | 81217 | 44 | 67 | 572 |
| Randomness: $r$ | 0.57 | 0.63 | 4.7 | 5.0 | $\infty$ | $\infty$ |
| Avg. In-Degree: $m$ | 4.6 | 5.0 | .84 | 3.5 | 2.7 | .83 |
| Avg. Clustering | .11 | .07 | .16 | .47 | .31 | - |

In all cases, the degree distribution fits the data quite closely (the $R^2$'s on the corresponding regressions vary from .94 to .99).

We see even in the data of Albert, Jeong and Barabási [8] for the Notre Dame web sites, that although the degree distribution appears to be scale-free to the naked eye, when we fit a model to the data, it is best fit by a mixture of more than 1/3 parts uniformly random to 2/3 parts scale-free ($r = .57$ has random/scale-free of more than 1/2).[21] When we get to some of the more purely social networks, we see parameters that indicate much higher levels of random link formation, which are very far from satisfying a power law. In fact, the degree distribution of the romance network among high school students is essentially the same as that of a purely random network.

So there are two lessons here. The first is that many social networks exhibit "fat tails" in that there are more nodes with relatively high and low degrees than would tend to arise if links were formed independently. The second is that it is hard to find networks that actually follow a strict power law. Even networks that are often cited as exhibiting power laws (e.g., the world wide web) are better fit by distributions that differ significantly from a power distribution.

### 3.2.4 Correlations and Assortativity

Beyond the degree distribution of a network, we can also ask questions about the correlation patterns in the degrees of connected nodes. For instance, do relatively high degree nodes have a higher tendency to be connected to other high degree nodes? This

[21] In terms of a comparison, the statistical fit of this model (in terms of $R^2$) with $r = .57$ is .99 while the fit of a scale-free distribution is only .86.

is termed positive assortativity.[22]

While there is little systematic study of assortativity, there is a hypothesis that positive assortativity is a property of many socially generated networks, and contrasts with the opposite relationship that is more prevalent in technological and biological networks. This hypothesis is put forth by Newman [480] (see p. 314) who examines the correlation in degree across linked nodes.[23] Newman [480] reports the following correlations among degrees in seven different applications:

Table 3.3: Correlations in Degrees

| | math co-authorship | physics co-authorship | email network | film actors | internet wiring | electric power grid | neural network |
|---|---|---|---|---|---|---|---|
| Correlation | .12 | .36 | .09 | .21 | -.19 | -.003 | -.23 |

Newman refers to the first four networks as social networks and the latter three as technological networks, and remarks the pattern of the positive correlations for the social networks and the negative correlations in the technological networks.

The network of reported friendships among prison inmates of MacRae [422] shows a similar positive assortative relationship as the social networks above. It has a correlation between a node's in-degree and the average in-degree of its neighbors of .58 (as reported by Jackson and Rogers [337]). However, one can also find exceptions. For example, the Ham radio network of interactions between amateur radio operators studied by Killworth and Bernard [372], has a negative correlation. There, the correlation between a node's in-degree and the average in-degree of its neighbors is -.26 (as reported by Jackson and Rogers [337]); although since it is a small network the correlation is not statistically significant. Once one examines networks such as a network of trading

[22] Even a finite network where links are formed completely independently, there can be correlation in degrees. For instance, consider the simplest possible setting with just two nodes, and where the probability of a link is 1/2. Here the degree of the two nodes is perfectly correlated, as they either both have degree 1 or both have degree 0. So, even when links are formed independently, the fact that a node has a high degree tells us something about which other nodes it is likely to be connected to based on their degrees. That correlation tends to disappear in a Poisson random network as the number of nodes grows, but this gives us an idea that such correlations will be delicate.

[23] Thus, the calculation for a network $g$ is simply $\frac{\sum_{ij \in g}(d_i - m)(d_j - m)}{\sum_{i \in N}(d_i - m)^2}$, where $m$ is average degree and $d_i$ is the degree of a node $i$.

relationships among countries, one ends up with structures that can be thought of as primarily economic in nature, and having some aspects of both social and technological relationships. For example, Serrano and Boguñá [557] find a negative correlation among the degrees of countries that trade with each other and suggest that the average degree of the neighbors of a given node is proportional to the inverse of the square root of that node's degree.[24] They describe the network as a "hub-and-spoke" system, where smaller countries (the spokes) have few partners and trade with larger countries (the hubs) which tend to have many more partners. While many larger countries trade with each other, one still sees a negative relationship overall.

Related to assortativity, studies of some social networks have also suggested "core-periphery" patterns (e.g., see Brass [**?**]), where there is a core of highly connected and interconnected nodes, and then a periphery of less connected nodes. Moreover, theories of structural similarity posit that people tend to use other people who are similar to themselves as a reference group (Festinger [**?**]). Studies building from this hypothesis (e.g., Burt [105]) have found that people with similar structural positions tend to have similar issues to deal with and that leads them to communicate with each other.

As the patterns of connections in a network can have a profound impact on things like the diffusion of behavior, information, or disease, it is important to develop a better understanding of assortativity and other characteristics that describe who tends to be connected to whom in a network.

### 3.2.5 Patterns of Clustering

There are other patterns in networks that help characterize overall structure. Beyond degree distributions and correlations in degrees, one can also examine how clustering is distributed across a network. Clustering measures such as average or overall clustering are simple summary statistics. While they give some insight, we can look at much more detailed information about how clustering varies throughout a network.

For instance, in the example of the Florentine marriages, we can keep track of the full distribution of individual clustering coefficients across nodes. In that network there are nine nodes with clustering coefficients of 0, five nodes with clustering coefficients of 1/3, one node with a clustering coefficient of 2/3, and one with a clustering coefficient

[24] Their network has a directed link from one country to another if the first exports to the second. The relationship they examine is similar whether or not one examines outdegree, indegree, an undirected version (with a link if there is a directed link in either direction), and a different undirected version where one only examines reciprocal links (where there are directed links both ways).

of 1/15. It is perhaps even more informative to see how the clustering relates to the degree of a node. In the Florentine Marriage example, all nodes with degree two or less have individual clustering coefficients of 0. Degree three nodes have on average a clustering of 4/15 (one has 0 and four have 1/3), degree four nodes have on average a clustering of 1/6 (one has 2/6 and the other 0), and the degree six node has a clustering of 1/15. Here there is some pattern. First, the low degree nodes have clusterings of 0, most simply by convention. But more interestingly, the rate of clustering among the higher degree nodes is decreasing in the degree. This sort of pattern has been noted in other applications as well. That is, the neighbors of a higher degree node are less likely to be linked to each other as compared to the neighbors of a lower degree node. For example, Goyal, van der Leij, and Moraga-González [286] observe that a network of co-authorship among 81217 economists in the 1990's had an overall clustering coefficient of .157, while averaging over the one hundred nodes with the highest degrees only yielded an average clustering of .043. Thus, the highest degree nodes tend to exhibit lower clustering than one sees on average across the whole network. A very simple way to see if a network might exhibit such a pattern is to compare the overall clustering to the average clustering. Overall clustering can be thought of a weighted averaging of clustering across nodes with weights proportional to the number of pairs of neighbors that the nodes have (so the weight on node $i$ is $d_i(d_i - 1)/2)$), while average clustering weights all nodes equally. Thus, overall clustering is weighting the higher degree nodes much more than average clustering is, and so if the overall clustering is significantly lower than the average clustering, then there is a sense in which the clustering is relatively lower for higher degree nodes. We can see this by comparing the overall clustering numbers for the networks reported in Table **??** to the average clustering numbers for the same networks. For instance, the ratios of overall clustering to average clustering are .09/.60 for biology, .15/.34 for math, and .45/.56 physics.[25] We can also simply examine the correlation between degree and the clustering in a node's neighborhood. Jackson and Rogers [337] calculate such correlations for the prison and ham radio networks mentioned above. They find a correlation of -.05 between a node's in-degree and the clustering in its neighborhood for MacRae's [422] friendship network among prison inmates, and a correlation of -.27 between a node's degree and the clustering in its neighborhood for Killworth and Bernard's [372] Ham radio network. However, these are both small networks and so

[25] The number is not reported in the economics co- authorship data set, but we already saw some aspect of this relationship in that network, as discussed above.

neither of these figures is statistically significant and are thus only suggestive.

It is not obvious whether this is a general pattern of social and/or other forms of networks, but it is at least exhibited in some observed networks, and is something that we will see exhibited by some models of network formation.

### 3.2.6 Homophily

Many social networks exhibit what is was named "homophily" by Lazarsfeld and Merton [404]. As we saw in Section 1.2.2, this refers to the fact that people are more prone to maintain relationships with people who are similar to themselves. This applies very broadly, as measured by age, race, gender, religion, profession and is generally a quite strong and robust observation (see McPherson, Smith-Lovin and Cook [439] for an overview of research on homophily). It was first noted by Burton [108] who coined the phrase "birds of a feather." For example, based on a national survey Marsden [429] finds that only 8 percent of people have *any* people of another race with whom they "discuss important matters." Homophily is an important aspect of social networks since it means that some social networks may be largely segregated. This, for instance, has profound implications in the access to job information (e.g., see Calvó-Armengol and Jackson [119]). It can also have profound implications for the spread of other sorts of information, behaviors, and so forth.

There is an important distinction between different forms of homophily. One is due solely to opportunity, while the other is due to choice. For instance, it is not surprising that most children have closest friends who are of a similar age as themselves. Much of this is due to the fact that they form friendships with other children with whom they regularly interact at school. This is the aspect that it is due to opportunity, which is constrained by the structure of classes within schools, among other things. Beyond this, even when presented with opportunities to form ties across age, there is still a tendency to form a disproportionate fraction with own-age individuals. This has been attributed to a number of factors including maturity and interests. One also sees this with respect to other factors such as race. For example, in middle school, less than 10 percent of "expected" cross-race friendships exist (Shrum et al [**?**]). That is, given the composition of schools in terms of race, if individuals form relationships in proportion to the relative numbers of people of various races that they encounter, there should be ten times more cross-race relationships than are observed. Thus, in addition to the substantial homophily one would expect due to the fact that most schools are biased in their racial composition and thus there is a bias in opportunity towards own-race

relationships, one also sees a very strong own-race bias in the relationships formed beyond that governed by relative population sizes.

To get a better impression of homophily, consider Table 3.4, which describes the frequency of friendships across different ethnicities in a Dutch high school. The data were collected by Baerveldt, Van Duijn, Vermeij, and Van Hemert [24].

| | Ethnicity of Students | | | | |
|---|---|---|---|---|---|
| Percent of Friends by Ethnicity: | Dutch n=850 | Moroccan n=62 | Turkish n= 75 | Surinamese n=100 | Others n=230 |
| Dutch | 79 | 24 | 11 | 21 | 47 |
| Moroccan | 2 | 27 | 8 | 4 | 5 |
| Turkish | 2 | 19 | 59 | 8 | 6 |
| Surinamese | 3 | 8 | 8 | 44 | 12 |
| Others | 13 | 22 | 14 | 23 | 30 |

Table 3.4: Friendship Frequencies (in percent) by Ethnicities in a Dutch High School; from Baerveldt et al [24].

For instance, the first column indicates that Dutch students form 79 percent of their friendships with other Dutch students, 2 percent of their friendships with Moroccan students, etc. Here we see "inbreeding" homophily through the high percentages occurring on the diagonals, which are higher than the relative percentages in population. This inbreeding can be due to biases in the interactions within the school that provide the opportunities to form friendships, and can also be influenced by the choices made by the students.[26] There are other factors influencing these tendencies, such as religious and economic background.[27]

### 3.2.7 The Strength of Weak Ties

The role of social networks in finding jobs was at the heart of some of the most influential research in the social networks area, which was conducted by Granovetter [289],

[26] See Currarini, Jackson and Pin [171] for evidence that both effects are present.

[27] See Baerveldt et al [24], Moody [458], Fong and Isajiw [234], Adamic and Adar [2], Fryer [244], and Currarini, Jackson and Pin [171] for more discussion and background on the factors influencing friendship formation.

[290]. He interviewed people in Amherst Massachusetts, across a variety of professions to determine how they found out about their jobs. He recorded not only whether they used social contacts in their employment searches, but also the strength of the social relationships as measured by frequency of interaction. He found that a surprising proportion of jobs were obtained through "weak ties" (as opposed to "strong" ones). There are various ways to measure the strength of a tie, but Granovetter's basic idea is that strength is related to the "amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie," (Granovetter [289] p. 1361). His measure of the strength of a tie was by the number of times that individuals had interacted in a past year (strong = at least twice a week, medium = less than twice a week but more than once a year, and weak = once a year or less). Out of the 54 people that Granovetter had detailed interviews with and who had found their most recent job through a social contact, he found that 16.7 percent had found their job through a strong tie, 55.7 percent through a medium tie, and 27.6 percent through a weak tie.[28]

Granovetter's idea was that individuals involved in a weak tie were less likely to have overlap in their neighborhoods than individuals involved in a strong tie. Such ties then are more likely to form bridges across groups that have fewer connections to each other, and can thus play critical roles in the dissemination of information. Granovetter concludes that weak ties are the glue that holds communities together, and in paradoxical contrast, strong ties lead to more local cohesion but then to more overall fragmentation.

There are numerous follow-up studies, including direct tests of some of the hypotheses put forth by Granovetter (e.g., see Friedkin [240]), as well studies of the roles of weak ties hypotheses in a variety of settings, from the diffusion of technological information to patterns of immigration. There remain many interesting and basic questions about the relative use of weak ties that are not fully answered or such that the answers vary across applications. For instance, given that individuals tend to have many weak ties (and for employed adults in large societies, conceivably orders of magnitude more weak ties than strong ties), how active should we expect weak ties to be in diffusing vital information? Are weak ties important solely because of their bridging behavior and information that they diffuse, or more generally because of other features they embody? And even more fundamentally, is it even generally true that individuals in-

[28] These are raw numbers and to keep these in perspective it is worth noting that we tend to have far more weak ties than strong ones.

volved in stronger ties are more likely to have strong overlap in their neighborhoods? Regardless of the answers to these and other related questions, Granovetter's work on the strength of weak ties makes it clear that the abstraction to simple 0-1 networks is a crude approximation of interaction structures, and that developing richer models capturing additional nuances of interaction frequency, duration, and heterogeneity, is important.

### 3.2.8 Structural Holes

Another important concept regarding the structure of networks is due to Burt [105] and concerns what he named "structural holes". A structural hole is a void in a social structure, and in terms of social networks refers to an absence of connections between groups. As Burt points out, this does not mean that the groups are unaware of each other, but instead that the lack of links between the groups leads to non-redundancies in the information between the groups and can also lead to a failure of diffusion between the groups. One of Burt's main points is that individuals who fill structural holes, by offering connections between otherwise separated or sparsely inter-connected groups, end up with power and control over the flow of information and favors between groups. For instance, Burt [106] offers evidence that filling structural holes leads to benefits in the form of promotion, bonuses, and other measures of performance networks of managers.

### 3.2.9 Social Capital

The term social capital has come to embody a number of different concepts related to how social relationships lead to individual or aggregate benefits in a society. The concept has been defined in many ways and applied many contexts, so that there is no tight and encompassing definition of social capital. For instance, in some incarnations it refers to the relationships that an individual has and the potential benefits that those relationships can bestow (e.g., Bourdieu [**?**]) and in others it refers to the aspects of broader social interaction (relationships, norms, trust,...) that facilitate cooperation (e.g., see Putnam [**?**]).[29] Although there is resonance in the idea that social networks and social relationships can translate into individual and societal benefits, and although various definitions of social capital can be useful in identifying the relationship

[29] Sobel [**?**] provides an insightful overview of the objectives, as well as the shortcomings, of the literature including some of the difficulties with various definitions.

between social structure and welfare, it is important to be precise in the definition and application of social capital, as the term has been so broadly and differentially used that it is not always clear what it means. While I will avoid the term social capital in most of what follows, many of the models that appear in this book, especially ones that directly relate social network structure to individual behavior and welfare, can be interpreted as capturing various aspects of social capital, operationally defined. There is much that remains to be done in terms of providing definitions and models of social capital that are incisive and yet still portable across applications.

### 3.2.10 Diffusion

One important role of social networks is as conduits of information. People often learn from each other, and this has important implications not only for how they find employment, but also about what movies they see, which products they purchase, how technology becomes adopted, whether or not they participate in government programs, whether they protest, and so forth. Many studies on the diffusion of innovation, including some classic early ones such as Ryan and Gross's [544] study of the diffusion of hybrid corn seed among Iowa farmers, and Hagerstrand's [**?**] examination of the diffusion of the telephone, have shown how important social contacts are in determining behavior.[30]

A classic study in this area that illustrates the relation between social structure and adoption of a new technology is that of Coleman, Katz, and Menzel [154]. They examined the adoption of a new drug by doctors in four cities over a period of fifteen months.[31] The adoption of the drug means that it was prescribed to a patient by the doctor, as found through pharmacists' records. The drug was first used in trials by a few "innovators" and subsequently was adopted by almost all of the doctors by the end of the study. Along with the information about adoption times, Coleman, Katz and Menzel interviewed the doctors to collect other information, including the type of each doctor's practice, his or her age, general prescription habits, etc.; as well as information about the doctor's social interaction. To get at the social network, Coleman, Katz and Menzel asked each doctor questions which, quoting Coleman, Katz, and Menzel [154], were: "To whom did he most often turn for advice and information?", "With whom did he most often discuss his cases in the course of an ordinary week?", and "Who were

[30] Rogers [536] provides a detailed overview of much of the research on diffusion.

[31] The data cover a period of seventeen months.

the friends, among his colleagues, whom he saw most often socially?". The interviewed doctors were asked to provide three names in response to each question.

Using the answers to these questions, and the time series data about adoption rates, Coleman, Katz and Menzel were able to deduce some things about how the time of adoption related to the social structure. For instance, they examined how the proportion of doctors who had adopted the drug depended on how many social contacts the doctors had.[32]

As summarized in the table below, after six months, among the 36 doctors who were not named as "friends" by any of the other doctors in their survey only one third had adopted the drug, while this ratio was just over one half for the 56 doctors named as friends by one or two of the other doctors, and the adoption ratio was over seventy percent for the 33 doctors names as friends by three or more other doctors. By ten months, the adoption rate among the doctors not named as friends was still just below fifty percent, while it as roughly seventy percent among doctors named as friends by one or two others, and ninety-four percent for the doctors named as friends by three or more others.

Table 3.5: Diffusion of Drug Among Doctors

| | Named as Friend: | | |
|---|---|---|---|
| Fraction adopting by: | by 0 others (36 subjects) | by 1 or 2 others (56 subjects) | by 3 or more others (33 subjects) |
| 6 months | .31 | .52 | .70 |
| 8 months | .42 | .66 | .91 |
| 10 months | .47 | .70 | .94 |
| 17 months | .83 | .84 | .97 |

As with any data, one has to be careful about inferring causation from correlations; but the differences in adoption rates do indicate that the level of social integration as measured through this survey is related to the speed of adoption. As we shall

[32] The particular explanation for this relationship is not obvious. It appears from the study that there was information about the drug widely available, and so one must rely on other sorts of explanations for such a peer effect, for example, such as some sort of validation: one is more willing to prescribe if one knows a colleague that has prescribed, or that experience from colleagues is more trusted than studies and marketing information, etc. See Section **??** for more discussion.

see in Chapters 7 and 8, this is to be expected for a variety of reasons relating to position in a network. As one should intuitively expect, nodes with greater numbers of connections are more likely to hear information more quickly and can serve as conduits of information. In terms of empirical work, in order to determine the role of social structure in influencing behavior, one has to carefully sort out other factors which might be correlated with position in a network and influence behavior. This is often a challenge, as in any sort of empirical work where critical variables are endogenous.

Just as an indicator of how wide the variety of applications is where diffusion is important, consider a recent study by Cristakis and Fowler [**?**]. They examined a network of 12,067 people during a period from 1971 to 2003, based on data including both social relationships and health outcomes. Given that the data included weight at different times for the same individuals, they were able to examine whether weight gain by one individual correlated with weight gains of that individual's friends, while controlling for other factors that might have influenced weight gain. They reported a significant increase in the probability of a weight gain due to a friend's weight gain, which is not present when looking at close geographic proximity. While this leaves many questions of causation and interpretation open, it does suggest that network structure is important in understanding various forms of diffusion.

With some definitions and empirical background in hand, let us now turn to modeling network formation.

# Chapter 4

# Random Graph-Based Models of Networks

In this chapter, I discuss some of the workhorse models of static random networks and some of the properties that they exhibit. As we saw in the introductory chapter, randomly generated networks exhibit a variety of features that we see in the data, and through examining the properties of these models we can begin to trace traits of observed networks to characteristics of the formation process.

Models of random networks find their origin in the studies of random graphs by Solomonoff and Rapoport [576], Rapoport [526] and Erdös and Rényi [211], [212], [213]. The canonical version of such a model, a Poisson random graph, was discussed in Section **??**. The next chapter is a "sister" to this one, where I discuss a series of recent models of growing random networks that have arisen in an attempt match more of the properties, such as those discussed in Chapter 3, that are exhibited by many observed networks. Indeed, random-graph-based models of networks have been a primary tool in analyzing various observed networks. For example the network of high school romances described in Section 1.2.2 has a number of features that are well-described by a random network model, such as having a single giant component and then a large number of much smaller components and a few isolated nodes. Such random models of network formation help tie observed social patterns back to the structure of the inherent randomness and the process of link formation.

Beyond their direct use in analyzing observed networks, random network models also serve as a platform for modeling how behaviors diffuse through a network. For instance, the spread of a disease depends on the contact that various individuals have with each other. That spread can be very different depending on how much interaction

there is on average (e.g., do people interact with a few others or hundreds of others) as well as how it is distributed throughout the population (e.g., does everyone interact with roughly the same number of people, or are there some people who have contact with very large numbers while others have contact with very few). In order to understand how such diffusion works, one has to have a tractable model of what the link structure within a society looks like, and random graph models provide such a base. These models are not only be useful in understanding the diffusion of a disease, but also in modeling things like the spread of information, or decisions that are heavily influenced by one's peers (e.g., whether or not to go to college), as we shall see in more detail in Chapters 7 and 8.

Let me reiterate that random models of network formation are largely context-free, in that the nodes and processes for link formation are often simply governed by some given probabilistic rules. Some of these probabilistic rules have stories behind them, but this is not true of all such models. As such, these models are generally missing the social and economic incentives and pressures that underlie network formation, as discussed more fully in Chapter **??**. Nevertheless, these models are still quite useful for the reasons mentioned above, and they also serve as useful benchmarks. By keeping track of the properties that random-graph models of networks exhibit, and which ones they fail to exhibit, we end up with a useful reference point for building richer models, and also for understanding the strengths and weaknesses of models of networks that are tied to social and economic forces influencing individual decisions to form and maintain relationships.

The chapter starts with the presentation of a series of fundamental random graph models that have been useful in various aspects of network analyis. This includes variations on the basic Poisson random graph model that include correlations between links and allow richer degree distributions to be generated. Once these models have been described, I turn to presenting some of the properties of the resulting networks. This includes understanding how small changes in underlying parameters can lead to large changes in the properties of the resulting graphs (thresholds and "phase transitions"), as well as understanding when it is that resulting networks are connected, have a giant component, and other properties such as their diameter and clustering. The chapter concludes with an illustration of how random networks can be used as a basis for understanding the spread of contagious diseases or behaviors in a society.

## 4.1 Static Random-Graph Models of Random Networks

The term "static" refers to the fact that a model can be thought of as having all nodes present at the same time and then having links drawn according to some probabilistic rule. Poisson random graphs constitute one such static model. This class of "static" models contrasts with processes where networks "grow" over time. In such models new nodes are introduced over time, and form links with existing nodes as they enter. Such growing processes can result in properties that are different from those of static networks, and allow different tools for analysis. They are also naturally suited to different applications, and are discussed in detail in Chapters 4 and **??**.

### 4.1.1 Poisson and Related Random Network Models

The Poisson random graph model is one of the most extensively studied models in the static family. Closely related models are the ones mentioned in Section **??**, where a network is randomly chosen from some set of networks. For instance, out of the all the possible networks on $n$ nodes, one could simply pick one completely at random, with each network having an equal probability. Alternatively, one could simply specify that the network should have $M$ links, and then pick one of those networks at random with equal probability (that is, with each $M$ link network having probability $\binom{N}{M}^{-1}$, where $N = \binom{n}{2}$ is the number of potential links among $n$ nodes). Some of these different models of random networks turn out to have remarkably similar properties. On an intuitive level, if we examine a network where each link is formed with an independent probability $p$, we expect to have $pn(n-1)/2$ links formed (where $n(n-1)/2$ is the potential number of links). While we might end up with more or fewer links, with a large number of nodes, an application of the law of large numbers tells us that we will not deviate too much from this expected number of links in terms of the percentage formed. This turns out to be enough to guarantee that a model where links are formed independently has many things in common with a model where we force the network to have the expected number of links.[1]

[1] Let $G(n,p)$ denote the Poisson random graph model on $n$ nodes with probability $p$ of any given link, and $G(n,M)$ denote the model where a network with $M$ links is chosen with a uniform probability over all networks of $M$ links on $n$ nodes. The properties of $G(n,p)$ and $G(n,M)$ are closely related for large $n$ when $M$ is near $pn(n-1)/2$. In particular, if $n^2p(1-p) \to \infty$, and a property holds for each sequence of $M$'s that lie within $\sqrt{p(1-p)n}$ of $pn(n-1)/2$, then it holds for $G(n,p)$. The

While these networks are static in the way that they are generated, much of the analysis of such random networks concerns what happens when $n$ becomes large. It is easy to understand why most results for random graphs are stated for large numbers of nodes. For example, in the Poisson random graph model, if we fix the number of nodes and some probability of a link forming, then every conceivable network has some positive probability of appearing. In order to talk sensibly about what might emerge, we want to make statements of the sort that networks exhibiting some property are (much) more likely to appear than networks that fail to exhibit that property. As such, most results in random graph theory concern the probability with which a network generated by one of these processes will have a given property as $n$ goes to infinity. For instance, what is the probability that a network will be connected and how does this depend on how $p$ behaves as a function of $n$? Many such results are proven by finding some lower or upper bound on the probability that a given property will hold, and then seeing if the bounds can be shown to converge to 0 or 1 as $n$ becomes large.

We shall examine some of these properties for a general class of static random networks below.

Let me begin, however, by describing some variations of static random graph models other than the Poisson model that provide a feeling for the variety of such models and the motivations behind their development.

### 4.1.2 "Small World" Networks

While random graphs can exhibit some of the features of observed social networks, (e.g., diameters that are small relative to the size of the network when average degree grows sufficiently quickly), it is clear that random graphs lack some of the features that are prevalent among social networks, such as the high clustering discussed in Sections **??** and **??**. To see this, consider the Poisson random network model, and let us ask what its clustering will be. Suppose that $i$ and $j$ are linked and $j$ and $k$ are linked. What is the frequency with which $i$ and $k$ will be linked? Since link formation is completely independent, it is simply $p$. Thus, as $n$ becomes large, if the average degree grows more slowly than $n$ (which would be true in most "large" social and economic networks where there are some bounds on the number of links that agents can maintain) then it must be that $p$ tends to 0 and so the clustering (both average and overall) will tend to 0.

converse holds for a rich class of properties (called convex properties). See Chapter 2 in Bollobas [80] for detailed definitions and results along these lines.

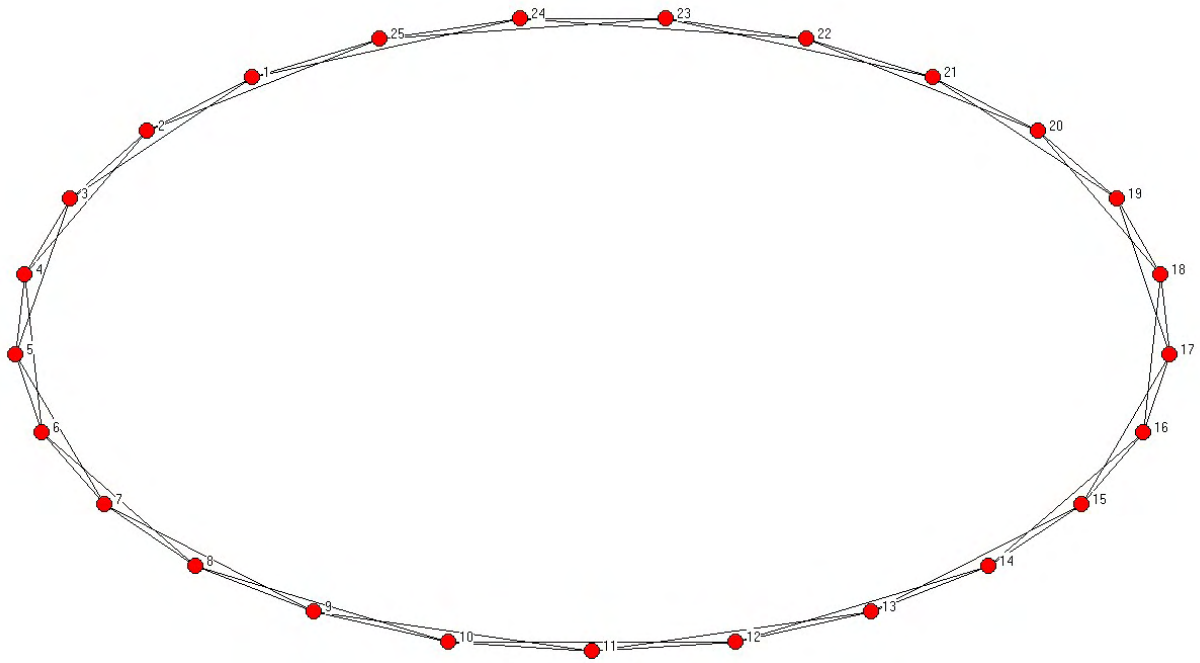



**Figure 4.1.2.** A Ring Lattice on 25 Nodes with 50 Links

With this in mind, Watts and Strogatz [623] developed a variation of a random network that showed that it only takes a small number of randomly placed links in a network to generate a small diameter. They combined this with a highly regular and clustered starting network in order to generate networks that simultaneously exhibit high clustering and low diameter, a combination observed in many social networks. Their point is easy to see. Suppose we start with a very structured network that exhibits a high degree of clustering. For instance, let us construct a large circle, but then connect a given node to the nearest four nodes rather than just its nearest two neighbors, as in Figure 4.1.2.

In such a network, each node's individual clustering coefficient will be 1/2. To see this, consider some set of consecutive nodes 1, 2, 3, 4, 5, that are part of such a network for a large $n$. Consider node 3, which is connected to each of nodes 1,2, 4 and 5. Out of all the pairs of 3's neighbors $(\{1,2\},\{1,4\},\{1,5\},\{2,4\},\{2,5\},\{4,5\})$, we see that half of them are connected $(\{1,2\},\{2,4\},\{4,5\})$. Here, as we let $n$ grow, the clustering (both overall and average) will stay at 1/2. By adjusting the structure of the local connections we can also adjust the clustering.

While this sort of regular network exhibits high clustering, it fails to exhibit some

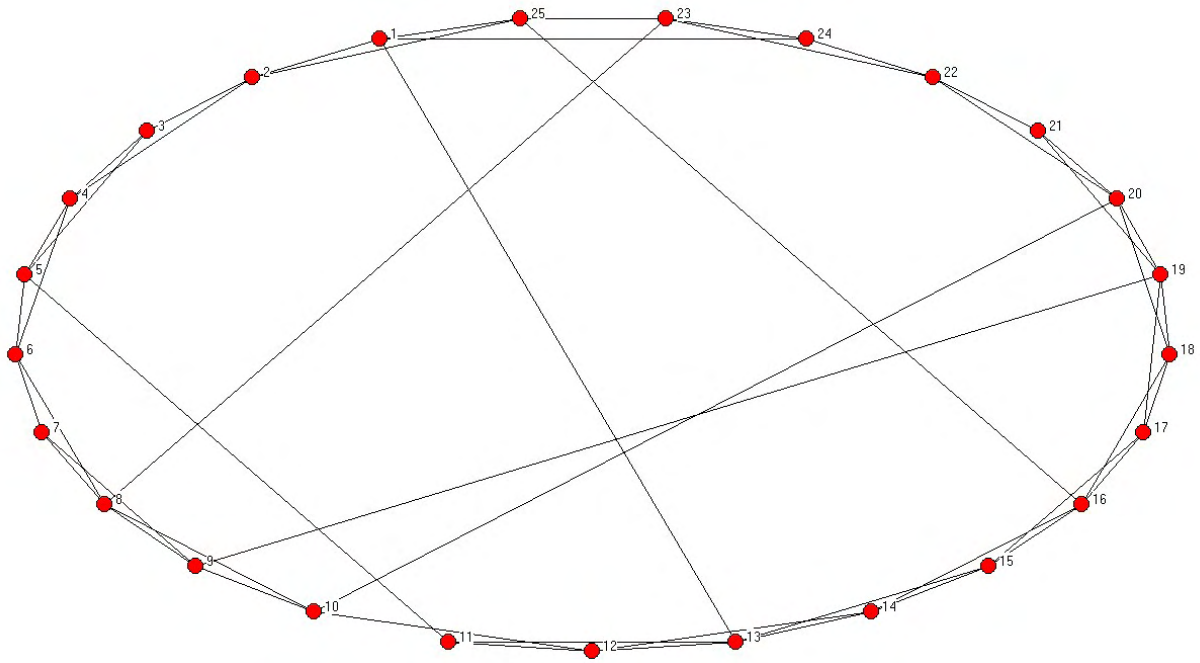



**Figure 4.1.2.** A Ring Lattice on 25 Nodes Starting with 50 Links and Rewiring 6

of the other features of many observed networks, such as a small diameteter and at least some variance in the degree distribution. The diameter of such a network is on the order of $n/4$. The main point of Watts and Strogatz [623] is that by randomly re-wiring relatively few links, we can end up with a network that has a much smaller diameter but still has substantial clustering. The re-wiring can be done by randomly selecting some link $ij$ and disconnecting it and then randomly connecting $i$ to another node $k$ chosen uniformly at random from nodes which are not already neighbors of $i$. Of course, as more such re-wiring is done, the clustering will eventually vanish. The interesting region is where enough re-wiring has been done to substantially reduce (average and maximal) path length, but not so much that clustering vanishes.

Here we see that after having rewired just six links the diameter of the network has decreased from 6 in the network pictured in Figure 4.1.2 to 5 in the network pictured in Figure 4.1.2, with minimal impact on the clustering. Note also that in the network in Figure 4.1.2 every node is at distance 6 from three other nodes (e.g., node 1 and nodes 13, 14, and 15), so it is not simply that the rewiring has shortened a few long paths, but rather that these new links shorten many paths in the network, as there are 39 pairs of nodes at a distance of 6 from each other in the original network, which are

all moved closer to each other by the rewiring. This example is suggestive, and Watts and Strogatz perform simulations to provide an idea of how this works for ranges of parameters.

This model of networks makes an interesting point in showing how clustering can be maintained in the presence of enough random link formation to produce a low diameter. The model also has obvious shortcomings, in particular in that the degree distribution is essentially a convex combination of a degenerate distribution with all weight on a single degree and a Poisson distribution. Such a degree distribution is fairly particular, and not often observed in social networks. I return to discuss alternative models that better match observed degree distributions in Sections **??** and **??**.

### 4.1.3 Markov Graphs and p* networks

Next, I describe a generalization of Poisson random graphs that has been useful in statistical analysis of observed networks and was introduced by Frank and Strauss [236]. They called this class of graphs "Markov graphs", and such random graph models were later imported to the social networks literature by Wasserman and Pattison [616] under the name of p* networks, and further studied and extended in various directions.[2] The basic motivation is to provide a model that can be statistically estimated, and still allows for specific dependencies between the probabilities with which different links form.

Again, one important aspect of introducing dependencies is related to clustering, since Poisson random networks with average degrees growing more slowly than the number of nodes have clustering ratios tending to zero, which are too low to match many observed networks. Having dependences in the model can produce nontrivial clustering.

Conditional dependencies can be introduced so that the probability of a link $ik$ depends on whether $ij$ and $jk$ are present. The obvious challenge is that such dependencies will tend to interact with each other in ways that could make it impossible to specify the probability of different graphs in a tractable manner. For instance, if the conditional probability of a link $ik$ depends on whether $ij$ and $jk$ are present, but also on any other adjacent pairs being present, and the conditional probability of $jk$ depends on other adjacent pairs being present, etc.; we end up with a complicated set

[2] For instance, see Pattison and Wasserman [509] for an extension to multiple interdependent networks on a common set of nodes.

of dependencies. The important contribution of Frank and Strauss [236] is to make use of a theorem by Hammersley and Clifford (see Besag [55]) to derive a simple log-linear expression for the probability of any given network in the presence of arbitrary dependencies.

One of the more useful results of Frank and Strauss [236] can be expressed as follows. Consider $n$ nodes, and keep track of the dependencies between links by another graph, $D$, which is a graph among all of the $n(n-1)/2$ possible links.[3] So, $D$ is not a graph on the original nodes, but a graph whose nodes are all the possible links. The idea is that if $ij$ and $jk$ are neighbors in $D$, then there is some sort of conditional dependency between them, possibly in combination with other links. Thus, $D$ captures which links are dependent on which others, possibly in quite complicated combinations. For example, the Poisson random graph model is one where $D$ is empty, as all links are independent. If instead, we wish to capture the idea that there might be clustering, then we would like the link $ik$ to depend on the presence of $ij$ and $kj$ for each possible $j$. Thus, $D$ would have $ik$ connected to each other link that contains either $i$ or $k$.

Let $C(D)$ be all the cliques of $D$; that is, all of the completely connected subgraphs of $D$ (where the singleton nodes are considered connected subgraphs). So, in the case of a Poisson random graph $C(D)$ would simply be the set of all links $ij$. In the case of the clustering dependence just mentioned above, the set $C(D)$ would include all individual links and also all of the triads (sets of the form $\{ij, jk, ik\}$). Given a generic element $A \in C(D)$, let $I_A(g) = 1$ if $A \subset g$ (viewing $g$ as a set of links), and $I_A(g) = 0$ otherwise. So, if $A$ is a triad $\{ij, jk, ik\}$, then $I_A(g) = 1$ if each of the links $ij$, $jk$ and $ik$ are in $g$, and $I_A(g) = 0$ otherwise. Then, Frank and Strauss show that Hammersley and Clifford's theorem implies that the probability of a given network $g$ depends only on which cliques of $D$ it contains, and that it can be written as

$$log(\Pr[g]) = \sum_{A \in C(D)} \alpha_A I_A(g) - c, \tag{4.1}$$

where $c$ is a normalizing constant, and the $\alpha_A$'s are other free parameters.

In general, given that $D$ can be very rich and that the $\alpha_A$'s can be chosen at will, this allows for an almost arbitrary probability specification. The difficulty and art in applying this type of model in practice is in specifying the dependencies sparingly and imposing restrictions on the $\alpha_A$'s so that the resulting probabilities are simple and

[3] This is easily adapted to directed links, by having $D$ be a graph on the $n(n-1)$ possible directed links.

practical. For certain kinds of dependencies, the expressions can be quite simple and useful (e.g., see Anderson, Wasserman and Crouch (1998)).

To see how the expressions can simplify, let us consider the clustering dependency we mentioned above. This means that $C(D)$ is just the set of all links and all triads (triplets of the form $\{ij, jk, ik\}$). To simplify things further, let us also suppose that there is a symmetry among nodes, so that the probability of any two networks that have the same architecture but possibly different labels on the nodes is identical. This means that the $\alpha_A$'s are the same across all $A$'s that correspond to single links, and the same across all $A$'s that correspond to triads. Thus, the expression in 4.1 simplifies substantially. Let $n_1(g)$ be the total number of links in $g$, and let $n_3(g)$ be the total number of completed triads in $g$. Then there exist $\alpha_1$, $\alpha_3$ and $c$ such that (4.1) becomes

$$log(\Pr[g]) = \alpha_1 n_1(g) + \alpha_3 n_3(g) - c.$$

This then provides us with a simple generalization of Poisson random graphs (which are the special case where $\alpha_3 = 0$), which will allow us to control the frequency of clusters. That is, we can adjust the parameters so that graphs that have more substantial clustering will be relatively more likely than graphs that have less clustering (for instance, by increasing $\alpha_3$).[4]

While such a model can be cumbersome as we try to capture more complicated dependencies, it still provides a powerful statistical tool for testing for the presence of some specific dependency.[5] One can test for significant differences between fits of a model where such dependencies are present and a model where such dependencies are absent. Obviously, the validity of the test depends on the appropriateness of the basic specification of the model, as it could be that the model is not a very good fit with or without the dependencies, and so the comparison is invalidated.[6]

### 4.1.4 The Configuration Model

While the Markov model of random networks allows for general forms of dependencies, it is hard to keep track of the degree distribution that it will generate, and to adjust

[4] See Park and Newman (2004) for some derivations of clustering probabilities for this example.

[5] There are other such models designed for statistical analysis, as well as associated Monte Carlo estimation techniques, as for instance in Handcock and Morris [303].

[6] There are some challenges in estimating such models. A useful technique is proposed by Snijders [572], based on a Monte Carlo style simulation of the model and sampling of those simulations and then using an algorithm to approximate the maximum likelihood fit.

that to match observed networks. In order to generate random networks with a given degree distribution, various methods have been proposed. One of the most widely used is what is referred to as the "configuration model," as developed by Bender and Canfield [49]. The model has been further elaborated on and used by Bollobás [80], Wormwald [**?**], Molloy and Reed [449], Newman et al [**?**], among others.

To see how the configuration model works, it is useful to work with degree sequences rather than degree distributions. That is, given a network on $n$ nodes, we end up with a list of the degrees of different nodes: $(d_1, d_2, \ldots, d_n)$, which is the *degree sequence*.

Now suppose that we have an idea of the degree sequence $(d_1, d_2, \ldots, d_n)$ that we wish to generate in a network of $n$ nodes. This is directly tied to the degree distribution, so that the proportion of nodes that have degree $d$ in this sequence is $P^n(d) = \#\{i : d_i = d\}/n$.

Construct a sequence where node 1 is listed $d_1$ times, node 2 is listed $d_2$ times, etc.:

$$\underbrace{1,1,1,1,1,1,1,1,1,1,1,1,1}_{d_1 \text{ entries}} \quad \underbrace{2,2,2,2,2,2}_{d_2 \text{ entries}} \quad \ldots \quad \underbrace{n,n,n,n,n,n,n,n,n,n,n,n}_{d_n \text{ entries}}$$

Now randomly pick two elements of the sequence and form a link between the two nodes corresponding to those entries. Delete those entries from the sequence in 4.1.4 and repeat.

There are a few things to note about this procedure. First, it is possible to have more than one link between two nodes. As such, it generates what is called a multi-graph (allowing for multiple links) instead of a graph. Second, self links are possible and may even occur multiple times, while we have generally been ignoring self links in our discussion of networks up to this point. Third, as a more minor point, the sum of the degrees needs to be even or else there will be a leftover entry at the end of the process.

Despite these difficulties, this process still has nice properties for large $n$. There are two different ways in which we can work with this from our perspective of understanding randomly generated networks. One is to work directly with multi-graphs instead of graphs, and then try to show that multi-graphs generated (under suitable assumptions on the degree sequence) will have essentially the same properties as a randomly selected graph with the same degree sequence. Another is to generate a multigraph, and then from it delete self-links and duplicate links between two nodes. This is then a graph, and if the proportion of links we needed to delete is suitably small, then we end up with a graph with a degree distribution close to what we started with.

With this in mind, let us make the idea of growing the sequence more explicit. We can begin with an infinite degree sequence $(d_1, d_2, d_3, \ldots)$ and then dicuss increasing portions of the sequence. Let $q_i^n$ denote the number of links that node $i$ ends up with that are either self-links or duplicate links when the configuration model is operated on the first $n$ nodes. Then let $Q_i^n$ denote the probability that under the configuration model node $i \leq n$ ends up with at least one self-link or duplicate link, so that $Q_i^n = \Pr[q_i^n > 0]$. We can then show the following.

PROPOSITION **4.1.1** *If a degree sequence* $(d_1, d_2, d_3, \ldots)$ *is such that* $\max_{i \leq n} d_i / n^{1/3} \to 0$*, then* $\max_{i \leq n} Q_i^n \to 0$.

This proposition is not true if we drop the restriction that $\max_{i \leq n} d_i / n^{1/3} \to 0$ (see Exercise **??**). The reason is that if some nodes have degrees that are too large relative to $n$ then nontrivial portions of the links involve these nodes, and then the probability of self-links and/or multiple links can be nontrivial. Thus, while the configuration model is a useful network model when the degrees of nodes are not growing too large relative to the number of nodes, one has to be careful about the degree sequences that are admitted in order to have the resulting multi-graph be "close" to a graph.

The proposition establishes that if $\frac{\max_{i \leq n} d_i}{(n\langle d \rangle)^{1/3}}$ tends to 0, then the chance that any given node (including the largest ones) has a duplicate or self-link tends to 0. From this proposition, we can deduce that if we delete multiple links and self-links from the resulting multi-graph, we end up with a network where the proportion of nodes having the correct degree approaches 1 as the number of nodes grows, and the resulting degree distribution converges to the desired degree distribution (pointwise, if there is an upper bound on degrees). However, this does not imply that the multigraph will be a graph. When one aggregates across many nodes, there will tend to be some duplicate and self-links in this process, except under more extreme assumptions on the degree sequences; but there will not be many of them relative to the total number of links. To explore this in more detail, let us consider different statements that we could envision about the probability of self-links or multiple links under the configuration model:

(1) Fixing a node and its degree, as the the number of nodes grows the probability that the given node has any self or multiple links vanishes. That is, $Q_i^n \to 0$.

(2) The maximum probability across the nodes of having any self or multiple links vanishes. That is, $\max_{i \leq n} Q_i^n \to 0$.

(3) The fraction of nodes that have self-links or duplicate links goes to 0. That is, for any $\varepsilon > 0$, $\Pr\left[\#\{i \leq n : q_i^n > 0\}/n > \varepsilon\right] < \varepsilon$ for large enough $n$.

(4) The fraction of links that are self-links or duplicate links goes to 0. That is, for any $\varepsilon > 0$, $\Pr\left[\sum_{i\leq n} q_i^n / \sum_{i\leq n} d_i > \varepsilon\right] < \varepsilon$ for large enough $n$.

(5) The probability of seeing any self or multiple links vanishes. That is, $\Pr\left[\sum_{i\leq n} q_i^n > 0\right] \to 0$.

It is easy to see that (1) is true for any degree sequence, presuming that the sequence includes an infinite number of nodes with positive degree. (2) is what is shown in this proposition, and this then implies (3) based on an argument that the when the probability across nodes of having self or duplicate links goes to 0 uniformly across nodes, then it is impossible to expect a nontrivial fraction of nodes to have self or multiple links (see Exercise **??**). A similar argument establishes (4) (see Exercise **??**). The statement that would make our lives easiest in terms of ending up with a graph instead of a multigraph, (5), is only true under extreme conditions. To see why this (5) will generally fail, consider a degree sequence of $(2, 2, 2, \ldots)$, which is about as well-behaved as one could want in terms of having a good chance of avoiding self and duplicate links. Let us argue that even for this regular degree sequence there is still a nontrivial limiting probability of having at least one self-link. Here the probability that any given link is not a self link is $1 - \frac{1}{2n-1}$. To see this, think of starting by connecting one end of the link to some node, and then there are $2n - 1$ equally likely entries in the full sequence of points where we can attach the other end of this link to under the configuration model (see diagram **??**). Only one of these leads to a self-link. As we form links, continue to think of them as being formed in this way: randomly pick an entry to be one end of the link, and then pick a second entry for the other end. Now, as we proceed, regardless of how things work out, there will be at least $n/2$ links where the initial node that we picked for the first end of the link does not yet have any link attached to it. For each of these links, an upper bound on the probability of not ending up with a self-link is $1 - \frac{1}{2n-1}$. So, we have an upper bound on the probability of not ending up with any self-links in the whole process which is $\left(1 - \frac{1}{2n}\right)^{n/2}$ which converges to $e^{-1/4}$.

The nice implication of (3) is that we work with a degree sequence that has a nice limiting degree distribution $P(d)$,[7] if we generate a network through the configuration

[7] There are various definitions of a limiting distribution that we could work with. For instance, it

model and delete duplicate and self links, then proportion of nodes that have degree $d$ converges almost surely to $P(d)$ (so $\Pr[\lim_n |p^n(d) - P(d)| = 0] = 1$, where $p^n(d)$ is the realized proportion of nodes with degree $d$ after the deletion of duplicate and self-links).

**Proof of Proposition 4.1.1:** Let $\widehat{d^n} = \max_{i \leq n} d_i$ be the maximum degree up to node $n$ and $\langle d \rangle^n = \frac{\sum_{i \leq n} d_i}{n}$ be the average degree through node $n$. We can find a bound for the probability that any given node ends up with a self-link or a duplicate link. First, instead of thinking of the configuration process as picking two entries at random and matching them and then iterating, imagine instead that we start by picking the first entry of the first element and randomly choosing a match for it, and then movie on to the second remaining entry, and so forth. It is not hard to see that this leads to the same distribution over matchings and thus of links. Consider the first node and its first link (isolated nodes can be discarded). The chance that the link is not a self-link or duplicate link (so far) is $1 - \frac{d_1 - 1}{n\langle d \rangle^n - 1}$, as we only need to worry about self-links. This is greater than $1 - \frac{\widehat{d^n}}{n\langle d \rangle^n - \widehat{d^n}}$. The chance that the second link (if it has degree above 1) is not a self-link or duplicate link (so far), presuming the first one is not a self-link, is then $1 - \frac{d_1 - 2}{n\langle d \rangle^n - 2} - \frac{d_i - 1}{n\langle d \rangle^n - 2}$, where $d_i$ is the degree of the node that the first link went to. This is greater than $1 - \frac{2\widehat{d^n}}{n\langle d \rangle^n - \widehat{d^n}}$. Continuing in this manner, we end up with a lower bound on the probability of self or duplicate links of

$$\prod_{j=1,\ldots,\widehat{d^n}} \left(1 - \frac{j\widehat{d^n}}{n\langle d \rangle^n - \widehat{d^n}}\right).$$

This is larger than

$$\left(1 - \frac{(\widehat{d^n})^2}{n\langle d \rangle^n - \widehat{d^n}}\right)^{\widehat{d^n}}.$$

If $\frac{\widehat{d^n}}{(n\langle d \rangle^n - \widehat{d^n})^{1/3}}$ tends to 0, then we can approximate the above expression by[8]

$$e^{-(\widehat{d^n})^3/(n\langle d \rangle^n - \widehat{d^n})}$$

which tends to 1 if (and only if) $\frac{\widehat{d^n}}{(n\langle d \rangle^n)^{1/3}}$ tends to 0. ∎

could be that $P_n(d)$ converges to $P(d)$ for each $d$, but that it takes much longer to get to the limit for some $d$'s compared to others. To make the above statement precise, consider a form of uniform convergence where $\max_d |P^n(d) - P(d)| \to 0$. We can also work with other (weaker) definitions of convergence, such as pointwise convergence and also to what is known as weak convergence, or convergence in distribution (e.g., see Billingsley [63]).

[8] We can approximate $(1 - \frac{r}{x})^x$, when $r \to 0$ and $x$ does not decrease, by $e^{-r}$. See Section **??** for approximating expressions.

### 4.1.5 An Expected Degree Model

Chung and Lu [145] [146] provide a different random model that also approximates a given desired degree sequence. The advantage of this process is that it forms a graph instead of a multigraph, although it still allows for self loops and does not result in the exact degree sequence, even asymptotically.

Once more, start with $n$ nodes and a desired degree sequence $\{d_1, \ldots, d_n\}$. Form a link between nodes $i$ and $j$ with probability $d_i d_j / (\sum_k d_k)$, where the degree sequence is such that that $(\max_i d_i)^2 < \sum_k d_k$, so that each of these probabilities is less than 1.

It is clear that any node $i$'s expected degree is indeed $d_i$ (when a self-link $ii$ is allowed to form with probability $d_i^2 / \sum_k d_k$).

To get a better feeling for the differences between the configuration model and the Chung-Lu process, consider a degree sequence where all nodes have the same number of links $k = \langle d \rangle$. Let us first consider the configuration model, where we delete self and duplicate links. As we argued above, the probability that any given node has no duplicate links or self links, and hence degree exactly $k$, converges to 1. From here it is not difficult to conclude that with a probability going to 1, the proportion of nodes with degree $k$ will also converge to 1. Under the Chung-Lu process, although the expected degree of any given node is $k$ (and approaches this if we exclude self links), the chance that it ends up with exactly $k$ links is bounded away from 1, regardless of whether we allow self links. To see this, note that the number of links to other nodes for any node follows a binomial distribution on $n-1$ draws with a probability of $k/n$. As the probability of self links vanishes, the probability that the degree is the same as the number of links excluding self links approaches 1. However, even as $n$ becomes large, a binomial distribution of $n-1$ draws with probability $k/n$ places a probability bounded away from 1 on having exactly $k$ links. In fact, this is effectively the same as having a Poisson random network! The probability of having exactly $k$ links can be approximated from a Poisson approximation (recall (1.4)), and we find a probability on the order of $\frac{e^{-k}(k)^k}{k!}$, which is maximized at $k = 1$ and always less than $1/2$. This tells us that the realized degree distribution will differ significantly from the distribution of the expected degree sequence, which places full weight on degree $k$.

While the configuration process (under suitable conditions) leads to a degree distribution more closely tied to the starting one, the Chung-Lu expected degree process is still of interest and more naturally relates to the Poisson random networks. Both are useful.

### 4.1.6 Some Thoughts about Static Random Network Models

The configuration model and the expected degree model are effectively algorithms for generating random networks with desired properties in terms of their degree sequences. They will generally lack the observed clustering and correlation patterns that were discussed in Chapter 3, as the links are formed without paying attention to anything except relative degrees. A node forms links to two other nodes who are connected to each other purely by chance, and not because of their relation to each other. They are also severely limited as models of how social and economic networks form, since they miss the incentives and forces that influence the formation of relationships; as the models describe a world governed completely and uniformly by chance. So, why study such random graph models? One of the biggest challenges in network analysis is developing tractable models. The combinatorial nature of networks that exhibit any heterogeneity makes them complex animals. Much of the theory starts by building up from simple models and techniques, and seeing what can be carried further. These two models represent important steps in generalizing Poisson random graphs, and we can see that some of the basic properties of Poisson random graphs do generalize to some richer degree distributions, and we get a better understanding of how degree distributions relate to other properties of networks. Although there are more things that we will introduce to the models, but there is still much to be learned from looking at these relatively simple generalizations of the Poisson model. As we shall see, these models will be workhorses in providing foundations for understanding diffusion in a network, among other things.

## 4.2 Properties of Random Networks

If we fix some number of nodes $n$, and then try to analyze the properties of a resulting random network, we run into some difficulties. For instance, if we examine the Poisson random network model, then each possible network has a postive probability of being formed. While some are much more likely than others, it is difficult to talk about what properties the resulting network will exhibit since everything is possible. We could try to sort out which properties are "likely" to hold and how this depends on the probability with which links are formed, but for a fixed $n$ the likelihood of a given property holding will often be a complicated expression that will be difficult to interpret. One technique for dealing with this issue is to resort to computer simulations where a large number of

random networks are generated according to some model to estimate probabilities of different properties being exhibited on some fixed number of nodes. Another technique is to examine the properties of the network at some limit, for instance as the number of nodes tends to infinity. If one can show that a property does (or does not hold) at the limit, then one can often conclude that the probability of it holding for a large network is close to 1 (or 0). Simulations are useful in a number of ways. For instance, it can be that even limiting properties are hard to ascertain analytically, and then simulations provide the only real tool for examining a property. It could also be that we are interested in a relatively small network, or that we want to see how the probability of a given property being exhibited varies with parameters and the size of the population. As simulation techniques are more straightforward, I illustrate them at different points in what follows. The alternative approach of examining the limiting properties of large networks requires the development of some tools and concepts which I now discuss.

### 4.2.1 The Distribution of the Degree of a Neighboring Node

In a variety of applications, one is faced with the following sort of calculation. Start at some node $i$ with degree $d_i$. Consider a neighbor $j$. How many neighbors do we expect $j$ to have? This is important in estimating the size of $i$'s expanding neighborhoods, in keeping track of contagion and transmission of beliefs, in estimating diameters, and many other calculations. Basically, any time that we consider some process that moves through the network and we wish to keep track of how many lins it expects to have to be able to follow at a next step, this is an important sort of calculation.

To understand such calculations, let us start by examining the following related calculation. Suppose that we randomly select a link from a network and then randomly pick one of the nodes at either end of the link. What is the conditional probability that describes that node's degree? If the network has a degree distribution described by $P$, the answer is *not* simply $P$. To understand this, let us start with a simple case where the network is such that $P(1) = \frac{1}{2} = P(2)$. So, one half of the nodes have degree 1 and one have have degree 2. For instance, a network on four nodes with links: $\{12, 23, 34\}$. While the degree distribution is $P(1) = \frac{1}{2} = P(2)$, it is easy to see that if we randomly pick a link and then randomly pick an end of it, there is a $\frac{2}{3}$ chance that we find a node of degree 2 and a $\frac{1}{3}$ chance that we find a node of degree 1. This just reflects the fact that higher degree nodes are involved in a proportionately higher percentage of the links. In fact, their degree determines relatively how many more links they are involved with. In particular, if we randomly pick a link and a node at the end of it,

and we consider two nodes of degrees $d_j$ and $d_k$, then node $k$ is relatively $d_k/d_j$ times more likely to be the one we find than node $j$. Extrapolating, it is easy to see that the distribution of degrees of a node found by choosing a link uniformly at random from a network that has degree distribution $P$ and then picking either one of the end nodes with equal probability is

$$\widetilde{P}(d) = \frac{P(d)d}{\langle d \rangle} \tag{4.2}$$

where $\langle d \rangle = E_P[d] = \sum_d P(d)d$ is the expected degree under the distribution $P$.

This means that simply randomly picking a node from a network, and finding nodes by randomly following the end of a link, are two very different exercises. One is much more likely to find high degree nodes by following the links in a network than by randomly picking a node.

Now let us return to the question we started with: let us start at a node $i$ with degree $d_i$ and examine the distribution of the degree of one of its randomly selected neighbors. If we consider either the configuration or expected degree models, and we let the number of nodes grow large and have the degree distribution converge (uniformly) to $P$, then the answer will converge to the $\widetilde{P}$ described in (4.2). This is true since the degrees of two neighbors are approximately independently distributed for large networks provided the largest nodes are not too large.[9] This is also true in the Poisson random networks of Erdös and Rényi. We can also directly deduce that the distribution of the *expected* degree of the node at a given end of any given link (including self-links) under the Chung-Lu process is exactly given by $\widetilde{P}$. However, this might not match the distribution of degree of the node at a given end of any given link. As an example, under the Chung-Lu process if we have $P(2) = 1$, so that all nodes have an expected degree of 2, some nodes will end up with more than two links and some with less. There, $\widetilde{P}$ places probability 1 on having degree 2. If we rewrite $P$ to be the realized degree distribution, then for large $n$, (4.2) provides a good approximation of the degree of a neighbor.

However, it is important to note that (4.2) does not provide the right calculation for many prominent models of growing random networks (such as those coming from preferential attachment) that are discussed in Chapter **??**. In those random networks there is nonvanishing correlation between the degrees of nodes, so that higher degree

[9] To see why this is only approximate, consider any given degree sequence and for the expected degree model. Say that there are $n_d$ nodes with degree $d$. One of those nodes can only be connected to $n_d - 1$ nodes with degree $d$, while a node with degree $d' \neq d$ can be connected to $n_d$ nodes with degree $d$. So, here we actually see a (slight) negative correlation in the degrees of neighboring nodes.

nodes tend to have neighbors with higher degrees than do lower degree nodes.

To see how correlation can change the calculations, consider two different methods of generating a network with a degree distribution such that half of the nodes have degree 1 and half have degree 2. First, generate such a network by operating the configuration model on a degree sequence of $(1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2 \ldots)$.[10] In this case, it is clear that picking any node and asking what the degree of a randomly selected neighbor is has an answer that converges to a $\frac{2}{3}$ chance that it is a node of degree 2 and a $\frac{1}{3}$ chance that it is a node of degree 1.

Second, consider the following very different way of generating a network with the same degree distribution. Start with eight nodes. Connect four of them in a square, so that they each have degree 2 and have two neighbors each with degree 2. Connect the other four of them in two pairs, so that each has degree 1 and has a neighbor with degree 1, as in Figure 4.2.1. Now replicate this process. We end up with the same degree distribution, so that half of the nodes are of degree 1 and half of degree 2, but nodes are segregated so that nodes with degree 1 are only connected to nodes of degree 1, and similarly nodes of degree 2 are only connected to nodes of degree 2. Here, the degree of a node's neighbor is perfectly correlated with that node's degree. Note also that if we examine the degree of a neighbor of a randomly picked node in Figure 4.2.1, we end up with an equal probability that it will have degree 1 or degree 2! That is, if we examine nodes 1 to 4, then any randomly selected neighbor will have degree 2, while if we examine nodes 5 to 8, then any randomly selected neighbor will have degree 1. So, the distribution of a node's neighbor's degree is quite different from $\widetilde{P}$, regardless of whether we condition on the starting node's degree or whether we simply pick a node uniformly at random and then examine one of its neighbors' degrees.

While this example is stark, it illustrates that we do need to be careful in keeping track of how a network was generated, and not only its degree distribution, in order to properly calculate things like the distribution of degrees of neighboring nodes.[11]

[10] For this calculation, let us work with the resulting multigraph, so that we allow for self-links and duplicate links, and so that the degree distribution is exactly realized when the number of nodes is a multiple of four.

[11] A caution here is that some of the literature proceeds with calculations as if there were no correlation between neighboring nodes, even though some of the models (like preferential attachment discussed in Chapter **??**) used to motivate the analysis generate significant correlation. Using a variation on the configuration model is one approach to avoiding such problems, but it does limit the scope of the analysis.

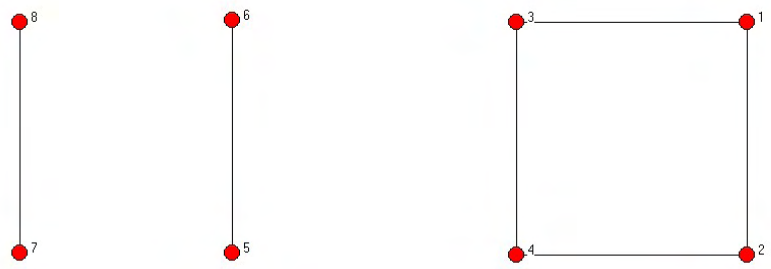



**Figure 4.2.1.** Forming Networks with Perfect Correlation in Degrees

### 4.2.2 Thresholds and Phase Transitions

When we examine random networks on a growing set of nodes and for some given parameters or structure, often properties hold with either a probability approaching 1 or a probability approaching 0 in the limit. So, while it may be difficult to figure out the precise probability that some property holds for some fixed $n$, it is often much easier to discern whether that probability is tending to 1 or 0 as we let $n$ approach infinity.

To see how this works, let us consider the Poisson random network model on a growing set of nodes $n$, where we index the probability of a link forming as a function of $n$, denoted $p(n)$. It is quite natural to have the probability of a link forming between two nodes vary with the size of the population. For example, if we imagine that people have on average several thousand acquaintances, then we might need $p$ to be on the order of one percent if we are dealing with an $n$ that includes a few hundred thousand nodes, but more on the order of a fraction of a percent if we are dealing with a population of millions of nodes. So, for instance if we want to keep average degree constant as the number of nodes grows then we need $p(n)$ to be proportional to $1/n$. With a $p(n)$ in hand, we can ask what the probability is that a given property holds as

$n \to \infty$. Interestingly, many properties hold with a probability that approaches either 0 or 1 as the number of nodes grows, and the probability that a property holds can shift sharply between these as we change the underlying random network process. For example, we can ask what the probability is that a network will have some isolated nodes. For some random network formation processes if the network is large then it will be almost certain that there will exist some isolated nodes, while for other network formation processes it will be almost certain that the resulting network will not have any isolated node. We will see that this sharp dichotomy will be true of a variety of properties such as whether the network has a giant component, or has a path between any two nodes, or has at least one cycle, etc. There are also many exceptions, in terms of properties that do not exhibit such convergence patterns. For instance, consider the property that a network has an even number of links. For many random network processes, the probability of this property holding will be bounded away from 0 and 1.

There are different ways of specifying a property, but an easy way is just to list the networks that satisfy it. Thus, properties are generally specified as a set of networks for each $n$, and then a property is satisfied if the realized network is in the set. Thus a property is a list of $A(N) \subset G(N)$ listing the networks that have the property when the set of nodes is $N$. So, for instance, the property that a network has no isolated nodes is

$$A(N) = \{g \mid N_i(g) \neq \emptyset \forall i \in N\}.$$

Most properties that are studied are referred to as *monotone or increasing properties*. Those are properties such that if a given network satisfies the property, then any supernetwork (in the sense of set inclusion) satisfies it. So a property $A(\cdot)$ is monotone if $g \in A(N)$ and $g \subset g'$ implies that $g' in A(N)$. The property of having an even number of links is obviously not a monotone property, while the property of being connected is a monotone property.

If we work in the Poisson model, then the model is completely specified by $p(n)$ where $n$ is the cardinality of the set of nodes $N$. In that case, a *threshold function* for some given property is a function $t(n)$ such that the property holds with a probability approaching 1 (i.e., $\Pr[A(N)|p(n)] \to 1$) if $p(n)/t(n) \to \infty$, while the property holds with a probability approaching 0 (i.e., $\Pr[A(N)|p(n)] \to 0$) if $p(n)/t(n) \to 0$. When such a threshold function exists, it is said that a *phase transition* occurs at that threshold.[12] Even when there are not sharp threshold functions, we can still often produce

[12] There are different sorts of probabilistic statements that one can make, analogous to differences between the weak and strong laws of large numbers. That is, it can be that the as $n$ grows the

lower or upper bounds so that we know that a given property holds for $p(n)$'s above or below those bounds.

This definition of a threshold function is tailored to the Erdös-Rényi or Poisson random network setting, as it is based on having a function $p(n)$ describe the network formation process. We can also define threshold functions for other sorts of random network models, but they will be relative to some other description of the random process, generally characterized by some parameter(s).

To get a better feeling for a threshold function, let us consider a relatively simple one. Let us consider the property that node 1 has at least one link; that is the property described by $A(N) = \{g \,|d_1(g) \geq 1\}$. In the Poisson model, the probability that node 1 has no links is $(1-p(n))^{n-1}$ and so the probability that $A(N)$ holds is $1-(1-p(n))^{n-1}$. To derive a threshold function, we just need to see for which $p(n)$ this tends to 0 and for which $p(n)$ this tends to 1. If we consider $t(n) = \frac{r}{n-1}$, then by a definition of the exponential function (see Section **??**), the limit of the probability that node 1 has no links is

$$\lim_n (1 - t(n))^{n-1} = \lim_n \left(1 - \frac{r}{n-1}\right)^{n-1} = e^{-r}. \tag{4.3}$$

So, if $p(n)$ is proportional to $\frac{1}{n-1}$, then there will be a probability that node 1 has at least one link that is bounded away from 0 and 1 in the limit. Thus, $t^*(n) = \frac{1}{n-1}$ is a function that could potentially serve as a threshold function. Let us check that $t^*(n) = \frac{1}{n-1}$ is in fact a threshold function. Suppose that $p(n)/t^*(n) \to \infty$. This implies that $p(n) \geq \frac{r}{n-1}$ for any $r$ and large enough $n$. Therefore, from (4.3) it follows that $\lim_n (1 - p(n))^{n-1} \leq e^{-r}$ for all $r$, and so $\lim_n (1 - p(n))^{n-1} = 0$. Similarly, if $p(n)/t^*(n) \to 0$, then an analogous comparison implies that $\lim_n (1 - p(n))^{n-1} = 1$. Thus, $t^*(n) = \frac{1}{n-1}$ is a threshold function for a given node having neighbors in the Poisson random network model.

Note that the threshold function is not unique here, as $t(n) = an + b$ for any fixed $a$ and $b$ will also provide the same conclusion. Moreover, threshold functions provide

probability of a property holding goes to one. This is the "weak" form of the statment. The stronger form of the statement reverses the order between the probability and the limit, stating that the probability that the property holds in the limit is one. This is also stated as having something hold *almost surely*. For many applications this difference is irrelevant, but in some cases it can be an important distinction. In most instances in this text, I will claim or use the weaker form, as that is generally much easier to prove and one can work with a series of probabilities, which keeps the exposition relatively clear, rather than having a probability defined over sequences. Nevertheless, many of these claims hold in their stronger form.

only conclusions about the how large or small $p(n)$ has to be in terms of its limiting order and only provide limiting conclusions. How large $n$ has to be in order for the property to hold with a high probability depends on more detailed information. For instance, $p(n) = e^{-n}$ and $p(n) = \frac{1}{n^{1.0001}}$ both lead to probabilities of 0 that node 1 will have any neighbors in the limit, but the second function gets there much more slowly. Determining smaller $n$ properties requires examining the probabilities directly, which is feasible in this example, but more generally may require simulations.

With regards to the Poisson random network model, there is much that is known about properties and thresholds. A very brief summary is as follows.

- At the threshold of $\frac{1}{n^2}$ the first links emerge, so that the network is likely to have no links in the limit for $p(n)$'s of order less than $\frac{1}{n^2}$, while for $p(n)$'s of order larger than $\frac{1}{n^2}$ the network has at least one link with a probability going to one.[13] (The proof of this is Exercise **??**.)

- Once $p(n)$ is at least $\frac{1}{n^{3/2}}$ there is a probability converging to one that the network has at least one component with at least three nodes.

- At the threshold of $\frac{1}{n}$ we see cycles emerge, and we also see the emergence of a "giant component," which is a unique largest component which contains a nontrivial fraction of all nodes.

- The giant component grows in size until the threshold of $\frac{\log(n)}{n}$, where the network becomes connected.

These various thresholds are illustrated in the following series of figures from randomly drawn networks. These are Poisson random networks generated on fifty nodes (using the program Ucinet). With $n = 50$, we have the first links emerging at $p = \frac{1}{n^2} = .0004$. The threshold where we see the first component emerge with more than two nodes is at $p = n^{-3/2} = .003$. Indeed, we see in the first network with $p = .01$ that we have a component with three nodes, but still the network is very sparse.

[13] Note that this does not contradict the calculations above, which were for the property that single node did/did not have any neighbors. The property here is that none of the nodes have any neighbors.

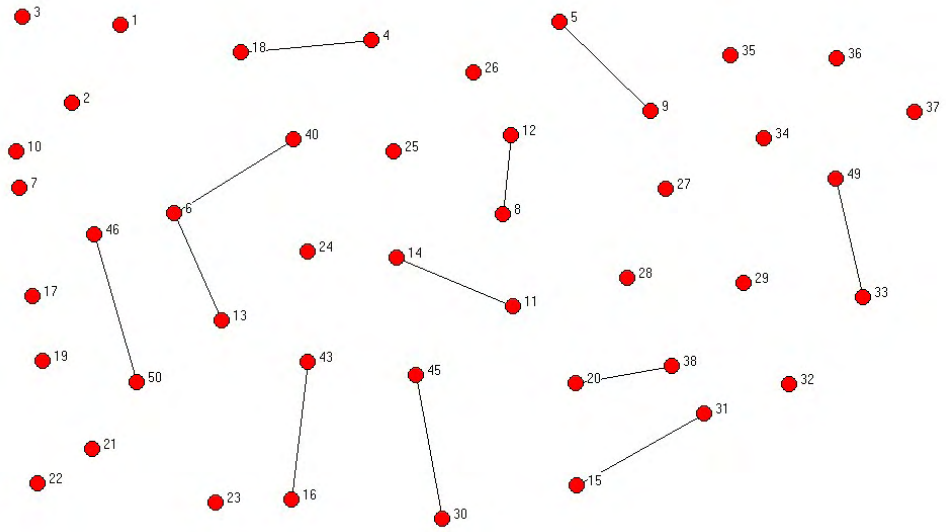



**Figure 4.2.2.** A First Component with More than Two Nodes: A Random Network on 50 Nodes with p=.01

At the threshold of $p = \frac{1}{n} = .02$ we should see cycles start to emerge. We see this in that the first network with $p = .01$ has no cycles, while the networks with $p = .03$ (or more) all have cycles. Moreover, this is also where we see the first signs of a giant component.

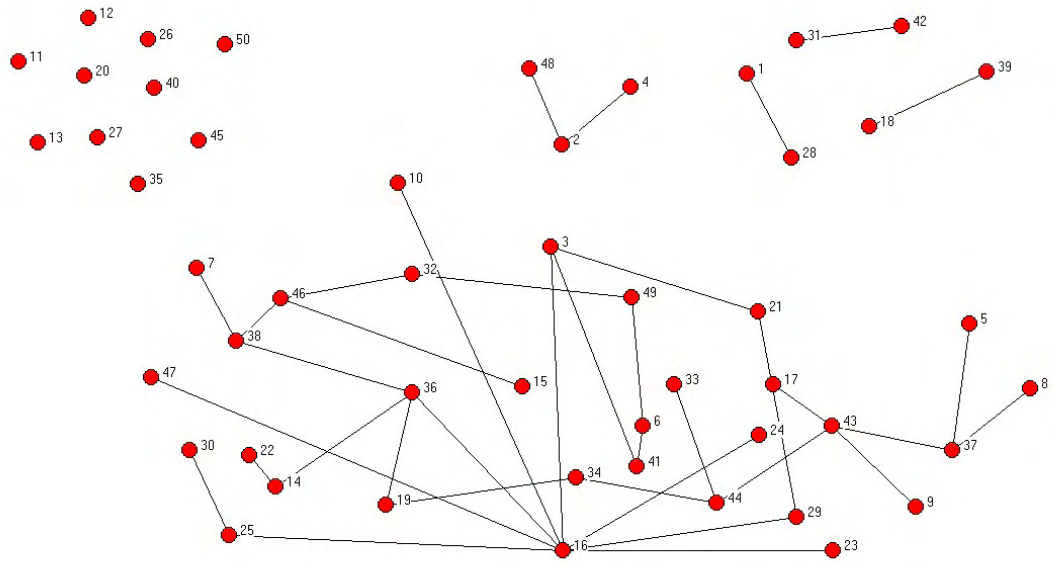



**Figure 4.2.2.** The Emergence of Cycles: A Random Network on 50 Nodes with p=.03

As we increase $p$ we see that the giant component starts to swallow more and more nodes.

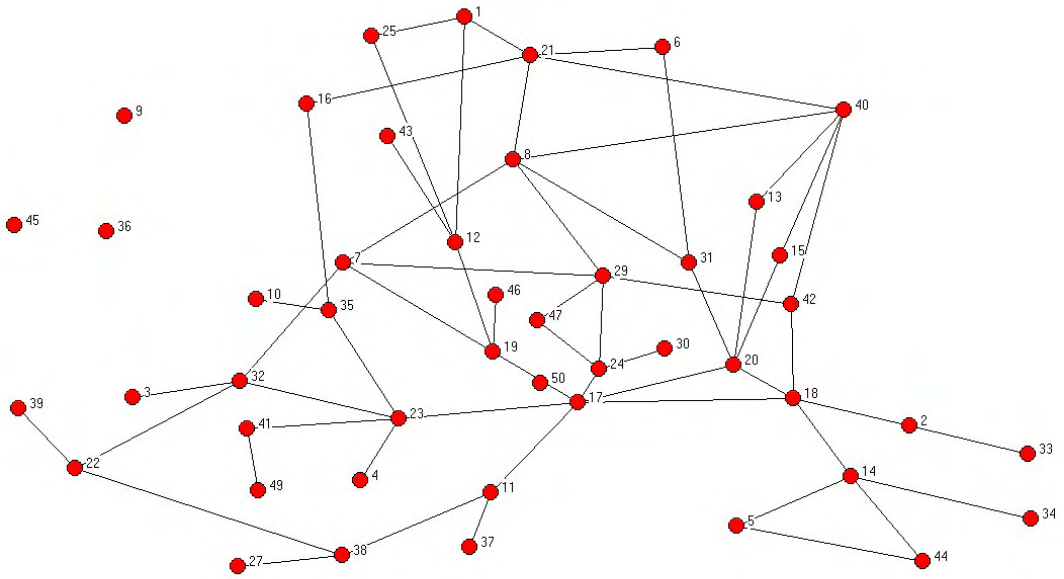



**Figure 4.2.2.** The Giant Component: A Random Network on 50 Nodes with p=.05

Eventually, at the threshold of $p = \frac{log(n)}{n} = .08$ we should see the network become connected. Again, this is seen in the random networks, as the networks generated with $p = .01$, $p = .03$, and $p = .05$ all have at least two components, while the network generated with $p = .10$ is connected.

To get a deeper understanding of how some of these thresholds work, let us start by examining the connectedness of a random network.

### 4.2.3 Connectedness

Whether or not a network is connected, and more generally what its component structure looks like, is important in the transmission and diffusion of information, behaviors, and diseases, as we shall see in Section **??**. Thus, it is important to understand how these properties relate to the network formation process.

The phase transition from a disconnected to a connected network was one of the many important discoveries of Erdös and Rényi [211] about random networks. Exploring this phase transition in detail is not only useful for its own sake, but also because

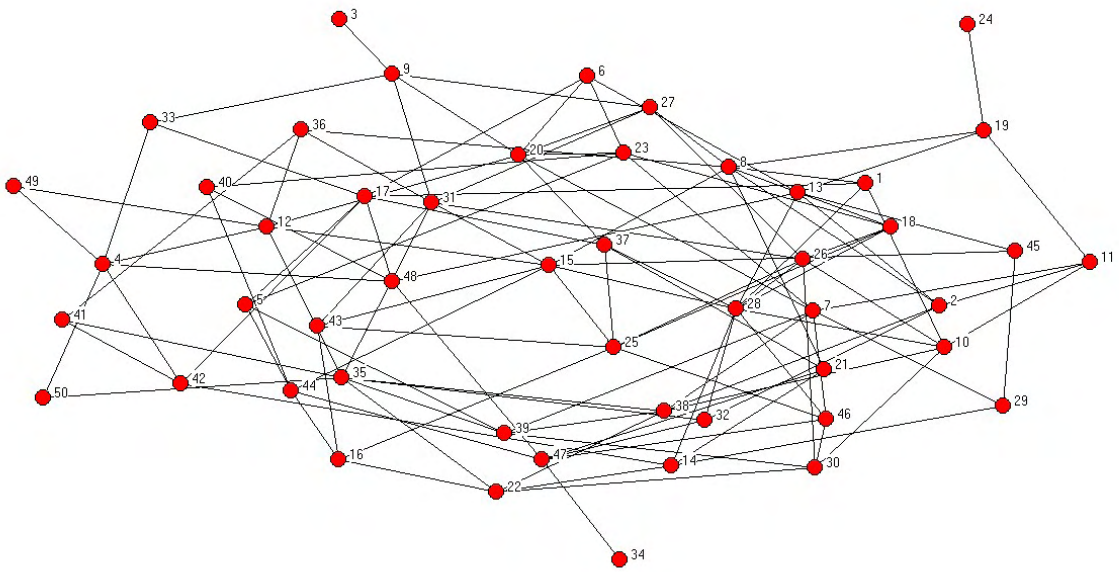



**Figure 4.2.2.** Emergence of Connectedness: A Random Network on 50 Nodes with p=.10

it helps illustrate the idea of phase transitions and provides some basis for extensions to other random network models.

THEOREM **4.2.1** [Erdös and Rényi] *A threshold function for the connectedness of the Poisson random network is* $t(n) = log(n)/n$.

The theorem thus shows that if the probability of a link is larger than $log(n)/n$, then the network is connected with a probability tending to one, while if it is smaller than $log(n)/n$ then the probability that it is not connected tends to one. This threshold corresponds to an expected degree of $log(n)$.

The ideas behind Theorem 4.2.1 are relatively easy to understand, and a complete proof is not too long, even though the conclusion of the theorem is profound. In order to show that a network is not connected, it is enough to show that there is some isolated node. It turns out that $t(n) = log(n)/n$ is not only the threshold for a network being connected, but also for there not to be any isolated nodes. To see why this is true, note that the probability that a given node is completely isolated is $(1 - p(n))^{n-1}$ or roughly $(1 - p(n))^n$. When working with a $p(n)$ near the threshold, $p(n)/n$ converges

to 0, and so we can approximate $(1-p(n))^n$ by $e^{-np(n)}$. Thus, the probability that any given node is isolated goes to $e^{-p(n)n}$, which evaluated at the threshold is $1/n$. When we have $n$ nodes, it is then not too hard to show that this is the threshold of having some of them be isolated, as below the threshold the chance of any node being isolated is significantly less than $1/n$ while above the threshold it is significantly above $1/n$. The proof then shows that above this threshold it is not only that there are no isolated nodes, but also no components of size less than $n/2$. The intuition behind this is that the probability of having a component of some small finite size is similar (asymptotically) to having an isolated node: there need to be no connections between any of the nodes in the component and any of the other nodes. Thus, one either has some isolated nodes, or else the smallest components must be approaching infinite size. However, the chance of having more than one component of substantial size goes to 0, as there are many nodes in each component and there cannot be any links between separate components, which is then a very low probability event. So, components roughly come in two flavors: very small or very large.

I now offer a full proof of the Theorem in order to give a rough idea of how some of the many results in random graph theory have been proven: basically by bounding probabilities and expectations and showing that the bounds have the claimed properties.

**Proof of Theorem 4.2.1:**[14]

Let us start by showing that $t(n) = log(n)/n$ is the threshold for having isolated nodes. First, we show that if $p(n)/t(n) \to 0$, then the probability that there are isolated nodes is tending to 1. This clearly implies that the network is not connected.

The probability that a given node is completely isolated is $(1-p(n))^{n-1}$ or roughly $(1-p(n))^n$ if $p(n)$ is converging to 0. Given that $p(n)/n$ converges to 0, we can approximate $(1-p(n))^n$ by $e^{-np(n)}$. Thus, the probability that any given node is isolated goes to

$$e^{-p(n)n}.$$

We can write $p(n) = \frac{log(n)-f(n)}{n}$, where $f(n) \to \infty$ and $f(n) < log(n)$, and then $e^{-p(n)n}$ becomes

$$\frac{e^{f(n)}}{n}.$$

The expected number of isolated nodes is then $e^{f(n)}$, which tends to infinity.

[14] This proof is adapted from two different proofs by Bollobas (Theorem 7.3 in [80] and Theorem 9 on page 233 of [79]).

While expecting a divergent number of isolated nodes in the limit is suggestive that there will be some isolated nodes, it does not prove that the probability of there being at least one isolated node converges to 1. We show this via Chebychev's inequality.[15] Let $X^n$ denote the number of isolated nodes. We have shown that $E[X^n] \to \infty$. If we can show that the variance of $X^n$, $E[(X^n)^2] - E[X^n]^2$, is no more than twice $\mu = E[X^n]$, then we establish the claim by applying Chebychev's inequality. In particular, we then can conclude that $\Pr[X^n < \mu - r\sqrt{2\mu}] < 1/r^2$ for all $r > 0$, which since $\mu \to \infty$ implies that the probability converges to 1 that $X^n$ will be arbitrarily large and so there will be an arbitrarily large number of isolated nodes. To obtain an upper bound on $E[(X^n)^2] - E[X^n]^2$, note that $E[X^n(X^n - 1)]$ is the expected number of ordered pairs of isolated nodes, which is $n(n-1)(1-p)^{2n-3}$ since a pair of nodes is isolated from the other nodes if none of the $2(n-2)$ links from either of them is present and the link between them is not present. Thus,

$$\begin{aligned}
E\left[(X^n)^2\right] - E\left[X^n\right]^2 &= n(n-1)(1-p)^{2n-3} + E[X^n] - E[X^n]^2 \\
&= n(n-1)(1-p)^{2n-3} + E\left[X^n\right] - n^2(1-p)^{2n-2} \\
&\leq E\left[X^n\right] + pn^2(1-p)^{2n-3} \\
&= E\left[X^n\right]\left(1 + pn(1-p)^{n-2}\right) \\
&\leq E\left[X^n\right]\left(1 + (log(n) - f(n))e^{-log(n)+f(n)}(1-p)^{-2}\right) \\
&\leq 2E\left[X^n\right].
\end{aligned}$$

To complete the proof that $t(n) = log(n)/n$ is the threshold for having isolated nodes, we need to show that if $p(n)/t(n) \to \infty$, then the probability that there are isolated nodes is tending to 0. It is enough to show this for $p(n) = \frac{log(n)+f(n)}{n}$, where $f(n) \to \infty$ but $f(n)/n \to 0$.[16] By a similar argument to the one above, we conclude that the expected number of isolated nodes is tending to $e^{-f(n)}$, which tends to 0. The probability of having $X^n$ be at least one then has to tend to 0 as well in order for $E[X^n] \to 0$.

To complete the proof of the Theorem, we need to show that if $p(n)/t(n) \to \infty$, then the chance of having any components of size 2 to size $n/2$ tends to 0. Let $X_k$ denote

[15] Chebychev's inequality (see Section 4.5.3) says that for a random variable $X$ with mean $\mu$ and standard deviation $\sigma$, $\Pr[|X - \mu| > r\sigma] < 1/r^2$ for every $r > 0$.

[16] Having no isolated nodes is clearly an increasing property, so that it holds for larger $p(n)$. The reason for working with $f(n)/n \to 0$ is to ensure that the approximation of $(1-p(n))^n$ by $e^{-np(n)}$ is valid asymptotically.

the number of components of size $k$, and write $p(n) = \frac{log(n)+f(n)}{n}$, where $f(n) \to \infty$ and $f(n)/n \to 0$.[17] It is enough to show that $E[\sum_{k=2}^{n/2} X_k] \to 0$.

$$\begin{aligned}
E\left[\sum_{k=2}^{n/2} X_k\right] &= E\left[\sum_{k=2}^{n/2} X_k\right] \\
&\leq \sum_{k=2}^{n/2} \binom{n}{k} (1-p)^{k(n-k)} \\
&= \sum_{k=2}^{n^{3/4}} \binom{n}{k} (1-p)^{k(n-k)} + \sum_{k=n^{3/4}}^{n/2} \binom{n}{k} (1-p)^{k(n-k)} \\
&\leq \sum_{k=2}^{n^{3/4}} \left(\frac{en}{k}\right)^k e^{-knp} e^{k^2 p} + \sum_{k=n^{3/4}}^{n/2} \left(\frac{en}{k}\right)^k e^{-knp/2} \\
&\leq \sum_{k=2}^{n^{3/4}} e^{k(1-f(n))} k^{-k} e^{2k^2 log(n)/n} + \sum_{k=n^{3/4}}^{n/2} \left(\frac{en}{k}\right)^k e^{-knp/2} \\
&\leq 3e^{-f(n)} + n^{-n^{3/4}/5},
\end{aligned}$$

which tends to 0 in $n$. ∎

The above proof used the specific structure of the Poisson random networks model fairly extensively. How far can we extend it to other random network models?

It is fairly clear that the argument we used to prove the above theorem is not well suited to the configuration model. Under the configuration model, under some reasonable bounds on degrees, each node will end up with its specified degree with a probability approaching 1. This renders the approach above inapplicable. It is clear that if the limiting degree distribution has a positive mass on nodes of degree 0 in the limit then it will not be connected, but otherwise it is not so clear what will happen. For instance, if the associated $\widetilde{P}$ has mass on nodes of some bounded degree in the limit, then there will be a non-vanishing probability that the network will be not be connected. However, requiring that the mass on nodes of some bounded degree vanish is not enough, as it is still possible to have the network have a nontrivial probability of being disconnected.

[17] Here again, we work with $p(n)$ "near" the threshold, as this will establish that the resulting network is connected with a probability going to one for such $p$'s, and then it holds for larger $p$'s.

The expected-degree model of Chung and Lu **??** is better suited for an analysis with regard to being connected, or at least we can make some progress with regard to the threshold for the existence of isolated nodes. This follows since it is essentially a generalization of the Poisson random network model that allows for different expected degrees across nodes (with the possibility of self loops).

Recall that in the expected-degree model, we work with degree sequences of an expected degree for each node $d_1, \ldots, d_n$. Let

$$Vol^n = \sum_{i=1}^{n} d_i, \tag{4.4}$$

denote the total expected degree of the network on $n$ nodes. The probability of a link between nodes $i$ and $j$ is then $\frac{d_i d_j}{Vol^n}$, and so the probability that node $i$ is isolated is

$$\prod_j (1 - \frac{d_i d_j}{Vol^n}).$$

The probability that a given node $i$ is isolated is then approximately $e^{-d_i \sum_j d_j / Vol^n} = e^{-d_i}$ for large $n$ (under the assumption that $max_i \frac{d_i^2}{Vol^n}$ converges to 0, which is maintained under the expected-degree model). The probability that no node is isolated is then

$$\prod_i (1 - e^{-d_i})$$

or approximately

$$e^{-\sum_i e^{-d_i}}.$$

This suggests a threshold such that if $\sum_i e^{-d_i} \to 0$ then there will be no isolated nodes, while if $\sum_i e^{-d_i} \to \infty$ then there will be isolated nodes.[18] As a double-check of this, let $d_i = d(n) = log(n) + f(n)$ for each $i$ the Poisson random network setting (where $p(n) = d(n)/n$). This leads to $\sum_i e^{-d_i} = e^{-f(n)}$, so if $f(n) \to \infty$ we end up with no isolated nodes (and a connected network) and if $f(n) \to -\infty$ then with a probability going to 1 there are isolated nodes. Indeed, this corresponds to the threshold we found in the Poisson random network model.

[18] I am not aware of results on this question or the connectedness of the network under the expected degree model. While it seems natural to conjecture that the threshold for the existence of isolated nodes will again be the same as the threshold for connectedness, the details need to be checked.

### 4.2.4 Giant Components

indexgiant component

In cases where the network is not connected, it will be interesting to know something about the component structure as there will generally be many components. In fact, we have already shown that if the network is not connected in the Poisson random network model then there should be an arbitrarily large number of components. We also know from Section **??**, that in this case there may still exist a giant component. Let us examine this in more detail and for a wider class of degree distributions.

In defining the size of a component, a convention is to call a component "small" if it has fewer than $\frac{n^{2/3}}{2}$ nodes, and "large" if it has at least $n^{2/3}$ nodes (e.g., see Chapter 6 in Bollobas **??**). The term "giant component" refers to the unique largest component if there is one. This may turn out to be a small component in some networks but we will generally be interested in giant components that involve non-vanishing fractions of nodes, which will necessarily be "large" components.

The idea of there being a unique largest component is fairly easy to understand, in the case where these are large components. It relates back to what we saw in the proof of Theorem 4.2.1: for any two large sets of nodes (each containing at least $n^{2/3}$ nodes) it is very unlikely that there will be no links between them, unless the overall probability of links is very small. For instance, in the Poisson random network model the probability of having no links between two given large sets of nodes is no more than $(1-p)^{n^{4/3}}$. If $pn^{4/3} \to 0$, then this expression is positive, but otherwise it tends to 0. Proving that the probability of not having two separate large components goes to 0 involves a bit more proof, but is relatively straightforward (see Exercise 4.7).

### 4.2.5 Size of the Giant Component in Poisson Random Networks

As we have already seen, it is not even clear whether each node will reach every other node. Unless $p$ is high enough relative to $n$, it is likely there will be pairs of nodes that are not path-connected. As such, diameter is often measured with respect to the largest component of a network.[19] But this also raises a question as to what the network looks like in terms of components. The answer is one of the deeper and more elegant results of Erdös and Rényi's work.

[19] This can result in some distortions, as, for instance, a network where each node has exactly one link has a diameter much smaller than a network that has many more links.

To get some impression as to the size of the largest component, generally referred to as the "giant component," let us do a simple heuristic calculation.[20] Form a Poisson random network on $n-1$ nodes with a probability of any given link being $p > 1/n$. Now let us add a last node, and again connect this node to each other node with an independent probability $p$. Let $q$ be the fraction of nodes in the largest component of the $n-1$ node network. As a fairly accurate approximation for large $n$, this will also be the fraction of nodes in the largest component of the $n$ node network. [The only possible exception to this is if the added node ends up connecting two large components that were not connected before. As argued above, the chance of having two components with large numbers of nodes that are not connected to each other goes to 0 in $n$, given that $p > 1/n$.] Now, the chance that this added node ends up outside of the giant component is the probability that none of its neighbors are in the giant component. If the new node has degree $d_i$ this probability is converging to $(1-q)^{d_i}$, as we let $n$ become large. As we can think of any node as having been added in this way, in a large network the expected frequency of nodes of degree $d_i$ that end up outside of the giant component is approximately $(1-q)^{d_i}$.[21] So, the overall fraction of nodes outside of the giant component, $1-q$, can then be found by averaging $(1-q)^{d_i}$ across nodes. This leads to[22]

$$1-q = \sum_d (1-q)^d P(d). \tag{4.5}$$

When we apply this to the Poisson degree distribution described by (1.4), the fraction of nodes outside of the giant component is then approximated by the solution of

$$1-q = \sum_d \frac{e^{-(n-1)p}((n-1)p)^d}{d!}(1-q)^d.$$

[20] The heuristic argument is based on Newman [480], but a very different and complete proof of the characterizing equation above the threshold for the emergence of the giant component can be found in Bollobas [80].

[21] There are steps omitted from this argument, as for any finite $n$ the degrees of nodes in the network are correlated, as are their chances of being in the largest component conditional on their degree. For example, for a node of degree 1, it is in the giant component if and only if its neighbor is. Then, if that neighbor has degree $d$, then it has $d-1$ chances to be connected to a node in the giant component. So, now the calculation begins to look like $(1-q)^{d-1}$ for the neighbor to be in the giant component. To see a fuller proof of this derivation, see Bollobás [80].

[22] Here take the convention that $0^0 = 1$, so that if $q = 1$, then the right hand side of this equation is $P(0)$.

Since $\sum_d \frac{((n-1)p(1-q))^d}{d!} = e^{(n-1)p(1-q)}$, an approximation is described by the solution to

$$q = 1 - e^{-q(n-1)p}. \tag{4.6}$$

There is always a solution of $q = 0$ to this equation. In the case where the average degree is larger than 1 (i.e., $p > 1/(n-1)$), and only then, there is also a solution for $q$ that lies between 0 and 1.[23] This corresponds to phase transition, in that the appearance of such a giant component comes above the threshold of $(n-1)p = 1$. That is, there is a marked difference in the structure of the resulting network depending on whether average degree is bigger or smaller than one. If the average degree is less than one, then there is essentially no giant component, but instead the network consists of many components which are all of small size relative to the number of nodes. If the average degree exceeds one, then there is a giant component which contains a non-trivial fraction of all nodes (approximately described by (4.6)).

Note that if we let $p(n-1)$ grow (so that the average degree is unbounded as $n$ grows), then the solution for $q$ tends towards 1. Of course, that requires the average degree to become large. In a random network where there is some bound on average degree, so that $p(n-1)$ is bounded, then $q$ will be somewhere between 0 and 1. If we look for a solution to $q = 1 - e^{-q(n-1)p}$ when $n = 50$ and $p = .08$, we are looking for a $q$ that roughly satisfies $q = 1 - e^{-4q}$, and such a $q$ is about .98. So, an estimate for the size of the giant component is 49 nodes out of 50 - which happens to match the realized network in Figure 1.2.3 exactly.

### 4.2.6 Giant Components in the Configuration Model

Understanding giant components more generally is especially important as they play a central role in various problems of diffusion, and a giant component gives an idea of the most nodes that one might possibly reach starting from a single node. With this in mind, let us examine giant components for more general random networks, using

[23] To see this let $f(q) = 1 - e^{-q(n-1)p}$. We are looking for points $q$ such that $f(q) = q$; known as a "fixed-point". Since $f(0) = 1 - e^0 = 0$, $q = 0$ is always a fixed point. Next, note that $f$ is increasing in $q$ with derivative $f'(q) = (n-1)pe^{-q(n-1)p}$ and strictly concave (as the second derivative is negative: $f''(q) = -((n-1)p)^2 e^{-q(n-1)p}$). Since $f(1) = 1 - e^{-(n-1)p} < 1$, we have a function that starts at 0 ends up with a value below 1 and which is increasing and strictly concave. The only way in which it can ever cross the 45 degree line is if it has a slope greater than 1 when it starts, otherwise it will always lie below the 45 degree line and 0 will be the only fixed-point. The slope at 0 is $f'(0) = (n-1)p$, and so there is a $q > 0$ such that $q = f(q)$ if and only if $(n-1)p$ is greater than 1.

the configuration model as a basis.[24] We will work with randomly formed networks according to the configuration model on $n$ nodes; and will then be looking at the limiting probability the resulting networks have a giant component when we let $n$ grow to be large. Consider a sequence of degree sequences, ordered by the number of nodes $n$, with corresponding degree distributions described by $P^n(d)$. Assume that these satisfy some conditions:

(i) the degree distributions converge uniformly to a limiting degree distribution $P$ that has a finite mean,

(ii) there exists $\varepsilon$ such that $P^n(d) = 0$ for all $d > n^{\frac{1}{4}-\varepsilon}$,

(iii) $(d^2 - 2d)P^n(d)$ converges uniformly to $(d^2 - 2d)P(d)$, and

(iv) $E_{P^n}[d^2 - 2d]$ converges uniformly to its limit (which may be infinite).

An important aspect of such sequences is that the probability of having cycles in any small component is tending to 0. Let us examine properties of the degree distribution that tell us when such networks exhibit a giant component. The following is a simple and informal derivation. A somewhat more complete derivation appears in Section 4.3.1.

The idea is to look for the threshold where starting at a random node there is some chance of finding a nontrivial number of other nodes through tracing out expanding neighborhoods. Indeed, if a node is in a giant component then exploring longer paths from the node should lead to the discovery of more and more nodes, while if it is in a small component then expanding neighborhoods will not result in finding many more nodes.

When we are looking for the threshold where the giant component just emerges, then at or below this threshold we will be working with components that are essentially trees, and so each time we search along a link that we have not traced before, we will find a node that we have not visited before. This allows us to analyze component structure up to the point where the giant component emerges as if the network were a collection of trees. The following argument (due to Cohen et al [151]) provides the idea behind there being negligible numbers of cycles below the threshold.[25] Consider

[24] Similar results hold for the expected degree model of Chung and Lu (see [145]), and under weaker restrictions on the set of admissible degree distributions, but follow a less intuitive argument.

[25] This is part of the informality of the derivation, and I refer the interested reader to Molloy and Reed [449] for a more complete proof.

any link in the configuration model on $n$ nodes. The probability that the link connects two nodes that were already connected in a component with $s$ nodes ( where $s$ is the size of some component ignoring that link) is the probability that both of its end nodes lie in that component, which is proportional to $\left(\frac{s}{n}\right)^2$. Thus, the fraction of links that end up on cycles is on the order of $\sum_i \left(\frac{s_i}{n}\right)^2$, where the $s_i$'s are to the sizes of each of the components in the network. This is less than $\sum_i \frac{s_i S}{n^2}$, where $S$ is the size of the largest component. Thus, since $\sum_i s_i = n$, we find that the proportion of links than lie on cycles is of an order no more than $S/n$. If we are at or below the threshold where the giant component is just emerging, then with probability one, $S/n$ is vanishing for large $n$.

Thus, when we consider a sequence of degree disributions at or below the threshold of the emergence of the giant component, the components are essentially trees.[26]

To develop an estimate of component size as the network grows, then let $\phi$ denote the limiting number of nodes that can be found on average by picking a link uniformly at random, picking with equal chance one of its end nodes, and then exploring all of the nodes that can be found via expanding neighborhoods from that end nodes. Given an absence of cycles, the number of new nodes reached by a link is the first node reached plus that node's degree minus one (as one of its links points back to the original node) times $\phi$, which indicates how many new nodes are expected to be reached from each of the first node's neighbors. Thus,

$$\phi = 1 + \sum_{d=1}^{\infty}(d-1)\widetilde{P}(d)\phi = 1 + \sum_{d=1}^{\infty}\frac{P(d)d}{\langle d\rangle}(d-1)\phi.$$

We can rewrite this as

$$\phi = 1 + \frac{(\langle d^2\rangle - \langle d\rangle)\phi}{\langle d\rangle}$$

or

$$\phi = \frac{1}{2 - \frac{\langle d^2\rangle}{\langle d\rangle}}. \tag{4.7}$$

Now we deduce the threshold for where a giant component emerges. If $\phi$ has a finite solution, then when we start from a node picked uniformly at random in the network and examine a the number of nodes we can reach from one of its links, we expect to find a finite number of nodes. This places the node in a finite component. If $\phi$ does not have a finite solution, then we expect at least some nodes that we find uniformly

[26] The more rigorous result proven by Molloy and Reed [449] establishes that almost surely, no component has more than one cycle.

at random to be in components that are growing without bound, which should be happening right at the threshold for the emergence of a giant component. In order for $\phi$ to have a finite solution it must be that $0 > \langle d^2 \rangle - 2\langle d \rangle$. Thus, if

$$\langle d^2 \rangle - 2\langle d \rangle > 0. \tag{4.8}$$

we end up with a giant component, and so the threshold is where $\langle d^2 \rangle = 2\langle d \rangle$.

In the case of a Poisson distribution $\langle d^2 \rangle = \langle d \rangle + \langle d \rangle^2$, and so the giant component emerges when $\langle d \rangle^2 > \langle d \rangle$, or $\langle d \rangle > 1$. Indeed the threshold for the existence of a giant component under the Poisson random network model is $t(n) = 1/n$, which corresponds to an average degree of 1.

In the case of a regular network, were the degree sequences have full weight on some degree $k$, if we solve for $\langle d^2 \rangle = 2\langle d \rangle$ we end up with $k^2 = 2k$ and so a threshold for a giant component at $k = 2$. Clearly, with $k = 1$ we just end up with a set of dyads (paired nodes) and no giant component.

In the case of a scale-free network, where the probability of degree $d$ is of the form $P_n(d) = cd^{-\gamma}$, we find that $\langle d^2 \rangle$ diverges, whenever $\gamma < 3$, and so there will generally be a giant component regardless of the specifics of the distribution.

In order to estimate the size of the giant component, the arguments underlying the derivation of (4.5) were not specific to a Poisson distribution, and so for the configuration model where the probability of loops is still negligible, we still have the approximation for the size of the giant component to be the largest $q$ that solves

$$1 - q = \sum_d (1 - q)^d P(d). \tag{4.9}$$

Using this expression, there is much that we can deduce about how the size of the giant component changes with the degree distribution. For instance, if we increase the distribution in terms of putting more weight on higher nodes (in the sense of first order stochastic dominance, and see Section 4.5.5 for definitions of stochastic dominance), then the right hand side expectation goes down for any value of $q$, and then the new value of $1 - q$ that solves (4.9) has to decrease as well, which corresponds to a larger giant component, as detailed in Exercise 4.11. This makes sense, since we can think of such a modification as effectively adding links to the network, which should increase the size of the giant component. Interestingly, providing a mean-preserving spread in the degree distribution has the opposite effect, decreasing the size of the giant component. This is a bit more subtle, but has to do with the fact that $(1 - q)^d$ is a convex function of $d$. So, spreading out the distribution leads to some higher degree nodes which have

a higher chance of being in a giant component, but also some lower degree nodes which have a much lower chance of being in the giant component. The key is that the convexity implies that there is more loss in probability from moving to lower degree nodes than gain in probability from the high degree nodes.

### 4.2.7 Diameter Estimation

Another important feature of a network is its diameter. This, as well as other related measures of distances between nodes, are important in understanding how quickly behavior or information can spread through a network, among other things.

To explore the diameter of a network, let us start by calculating the diameter of a network which makes such calculations relatively easy. Suppose that we examine a component that we know to be a tree so that there are no cycles. A method of obtaining an upper bound on diameter is to pick some node and then successively expanding its neighborhood by following paths of length $\ell$, where we increase $\ell$ until the paths are long enough so that we reach all nodes. Then we know that every node is at distance at most $\ell$ from our starting node and that no two nodes can be at a distance of more than $2\ell$ from each other; and so the diameter is bounded below by $\ell$ and above by $2\ell$.[27] What this diameter works out to be will depend on the shape of the tree.

Let us explore a particularly nicely behaved class of trees. Consider a tree such that every node either has degree $k$ or degree 1 (the "leaves"), and such that there is a "root" node that is equidistant from all of the leaves. Let us start from that "root" node.[28] If we then move out by a path of 1, we have reached $k$ nodes. Now, by traveling on all paths of length 2, we will have reached all of the nodes in the immediate neighborhoods of the nodes in the original node's neighborhood. We will have reached $k + k(k-1)$ or $k^2$ nodes. Extending this reasoning, by traveling on all paths of length $\ell$, we will have reached

$$k + k(k-1) + k(k-1)^2 + \ldots + k(k-1)^{\ell-1}$$

This can be rewritten (see the appendix) as

$$k\frac{(k-1)^\ell - 1}{k-1-1} \;=\; \left(\frac{k}{k-2}\right)\left((k-1)^\ell - 1\right).$$

[27] Note that we can easily see that both of these bounds could be reached. If the network is a "line" with an odd number of nodes and we do this calculation from the middle node then the diameter is exactly $2\ell$, while if we start at one of the end nodes, then the diameter is exactly $\ell$.

[28] Trees where all nodes have either degree $k$ or degree 1 are known as *Cayley trees*.

We can thus find the neighborhood size needed to reach all nodes by finding the smallest $\ell$ such that

$$\left(\frac{k}{k-2}\right)\left((k-1)^{\ell}-1\right) \geq n-1.$$

Approximating this equation provides us with a fairly accurate estimate of the neighborhood size needed to reach all nodes of

$$(k-1)^{\ell} = n-1,$$

or

$$\ell = \frac{log(n-1)}{log(k-1)},$$

and then the estimated diameter for this network is $2\frac{log(n-1)}{log(k-1)}$.

Newman, Watts and Strogatz [**?**] follow a reasoning similar to this to develop a very rough estimate of the diameter of more general sorts of random networks by examining the expansion in the neighborhoods. The calculation presumes a tree structure, which in the Poisson random network setting we know not to be valid beyond the threshold where the giant component emerges, and so it only gives us an order of magnitude approximation near the threshold. Generally, obtaining bounds on diameters is a very challenging problem. We will see a number of other points where there are potential problems with the calculation as we proceed.

A randomly picked node $i$ has an expected number of neighbors of $\langle d\rangle$ (recalling the $\langle\cdot\rangle$ notation for the expectation operator). If we presume that nodes' degrees are approximately independent, then each of these nodes has a degree described by the distribution $\widetilde{P}(d)$ from (4.2). Thus, each of these nodes has an expected number of neighbors (besides $i$) of $\sum_d (d-1)\widetilde{P}(d)$ or $\frac{\langle d^2\rangle - \langle d\rangle}{\langle d\rangle}$. So, the expected number of $i$'s second neighbors (who are at a distance of 2 from $i$) is *very* roughly $\langle d\rangle \frac{\langle d^2\rangle - \langle d\rangle}{\langle d\rangle}$.[29] Iterating, the expected number of $k$-th neighbors is estimated by

$$\langle d\rangle \left(\frac{\langle d^2\rangle - \langle d\rangle}{\langle d\rangle}\right)^{k-1}.$$

[29] So, we see several approximations. The tree structure is implicit in the assumption that each of these "second neighbors" are not already first neighbors. There is an assumption about the correlation in neighbors' degrees implicit in the use of $\widetilde{P}$ to calculate neighbors' degrees. Next, the expected number of second neighbors is found by multiplying first neighbors times expected number of their neighbors, again embodying some independence to have the expectation of a product equal the product of the expectations.

This means that expanding out to a $\ell$-th neighborhood reaches:

$$\sum_{k=1}^{\ell} \langle d \rangle \left( \frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle} \right)^{k-1} \tag{4.10}$$

nodes. When this sum is equal to $n-1$, then we have an idea of how far we need to go from a randomly picked node to hit all other nodes. This gives us a *very* rough estimate of the diameter of the largest component. Substituting for the sum of the series in (4.10) (see Section 4.5 for some facts about sums of series), we obtain an estimate of diameter as the $\ell$ that solves

$$\langle d \rangle \left( \frac{\left( \frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle} \right)^{\ell} - 1}{\left( \frac{\langle d^2 \rangle - \langle d \rangle}{\langle d \rangle} \right) - 1} \right) = n - 1,$$

or

$$\ell = \frac{\log\left[(n-1)\left(\langle d^2 \rangle - 2\langle d \rangle\right) + \langle d \rangle^2\right] - \log\left[\langle d \rangle^2\right]}{\log\left[\langle d^2 \rangle - \langle d \rangle\right] - \log\left[\langle d \rangle\right]} \tag{4.11}$$

In cases where $\langle d^2 \rangle$ is much larger than $\langle d \rangle$, this is approximately

$$\ell = \frac{\log[n] + log\left[\langle d^2 \rangle\right] - 2\log\left[\langle d \rangle\right]}{\log\left[\langle d^2 \rangle\right] - \log\left[\langle d \rangle\right]} = \frac{\log[n/\langle d \rangle]}{\log\left[\langle d^2 \rangle / \langle d \rangle\right]} + 1. \tag{4.12}$$

although here we have to be careful, as we are ignoring any cycles and when we are above the threshold for a giant component to exist and include many nodes (e.g., when $\langle d^2 \rangle$ is much larger than $2\langle d \rangle$), then there can be nontrivial clustering for some degree sequences.

If we examine (**??**) for the case of a Poisson random network, then $\langle d^2 \rangle / \langle d \rangle = 1 + \langle d \rangle = 1 + (n-1)p$, and then

$$\ell = \frac{\log\left((n-1)\frac{\langle d \rangle - 1}{\langle d \rangle} + 1\right)}{\log(\langle d \rangle)}.$$

In cases where $\langle d \rangle$ is subsantially above 1, this is roughly $\log(n)/\log(\rangle d\langle)$, which is very similar to the regular the tree example. If $p$ is held constant, then as we increase $n$, $\ell$ decreases and converges to 1 from above. In that case we would estimate the diameter to be 2. In fact, it can be shown that for a constant $p$, this crude approximation is right on the mark in that the diameter of a large random graph with such a $p$ is 2 with a probability tending to one (see Corollary 10.11 Bollobás [80]). Next, let us consider the case where average degree is not exploding, but instead average degree is

held constant so that $p(n-1) = \langle d \rangle > 1$. Then our estimate for diameter is on the order of $log(n)/log(\langle d \rangle)$. Here the estimate is not as accurate.[30] Applying this to the network generated in Figure 1.2.3 where $n = 50$, $p = .08$ and average degree is roughly $\langle d \rangle = 4$, we get an estimated diameter of 2.8. While this is not precise, it is not far off for the largest component in Figure 1.2.3.

Developing accurate estimates for diameters, even for such completely random networks, turns out to be a formidable task that has been an active area of study in graph theory for the past four decades.[31] Nevertheless, the above approximations reflect the fact that the diameter of a random network is likely to be "small" in the sense that it is significantly smaller than the number of nodes.

## 4.3 An Application: Contagion and Diffusion

To get a feeling for how some of the derivations from random networks might be useful, consider the following application. There is a society of $n$ individuals. One of them is initially infected with a contagious virus (possibly even a computer virus). Let the network of interactions in the society be described by a Poisson random network with link probability $p$.

The initially infected person interacts with each of his or her neighbors. Some of the neighbors are immune to the virus, while others are not. Let any given individual be immune with a probability $\pi$. For instance, this might represent a natural immunity, a percentage of people who have been vaccinated, or the percentage of people whose computers are not susceptible to the given disease. This is a variation on what has is known as the Reed-Frost model in the epidemiology literature (see Bailey [25], as the work of Reed and Frost was never published), and is is discussed in more detail in Section **??**.

The eventual spread of the disease can then be modeled by:

- generating a Poisson random network on $n$ nodes with link probability $p$,
- deleting $\pi n$ of the nodes (uniformly at random) and considering the remaining network,

[30] In this range of $p$, the network will generally have a giant component, but will most likely not even be completely connected.

[31] See Chapter 10 in Bollobás [80] for a report on some of the results and references to the literature.

- identifying the component that the initially infected individual lies in on this subnetwork.

This calculation is equivalent to examining a network on $(1-\pi)n$ nodes with a link probability of $p$, and then examining the size of the component containing a randomly chosen node. Thus, given that the threshold for the emergence of a giant component is at $p(1-\pi)n = 1$, then if $p(1-\pi)n < 1$, we expect the disease to die out and only infect a negligible fraction of the population. In contrast, if $p(1-\pi)n > 1$, then we there is a nontrivial probability that it will spread to a some fraction of the originally susceptible population. In particular, from (4.6) we know that for large $n$, if an agent in the giant component of the susceptible population is infected, then the expected size of the epidemic as a percentage of the nodes that are susceptible is approximated by the nonzero $q$ which solves

$$q = 1 - e^{-q(1-\pi)np}. \tag{4.13}$$

Furthermore, from Theorem 4.2.1 we also know that if $p > \frac{log((1-\pi)n)}{(1-\pi)n}$, then with a probability approaching 1 (as $n$ grows) the network of susceptible agents will be connected and so all of the susceptible population will be infected.

While the equation above is difficult to solve directly for $q$, we can rewrite the equation as

$$(1-\pi)np = \frac{\log(1-q)}{q}. \tag{4.14}$$

Then putting in different values of $q$, we find the corresponding levels of $(1-\pi)np$ that lead to those $q$'s. This leads to the following figure:

The figure provides us with an initial threshold of $(1-\pi)np = 1$, and then we see that nearly the entire population of susceptible individuals is connected once we approach $(1-\pi)np = 5$. So, for instance, if half the population is susceptible, and average degree is above 10, then nearly all of the susceptible agents are interconnected, and so the probability of them all becoming infected from a tiny initial seed is quite high.

It is also worth emphasizing that this can also capture diffusion of various behaviors. For instance, let susceptible indicate someone who would buy a certain product if made aware of it. Then $(1-\pi)$ can be interpreted as the percentage of the population who would buy the product if everyone was aware of it. The size of the giant component from these calculations indicates the potential impact of the reach of informing a few agents in the population about the product, when they communicate by word of mouth with others and are sure to learn about the product from any neighbor who buys it.

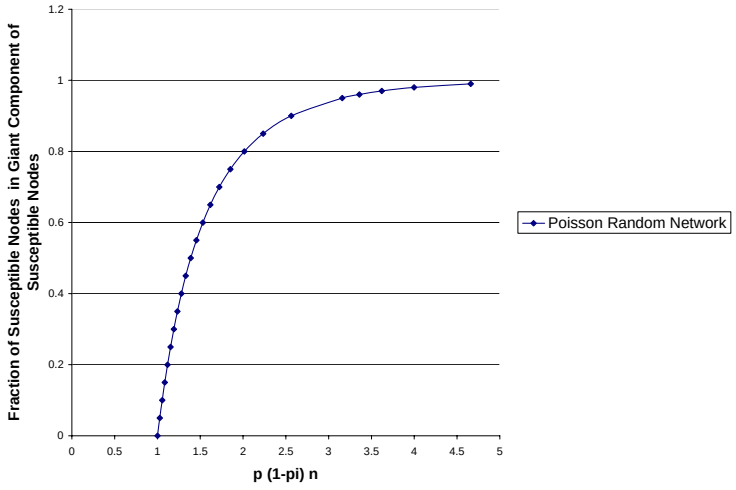



**Figure 4.3.** Fraction of the Susceptible Population in the Largest Component of a Poisson Random Network, as a Function of the Proportion of Susceptible Nodes (1-$\pi$) times the Link Probability, $p$, times the Population Size, $n$.

This analysis is built on contagion taking place with certainty between any infected and susceptible neighbors. When the transmission is probabilistic, which is the case in some applications, then the analysis needs to account for that. Such diffusion is discussed in greater detail in Chapter **??**.

### 4.3.1 Distribution of Component Sizes*

The derivations in Section 4.2.6 provide an idea of when a giant component will emerge, and its size, but we might be interested in more information about the distribution of component sizes that emerge in a network. Again, we will see how important this is when we examine network-based diffusion in more detail in Chapter **??**. Following Newman, Watts and Strogatz [**?**], we can use probability generating functions to examine the component structure in more detail. (For those not familiar with generating functions, it will be useful to read the appendix in Section 4.5.9 before preceding with this section.)

This analysis presumes that adjacent nodes have independent degrees, and so it is best to fix ideas with respect to the Configuration Model, where approximate independence holds for large $n$. Let the degree distribution be described by $P$.

Consider the following question. What is the size of the component of a node picked uniformly at random from the network? We find this by starting at a node, picking one of its edges and examining the neighboring node, and then following the edges from that neighboring node and seeing how many additional nodes we find. Then summing across edges leaving the initial node, we have an idea of the expected size of the component. This method presumes a tree structure, and is thus only a good approximation when the degree distribution is such that the number of cycles in the network is negligible.

So, let us first examine how many nodes we find when pick a link at random from the network, and then follow one of its nodes and count all of its further neighbors and so forth. In particular, let $Q$ denote the distribution of the number of nodes that can be found by picking an edge uniformly at random from the network, then picking one of its nodes uniformly at random, and then counting that node plus all of the nodes that are found by following all paths from that node that do not use the original link. Let $G_Q(x)$ denote the generating function associated with this distribution.

Note that $Q$ can be thought of in the following way. There is a probability of $\widetilde{P}(d)$ that the node at the end of the randomly selected edge will have degree $d$. In that case, it will have $d-1$ edges emanating from it. The number of additional nodes that can be

found by starting from each such edge is a random variable $Q$.[32] We now use some facts about generating functions to deduce the generating function of $Q$. In Section 4.5.9 (see (4.25)) it is shown that the generating function of the sum of $d-1$ independent draws from the distribution of $Q$ is the the generating function of $Q$ raised to the power $d-1$, so the generating function of additional nodes found through the node if it happens to have degree $d$ is $[G_Q(x)]^{d-1}$. The overall distribution of the number of nodes found through the additional node, is then given by a mixture of distributions in the following sense, firt pick some random $d$ according to $\widetilde{P}(d)$, and then draw a random variable from a distribution having generating function $[G_Q(x)]^{d-1}$ (see (4.26)). So, the generating function of the distribution of the additional nodes found past the first one is $\sum_d \widetilde{P}(d)\,[G(q)]^{d-1}$. Finally, we need to add one node for the first one found, and the generating function of a distribution of a random variable plus one is just $x$ times the generating function of the random variable discussed above (see (4.27)). So, the distribution function of the number of nodes found from one side of an edge picked uniformly at random is

$$G_Q(x) = x \sum_d \widetilde{P}(d)[G_Q(x)]^{d-1}.$$

Noting that $G_{\widetilde{P}}\left(G_Q(x)\right) = \sum_d \widetilde{P}(d)[G_Q(x)]^d$, we rewrite the above as[33]

$$G_Q(x) = x\frac{G_{\widetilde{P}}\left(G_Q(x)\right)}{G_Q(x)}.$$

or

$$G_Q(x) = \left(xG_{\widetilde{P}}\left(G_Q(x)\right)\right)^{1/2}. \tag{4.15}$$

As Newman, Strogatz, and Watts [**?**] point out, finding a solution to (4.15) is in general impossible with knowing something more about the structure of $P$. However, we can solve for the expectation of $Q$, as that is simply $G'_Q(1)$.

$$G'_Q(x) = \frac{1}{2}\left(xG_{\widetilde{P}}\left(G_Q(x)\right)\right)^{-1/2}\left(G_{\widetilde{P}}\left(G_Q(x)\right) + xG'_{\widetilde{P}}\left(G_Q(x)\right)G'_Q(x)\right).$$

Thus, recalling that $G(1) = 1$ for any generating function we find that:

$$G'_Q(1) = \frac{1}{2}\left(1 + G'_{\widetilde{P}}\left(1\right)G'_Q(1)\right). \tag{4.16}$$

[32] Here, we need to be working with a large number of nodes to have this be an accurate approximation, as otherwise with a small $n$ we are working with fewer potential nodes to explore.

[33] This appears different from (25) in Newman, Strogatz, and Watts [**?**], but in fact $\frac{G_{\widetilde{P}}(\cdot)}{G_Q(x)}$ is the same as their $G_1$, and allows for an easy derivation of (4.15).

Then since $G'_{\widetilde{P}}(1) = E_{\widetilde{P}}[d] = \frac{\langle d^2 \rangle}{\langle d \rangle}$ it follows from (4.16) that in cases where the expectation of $Q$ does not diverge it must be that

$$G'_Q(1) = \frac{1}{2 - \frac{\langle d^2 \rangle}{\langle d \rangle}}. \tag{4.17}$$

If $\langle d^2 \rangle \geq 2\langle d \rangle$, then the expectation of $Q$ diverges and so (4.17) is no longer valid, and indeed we see that this expression grows as $\langle d^2 \rangle$ approaches $2\langle d \rangle$. This is consistent with our earlier calculation in (4.7).

Now, we can calculate the average size of a component. Let $H$ be the distribution of the size of the component of a node picked uniformly at random. Then, starting from a node picked uniformly at random, the degree is governed by $P(d)$, the extended neighborhood size has generating function $[G_Q(x)]^d$, and we have to account for the initial node as well.[34] Thus, the generating function for $H$ is:

$$G_H(x) = x \sum_d P(d)[G_Q(x)]^d = xG_P\left(G_Q(x)\right). \tag{4.18}$$

Thus, the average size of the component that a randomly selected node lies in is (in situations where the average under $Q$ does not diverge):

$$G'_H(1) = 1 + G'_P(1)G'_Q(1) = 1 + \frac{\langle d \rangle^2}{2\langle d \rangle - \langle d^2 \rangle}. \tag{4.19}$$

If we examine Poisson random networks, we know that it must be that $\langle d \rangle = (n-1)p < 1$ in order for the network not to become connected, and so this must also hold in order to have the average size of the component not diverge. Indeed, in the Poisson random network model substituting for $\langle d \rangle = (n-1)p$, and $\langle d^2 \rangle = \langle d \rangle^2 + \langle d \rangle$, we find that average component size of a node picked uniformly at random is

$$G'_H(1) = 1 + \frac{1}{1 - (n-1)p}.$$

So, for instance, if $(n-1)p = 1/2$ then the average component size is 3, if $(n-1)p = 9/10$ then the average is 11.

If we examine scale-free networks, then $\langle d^2 \rangle$ is generally large relative to $\langle d \rangle$, and diverges as we let $n$ grow. In that case, the expected component size diverges. The

[34] The derivation of the distribution for $Q$ was based on randomly picking a node at either end of an edge. Here, we are working out from a given node, but given (approximate) independence in degrees, the calculation is ok.

intuition behind this is as follows. Even though average degree might be low, when we examine a given node, if it has a neighbor, that neighbor is very likely to have a large degree (as $\widetilde{P}(d) = P(d)d/\langle d\rangle$, which in a scale free network places very high weight on the highest degree nodes), and then it is even more likely to have additional high degree neighbors, and so forth.

## 4.4 Exercises

EXERCISE **4.1** *Self and Multiple Links in the Configuration Model**

Show that in the Configuration model that if $\max_{i\leq n} Q_i^n \to 0$, then the fraction of nodes that experience self- or multi-links is vanishing as the population size $n$ grows; or more specifically, for any $\varepsilon > 0$, $\Pr\left[\#\{i \leq n : q_i^n > 0\}/n > \varepsilon\right] < \varepsilon$ for large enough $n$.

EXERCISE **4.2** *A Degree Sequence that always has Large Nodes.*

Consider the degree sequence (1,1,2,4,8,16,...). Show that in the Configuration model any fixed node has a probability of having any self-link or multiple links going to 0 as $n$ becomes large, but for each $n \geq 2$ there is some node with a significant probability of having a self-link or multiple links. That is, show that $Q_i^n \to 0$; but that $Q_n^n \to 1$ for all $n$.

EXERCISE **4.3** *A Degree Sequence for the Power Distribution in the Configuration Model*

Find a degree sequence that converges to a power distribution and has $\frac{\widehat{d^n}}{(n\langle d\rangle)^{1/3}}$ tend to 0.

EXERCISE **4.4** *The Distribution of Neighbors' Degrees in the Configuration Model and Expected Degree Model*

Consider a constant degree sequence $(d, d, d \ldots)$. Form a random network by operating the configuration model and form another random network by operating the expected degree model. Provide an expression for the resulting degree distributions in the limit as $n$ grows (working with a resulting multigraph in the configuration model). Provide an expression for the limiting distribution $\widetilde{P}$ of the degree of a node found at either end of a uniformly randomly chosen link.

EXERCISE **4.5** *The Distribution of Neighbors' Degrees*

Consider the Poisson random network model on $n$ nodes with a link probability of $p$. Consider a node $i$ and a node $j$, which are fixed in advance. Conditional on the link $ij$ being present, what is the distribution of $j$'s degree?

Consider a node $i$. Conditional on it having at least one link, randomly pick one of its neighbors (with equal probability on each neighbor). Argue that the conditional distribution of the node's degree is different from the conditional distribution for $j$'s degree that you found above. What does this distribution converge to as we let $n$ grow if $p$ is set to keep average degree constant (so that $p = m/(n-1)$ for some fixed $m > 0$)?

Explain the difference between these two distributions.

EXERCISE **4.6** *A Threshold for Links in the Poisson Random Network Model*

Show that $t(n) = 1/n^2$ is a threshold function for there being at least one link a network relative to the Poisson random network model.

EXERCISE **4.7** *There is at Most One Giant Component in the Poisson Random Network Model**

Consider the Poisson random network model when $p$ (as a function of $n$) is such that there exists $m > 0$ such that $pn \geq m$ for all $n$. Show that the probability of having more than one giant component vanishes as $n$ grows.

EXERCISE **4.8** *The size of the Giant Component.*

Show that there is a solution to (4.9) of $q = 1$ if and only if $P(0) = 0$.

Find a nonzero solution to (4.9) when $P(0) = 1/3$ and $P(2) = 2/3$.

EXERCISE **4.9** *Estimating the Extent of an Infection in an Exponential Random Network Model**

Consider a degree distribution given by $P(d) = \frac{e^{\frac{-d}{(1-\pi)m}+1}}{m}$ with support from $(1-\pi)m$ to $\infty$, which has a mean of $2(1-\pi)m$ (which is derived in Section **??** as the distribution corresponding to a uniformly random network where the number of nodes grows over time). Use (4.9) to estimate the percent of susceptible nodes that will be infected when a random selection $\pi$ of nodes are immune. Hint: See the Section 4.5 for helpful formulas of sums of series.

EXERCISE **4.10** *Estimating the Diameter in an Exponential Random Network Model*

Consider a degree distribution given by $P(d) = \frac{e^{\frac{-d}{m}+1}}{m}$ with support from $m$ to $\infty$, Use (**??**) to estimate the diameter.

EXERCISE **4.11** *First Order Stochastic Dominance and Increasing Giant Components*

Consider two degree distributions $\widehat{P}$ and $P$, such that $P$ first order stochastically dominates $\widehat{P}$ (see Section 4.5.5 of the appendix if this definition is unfamiliar). Show that if $q'$ and $q$ are interior solutions to (4.9) relative to $\widehat{P}$ and $P$, respectively, then $q \geq q'$.[35]

If $\widehat{P}$ is a mean-preserving spread of $P$ and $q'$ and $q$ are interior solutions to (4.9) relative to $\widehat{P}$ and $P$, respectively, how are $q'$ and $q$ ordered?

EXERCISE **4.12** *Mean Preserving Spreads and Decreasing Diameters*

Consider two degree distributions $\widehat{P}$ and $P$, such that $\widehat{P}$ is a mean preserving spread of $P$ (see Section 4.5.5 of the appendix if this definition is unfamiliar). Show that the solution to (**??**) under $\widehat{P}$ is lower than that under $P$. Show that if we change "is a mean preserving spread of" to "first order stochastically dominates" then we cannot order the solutions to (**??**).

EXERCISE **4.13** *First Order Stochastic Dominance and Decreasing Diameters**

Consider two degree finite degree sequences in the expected degree model of Section **??**, with corresponding distributions $\widehat{P}$ and $P$, such that $P$ first order stochastically dominates $\widehat{P}$. Show that the diameters of the random networks associated with $\widehat{P}$ have higher diameters in the sense of first order stochastic dominance of the realized network diameters compare to those associated with $P$.

EXERCISE **4.14** *Component Sizes for A Family of Degree Disributions**

Calculate $\langle d^2 \rangle$ under the degree distribution that has a distribution function of

$$F(d) = 1 - (rm)^{1+r} (d + rm)^{-(1+r)},$$

[35] To offer a complete answer to this, note that (4.9) can be written as a function $1 - q = H(1 - q)$, where you can show that $H(\cdot)$ is increasing and strictly convex and is such that $H(1) = 1$. Thus, you can show that it has at most one solution other than $q = 0$. Drawing a picture will help.

from (3.2), using this continuous distribution as an approximation for distributions with large $n$.

Show that $\langle d^2 \rangle$ diverges when $r < 1$. Use the expression for $\langle d^2 \rangle$ and (4.19) to estimate the expected component size in large networks with such a degree distribution when $r > 1$ and for $m = \langle d \rangle$ such that $\langle d^2 \rangle < 2\langle d \rangle$.

## 4.5 Appendix: Useful Facts, Tools, and Theorems

This appendix contains a few mathematical definitions, formulas, theorems, and approximations that are useful in working with random networks.

### 4.5.1 Sums of Series

A geometric series is one where we sum a series of powers of $x$ where $x \neq 1$:

$$\sum_{i=m}^{n} ax^i = a\frac{x^m - x^{n+1}}{1-x}.$$

Thus,

$$\sum_{i=0}^{n} ax^i = a\frac{1 - x^{n+1}}{1-x}.$$

For $x < 1$ it follows that

$$\sum_{i=1}^{\infty} ax^i = \frac{ax}{1-x}$$

and

$$\sum_{i=0}^{\infty} ax^i = \frac{a}{1-x}.$$

Another series of interest (especially for scale-free degree distributions) is

$$\sum_{i=1}^{\infty} a\frac{1}{i^\gamma}.$$

This is the Riemann Zeta Function, $z(\gamma) = \sum_1^\infty \frac{1}{\gamma}$, which is convergent whenever $\gamma$ is greater than 1.

A special case of this occurs if $\gamma = 1$, when we can look at a truncated series

$$\sum_{i=1}^{n} \frac{1}{i} = H_n \tag{4.20}$$

is known as a *harmonic number* and has various approximations. For large $n$, an approximation of $H_n$ is $\gamma + log(n)$, where the $\gamma$ of roughly .577 is the Euler-Mascheroni constant; and the difference between this approximation and $H_n$ tends to 0. This is useful in approximating some sequences such as

$$\frac{1}{i+1} + \frac{1}{i+2} + \cdots + \frac{1}{t}, \tag{4.21}$$

which can be written as $H_t - H_i$. For large $t$, this is approximately $log(t) - log(i)$, or $log\left(\frac{t}{i}\right)$.

### 4.5.2 e and Stirling's Formula

The exponential function can be defined in various ways that provides useful formulas. Fixing $x$ (at any positive, negative or complex value)

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x.$$

Another definition of $e$ is given by

$$e^x = \sum_{i=0}^{\infty} \frac{x^n}{n!}.$$

Stirling's formula for large $n$ is that

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

### 4.5.3 Chebyshev's Inequality and the Law of Large Numbers

Chebyshev's inequality says that for a random variable $X$ with mean $\mu$ and standard deviation $\sigma$,

$$\Pr[|X - \mu| > r\sigma] < 1/r^2$$

for every $r > 0$. This is easy to prove directly from the definition of standard deviation. Letting $r = \frac{x}{\sigma}$, we can also write this as

$$\Pr[|X - \mu| > x] < \sigma^2/x^2$$

for every $x > 0$.

Chebyshev's inequality leads to a very easy proof of a version of the Weak Law of Large Numbers:

THEOREM **4.5.1** *[The Weak Law of Large Numbers] Let $(X_1, X_2, \ldots)$ be a sequence of independently distributed random variables such that $E[X_i] = \mu$ for all $i$ and there is a finite bound $B$ so that $Var(X_i) \leq B$ for all $i$. Then*

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| > \varepsilon\right] \to_n 0$$

*for all $\varepsilon > 0$.*

**Proof of Theorem 4.5.1:** Let $S_n = \sum_{i=1}^{n} \frac{X_i}{n}$. Then $Var(S_n) = \sum_i \frac{Var(X_i)}{n^2} \leq \frac{B}{n}$. Thus, $Var(S_n) \to 0$. By Chebyshev's inequality, fixing any $\varepsilon > 0$

$$\Pr\left[\left|\sum_{i=1}^{n} \frac{X_i}{n} - \mu\right| > \varepsilon\right] \leq \frac{Var(S_n)}{\varepsilon^2} \to 0,$$

which establishes the claim. ∎

There is also a stronger conclusion that is possible. The weak law of large numbers just states that the probability that a sequence of observed sample means deviates from the true mean of the process tends to 0. This does not directly imply that there is a probability 1 that the sequence will converge. The strong law of large numbers provides this stronger conclusion. For a proof, see Billingsley [64].

THEOREM **4.5.2** *[The Strong Law of Large Numbers] Let $(X_1, X_2, \ldots)$ be a sequence of independently and identically distributed random variables such that $E[X_i] = \mu$ for all $i$. Then*

$$\Pr\left[\lim \frac{\sum_{i=1}^{n} X_i}{n} = \mu\right] = 1.$$

### 4.5.4 The Binomial Distribution

There are many situations where we need to make use of the binomial distribution, and this also provides us with an illustration of the law of large numbers.

Consider flipping a coin repeatedly, but in a situation where the coin is not a "fair" coin, but instead has a probability of $p$ of coming up heads and a probability of $1 - p$ of coming up tails. A single flip is called a "Bernoulli trial." In many instances we are interested in a whole set of flips. If we ask the probability of a particular sequence of heads and tails being realized, say heads, tails, tails, heads, heads, ..., where there are $m$ heads out of $n$ flips, then its probability is $p^m(1-p)^{n-m}$. Noting that there are $\binom{n}{m}$

(read "$n$ choose $m$," where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$) different orderings that have $m$ heads out of $n$ flips of the coin, the probability that there are $m$ heads if we flip it $n$ times is

$$\binom{n}{m} p^m (1-p)^{n-m}.$$

The expected number of heads out of $n$ flips is simply $pn$, while the standard deviation is $\sqrt{np(1-p)}$.

Note also that the expected fraction of flips of the coin that come up heads is simply $p$, and the standard deviation of this fraction out of $n$ flips is $\sqrt{\frac{p(1-p)}{n}}$.

Then applying Chebyshev's inequality, letting $X$ be the realized fraction of flips that come up heads,

$$\Pr\left[|X-p| > r\sqrt{\frac{p(1-p)}{n}}\right] < \frac{1}{r^2}.$$

So, if we let $r = n^{1/4}$, then we see that

$$\Pr\left[|X-p| > \frac{\sqrt{p(1-p)}}{n^{1/4}}\right] < \frac{1}{n^{1/2}},$$

and so with a large number of flips of the coin it is very unlikely that the realized fraction of heads will differ from $p$ by very much, just as the law of large numbers tells us.

### 4.5.5 Stochastic Dominance and Mean-Preserving Spreads

Consider discrete distributions $\widehat{P}$ and $P$ with support on $\{0, 1, 2, \dots\}$.[36]

The concept of first order stochastic dominance captures the idea that $P$ is obtained by shifting mass from $\widehat{P}$ to place it on higher values. The following are equivalent

- $\sum f(d)P(d) \geq \sum f(d)\widehat{P}(d)$ for all nondecreasing functions $f$,
- $\sum_0^x P(d) \leq \sum_0^x \widehat{P}(d)$ for all $x$,
- $\sum_x^\infty P(d) \geq \sum_x^\infty \widehat{P}(d)$ for all $x$

and if they hold we say that

*$P$ first order stochastically dominates $\widehat{P}$.*

[36] The extension of these definitions is straightforward to the case of more general probability measures, simply substituting $\int \cdot dP$ in the place of sums with respect to $P$.

We say that the dominance is strict if the inequalities above hold strictly for some $x$ (or $f$). Note that if strict dominance holds, then it must be that $\sum f(d)P(d) > \sum f(d)\widehat{P}(d)$ for any strictly increasing $f$.

An example of a degree distribution that (strictly) first order stochastic dominates another is pictured in Figure 7.2.5.

The last two items above are clearly equivalent and capture the idea that $P$ places less weight on low values, and thus more weight on higher values than $\widehat{P}$. The idea that stochastic dominance provides higher expectations for all nondecreasing functions is not difficult to prove as $P$ is shifting weight to higher values of the function $f$, and the converse is easily seen using the last item above and a simple step function that has value 0 up to $x$ and then 1 from $x$ onwards.

Often when people refer to first order stochastic dominance, the "first order" is omitted and it is simply said that $P$ stochastically dominates $\widehat{P}$.

The idea of second order stochastic dominance is a less demanding relationship than first order stochastic dominance and so it orders more pairs of distributions. It is implied by first order stochastic dominance. Instead of requiring a higher expectation relative to all nondecreasing functions, it only requires a higher expectation relative to all nondecreasing functions that are also concave. This has deep roots in foundations of decision making and risk aversion, although for us it quite useful in comparing degree distributions of different networks.

THEOREM **4.5.3 (Rothschild and Stiglitz [541])** *The following are equivalent*

- $\sum f(d)P(d) \geq \sum f(d)\widehat{P}(d)$ *for all nondecreasing, concave functions* $f$*,*
- $\sum f(d)P(d) \leq \sum f(d)\widehat{P}(d)$ *for all nonincreasing, convex functions* $f$*,*
- $\sum_{z=0}^{x}\sum_{d=0}^{z} P(d) \leq \sum_{z=0}^{x}\sum_{d=0}^{z} \widehat{P}(d)$ *for all* $x$*,*

*and when they hold we say that* $P$ second order stochastically dominates $\widehat{P}$*. If* $P$ *and* $\widehat{P}$ *have the same mean then the above are also equivalent to*

- $\widehat{P}$ *is a* mean-preserving spread *of* $P$*,*[37]
- $\sum f(d)P(d) \geq \sum f(d)\widehat{P}(d)$ *for all concave* $f$*.*

[37]This indicates that the random variable described by $\widehat{P}$ can be written as the random variable described by $P$ plus a zero mean random variable.

Again, the dominance (or mean-preserving spread) is strict if the inequalities listed above hold strictly for some $f$ (or $x$). In that case, $\sum f(d)P(d) > \sum f(d)\widehat{P}(d)$ for any strictly increasing and strictly concave functions $f$.

So, if $P$ and $\widehat{P}$ have the same average, then $P$ second order stochastically dominates $\widehat{P}$ if and only if $\widehat{P}$ is a mean preserving spread of $P$. This implies that $\widehat{P}$ has a (weakly) higher variance than $P$, but also requires a more structured relationship between the two. Having a higher variance and identical mean is not sufficient for one distribution to be a mean preserving spread of another.

### 4.5.6 Domination

There are also definitions of domination for distributions on several dimensions.

Consider two probability distributions $\mu$ and $\nu$ on $\mathbb{R}^n$.

$\mu$ *dominates* $\nu$ if

$$E_\mu[f] \geq E_\nu[f]$$

for every non-decreasing function $f : \mathbb{R}^n \to \mathbb{R}$. The domination is *strict* if strict inequality holds for some non-decreasing $f$.

Domination captures the idea that "higher" realizations are more likely under $\mu$ than under $\nu$. In the case where $n = 1$, domination reduces to first order stochastic dominance.

### 4.5.7 Association

Beyond comparing two different distributions, we will also be interested in knowing when it is that a joint distribution of a set of random variables exhibits relationships between the variables. Concepts like correlation and covariance can address two random variables, but when working with networks we will often work with groups of variables at the same time. A notion that captures such relationships is association, a definition due to Esary, Proschan, and Walkup [**?**].

Let $\mu$ be a joint probability distribution describing a random vector $S = (S_1, \ldots S_n)$, where each $S_i$ is real-valued.

$\mu$ is *associated* if

$$\mathrm{Cov}_\mu(f, g) = E_\mu[f(\mathbf{S})g(\mathbf{S})] - E_\mu[f(\mathbf{S})]E_\mu[g(\mathbf{S})] \geq 0,$$

for all pairs of non-decreasing functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$.

If $S_1, \ldots, S_n$ are the random variables described by a measure $\mu$ that is associated, then we say that $S_1, \ldots, S_n$ are associated.

Association of $\mu$ implies that $S_i$ and $S_j$ are non-negatively correlated for any $i$ and $j$, and it entails that all dimensions of $\mathbf{S}$ are non-negatively interrelated.[38]

To establish strictly positive relationships, as opposed to non-negative ones, Calvó-Armengol and Jackson [119] define a strong version of association.

A partition $\Pi$ of $\{1, \ldots, n\}$ captures which random variables are positively related (for instance, the components of nodes in a network).

A probability distribution $\mu$ describing the random variables $(S_1, \ldots S_n)$ is *strongly associated* relative to the partition $\Pi$ if it is associated, and for any $\pi \in \Pi$ and nondecreasing functions $f$ and $g$

$$\mathrm{Cov}_\mu(f, g) > 0$$

whenever $f$ is increasing in $s_i$ for all $s_{-i}$, $g$ is increasing in $s_j$ for all $s_{-j}$, and $i$ and $j$ are in $\pi$.

An implication of strong association is that $S_i$ and $S_j$ are positively correlated for any $i$ and $j$ in $\pi$.

### 4.5.8 Markov Chains

There are many settings where one considers a random process over time, the world can be described by a state, and the transition from one state to another depends only on the current state of the system and not how we got there. For the applications in this book, we will mainly be concerned with finite-state systems. For instance, states could be the network that is presently in the society. Alternatively, a state might instead describe something that the agents in a network are doing.

Let the finite set of states be denoted $S$. If the state of the system is $s_t = s$ at time $t$, then there is a well-defined probability that the system will be in state $s_{t+1} = s'$ at time $t + 1$. What is critical is that are well-defined probabilities of being in each state tomorrow as a function only of the state today. Let $\Pi$ be the $n \times n$ matrix describing these *transition probabilities* with entries

$$\Pi_{ss'} = \Pr(s_{t+1} = s' \mid s_t = s).$$

[38] This is a weaker concept than "affiliation," which requires association for when conditioning on various events. The weaker concept is useful in many network settings where the states of nodes of a network will be associated, but are not affiliated. See Calvó-Armengol and Jackson [119] for an example and discussion.

This results in what is known as a (finite state) Markov chain, where "Markov" refers to the property that the ditribution of what will happen in the future of the system only depends on the current state, and not how we got to the current state. Markov chains have a number of applications and very nice properties, so that they have been studied extensively.

There are some basic facts about Markov chains that are quite useful.

The Markov chain is said to be *irreducible* when for any two states $s'$ and $s$, if the system starts in state $s'$ in some period, then there is a positive probability that it will reach $s$ at some future date.

Irreducibility corresponds to the strong connectedness of sthe associated directed graph where the nodes are the states and $s$ points to $s'$ if $\Pi_{s's} > 0$.

An irreducible Markov chain is *aperiodic* if the greatest common divisor of its cycle lengths is one, where the cycles are in the associated directed graph just described. Checking whether a system is aperiodic is equivalent to asking the following. Start in some state $s$ at time 0 and list all the future dates where there is a positive probability of being in this state again. If the answer for a state $s$ a list of dates with a greatest common divisor greater than 1 then that state is said to be periodic. If no state is periodic, then the Markov chain is aperiodic.[39]

Noting that the probability of starting in state $s$ and ending in state $s'$ in two periods is simply $\Pi^2$, we see by similar reasoning that the probability of starting in state $s$ and ending in state $s'$ in $t$ periods is $\Pi^t$. If $\Pi^t$ has all entries greater than 0 for some $t$, then it is clearly both irreducible and aperiodic as it will then have all positive entries for all times thereafter. In contrast, if it never has all positive entries, then it either fails to be irreducible, or it is periodic for some states.

An important theorem about Markov chains states that an irreducible and aperiodic finite-state Markov chain has what is known as a *steady-state* distribution (e.g., see Billingsley [64]). The steady-state of the Markov process is described by a vector $\mu$ with dimension equal to the number of states, where $\mu_s$ is the probability of state $s$. The steady-state condition is that if the process is started at time 0 by randomly drawing the state according to the steady state distribution, then the distribution over

[39] For those who want to be sure to master all of the definitions, verify that if a Markov chain has a finite number of states and is irreducible, then one state is periodic if and only if all states are periodic, and in that case they all have the same period (greatest common divisor of dates at which they have a probability of recurring).

the state at time 1 will be given by the same distribution. That is,

$$\mu_{s'} = \sum_s \mu_s \Pi_{ss'},$$

or

$$\mu = \mu\Pi.$$

We can find the steady-state distribution as a left-hand unit eigenvector, noting that $\Pi$ is a row-stochastic matrix (that is, the elements of each row sum to 1).

Other useful facts about the steady-state distribution of a finite-state, irreducible, and aperiodic Markov chain include that it provides the long-run limiting average fraction of periods that the process will spend in each state regardless of the starting state; and regardless of where we start the system the probability of being in state $s$ at time $t$ as $t$ grows goes to $\mu_s$.

Thus, in situations where behavior can be described by a Markov chain, we have sharp predictions about behavior over the long run.

### 4.5.9 Generating Functions

Generating functions (also known as probability generating functions[40]) are useful tools for encapsulating the information about a discrete probability distribution and also for calculating moments of the distribution and various other statistics associated with the distribution.

Let $\pi(\cdot)$ be a discrete probability distribution, which for our purposes has support in $\{0, 1, 2 \ldots\}$. The *generating function* associated with $\pi$, denoted $G_\pi$, is defined by

$$G_\pi(x) = \sum_{k=0}^{\infty} \pi(k) x^k = E_\pi \left[x^k\right]. \tag{4.22}$$

Note that since $\pi(\cdot)$ is a probability distribution, $G_\pi(1) = 1$.

Moreover, $G_\pi$ has a number of useful properties. Taking various derivatives of it helps us to recover the various expectations with respect to $\pi$.

$$G'_\pi(x) = \sum_{k=0}^{\infty} \pi(k) k x^{k-1}. \tag{4.23}$$

[40] These are distinct from moment generating functions, which are defined by $\sum_{k=0}^{\infty} \pi(k) e^{xk} = E_\pi \left[e^{xk}\right]$.

Thus,

$$G'_\pi(1) = \sum_{k=1}^{\infty} \pi(k)k = E_\pi[k] = \langle k \rangle.$$

More generally,[41]

$$\left(x\frac{d}{dx}\right)^m G_\pi = \sum_{k=1}^{\infty} \pi(k)k^m x^k, \tag{4.24}$$

and so

$$E\left[k^m\right] = \langle k^m \rangle = \left(x\frac{d}{dx}\right)^m G_\pi \mid_{x=1}$$

Next, suppose that we consider two independent draws of the random variable $k$ and we want to know the sum of them. The probability that the sum is $k$ is given by $\sum_{i=0}^{k} \pi(i)\pi(k-i)$. This new distribution of the sum, denoted $\pi_2$, is then such that $\pi_2(k) = \sum_{k=0}^{k} \pi(i)\pi(k-i)$. It has an associated generating function

$$G_{\pi_2}(x) = \sum_{k=0}^{\infty} \pi_2(k)x^k = \sum_{k=0}^{\infty}\sum_{i=0}^{k} \pi(i)\pi(k-i)x^k.$$

Note that

$$[G_\pi(x)]^2 = \left[\sum_{k=0}^{\infty} \pi(k)x^k\right]^2 = \sum_{i,j} \pi(i)\pi(j)x^{i+j} = G_{\pi_2}(x).$$

This extends easily to higher powers (simply iterating gives even powers) and so the generating function associated with the distribution $\pi_m$ of a sum of $m$ independent draws of $k$ from $\pi$ is given by

$$G_{\pi_m}(x) = [G_\pi(x)]^m. \tag{4.25}$$

Another useful observation is the following. Consider a distribution $\pi$ which is derived by first randomly picking a distribution from a series of distributions $\pi_1, \pi_2, \ldots, \pi_i, \ldots$, picking each with corresponding probability $\gamma_k$, and then drawing from the chosen distribution. Then it follows almost directly that

$$G_\pi = \sum_i \gamma_i G_{\pi_i}. \tag{4.26}$$

[41] The notation $\left(x\frac{d}{dx}\right)^m G_\pi$ indicates taking the derivative of $G_\pi$ with respect to $x$ and then multiplying the result by $x$, and then taking the derivative of the new expression and multiplying it by $x$, and so forth for $m$ iterations.

Finally, there are many situations where we have a variable $k$ with distribution $P$ and we want to work with the distribution of $k+1$. The distribution $\overline{P}$ of $k+1$ is described by $\overline{P}(k) = P(k-1)$, where $k \geq 1$. Thus, it has a generating function of

$$G_{\overline{P}}(x) = \sum_{k=1}^{\infty} P(k-1)x^k = xG_P(x). \tag{4.27}$$

To use generating functions in the context of degree distributions, let us begin with a degree distribution $P$. Let it have an associated generating function $G_P$, defined as under (4.22). Suppose that we are also interested in the generating function $G_{\widetilde{P}}$ associated with the distribution of neighboring degrees under the configuration model, denoted $\widetilde{P}$. Recalling from (**??**) that $\widetilde{P}(d) = \frac{d\ P(d)}{\langle d \rangle}$, it follows that

$$G_{\widetilde{P}}(x) = \sum_{k=0}^{\infty} \widetilde{P}(d)x^d = \sum_{k=0}^{\infty} \frac{P(d)d}{\langle d \rangle} x^d = \frac{xG'_P(x)}{G'_P(1)}. \tag{4.28}$$