Topics on Social Networks

Aris Anagnostopoulos

March 6, 2024

Chapter 1

Introduction

A social network is a network where nodes correspond to individuals (usually, although they may be different entities such as animals, groups, companies, etc.) and links indicate some relationship between the individuals. The relationship can be one-directional or multi-directional.

There are offline social networks, such as the network of friendships, the network of actors, networks of professionals for a given professions and so on. Furthermore, we have online social networks such as facebook or instant messaging software. We will elaborate on both of them.

It is not hard to realize the importance played by social networks. When we search for a job, our professional contacts are one of the most defining factors (sometimes the most important) for the type of interviews and even job that we will get. The friends that we have directly impact our lifestyle; as the old saying goes, "show me your friends and I'll show you who you are." For example, probably the most important factor for teenager smoking is the smoking by their peers. Viruses spread through social networks, therefore if we wish to prevent large-scale epidemics we should understand the behavior of spreading through the social networks. The examples demonstrating the ubiquitousness of social networks and the importance of their analysis are countless.

Social networks have many structural resemblances to other types of networks that scientists have studied such as the web graph, computer networks, network of citations, or biological networks. We call all these types of networks *complex networks*. While here we are mostly interested in social networks, we will also study and review results that were obtained for various types of complex networks; a lot of the empirical observations on those complex networks hold for social networks as well and some of the models developed for them are also good models for social networks. In Chapter 3 we will see some of the characteristics shared among all types of complex networks.

1.1 Offline Social Networks

We all are parts of several offline social networks and they impact a lot our lives. Our friends form the set of neighbors in the friendship social network. The structure of the network around us is an indication of what type of lifestyle we have. The same holds for the network of sexual partners, where a link between two persons indicates that they had a sexual relationship.

The network of professional contacts is another example of a very important network. When searching for a job our contacts, as well as the contacts of our contacts can have a tremendous effect in the types of interviews or even job offers that we have.

Another network that we can consider is the network of contacts in a given profession. For example, lawyers form a social network and someone's contacts in that network are those that will recommend her for a particular case. The same is true for doctors. In addition, the social

network corresponding to a profession can affect the behavior of the professionals, for example it may impact the adoption of new techniques or the extent to which they will obey to the governmental policies [].

A link on a social network can also mean that a person has communicated with someone. For example we can think of the network of people and of telephone calls where a link from a user to another exists if the first one has called the latter in a given time period.

Another type of relationship that a link can represent is the participation of two entities in the same action. For example we can define the network of collaborations of scientists, where two scientists are connected if they have coauthored a work together. Or the network of actors who are connected if they have participated in the same film. Similarly for the network of musicians who are connected if they have been part of the same band. Or the network of executives who are connected if they have been part of the same company board. All these are examples of social networks that scientists have studied and it is those networks that have given rise to the "Erdős" and the "Kevin Bacon" numbers.[1]

1.2 Online Social Networks

In the past few years we have been witnesses of what could be called an online social revolution. There is a large number of online social networks to which we belong and become increasingly more important to our lives.

Some of them are explicit. As of March 2010 facebook has become the number one social network with more than 400 million users. It started as a network of college students at Harvard, it extended to all the US schools and now it has spread throughout the world, and people use it for performing several tasks that they were previously doing online: catching up, exchanging birthday wishes, organizing events, sharing photographs, and so on. MySpace is currently the second largest network and several music groups rely on it for communicating with their fans to announce new albums, where the next concert is going to be, to distribute their music and so on. Examples of other popular social networking sites are LinkedIn, Twitter, Orkut and hi5, and depending on the focus of the network or on the geographical location, one might be more popular than the other (LinkedIn focuses on professional contacts, Twitter on *microblogging*, that is, sharing instantly information with your friends usually using mobile phone for example, Orkut is the most popular online social network in Brazil, hi5 is very popular in Mexico).

In addition to all those sites whose explicit goal and focus is the maintenance of the social network, social networks underlie other services as well. Instant messaging services (e.g., MSN messenger, Yahoo! messenger, Google Talk) form social networks, where a link is a "buddy" in those systems. Similarly one can define a social network over other communication systems such as Skype. We can even imagine the social network on top of email: a link from a user to another might exist if the first user sent an email message to the second during the last year, for example.

Another class of social networks consists of the networks hidden in several online content providing systems. YouTube, a service for sharing videos online, supports the notion of friends and subscribers, which make easier to see when a user has uploaded new material and facilitates sharing. Similarly for flickr, a photo sharing service. digg and delicious, which are services for

[1]Someone's Erdős number is the distance from the mathematician Paul Erdős in the coauthorship network. For example Andras Sarkozy has an Erdős number of 1 because he has coauthored a paper with Erdős (he has actually co-authored the highest number of papers, 62). Henryk Iwaniec has Erdős number 2 because he has coauthored a paper with Sarkozy but not with Erdős directly. Similarly, the Kevin Bacon number is the distance from the actor Kevin Bacon in the social network where a link corresponds to having participated in the same movie.

sharing discovered websites can also be assigned into this category.

We also have social networks that are used for creating new content. Wikipedia is such an example: users are connected if they have worked on the same article. We could also assign to this category the Yahoo! answers service, a link between two users exists if they have answered the same question or if one of them answered the other one's question.

Yet a network can even be defined on entirely fictional entities. There are several online games such as Second Life, World of Warcraft, or Age of the Empires, where players create characters who interact with other characters in the game. The analysis of a social network that exists in such a game can potentially lead to a lot of conclusions, since games record a lot of details in characters' actions which are unavailable in the offline world. To what extent the conclusions drawn by the study of such a network correspond to the real world remains to be seen.

1.3 Importance of Online Social Networks

We have hopefully convinced you in Section 1.1 that the social networks to which we belong have a great importance in our lives, and this has been the case since the beginning of civilization. In the last 5–10 years, however, with the wide spread of the Internet worldwide and with the development of all those so called Web 2.0 services, such as those of the previous section, we have been witnesses of this online social revolution, the effects of which are already present in our lives. As a concrete example, on December 6, 2008 there was a police-shooting incident in Greece that lead to the death of a teenager. The shooting took place on a Saturday, at 10.00pm. Monday morning, slightly more than a day later, there were protests by tens of thousands of Greek students in most of Greek cities. While there might have presumably been some central kernels that organized the protests—although even that is not clear, the coordination and the spread of the message happened through facebook, blog sites, as well as mobile phone messages. Such a fast collective reaction, would have been unthinkable a few years back.

One effect is that we move a lot of our interactions from offline to online: we send messages to give wishes than in person, we organize events online, we use instant-messaging software to remain in touch, we exchange photographs online, these are just some action that we nowadays do more and more online as opposed to just a few years ago. It is then expected that sociologists want to understand what effects can this have to the functioning of society.

For example, one of the results is that it increases the number of contacts that we can keep track of. In the early 1980s anthropologist Robin Dunbar suggested that the number of peers that a primate can keep track of is proportional to the size of the neocortex, a part of the brain []. After analyzing several data he concluded that for humans this number should be around 150, and that it is higher than that of other primates. Furthermore, the structure of our society is more complex and this might be caused by this mere fact, to some extent. Later data from other areas seem to indicate that this number of 150 is fairly accurate. With online social networks we are currently able to keep track of many more contacts. facebook, for example, not only allows us to find out recent details about a given contact if we want, it actually gives it to us automatically through the news feed. On the other hand, while indeed we can have many contacts online, it is not clear whether we as humans actually do keep track of all or most of them, or whether we are actually confined to a much smaller number of really close friends. Time will show which is actually the case.

Another big change that social networks have brought is the redefinition of the notion of privacy. Parts of our lives that we considered private before the existence of social networks are now exposed online. To give a few examples, the list of our friends, our political or religious views, relationship status, photographs from last week's party, all those are often available to

our contacts or even to a larger circle. This has changed our mindset and we are currently willing to expose a lot of details about our lives. Again, time will show if and what the effects are in the long term.

Another novelty of online social networks is that they may change the way that we look for information. While when we want to search for information online most of us refer to a search engine, for some types of information we might have better results if we use a social-networking service. Wikipedia, thanks to the collaborative effort is an organized source that can cover a large fraction of our informational needs. For some more personalized types we might be better off using a service such as Yahoo! answers. A query of the type "what is a good vegetarian restaurant to take my inlaws in San Francisco" is much more probable to be answered well using such a service, than a search engine. In the search of websites, for some types of queries a service such as delicious, and the use of the underlying social network is preferable. The same holds for image or video search where one can take advantage of the flickr and YouTube social networks. Finally, there has been discussion and theoretical work [**?**] about directing our queries to our social network as opposed to a search engine, and it seems that Twitter may provide a such a search service, which several journalists have rushed to argue that this will be the next revolution in search. Once again, time will show if this assessment is correct.

For social scientists, online social networks are a large source of data for the study of human behavior. Before, the standard way to obtain data was through surveys, thus the size of data was rather limited; it usually consisted of a few hundreds, in the best case a few thousands individuals. Furthermore users would provide a limited amount of information through a specific set of questions. Instead, social networking sites log information about all user actions while they are using the service, thus we can obtain a much more fine-grain and probably less biased view of human behavior. In addition, the number of users can be up to hundreds of million or even higher, and this can allow for data mining that can lead to more robust results and to the discovery of patterns and trends that have very low probability to be present if the user sample is significantly smaller.

1.4 Complex Networks

Until now we have been talking about all different types of offline and online social networks. For some of their aspects we often study them in the more broader context of what are called *complex networks*. Roughly, complex networks are all the different networks that we find in various different areas, that are usually created through a complicated and decentralized process that somehow creates networks with some similar structural characteristics.

Apart from social networks, other examples are the Internet (nodes are hosts and edges are connections between the hosts), the phone network, the electricity power grid, the world-wide web (nodes are web pages and edges are hyperlinks), citation networks (nodes are scientific papers and edge exists if a paper references another one), the airline schedules (nodes are cities and edges are flight connections), and several types of networks that appear in biology, such as neural networks.

What might be initially surprising is that while a lot of these seem to be completely unrelated, they apparently share a lot of common characteristics, some of which we describe later in Chapter 3. Therefore a lot of the study, on the structure of social networks, is performed in the more general context of complex networks. We will see that some of the rules that govern the evolution of social networks govern the creation of those other as well (such as the rich-get-richer phenomenon) and this creates a lot of the structure similarities.

Chapter 2

Graph Theory and Other Mathematical Preliminaries

In this chapter we start by giving some basic definitions from graph theory. This will serve as a refresher and will establish notation for the rest of the text. We then give some definitions that are important for the analysis of social networks. Finally we describe the power-law distribution as it appear in several occasions in social-network analysis.

2.1 Graph Theory for Social Networks

In this section we first give some basic definitions of graph theory. We assume that the reader knows basic graph theory and this section is only for reference of the terms and to define notation. Then we define some of the terms that are often used in social network analysis.

A *graph* $G = (V, E)$ consists of a set of *nodes* V, and a set of *edges* $E \subset V \times V$. Unless specified otherwise, we assume that $|V| = n$ and that $|E| = m$. Depending on the literature, a node is also called *vertex*, *site*, *actor*, or *agent*. An edge is also called *bond*, *link*, *connection*, or *tie*. A graph can be *directed* or *undirected*. For simplicity, for the rest of the section we deal with undirected graphs, although the definitions can be extended to directed graphs as well. If graph G is undirected then an edge (u, v) is considered an unordered pair, in other words we assume that (u, v) and (v, u) are the same edge. If G is directed then (u, v) and (v, u) are different edges.

If an edge $e = (u, v) \in E$ we say that nodes u and v are *adjacent* or *neighboring*, and that nodes u and v are *incident* with the edge e. Informally, we will often call two adjacent nodes *friends*, or *peers*, or *neighbors*.

A *loop* is an edge from a node to itself: (v, v). Two or more edges that have the same endpoints (u, v) are called *multiple edges*. The graph is called *simple* if it does not have any loops or multiple edges. We will be dealing almost exclusively with simple graphs.

A *path* of length k is a sequence of nodes $(v_0, v_1, \ldots, v_k)$, where we have $(v_i, v_{i+1}) \in E$. If $v_i \neq v_j$ for all $0 \leq i < j \leq k$ we call the path *simple*. If, $v_i \neq v_j$ for all $0 \leq i < j < k$ and $v_0 = v_k$ the path is a *cycle*. A path from node u to node v is a path $(v_0, v_1, \ldots, v_k)$ such that $v_0 = u$ and $v_k = v$.

A *subgraph* G' of a graph $G = (V, E)$ is a graph $G' = (V', E')$ where $V' \subset V$ and $E' \subset E$.

For an undirected graph, the *degree* of a node v (sometimes called *connectivity* in the sociology literature) is the number of edges incident with v and is denoted by d_v. For a directed graph we have the *indegree*, d_v^-, which is the number of edges that go into node v, and the *outdegree*, d_v^+, which is the number of edges that go out of node v.

A *triangle* or a *triad* in an undirected graph is a triplet (u, v, w), where $u, v, w \in V$ such that $(u, v), (v, w), (w, u) \in E$.

Two nodes u and v are *connected* if there is a path from u to v. A graph G is *connected* if each pair of nodes is connected, otherwise we say that the graph is *disconnected*. Any graph can be decomposed into a set of one or more *connected components*, where each connected component is a maximal connected subgraph of G.

A simple graph that does not contain any cycles is called a *forest*. A forest that is connected is called a *tree*. A tree has $n - 1$ edges. Actually any two of the following three statements imply that the graph is a tree (and thus they also imply the third one):

1. The graph has $n - 1$ edges.
2. The graph does not contain any cycles.
3. The graph is connected.

A *shortest path* (sometimes also called *geodesic path*, or *degree of separation*) between nodes u and v is a path from u to v of minimum length. The *distance* $d(u, v)$ between nodes u and v is the length of a shortest path between u and v. If u and v are in different connected component then $d(u, v) = \infty$.

The *diameter* D of a connected graph is the maximum (over all pairs of nodes in the graph) distance. If a graph is disconnected then we define the diameter to be the maximum of the diameters of the connected components. In other words we define

$$D = \max_{(u,v):u,v \text{ are connected}} d(u, v).$$

The *average diameter* of graph G is the average distance between all the connected nodes of G. Some authors use the term diameter to call this quantity but we avoid that here.

The *effective diameter* is the smallest distance that is larger than 90% of the distances between connected nodes. In other words, it is computed according to the following process: compute the distances between all connected nodes in G, ignore the 10% largest distances, and look at the maximum distance left. This is a quantity often used instead of the diameter as it is more robust with respect to outliers.

Another notion important in the analysis of social networks is the *correlation coefficient*, which is a measure of transitivity, that is, a measure of how much do friends of friends tend to be friends. There are a few different variations of the correlation coefficient that capture this concept, but the most commonly used is the following. We define the clustering coefficient of node v C_v to be the ratio of all the edges that exist between the friends of v over all the edges that could possibly exist between the friends of v (see Figure 2.1). Formally, let us define $\hat{d}_v$ to be the number of nodes different than v that are adjacent to node v; note that for a simple graph $\hat{d}_v$ is just the degree d_v. Then the clustering coefficient (recall that we consider the graph to be undirected) is defined as

$$C_v = \frac{|\{(u, w) \in E : u, w \text{ are adjacent to } v\}|}{\binom{\hat{d}_v}{2}}.$$

Note that if the graph is simple then the denominator equals $\binom{d_v}{2}$, and we have

$$C_v = \frac{2\,|\{(u, w) \in E : u, w \text{ are adjacent to } v\}|}{d_v(d_v - 1)}.$$

The clustering coefficient of graph G is denoted by C and is the average clustering coefficient among all the nodes:

$$C = \frac{1}{n} \sum_{v \in V} C_v.$$

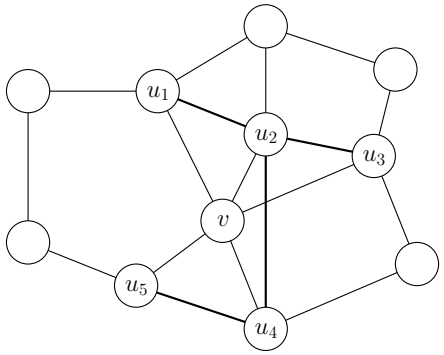


Figure 2.1: Clustering coefficient of node v. Node v has 5 neighbors and there are 4 edges between those neighbors (the bold edges). Therefore the clustering coefficient of node v is $C_v = \frac{4}{\binom{5}{2}} = 0.4$.

2.2 Centrality*

Another notion that is often used by sociologists at the study of social networks is that of *centrality*. As the name implies, centrality measures how central is the node in the graph. Depending on what we mean by "central," there are versions of centrality measure. The most common ones are the *degree centrality*, the *closeness centrality*, and the *betweeness centrality*. We will not be using them here but we describe them for completeness, as they are sometimes found in papers.

The *degree centrality* is the simplest one. The unnormalized one equals to the number of neighboring nodes (the degree in the case of a simple graph). To be able to compare values between different graphs, we define the normalized version, which is normalized by the maximum possible value, $n - 1$. So we have

$$\text{Degree centrality of node } v = \frac{d_v}{n-1}.$$

The second notion of centrality, the *closeness centrality*, or just *closeness*, measures how close is the node to the rest of the network. The total distance of node v to the rest of the nodes equals

$$\sum_{u \in V} d(v, u),$$

and since we want the centrality to be large when the distance is small (intuitively a node is central if its distance from the other nodes is small) we take the reciprocal of that. Furthermore, we again normalize so that the value ranges between 0 and 1 by dividing by the maximum possible value, $(n-1)^{-1}$ (which is the value when a single node is connected with $n-1$ other nodes). In other words we define

$$\text{Closeness centrality of node } v = \frac{\frac{1}{\sum_{u \in V} d(v,u)}}{\frac{1}{n-1}} = \frac{n-1}{\sum_{u \in V} d(v,u)}.$$

The third notion of centrality that we define here is the *betweenness centrality*, or just *betweenness*. Assume that two nodes, u and w need to communicate with each other. Then they will ideally use a shortest path. Any node v that is in that path has the ability to affect the communication by distorting it or slowing it down, for example. A node that belongs to a lot of such paths therefore is central in the sense that it can be in the middle and can affect a lot of such communications. That is what betweeness measures.

To define it formally, assume that nodes u and w have g_{uw} shortest paths that connect them (not necessarily disjoint). Then the probability that they use a particular one when they need to communicate is $1/g_{uw}$, assuming that they choose a shortest path uniformly at random among all shortest paths. For a node v define g_{uw}^v to be the set of those shortest paths between u and w that contain node v. Then the absolute centrality can be defined as

$$\sum_{u,w \in V \setminus \{v\}} \frac{g_{uw}^v}{g_{uw}}.$$

To make it a value between 0 and 1 we normalize it with the maximum value that it can take, which is for the center of the star graph (the graph where a node is connected with the rest $n-1$ nodes and there are no more connections), and in which case the value is $\binom{n-1}{2} = \frac{n^2-3n+2}{2}$. (This is the number of pairs of vertices not including node v, and in the star graph there is a unique shortest path between two nodes and it has to go through the center.) Thus we can define the relative betweeness of a node v as

$$\text{Betweeness centrality of node } v = \frac{\sum_{u,w \in V \setminus \{v\}} \frac{g_{uw}^v}{g_{uw}}}{\binom{n-1}{2}} = \frac{2 \sum_{u,w \in V \setminus \{v\}} \frac{g_{uw}^v}{g_{uw}}}{n^2 - 3n + 2}.$$

This quantity can be computed in polynomial time, the fastest algorithm currently being by Brandes [] and having running time $O(nm)$, for computing the betweeness of all nodes.

To compare the different version of centrality, degree centrality measures the ability of a node to develop communication. The closeness centrality measures the proximity of a node to the rest of the network, while the betweeness centrality measures in a sense the extent to which a node can control communications in the network.

2.3 The Power-Law Distribution

A very interesting phenomenon observed in the study of networks is that a lot of the quantities measured follow the *power-law distribution*. In this section we briefly describe it.

We say that a random variable follows a power law distribution with exponent $\gamma > 0$ if the probability that it obtains a value x is proportional $x^{-\gamma}$. Note that the probability to obtain a given value goes down only polynomially in the value and thus the power-law distribution belongs to the class of what are called *heavy-tail distributions*, which are the distributions for which the density function decays slower than that of the exponential distribution as x increases. The parameter γ specifies the "heaviness" of the tail: the larger it is the faster the probability decreases and the thinner is the tail; for example, as we see later, the variance decreases as γ increases.

In Figure 2.2 we see the distribution function of the exponential distributions, while in Figure 2.3 we can see the distribution function of the power-law distribution.

The power-law distribution can be discrete, where the probability of obtaining a value x is

$$\Pr(X = x) = C \cdot x^{-\gamma},$$

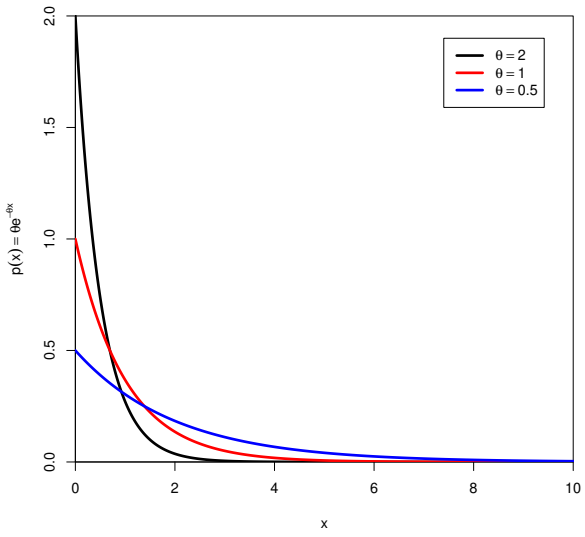


Figure 2.2: Density function for the exponential distribution

for a normalizing constant C, or continuous (as we depicted in Figure 2.3), where now the density function is given by

$$p(x) = C \cdot x^{-\gamma};$$

in our case we are mostly interested in the discrete case. Note that since for $x = 0$ the expression $C \cdot x^{-\gamma}$ becomes infinite, the power-law distribution is defined for values of x bounded away from 0. For simplicity we assume that $x \geq 1$. For the continuous version, and for $\gamma > 1$, the constant C is given by solving

$$\int_1^\infty C \cdot x^{-\gamma} \mathrm{d}x = 1 \Longrightarrow C = \frac{1}{\int_1^\infty x^{-\gamma} \mathrm{d}x} = \frac{1}{\frac{1}{1-\gamma} x^{1-\gamma} \Big|_1^\infty} = \gamma - 1.$$

For $\gamma \leq 1$ the integral diverges. We can, however, still define it if we assume that x can take values in $[1, M]$ for some finite number M, which is what we did in Figure 2.3.

For the discrete version, we just replace the integral with summation:

$$C = \left(\sum_{x=1}^{\infty} x^{-\gamma} \right)^{-1}.$$

The expectation equals

$$\sum_{x=1}^{\infty} x C x^{-\gamma} = C \sum_{x=1}^{\infty} \frac{1}{x^{\gamma-1}},$$

and notice that it is finite only for $\gamma > 2$.[1] For the continuous version we have

$$\int_1^\infty x C x^{-\gamma} \mathrm{d}x = C \int_1^\infty \frac{1}{x^{\gamma-1}} \mathrm{d}x = \frac{C}{\gamma - 2} = \frac{\gamma - 1}{\gamma - 2},$$

for $\gamma > 2$. Similarly, the second moment equals

$$\sum_{x=1}^{\infty} x^2 C x^{-\gamma} = C \sum_{x=1}^{\infty} \frac{1}{x^{\gamma-2}},$$

[1] The sum $\sum_{i=1}^{\infty} \frac{1}{x^\gamma}$ is finite only for $\gamma > 1$.

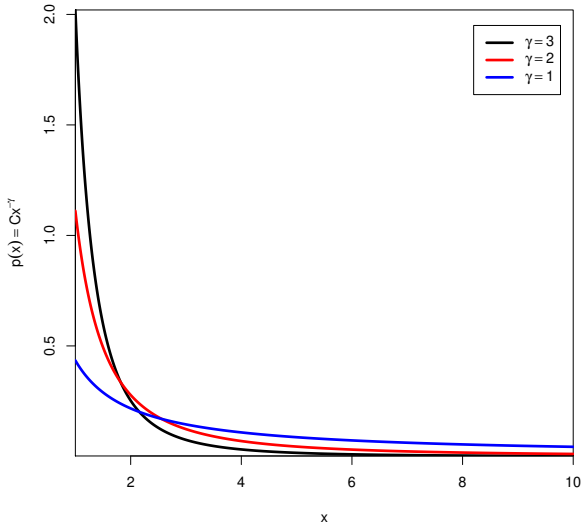


Figure 2.3: Density function for the power-law distribution

or for the continuous case

$$C \int_1^\infty \frac{1}{x^{\gamma-2}} \mathrm{d}x,$$

which is finite for $\gamma > 3$. More generally, the kth moment is finite only for $\gamma > k+1$. Of course, all the moments are finite if we are dealing with the truncated version (where $x \in [1, M]$).

For the continuous distribution, we can compute the variance in a closed form. If X follows a power law we have

$$\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 = \frac{\gamma-1}{\gamma-3} - \left(\frac{\gamma-1}{\gamma-2}\right)^2 = \frac{\gamma-1}{(\gamma-2)^2(\gamma-3)}.$$

Notice that as γ increases the variance (and the various moments) decreases. This is expected since the tail becomes thinner and the probability mass moves towards smaller values.

Consider now the density function $p(x) = Cx^{-\gamma}$. If we take the logarithm we obtain

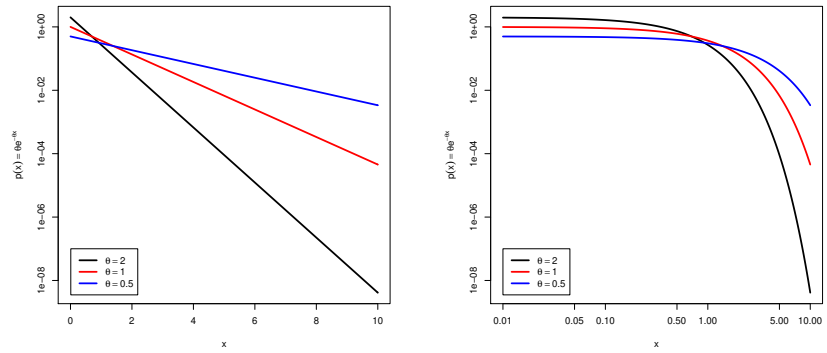
$$\ln p(x) = -\gamma \ln x + \ln C,$$

so the logarithm of the density function is linear to the logarithm of x. This means that if we plot the density function in a log-log scale the graph is a straight line, whose slope equals $-\gamma$. Similarly for the exponential distribution, whose density function is $p(x) = \theta e^{-\theta x}$, we obtain

$$\ln p(x) = -\theta x + \ln \theta.$$

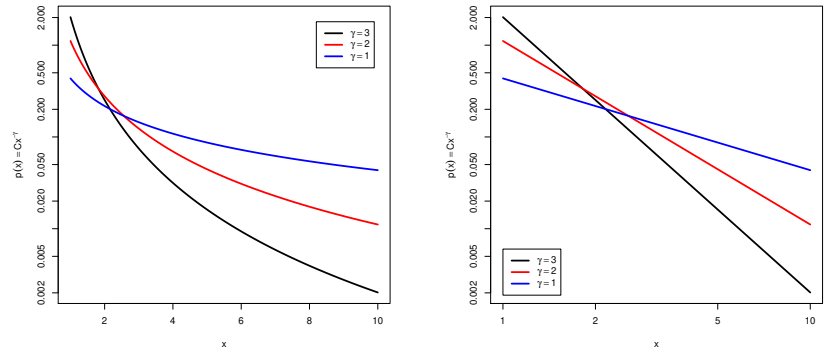
We can see that the logarithm of the density function is linear this time directly with x. Thus, if we plot the exponential distribution with only the y axis scaled logarithmically the graph is a straight line. These facts are shown in Figures 2.4 and 2.5. When we examine real data, if we plot then using those two scales we can obtain an idea about whether the data seem to follow a power-law or an exponential distribution.

The power-law distribution is also known as *scale-free* or *scale-invariant* distribution, because its density function is a scale-free function. A function f is called scale free if it is the case



(a) log scale (b) log-log scale

Figure 2.4: Density function for the exponential distribution in logarithmic scales.



(a) log scale (b) log-log scale

Figure 2.5: Density function for the power-law distribution in logarithmic scales.

that $f(ax) = bf(x)$, for a, b independent of x. Note that for the power-law density function we have

$$p(ax) = \frac{C}{(ax)^\gamma} = \frac{1}{a^\gamma}\frac{C}{x^\gamma} = \frac{1}{a^\gamma}p(x).$$

The power-law distribution is the only distribution satisfying this property, therefore the terms power law and scale free are equivalent.

The power-law distribution appears in a lot of seemingly unrelated instances. As we will see the distribution of the degrees of most complex networks follows follows a power law. Other examples include the populations of cities, the number of citations of publications, the number of occurrences of words in texts, name frequencies, and people's net worth [7]. On of the explanation for this universality is the rich-get-richer phenomenon: the most more money you have the more likely you are to obtain more; the more one's paper is cited the more likely get new citations, as more people are becoming aware of it.

Chapter 3

Structure of Social Networks

An important element in understanding the functionality of social networks is their structures. How many friends do people have on average? How much does this vary? Oftentimes we say "what a small world!" when we meet someone randomly and we happen to have common friends; how probable is that? How are we organized, do we form communities? What are the consequences of that? How many connections separate a random person from Jim Carey? How about from a random person of the opposite part of the world?

The questions mentioned above and several others have attracted the interest of several sociologists in the past years. The answers to those questions might seem sometimes surprising in the beginning but make sense after some thought and as our understanding of social networks grows. Thus, scientists have looked into several offline and online social networks and studied their structural properties. One of the most exciting findings is that for most of the networks the structure is very similar; they all possess some particular characteristics, whether they are large or small, those characteristics seem to be present. In this chapter we describe some of them. As we have already mentioned those structural properties are found in other types of complex networks, so we will also show some examples from them.

3.1 One Giant Component

The first fact that one notices when looking at a social network is the existence of a giant connected component. This usually contains a large fraction of the nodes and in some cases it contains the large majority of them. The second component size is much smaller. Finally (again depending on the type of social network), there is often a large number of singleton nodes (nodes with degree 0).

In Figure 3.1 we can see the distribution of the connected components of the MSN instant messenger communication graph. We can see that the largest component contains more that 10^8 nodes, the second one less than 1000, and there are about $100K$ singleton nodes.

3.2 Heavy-Tailed Degree Distribution

The next characteristic that becomes immediately eminent in social and complex networks is the heavy-tailed degree distributions. As a matter of fact most of the times we observe that the degrees follow a power law distribution (for information about the power-law distribution see Section 2.3).

In Figure 3.2 we can see the indegree and the outdegree of the web graph. Notice that the degrees seem to follow power-law distributions with exponents 2.1 and 2.7. Even in smaller scales

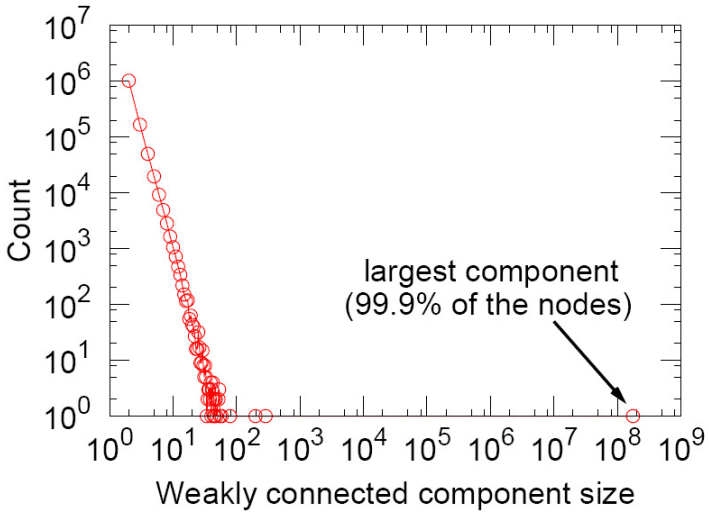
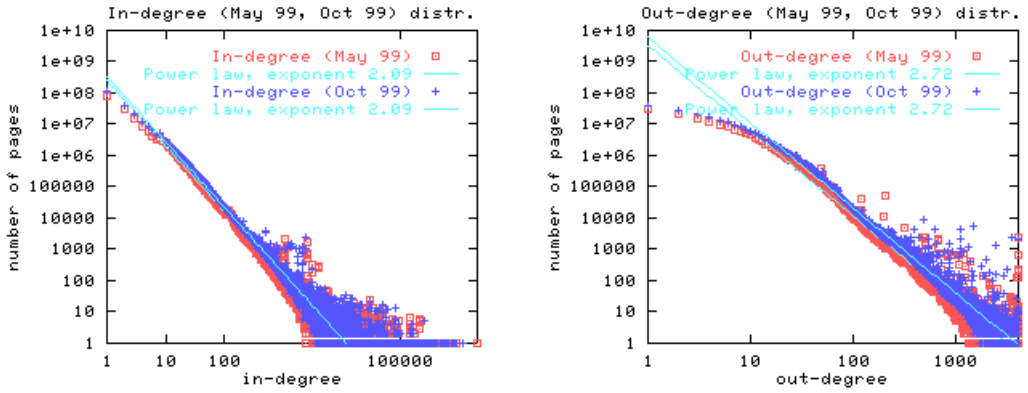


Figure 3.1: Distribution of the connected components of the MSN instant messenger communication graph.

we still find power-law distributions. In Figure 3.3 we see the corresponding plot restricted to the *.brown.edu domain, the domain of Brown University. We can see the power-law distributions of the indegrees and the outdegrees, and what is even more interesting is the fact that the exponents are the same as for the entire web.

We can, finally, see the indegree and outdegree of the flickr social network a few years ago in Figure 3.4. The power-law distribution is again clear.



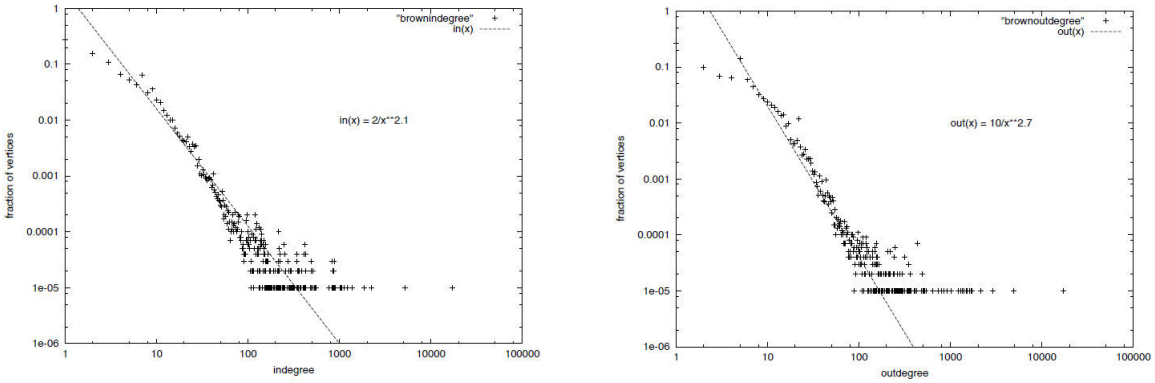
(a) Indegree.

(b) Outdegree.

Figure 3.2: Distribution of the indegree and outdegree of the webgraph in two different time periods. We can see that both of them follow a power-law distribution with different exponents. From [2].

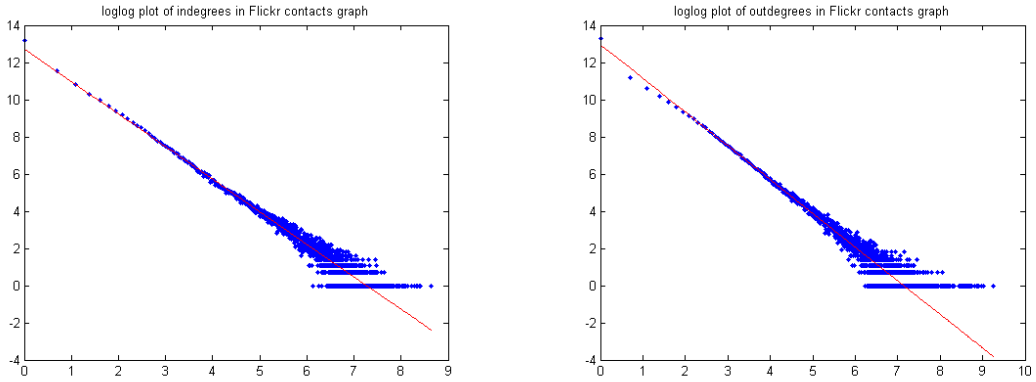
3.3 Small World

One of the most surprising, in the beginning, facts about social networks is that two individuals are not far from each other as nodes in the social network graph. The term six degrees of separation has been coined to refer to such networks, after a series of experiments by the Yale



(a) Indegree. (b) Outdegree.

Figure 3.3: Distribution of the indegree and outdegree of the webgraph inside *.brown.edu (Brown university). Notice that both distributions are power laws and with even the same exponent as for the entire web graph. From [8].



(a) Indegree. (b) Outdegree.

Figure 3.4: Distribution of the indegree and outdegree of the flickr social network.

sociologist Stanley Milgram, who discovered that the average distance between two people in the United States is around 6.

Stanley Milgram who is also famous for the so called Milgram experiment [], which dealt with issues of obedience and authority, conducted a series of experiments the most complete of which is the following study, which he performed along wih Jeffrey Travers [9] in the end of the 1960s. He selected an individual from Boston, Massachusetts as a "target" who was a stockbroker and 296 individuals as follows:

- 100 were a sample from residents in Boston
- 96 were a sample from Nebraska, about $2,700$km far from Massachusetts
- 100 were a sample of the share-owners in Nebraska

Each of these individuals was given a letter that they were supposed to send it to the stockbroker in Boston. They were told that he is a stockbroker and that he is situated in Boston, and the rules for sending the message were the following:

- If they know the stockbroker in a first-name basis then they send the letter directly to him.
- If not, then they send the letter to a person that they know in a first-name basis that they believe is closer to the stockbroker, along with these rules.

The choice of the three groups was in order to determine what difference does the proximity (geographic or professional) make to the path lengths. In the instructions given to the individuals there were included cards to mail back to Travis and Milgram to keep track of the message routes and to gather statistics.

From the 296 people, 217 proceeded with the experiment and from them 64 letters (about 30%) arrived at their destination. The lengths of the chains corresponding to letters that were successfully delivered are shown in Figure 3.5. The average length is about 5.2 and this is what lead to the term six degrees of separation. Other conclusions of the study were that some people used mostly the profession to find the target and others the geography and this is the reason for the bimodality (the two peaks) in Figure 3.5; the paths that were controlled by the geography were slightly longer due to the fact that the message would arrive to the area but then wander around until some acquaintance of the target was found.

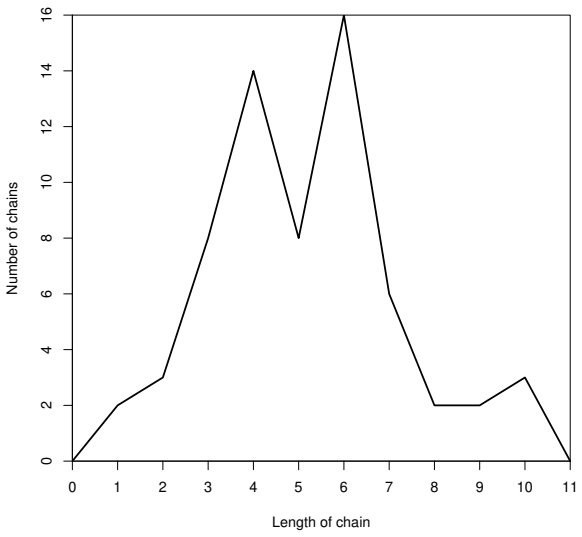


Figure 3.5: The lengths of the chains of the letters that reached their destination

An attempt to replicate Milgram's experiment in a larger scale was performed by Dodds, Muhamad, and Watts [3]. Their experiment was based on email messages and participants were trying to reach 18 targets, ranging from a university professor in the US to a policeman in Australia. About 24 thousand chains were initially created, and finally 384 of them reached their targets. Among the findings is the conclusion that the typical chain length (median) is about 7.

Researchers have also studied various other complex networks, all of them confirming that the average distance and the diameter are small. For example, in the MSN communication network of about 180 million nodes the average distance was found to be about 6.6 [5] and the effective diameter (Section 2.1) was about 7.8. If we look at the WWW graph (back in 1999), where nodes are web pages and an edge from a page to another one exists if there is a hyperlink from the first page to the second, the average distance between two pages is 16 (6 if links can

be followed backwards) [2]. Generally all the online and offline networks that we have studied show reveal a small average distance between two nodes and a small diameter.

3.4 Globally Sparse, Locally Dense

Another universal observation is that social (and complex networks in general) are globally sparse, yet they are dense locally. Globally sparse means that there are only a few edges in total, compared with the number of nodes. For example, facebook, as of March 2010 has more than 400 million users, which means that the total possible number of connections is about $8 \cdot 10^{16}$. Yet the average degree is about 120, therefore only $2.4 \cdot 10^{10}$ edges exist, or 3 in a billion. On the other hand, facebook users observe see that many if not most are connected with each other, that is then chance of an edge in the neighborhood is much larger than $3/1,000,000,000$.

Chapter 4

Graph Models for Social Networks

An important part of understanding phenomena that we observe or some data that we have gathered is that of modeling. *Modeling* is generally a way to capture a phenomenon in the real world. There are several different ways to model a phenomenon, and there are probabilistic, game-theoretic, agent-based, statistical, and other types of models. Here we will see some examples of statistical models, which can also be seen as generative models. A good model has a few desired properties:

- It creates data that have similar statistical properties with the ones observed in reality.
- The generation process seems natural and corresponding in some sense to what actually happens in reality.
- It is amenable to analysis.

These properties are rather vague and not always possible. For instance, very often a model to create data that are realistic it has to become complicated; this means that it will probably be hard to analyze it. Thus, we often consider models that replicate only *some* of the characteristics of the underlying data, precisely the characteristics that we are interested in understanding in the given time. This is why the whole process of modeling is usually hard, requires experience, and it is often both a science and an art.

Why are we interested in designing models? First, designing a good model can help us *making sense* of the phenomenon that we observe, and the process with which the data that we have collected are created. For instance, the Erdös–Renyi random graph model and the model of Watts and Strogatz [10], even though it is not realistic, can help us explain the *small-world phenomenon* in networks. The Barabási–Albert preferential-attachment model [1], can explain the power law observed in the degree distribution.

Second, a good model can help us *make predictions*. Assume that we have a history of data, say 100 days of transportation information. What we will often do is design a model (i.e., come up with a family of models, which are characterized by a set of *parameters*) and use the largest part of the past data, say 90 days, to *train* the model, that is, find the values of of the parameters of the model. Then using the part of the data that was left out we will evaluate the model: we will see if it can create data for the 10 days that are somehow similar to what the real data are in these 10 days. Or we will check what probability does the model give to the real data in these 10 days. This process (which is often iterative, and usually more complicated than what was just described) allows us to evaluate the model and see whether the family of the models that we have chosen and the parameters that we have estimated are proper. Assuming that this is the case, then we can use the model to make predictions for the future: if the model that was trained in 90 days in the past is able to predict the last 10 days,

then we assume (hope!) that it can predict the next 10 days as well. This of course requires several assumptions for our real-life scenarios.

Taking this one step further, we can now *take decisions* for the future. For instance, if the model predicts that we will have a lot of people wanting to go from Colosseum to San Peter's Cathedral, we can increase the number of buses that make this route. Or, if a particular flu strain is predicted to infect a lot of people on February, a country may try to increase its stock of flu vaccines and make a campaign vaccination. Or even, go to more drastic measures, such as reducing movement, closing schools, or even ordering a lockdown, as we have all observed.

Finally, a model can also be used to *design tools*. For instance, *PageRank*, the algorithm first used by Google (along with many other features) to assign scores to web pages, is actually the value obtained in the *random surfer model*: In this model, we consider a user who visits pages and follows outgoing links randomly, whereas with some probability he stops and restarts browsing at a random page; then the pagerank score of a given web page p is the frequency, in the long term, that this user visit page p.

Throughout the years, scientists have made efforts to create models that generate graphs with the same characteristics of complex networks. While not all of them have focused particularly in social networks, for example, a lot of the models proposed have targeted biological networks or the world wide web, nevertheless they generate some of the statistical properties that social networks share with all those graphs, presented in Section 3.

First we describe the classical Erdős and Rényi random-graph $G_{n,p}$ and $G_{n,m}$ models. While, as we will see, they are not appropriate models for social networks, they are the easiest to analyze and they will give us intuition and form the basis for analyzing some of the more complex models. Subsequently we study the small-world model proposed by Watts and Strogatz, which captures the idea that most of our contacts are in our neighborhood (geographically, professionally, etc.) but we also have a few apparently random acquaintances and those actually lead to the small-world phenomenon that we described in Section 3.3. Finally, we present a simple example of a random graph process and we show how what is called preferential attachment can lead to power-law distributions.

4.1 The Erdős-Rényi Random-Graph Models

The Erdős-Rényi random-graph model was first introduced by [] and subsequently studied by Erdős and Rényi []. There are actually two random graph models, which differ slightly with each other. The most common one is the $G_{n,p}$ model which is parametrized by two parameters, n the number of nodes, and some value $p \in [0, 1]$ which is the probability that an edge exists; p can in general be a function of n. In particular, a graph is created according to the following process. It has n nodes and each of the $\binom{n}{2}$ possible edges exists with probability p, each edge existing independently of the others. Note that there are $2^{\binom{n}{2}}$ possible graphs and if a graph contains m edges then it has probability $p^m(1-p)^{\binom{n}{2}-m}$ to be constructed. For large values of n we expect the number of edges to be very close to $\binom{n}{2}p$. In Figure 4.1 we can see how a random graph might look like.

The second random-graph model is the $G_{n,m}$ model. A graph created according to the $G_{n,m}$ model contains n nodes and exactly m edges, and the process gives to each graph of m edges the same probability to be created. Note that for given values of n and m the number of graphs that can be created according to the $G_{n,m}$ model is equal to

$$\binom{\binom{n}{2}}{m},$$

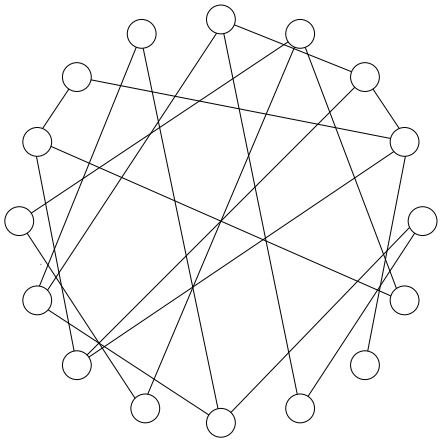


Figure 4.1: A random graph.

and each of the graphs has the same probability to be created. In most of the cases this random graph model is a bit harder to work with as there are dependencies among the edges (even though they are small): if an edge exists, this reduces slightly the probability that some other edge exists. Therefore we mostly work with the $G_{n,p}$ model. Note though that usually we are interested in studying the behavior of the graphs for large values of n and in this case, as we mentioned, the number of edges that will appear in the $G_{n,p}$ model will be sharply concentrated around the expected value $\binom{n}{2}p$. Therefore, the two models behave very similarly, and one can prove that properties that hold for one model can be carried to the other. In any case, for the rest of this section we will consider the $G_{n,p}$ random-graph model.

4.1.1 Graph Properties and Threshold Phenomena

In general we will be studying properties of random graphs such as connectivity, the size of the largest component, and so on. We usually fix the value of p to some particular function of n and then study properties of the graph as $n \to \infty$. We will see that properties hold with high probability, for example we will see that a if $p > \ln n/n$ then the graph is connected with high probability. When we say *with high probability* (abbreviated as whp.) we mean that it holds with probability that tends to 1 as $n \to \infty$. We can also say that it holds for *almost all graphs* to express the same thing.

Many of the properties that we will study are what we call *monotone* properties. A property is monotone if adding edges to the graph does not destroy it. In other words, if it holds for a graph $G = (V, E)$ then it also holds for a graph $G' = (V, E \cup E')$. An example of a monotone property is connectivity: if a graph is connected then by adding edges it remains connected. Another monotone property is the containment of a hamiltonian cycle.

One can see (and it is easy to prove) that if a monotone property hold whp. for $G_{n,p}$ then it also holds whp. for $G_{n,p'}$ for $p' > p$. A very interesting phenomenon that we will observe with monotone properties is the following: Some property will not hold whp. for small values of p until some value p^* (that depends on the property). Then, as soon as $p > p^*$ the property will hold whp. In other words we have a sharp threshold at the value p^*. We also say (borrowing the term from the study of physical systems) that we have a *phase transition* at the value p^*. For example we will see that a graph is not connected whp. if $p < \ln n/n$ and connected if $p > \ln n/n$. Similarly, we will see that the size of the largest component is $O(\ln n)$ if $p < 1/n$ while it becomes $\Theta(n)$ for $p = c/n$ for a constant $c > 1$.

4.1.2 Degree Distribution

The easiest property to study is the degree distribution. A given node v is incident with $n-1$ potential edges, and each of them exists with probability p independently of each other. Therefore the degree distribution follows a binomial distribution Binomial$(n-1,p)$:

$$\mathbf{Pr}(d_v = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

and therefore the expected degree is $(n-1)p \approx np$. Recall that in Section 3.2 we mentioned that in social networks we observe heavy-tailed degree distributions. This shows one of the main problems of the Erdős-Rényi random graph model for modeling social networks. In other words in the Erdős-Rényi model the nodes are very similar to each other; instead in reality we observe that nodes can be very different from each other.

4.1.3 Connectedness and Giant Component

An interesting and important property to study is the size of the largest component as the value of p increases. Here we only state the most important cases and for more details one can refer to one of []. In all the following statements hold whp. We will focus on a few ranges for p.

- $p < 1/n$,
- $p \sim 1/n$,
- $p = c/n$, where $c > 1$, and
- $p = c \ln n/n$, where $c \geq 1$.

When $p < 1/n$ the graph has a lot of small connected components. They are all trees, or trees with and additional edge. The size of the largest component is $O(\ln n)$.

At the value of $p \sim 1/n$ we have a phase transition. Suddenly we have the emergence or a large connected component of size $n^{2/3}$. The rest of the components are of size $O(\ln n)$.

When $p = c/n$ the size of the largest component becomes linear $\Theta(n)$, where the constant hidden in the Θ notation depends of course on c.

At the value $p = \ln n/n$, we have another phase transition. When $p = \ln n/n$ there is a giant component that contains all but a constant number of isolated nodes. Finally, for $p = c \ln n/n$, where $c > 1$ the graph becomes connected.

4.1.4 Diameter

The small diameter, even when the probability p is small, is one of the attractive properties of this model. The main result is that the diameter is always bounded by $O(\ln n)$, more precisely it is of the order of $\frac{\ln n}{\ln np}$.

To get an intuition of the proof for that, we consider a node u and we want to show that in a small number of steps we can reach all the nodes in the graph. Node u is expected to be connected to np other nodes. Similarly, each of them is connected to about np other nodes so there are about $(np)^2$ nodes of distance 2 from u. More generally there are about $(np)^t$ nodes at distance t. When $(np)^t \approx n$ it will have covered all the nodes. Solving for t we get that in about

$$\frac{\ln n}{\ln np}$$

steps it can reach all the nodes, and since this holds for each node the diameter is of the order of $\ln n / \ln np$.

There are two hand-waving parts in the above argument. First it assumes that there are no overlaps in the nodes that we find. For example, there are slightly less than $(np)^2$ nodes at distance 2 from u since the neighbors of u might share some neighbors. However if p is small enough the overlap is small. Second, during step t, when we have seen $\ell \geq (np)^t$ nodes the number of potential new nodes are not n but $n - \ell$. However, if we continue this argument as long as $\ell \leq n/2$ there is always a sufficient number of new nodes and in $\Theta(\ln n / \ln np)$ steps node u can reach half of the graph. But the same holds for every other node, so the diameter is at most double the quantity that we computed previously.

4.1.5 Clustering Coefficient

The clustering coefficient is easy. Consider a node v with degree d_v. There are $\binom{d_v}{2}$ potential links between the d_v neighboring nodes and in the $G_{n,p}$ model each of them exists with probability p. Therefore, for any degree d_v the clustering coefficient of v is $C_v = p$. Therefore, the clustering coefficient of the graph is $C = p$. Since p is usually small the clustering coefficient is also small which is another problem of this model since in reality networks are much more clustered. This motivates the model that we study next.

4.2 The Watts-Strogatz Small-World Model

The Erdős-Rényi random graph model is clearly not a good model for social networks: in reality the probability for a link to a person in the neighborhood is much higher than the probability of a link to a person on the other side of the world. Usually we have several links to people that we are close with (geographically, jobwise, etc.) but we also have some few additional "long-range" contacts, for example a person we met while traveling, the friend of a friend of a friend that we met at a party, and so on.

This is what motivated Duncan Watts and Steven Strogatz to come up with their model in 1998 [10]. On one hand we tend to be parts of communities and so our friends tend to be friends with each other. (This leads to a large clustering coefficient.) However a completely clustered network has long diameter: to reach to a person that is far one has to move from cluster to cluster and this might require a large number of steps.

One of the nice properties of the Erdős-Rényi random graph model, as we saw in Section 4.1.4 is that it creates graphs with small diameter. These random links suffice to bring the diameter to about $\ln n$. However, they do not create the clustered structure and this is demonstrated by the small clustering coefficient, as we mentioned in Section 4.1.5

The Watt-Strogatz model tries to get the best of both worlds. In high level it creates a graph by taking a structured graph such as a ring or a grid and then with some small probability replace each edge with a random one. To be more concrete, the simplest way is to start with a ring where a node is connected with the k closest nodes in the ring, $k/2$ in each side (see the left graph in Figure 4.2). Then each node u considers its $k/2$ edges on the left and with probability p it replaces the edge with a random edge with u as an endpoint. Note that duplicated edges are forbidden. In Figure 4.2 we see three realizations of this procedure for $p = 0$, for $p = 1$, and for a value of p between them. For $p = 0$ no edge is being rewired so the final graph is the initial regular graph that we started with. For $p = 1$ all the edges are being rewired, so we have a random graph (note though that this is not an Erdős-Rényi random graph as, for example, here each node has at least k edges something not guaranteed in the Erdős-Rényi model; the behavior, however, is similar). Note that this procedure might lead to a disconnected graph so

to avoid that we assume that $k \gg \ln n \gg 1$, in which case the graph will be connected with high probability. Later on we see a slight variation of this model that avoids this problem.

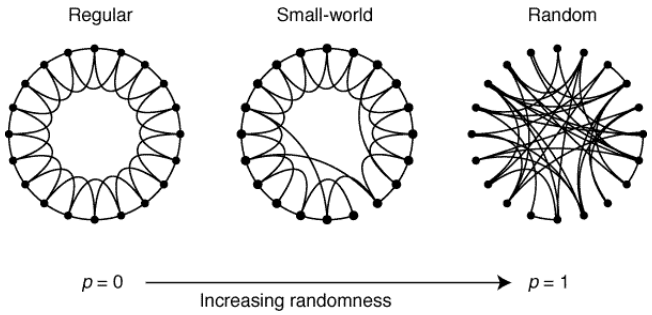


Figure 4.2: The Watts-Strogatz model as p varies from 0 to 1.

Let us look now at the structure of the graphs and in particular let us measure the diameter and the clustering coefficient. For $p = 0$ it is not hard to compute that the diameter is $D = n/k$ and the clustering coefficient approaches $3/4$ as $n, k \to \infty$. On the other hand, for $p = 1$ the diameter can be shown to be $\Theta(\ln n / \ln k)$ (similarly to the diameter of the Erdős-Rényi model with the same average degree) while the clustering coefficient is k/n. Thus in both cases we see that a large diameter goes with a large clustering coefficient, while a small diameter goes with a small clustering coefficient. One therefore might wonder whether this is always the case as $p \in (0, 1)$ or whether we can achieve a large clustering coefficient and a small diameter at the same time, as we observe in reality. It turns out that the answer is yes. If we take a value of p that is small, then the clustering coefficient does not change a lot: only a few edges are rewired, therefore most of the links between a node's neighbors will continue to exist. On the other hand, however, this small number of random links suffices to reduce the diameter significantly. For example if $k = \Theta(\ln n)$ and p is some small constant, each node will have at least $\Theta(\ln n)$ random links and thus the diameter will be $O(\ln n)$, as we saw in Section 4.1. This argument is of course hand waving, for example we are not dealing exactly with the Erdős-Rényi model, it conveys however the right intuition.

In Figure 4.3 we can see how the diameter and the clustering coefficient change as p varies from 0 to 1. Notice that the x axis is in a logarithmic scale and that even for small values of p the diameter drops immediately while the clustering coefficient remains high.

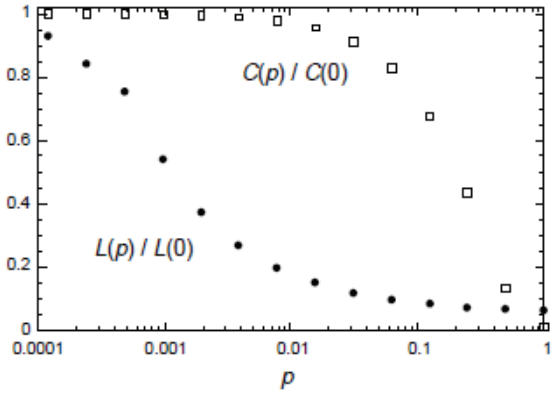


Figure 4.3: The change of the clustering coefficient and the diameter as p increases from 0 to 1. From [10].

A similar model that researchers have looked at [] is when instead of rewiring the edges the underlying structured graph remains as it is but in addition we add some shortcuts. In particular, each edge is shortcuted with an appropriate probability so that at the end the graph has an expected number of $Lpk/2$, similarly to the original model (therefore every edge is shortcuted with probability $pk/(L-k-1)$; why is that?). The advantage is now the network remains connected, but also that it resembles more the Erdős-Rényi model and thus it is analytically more tractable.

We also mentioned that the ring is not the only option. For example we can use a grid as the base model and this is actually what we will do in the next section.

4.3 Navigating in Social Networks

The first striking conclusion from Milgram's experiment is that we are all close to each other in a social network. However, another conclusion observed by Jon Kleinberg [4] is that actually people were able to find the path even though they had to route the message having only local information. Finding the shortest paths is easy to do if one has complete information of the entire network, for example by performing a breadth-first search. However, when an individual receives the letter he did not have global information of the network; instead he has information about his contacts and maybe some limited information about their contacts; furthermore, he had some information about the location and the occupation of the stockbroker. Nevertheless, the letters that arrived, did arrive quickly. Thus, after realizing this fact, a natural question is the following: Does the Watts-Strogatz model allow for routing algorithms that can route the message successfully with only local information? And if not, what can be a model that does allow decentralized algorithms to succeed?

Kleinberg addressed both of these questions. He first showed that the Watts-Strogatz model, while it create short paths, it does not allow for local routing. Furthermore, he considered a family of models that generalizes the Watts-Strogatz model and he found the models of this family for which the diameter is small and also a decentralized algorithm can find a short path between two nodes.

We now go ahead and describe the model in some detail. We start with a grid as in Figure 4.4 (a) (other structures are possible as well, but for simplicity we consider the grid like the original paper), and each node adds an additional random node (Figure 4.4 (b)). Note that we assume that the graph is directed and, in particular, each node's random connection can be used only one-way. Similar results hold if the links are undirected, however the analysis would be slightly more complicated.

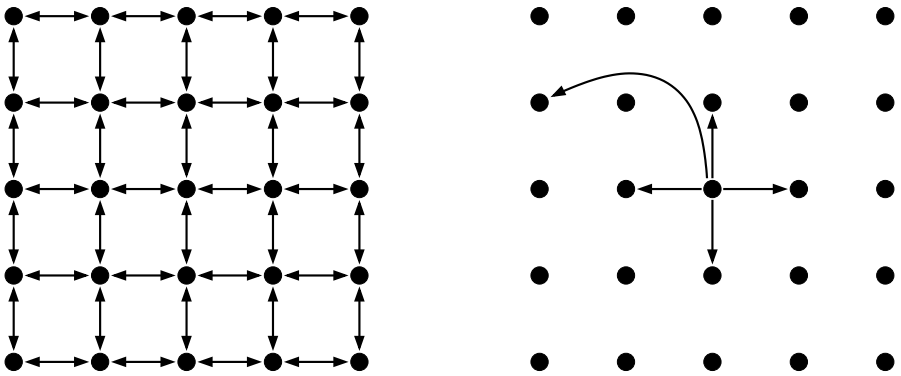
The grid is of size $n \times n$, so it has n^2 nodes; let V be the set of nodes. For each node v we can define its coordinates (v_x, v_y) in the straightforward way, where the node $(0,0)$ is, say, the bottom left. The distance of two nodes $u = (u_x, u_y)$ and $v = (v_x, v_y)$ is equal to the ℓ_1 distance, also known as Manhattan distance, on the grid:

$$d(u,v) = |u_x - v_x| + |u_y - v_y|.$$

The model is parametrized by some value $r \geq 0$ and this gives a family of models. Given r the probability that the random connection from node u to some other node v is proportional to $d(u,v)^{-r}$, which means that it is selected with probability

$$\frac{d(u,v)^{-r}}{\sum_{w \in V \setminus \{u\}} d(u,w)^{-r}}.$$

This means that nodes that are closer have higher probability to be selected, and how much higher is determined by the parameter r. For $r = 0$ the probability is uniform and this case



(a) Initial grid.

(b) A node with the four grid links an a long-range link.

Figure 4.4: Kleinberg's small-world model.

is similar to the Watts-Strogatz small-world model. As r increases the distribution puts more weight on points that are close to u.

The question that we now ask is for what range of values r does the model allow for a local routing protocol. To be more precise we consider a source node s and a target node t and we want to route a message from s to t in as few steps as possible, through a *local* or *decentralized* protocol. We say that a protocol is local or decentralized when each node has only the following knowledge when deciding how to route a message that it has received:

1. the set of all of its contacts;
2. the location of the target node t;
3. the location of all the neighbors of all the nodes that have come into contact with the message.

It might seem that the third condition gives too much power to the protocol. However, it is only used in the lower bounds so it just strengthens the results. In other words, for the values of r that allow local routing, the protocol that we will present only makes use of the first two conditions. On the other hand, we will show that for the rest of the values of r no protocol can route a message efficiently even if it makes use of the entire history of the message.

We are interested in studying the expected distance between a source and a uniformly selected random target. The expectation is over the selection of the target and over the structure of the network, in particular, over all the possible placements of the long-range links.

We are now ready to present the main result.

Theorem 1. *Consider a network with parameter r.*

1. *Let $0 \leq r < 2$. The expected delivery time of any decentralized algorithm is $\Omega(n^{(2-r)/3})$.*
2. *Let $r > 2$. The expected delivery time of any decentralized algorithm is $\Omega(n^{(r-2)/(r-1)})$.*
3. *Let $r = 2$. There is a decentralized algorithm with expected delivery time $O((\ln n)^2)$. $\Omega(n^{(r-2)/(r-1)})$.*

Note that for $r = 0$, the case that essentially corresponds to the Watts-Strogatz model, the theorem gives a lower bound of $\Omega(n^{2/3})$.

Here we will only try to give the main ideas behind the theorem by sketching the proofs for the case of $r = 0$ and $r = 2$. The details and the proofs for all the cases can be found in the original paper [4].

Let us begin with the case that $r = 0$ and show that the expected path length is at least $\Omega(n^{2/3})$. The main idea is the following: We consider a ball of radius $n^{2/3}$ around the target t (see Figure 4.5). If the source is outside the ball and as the message is routed towards t no long-range links are from the path to a node inside the ball (such as e_1 or e_2 in Figure 4.5), then the message must spend at least $n^{2/3}$ steps since from the time it enters the ball it uses only the grid links. It turns out that this happens with sufficiently large probability due to the fact that the ball is small compared to the entire network.

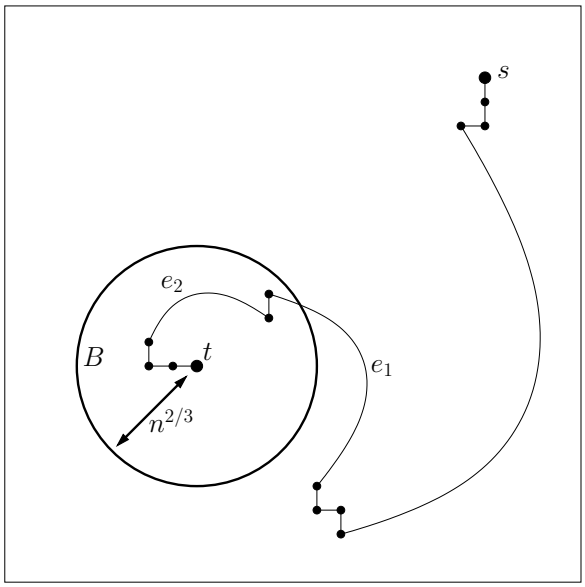


Figure 4.5: Lower bound for $r = 0$.

Let us now work out the details. Let B be the set of nodes that have distance at most $n^{2/3}$ from node t (Figure 4.5); we say that B is a ball of radius $n^{2/3}$ centered at node t. We will need an upper bound on the size of B; we omit the details but some thought shows that it is of size $O((n^{2/3})^2)$, that is

$$|B| \leq cn^{\frac{4}{3}}, \tag{4.1}$$

for some constant $c \geq 1$. Recall that node t is distributed randomly on the grid and assume for now that node s is outside of B. For a constant δ to be determined later, we consider the first $a = \delta n^{2/3}$ steps of the route of the message from node s to node t; if the message arrives at t in less than $\delta n^{2/3}$ steps we assume that the routing continues and that at each step from then on the message visits node t. We will make use of the *principle of deferred decisions* [6], which is often used in the analysis or randomized algorithms, and is the following, rather intuitive idea: although the network has been constructed a priori through the random process in which each nodes selects the random link, for the sake of the analysis, we may assume that each node that receives the message creates the random link at the moment that it receives it. This assumption can allow us to argue only about nodes that are in the route of the message since the links of the other nodes do not affect the route of the message.

Let us consider a node u that receives the message. Since $r = 0$ each node in the graph has the same probability $1/n^2$ to be the endpoint of the long-range link (recall that the probability for some node v being the endpoint is proportional to $d(u, v)^{-r}$ and this value is independent

of the distance for $r = 0$). Therefore, the probability that node u's long-range link is towards a node inside B (such as e_1 or e_2 in Figure 4.5) is

$$\frac{|B|}{n^2} \leq \frac{cn^{4/3}}{n^2} = cn^{-\frac{2}{3}},$$

by Equation (4.1). By the union bound we can conclude that the probability that one of the first $a = \delta n^{2/3}$ nodes in the path from s to t has a long-range link towards a node inside ball B is bounded by

$$\delta n^{\frac{2}{3}} \cdot cn^{-\frac{2}{3}} = \delta c = \frac{1}{2},$$

if we let $\delta = 1/2c$.

Now we can tie things together. If there are no long-range links towards a node inside B then even if the path reaches B it can only use the grid links (or it uses a long-range link which takes it out of B again). We have assumed that $s \notin B$ and since we have that $\delta < 1$ we can conclude that in the first $\delta n^{2/3}$ nodes of the path we cannot have reached node t. Therefore, with probability at least $\delta c = 1/2$ the length of the path from s to t is at least $\delta n^{2/3}$, in other words the expected length of the path is at least

$$\frac{\delta}{2} n^{2/3}. \tag{4.2}$$

Until now we have assumed that node $s \notin B$. To finish the proof, note that node s is distributed uniformly at random (by our assumption when we try to compute the expected length), therefore the probability that $s \notin B$ is at least

$$\frac{n^2 - |B|}{n^2} \geq \frac{n^2 - cn^{4/3}}{n^2} > \frac{1}{2}, \tag{4.3}$$

for sufficiently large n. We let $\mathcal{E}$ be the event that "$s \notin B$" and if we let L to be the length of the path from node s to node t we have

$$\begin{aligned}
\mathbf{E}[L] &= \mathbf{E}[L|\mathcal{E}] \cdot \mathbf{Pr}(\mathcal{E}) + \mathbf{E}[L|\mathcal{E}^c] \cdot \mathbf{Pr}(\mathcal{E}^c) \\
&\geq \mathbf{E}[L|\mathcal{E}] \cdot \mathbf{Pr}(\mathcal{E}) \\
&\geq \frac{\delta}{2} n^{2/3} \cdot \frac{1}{2} \\
&= \Omega(n^{2/3}),
\end{aligned}$$

where the last inequality follows from Equations (4.2) and (4.3).

Now we turn to the case that $r = 2$. First, we need to give the algorithm. It is the straightforward greedy algorithm: when a node receives the message it sends it along the link (among the five links starting from the node) that will bring it closer to the target t (with respect to the distance on the grid). In case of ties any link that minimizes the distance is fine.

The high-level idea of why the algorithm is successful is the following: We consider $\Theta(\ln n)$ balls around the target t of increasing radius that cover the entire grid (see Figure 4.6). We then show that to go from one ball to the immediately smaller we need $O(\ln n)$ steps in expectation. This will be due to the fact that the points between the two balls are sufficiently close, therefore the probability for a long-range link to go to a point inside the smaller ball is sufficiently large, $\Omega(1/\ln n)$. Therefore in total we can route the message in $O((\ln n)^2)$ steps in expectation.

Now we give the details. Let B_i be the ball centered on node t and has a radius of 2^i. The number of point in B_i is proportional to $(2^i)^2$, so there is a constant c such that

$$|B_i| \geq c2^{2i}. \tag{4.4}$$

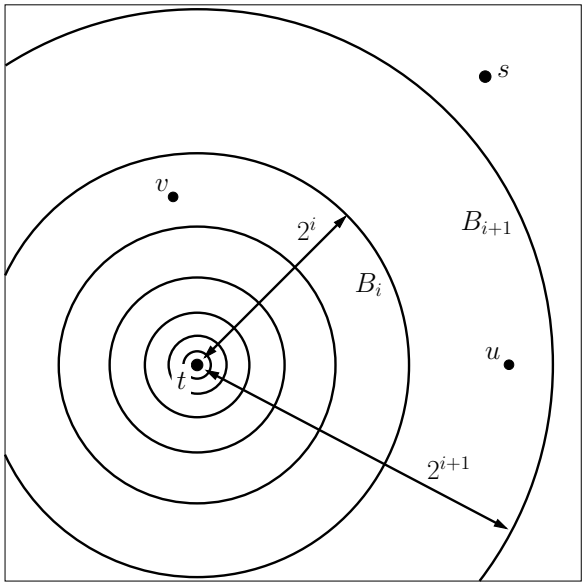


Figure 4.6: Upper bound for $r = 2$.

Note that B_i might not be a complete ball as the grid might be smaller, yet the number of points inside is still $\Omega(2^{2i})$. Since the maximum distance between two nodes in the grid is bounded by $2n$, the total number of balls needed to cover all the points is bounded by $\log(2n)$. Again we use the principle of deferred decisions. Consider a node $u \in B_{i+1} \setminus B_i$. The probability that u has a long range link to a node $v \in B_i$ equals

$$\frac{d(u,v)^{-2}}{\sum_{w \in V \setminus \{u\}} d(u,w)^{-2}},$$

so the probability that there is a long-range link from node u to some node inside ball B_i equals

$$\sum_{v \in B_i} \frac{d(u,v)^{-2}}{\sum_{w \in V \setminus \{u\}} d(u,w)^{-2}}. \tag{4.5}$$

We already have a lower bound for the size of B_i (Equation (4.4)), and we will try to lower bound the numerator and upper bound the denominator. Let us start with the latter. One can easily show that the number of points in the grid at distance j from a given node is bounded by $4j$ (it is equal to $4j$ before we reach the borders of the grid). Making use of this, we can compute[1]

$$\sum_{w \in V \setminus \{u\}} d(u,w)^{-2} \leq \sum_{j=1}^{2n} 4j \cdot j^{-2} = 4 \sum_{j=1}^{2n} \frac{1}{j} \leq 4 \ln 2n. \tag{4.6}$$

For $d(u,v)^{-2}$ notice that it achieves its minimum when $d(u,v)$ becomes maximum. We have

$$d(u,v) \leq d(u,t) + d(t,v) \leq 2^{i+1} + 2^i < 2^{i+2},$$

where the first inequality follows from the triangle inequality. Therefore, we get

$$d(u,v)^{-2} \geq 2^{-2i-4}. \tag{4.7}$$

[1]The sum $\sum_{i=1}^{n} \frac{1}{i}$ is called the nth *harmonic number* and it is denoted by H_n. It satisfies $H_n = \ln n + \Theta(1)$.

Combining Equations (4.4), (4.5), (4.6), and (4.7) we get that the probability that there is a long range from a given node $u \in B_{i+1} \setminus B_i$ to some node in B_i is at least

$$\sum_{v \in B_i} \frac{2^{-2i-4}}{4 \ln 2n} \geq c 2^{2i} \cdot \frac{2^{-2i-4}}{4 \ln 2n} = \frac{c}{64 \ln 2n} = \Omega\left(\frac{1}{\ln n}\right).$$

Therefore, the number of steps until the message reaches a node in B_{i+1} with a link to B_i is stochastically dominated by a geometric random variable with probability $\Omega(1/\ln n)$, so the expected number of steps needed to move from B_{i+1} into B_i is $O(\ln n)$.

Since the total number of balls is bounded by $\log(2n)$, as we mentioned previously, we can conclude that the expected number of steps for the message to reach t from s is $O((\ln n)^2)$.

4.4 The Barabassi-Albert Preferential Attachment Model

Another graph model that has been studied in parallel with the small-world model, actually an entire family of models, is the preferential attachment model. As a matter of fact, more than a simple graph model is what we call a *model for network growth* because it attempts to model the process with which the network is created. That is, initially it contains a small number of nodes and as time goes on new nodes and edges are added to the graph.

As we already mentioned, one of the problems of the Watts-Strogatz model is that it is to symmetric in the sense that nodes look very similar, especially as n and k grow. In particular, in the limit each node has about the same degree. The model proposed by Barabassi and Albert [1] is one of the first ones that tries to avoid this and instead creates graphs where the node degrees follow a power-law distribution, something that we have observed in practice, as we said in Section 3.2.

The main characteristic of the model is the so called *preferential attachment*, which is essentially the following concept: when a new edge is created, the probability to have node v as an endpoint is proportional to the degree d_v of node v. Thus the higher the degree already is, the higher chance it has to increase. This is what is also called the *rich-get-richer* phenomenon, and as we will soon see, it leads to power-law degree distributions. We mentioned in Section 2.3 that in many phenomena we observe power-law distributions, including the net worth of people, the number of paper citations, or city populations. If we reflect upon for a while we can see that in several of these cases we do indeed have a rich-get-richer behavior: the more property someone has the better investments he can make and the higher chances he has to increase his property; the more citations a publication has the higher chances it has to become discovered by some author and the higher is the probability for the author to cite it; cities with higher population generally offer more opportunities for labor and thus they attract more people. Therefore, this might be an explanation that many of those quantities follow power-law distributions.

Regarding social networks, we can also imagine a rich-get-richer behavior: The more friends somebody has, the more new people he potentially meets, thus the higher is the probability that he obtains more friends. In other words, nodes (people) with high degree (number of friends) tend to increase their degree more than nodes with smaller degree. It is no surprise then that in practice we observe power laws in the degree distributions, thus, naturally, scientists have studied generative models that follow the rich-get-richer principle.

The model proposed by Barabassi and Albert is essentially the following model, in which we have made some convenient assumptions when dealing with parameters that are left unspecified in the original paper [1]. We have discrete time steps, $t = 1, 2, 3, \ldots$ and a property that we maintain is that at time t the graph has t nodes and $t\ell$ edges.

- We start initially with a graph of 2 nodes, v_1 and v_2, and 2ℓ edges between them.

- At each time step $t = 3, 4, \ldots$

  1. A new node v_t is added to the graph.
  2. ℓ new edges are added. For each edge, one endpoint is v_t and the other one is selected with probability proportional to the degree at time $t-1$. In particular, node u is selected with probability
  $$\frac{d_u}{\sum_{w \in V_{t-1}} d_w} = \frac{d_u}{2(t-1)\ell}.$$

V_t is the set of nodes at time t, so $V_2 = \{v_1, v_2\}$ and $V_t = V_{t-1} \cup \{v_t\}$. To see why the equality holds, notice that at time $t-1$ we have $(t-1)\ell$ edges, so the sum of the degrees of all the nodes (except for the new node v_t) is $2(t-1)\ell$.

We now show that as $t \to \infty$ the degree distribution is a power law with exponent 3. We actually give only a heuristic argument to avoid technical details. First we define

- $n_k(t)$: mean number of nodes at time t with degree k.
- $p_k(t) = \frac{n_k(t)}{t}$: mean ratio of nodes at time t with degree k.

We will try to write a recursive equation for the values $p_k(t)$. Notice that when a new node is created, say at step $t+1$, we have ℓ new edges, and assuming that no two edges will end on the same node (clearly, this is not true especially in the beginning; however the probability of this goes to 0 as $t \to \infty$), the probability that a node, say u, of degree k increases to $k+1$ is

$$\ell \frac{d_u}{\sum_{w \in V_t} d_w} = \ell \frac{d_u}{2t\ell} = \frac{k}{2t}. \tag{4.8}$$

For degrees $k < \ell$ we have $n_k(t) = 0$ since every node has degree at least ℓ. For $k > \ell$ we can write

$$n_k(t+1) = n_k(t) + n_{k-1}(t)\frac{k-1}{2t} - n_k(t)\frac{k}{2t}. \tag{4.9}$$

To see why this is the case, note that at time t there are $n_k(t)$ nodes of degree k and $n_{k-1}(t)$ nodes of degree $k-1$. By Equation (4.8), at time $t+1$ the average number of nodes that will increase the degree to k is $n_{k-1}(t)\frac{k-1}{2t}$. Similarly, the average number of nodes that will change their degree from k to $k+1$ is $n_k(t)\frac{k}{2t}$. Now we can rewrite Equation (4.9) to

$$(t+1)p_k(t+1) = tp_k(t) + p_{k-1}(t)\frac{k-1}{2} - p_k(t)\frac{k}{2}. \tag{4.10}$$

For $k = \ell$ note that at time $t+1$ the number of nodes of degree ℓ increases by 1 due to the new node and it decreases, on average, by $n_\ell(t)\frac{\ell}{2t}$. Therefore we have

$$n_\ell(t+1) = n_\ell(t) + 1 - n_\ell(t)\frac{\ell}{2t},$$

or

$$(t+1)p_\ell(t+1) = tp_\ell(t) + 1 - p_\ell(t)\frac{\ell}{2}. \tag{4.11}$$

We want the distribution of $p_k(t)$ in the limit as $t \to \infty$, when it has converged to some value independent of t, say p_k. Then we have $p_k(t+1) = p_k(t) = p_k$ and then we can obtain from Equation (4.11)

$$(t+1)p_\ell = tp_\ell + 1 - p_\ell\frac{\ell}{2},$$

which means that

$$p_\ell = \frac{2}{\ell + 2}.$$

Similarly, Equation (4.10) gives

$$(t+1)p_k = tp_k + p_{k-1}\frac{k-1}{2} - p_k\frac{k}{2},$$

or, by rearranging the terms,

$$p_k = p_{k-1}\frac{k-1}{k+2}.$$

By induction, we can then show that for $k \geq \ell$ we have that

$$p_k = \frac{2\ell(\ell+1)}{k(k+1)(k+2)}.$$

Notice now that the numerator is a constant with respect to k, while the denumerator is approximately k^3, for large k. Therefore, the degree distribution follows approximately a power-law distribution with exponent $\gamma = 3$.

There are two obvious limitations of the model. The first is that the degree of each node is larger than ℓ, a very unrealistic assumption. Secondly, the exponent of the power law is always 3; in practice we have observed several different exponents. Therefore, we would like some a parametrized family of models that can be generate graphs with a variety of exponents. In any case the important characteristic is that the model demonstrates how a natural process can create graphs with power-law degree distributions. Furthermore, the generative process can be generalized to create graphs with arbitrary exponents and indeed there has been a large line of research that studies such generalizations.

One example of the possible generalizations is the following. Assume that the graph is generated according to the same process but with the difference that a node is selected with probability that is proportional to its degree plus a constant, that is, for some constant c (where $c \geq -\ell$), at time t, node u is selected with probability

$$\frac{d_u + c}{\sum_{w \in V_{t-1}}(d_w + c)} = \frac{d_u + c}{(t-1)(2\ell + c)}.$$

Then a similar analysis as above shows that the degree distribution follows a power-law distribution with exponent $3 + c/\ell$.

Bibliography

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[2] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[3] P. S. Dodds, R. Muhamad, and D. J. Watts. An experimental study of search in global social networks. *Science*, 301:827–829, Aug. 2003.

[4] J. Kleinberg. The small-world phenomenon: An algorithm perspective. In ACM, editor, *Proc. of the 32nd annual ACM Symposium on Theory of Computing (STOC 2000)*, pages 163–170. ACM Press, 2000.

[5] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, page 915–924, New York, NY, USA, 2008. Association for Computing Machinery.

[6] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, England, June 1995.

[7] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, Sept.-Oct. 2005.

[8] G. Pandurangan, P. Raghavan, and E. Upfal. Using pagerank to characterize web structure. *Internet Mathematics*, 3(1):1–20, 2006.

[9] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969.

[10] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.