Notes on Social Networks

Aris Anagnostopoulos

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:

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

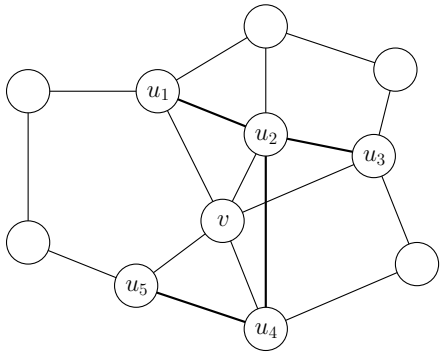


Figure 2.1: Clustering coefficient of node v. Node v has 5 neighbors and there are 4 edges between those neighbors (the bold edges). Therefore the clustering coefficient of node v is $C_v = \frac{4}{\binom{5}{2}} = 0.4$.

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

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

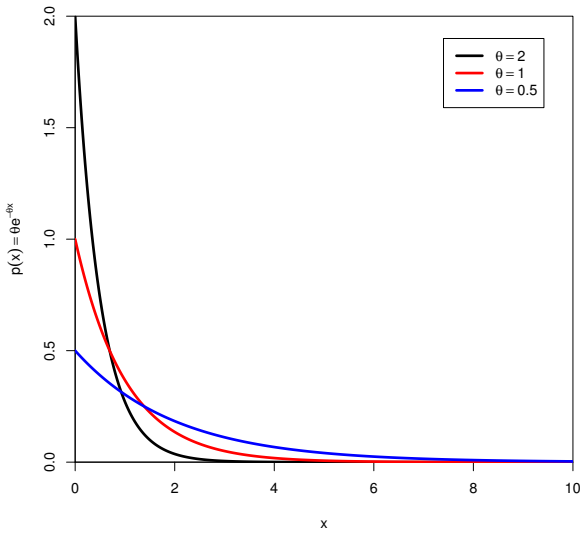


Figure 2.2: Density function for the exponential distribution

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 $\gamma > 1$.

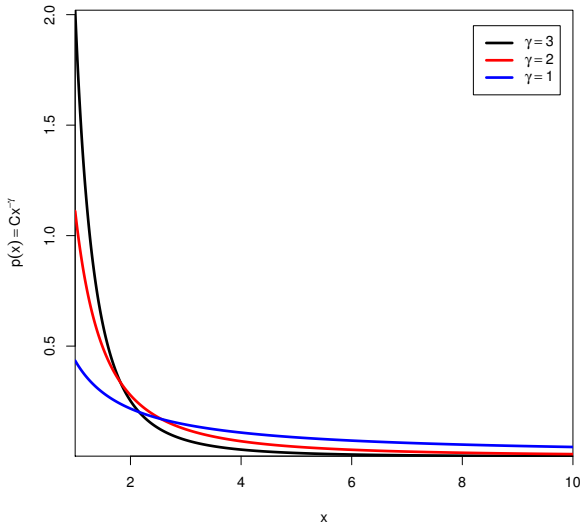


Figure 2.3: Density function for the power-law distribution

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

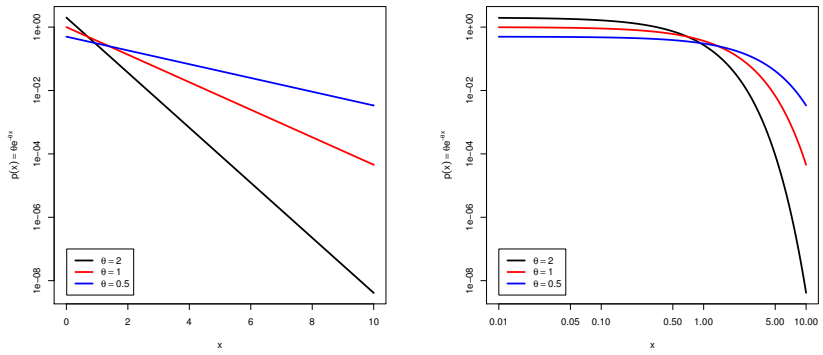
$$\ln p(x) = -\gamma \ln x + \ln C,$$

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

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

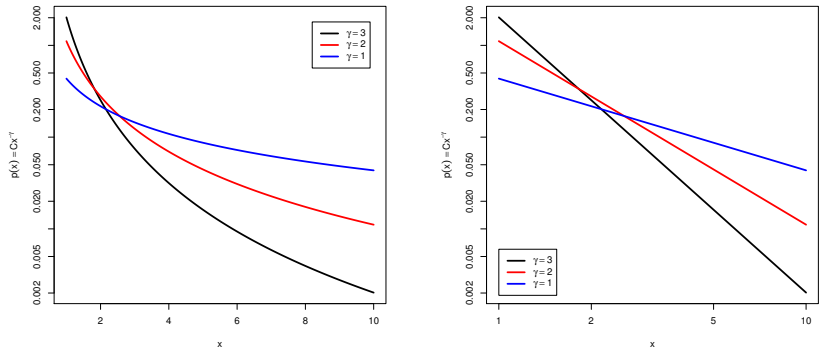
We can see that the logarithm of the density function is linear this time directly with x. Thus, if we plot the exponential distribution with only the y axis scaled logarithmically the graph is a straight line. These facts are shown in Figures 2.4 and 2.5. When we examine real data, if we plot then using 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 [5]. 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

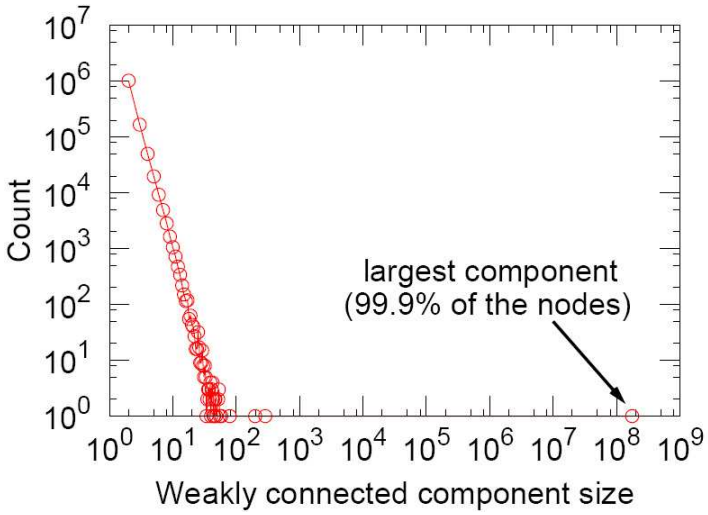
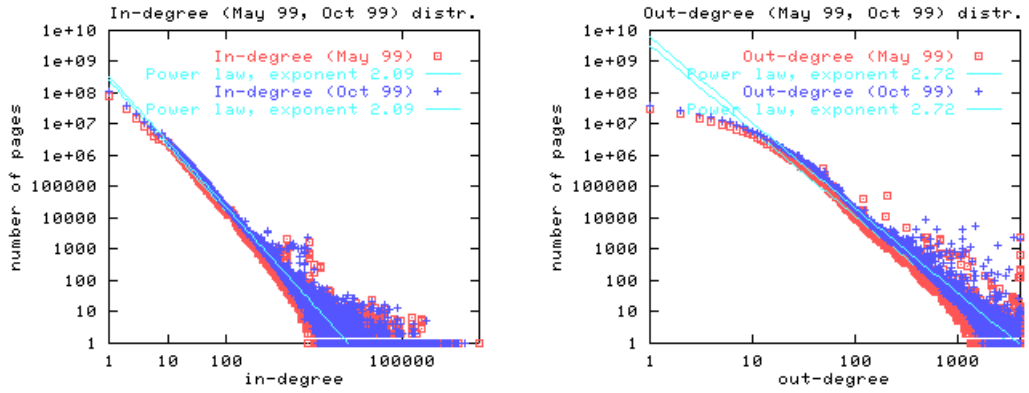


Figure 3.1: Distribution of the connected components of the MSN instant messenger communication graph.

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.



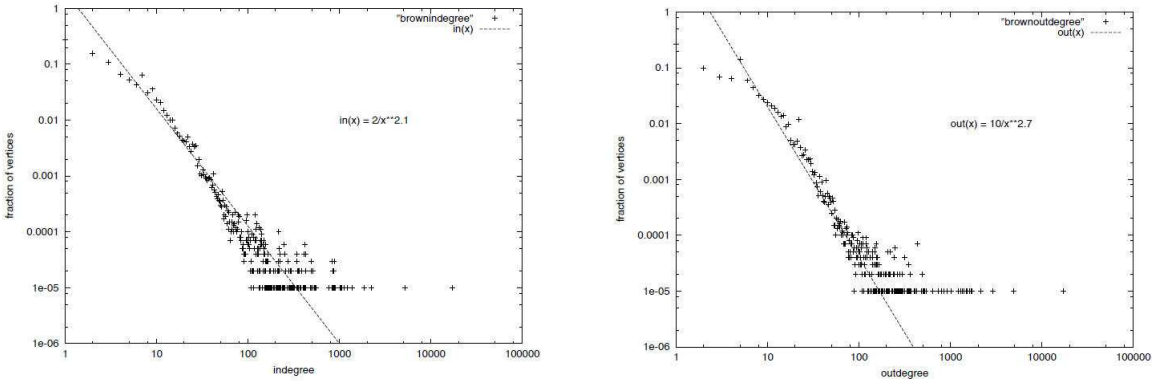
(a) Indegree.

(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].

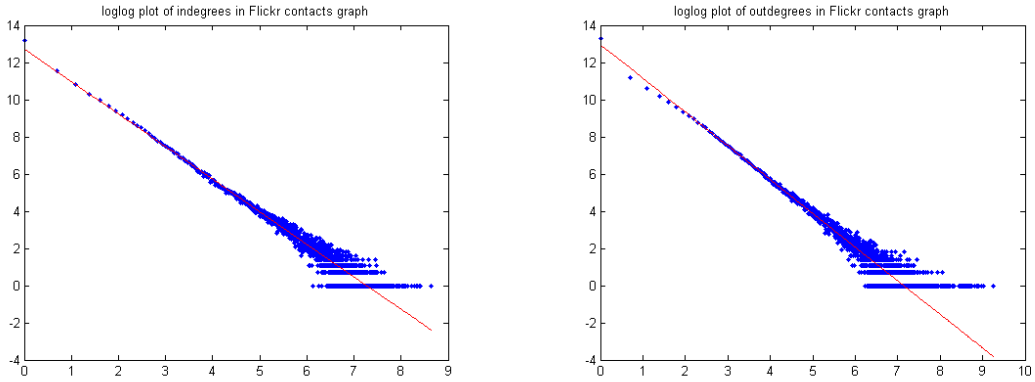
3.3 Small World

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 2,700km 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.

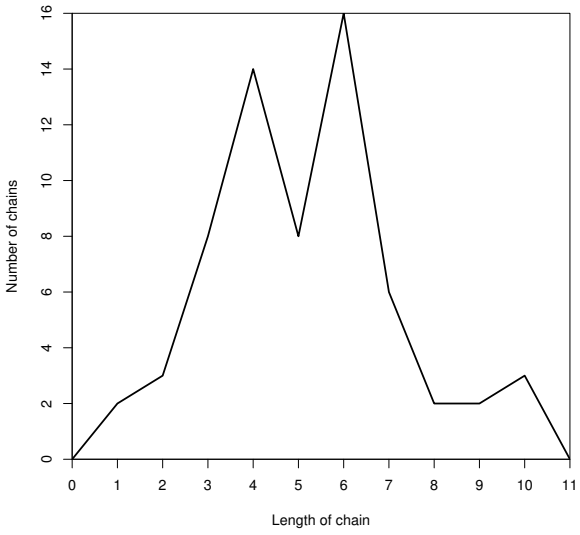


Figure 3.5: The lengths of the chains of the letters that reached their destination

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 [**?**] 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 $3/1,000,000,000$.

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] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

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

[6] P. E. os and A. Rényi. On random graphs i. *Publicationes Mathematicae*, 6:290–297, 1959.

[7] P. E. os and A. Rényi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[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.