1

Introduction

In this first chapter, we give an introduction to random graphs and complex networks. We discuss examples of real-world networks and their empirical properties, and give a brief introduction to the kinds of models that we investigate in the book. Further, we introduce the key elements of the notation used throughout the book.

1.1 Motivation

The advent of the computer age has incited an increasing interest in the fundamental properties of real-world networks. Due to the increased computational power, large data sets can now easily be stored and investigated, and this has had a profound impact on the empirical studies of large networks. As we explain in detail in this chapter, many real-world networks are small worlds and have large fluctuations in their degrees. These realizations have had fundamental implications for scientific research on networks. Network research is aimed to both to understand *why* many networks share these fascinating features, and also to investigate what the *properties* of these networks are in terms of the spread of diseases, routing information and ranking of the vertices present.

The study of complex networks plays an increasingly important role in science. Examples of such networks are electrical power grids and telecommunication networks, social relations, the World-Wide Web and Internet, collaboration and citation networks of scientists, etc. The structure of such networks affects their performance. For instance, the topology of social networks affects the spread of information or disease (see, e.g., Strogatz (2001)). The rapid evolution and success of the Internet have spurred fundamental research on the topology of networks. See Barabási (2002) and Watts (2003) for expository accounts of the discovery of network properties by Barabási, Watts and co-authors. In Newman et al. (2006), you can find some of the original papers detailing the empirical findings of real-world networks and the network models invented for them. The introductory book by Newman (2010) lists many of the empirical properties of, and scientific methods for, networks.

One common feature of complex networks is that they are *large*. As a result, a global description is utterly impossible, and researchers, both in the applications and in mathematics, have turned to their *local description*: how many vertices they have, by which *local rules* vertices connect to one another, etc. These local rules are *probabilistic*, reflecting the fact that there is a large amount of variability in how connections can be formed. Probability theory offers a highly effective way to deal with the *complexity* of networks, and leads us to consider *random graphs*. The simplest imaginable random graph is the Erdős-Rényi random

graph, which arises by taking n vertices, and placing an edge between any pair of distinct vertices with a fixed probability p, independently for all pairs. We give an informal introduction to the classical Erdős-Rényi random graph in Section 1.8. We continue with a bit of graph theory.

1.2 Graphs and Their Degree and Connectivity Structure

This book describes random graphs, so we start by discussing graphs. A graph $G = (V, E)$ consists of a collection of vertices, called vertex set, V, and a collection of edges, called edge set, E. The vertices correspond to the objects that we model, the edges indicate some relation between pairs of these objects. In our settings, graphs are usually *undirected*. Thus, an edge is an unordered pair $\{u, v\} \in E$ indicating that u and v are directly connected. When G is undirected, if u is directly connected to v, then also v is directly connected to u. Thus, an edge can be seen as a pair of vertices. When dealing with social networks, the vertices represent the individuals in the population, while the edges represent the friendships among them.

Sometimes, we also deal with *directed* graphs, where edges are indicated by the *ordered* pair (u, v). In this case, when the edge (u, v) is present, the reverse edge (v, u) need not be present. One may argue about whether friendships in social networks are directed or not. In most applications, however, it is clear whether edges are directed or not. For example, in the World-Wide Web (WWW), where vertices represent web pages, an edge (u, v) indicates that the web page u has a hyperlink to the web page v, so the WWW is a directed network. In the Internet, instead, the vertices correspond to routers, and an edge $\{u, v\}$ is present when there is a physical cable linking u and v. This cable can be used in both directions, so that the Internet is undirected.

In this book, we only consider *finite* graphs. This means that V is a finite set of size, say, $n \in \mathbb{N}$. In this case, by numbering the vertices as $1, 2, \ldots, n$, we may as well assume that $V = [n] \equiv \{1, \ldots, n\}$, which we will do from now on. A special role is played by the *complete graph* denoted by K_n, for which the edge set is every possible pair of vertices, i.e., $E = \{\{i, j\}: 1 \leq i < j \leq n\}$. The complete graph K_n is the most highly connected graph on n vertices, and every other graph may be considered to be a subgraph of K_n obtained by keeping some edges and removing the rest. Of course, infinite graphs are also of interest, but since networks are finite, we stick to finite graphs.

The *degree* d_u of a vertex u is equal to the number of edges containing u, i.e.,

$$d_u = \#\{v \in V: \{u, v\} \in E\}. \tag{1.2.1}$$

Sometimes the degree is called the *valency*. In the social networks context, the degree of an individual is the number of her/his friends. We will often be interested in the structural properties of the degrees in a network, as indicated by the collection of degrees of all vertices or the degree sequence $\boldsymbol{d} = (d_v)_{v \in [n]}$. Such properties can be described nicely in terms of the *typical degree* denoted by $D_n = d_U$, where $U \in [n]$ is a vertex chosen uniformly at random from the collection of vertices. In turn, if we draw a histogram of the proportion of vertices having degree k for all k, then this histogram is precisely equal to the probability mass function $k \mapsto \mathbb{P}(D_n = k)$ of the random variable D_n, and it represents the *empirical distribution* of the degrees in the graph.

We continue by discussing certain related degrees. For example, when we draw an edge uniformly at random from E, and choose one of its two vertices uniformly at random, this corresponds to an individual engaged in a *random friendship*. Denote the degree of the corresponding random vertex by $D_n^\star$. Note that this vertex is *not* chosen uniformly at random from the collection of vertices! In the following theorem, we describe the law of $D_n^\star$ explicitly:

Theorem 1.1 (Friends in a random friendship) *Let $G = ([n], E)$ be a finite graph with degree sequence $\boldsymbol{d} = (d_v)_{v\in[n]}$. Let $D_n^\star$ be the degree of a random element in an edge that is drawn uniformly at random from E. Then*

$$\mathbb{P}(D_n^\star = k) = \frac{k}{\mathbb{E}[D_n]}\mathbb{P}(D_n = k). \tag{1.2.2}$$

In Theorem 1.1,

$$\mathbb{E}[D_n] = \frac{1}{n}\sum_{i\in[n]} d_i \tag{1.2.3}$$

is the average degree in the graph. The representation in (1.2.2) has a nice interpretation in terms of *size-biased random variables*. For a non-negative random variable X with $\mathbb{E}[X] > 0$, we define its size-biased version $X^\star$ by

$$\mathbb{P}(X^\star \le x) = \frac{\mathbb{E}[X\mathbb{1}_{\{X\le x\}}]}{\mathbb{E}[X]}. \tag{1.2.4}$$

Then, indeed, $D_n^\star$ is the size-biased version of D_n. In particular, note that, since the variance of a random variable $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ is non-negative,

$$\mathbb{E}[D_n^\star] = \frac{\mathbb{E}[D_n^2]}{\mathbb{E}[D_n]} = \mathbb{E}[D_n] + \frac{\mathrm{Var}(D_n)}{\mathbb{E}[D_n]} \ge \mathbb{E}[D_n], \tag{1.2.5}$$

the average number of friends of an individual in a random friendship is at least as large as that of a random individual. The inequality in (1.2.5) is strict whenever the degrees are not all equal, since then $\mathrm{Var}(D_n) > 0$. In Section 2.3, we further investigate the relation between D_n and $D_n^\star$. There, we show that, in some sense, $D_n^\star \ge D_n$ with probability 1. We will make the notion $D_n^\star \ge D_n$ perfectly clear in Section 2.3.

We next extend the above result to a random friend of a random individual. For this, we define the random vector (X_n, Y_n) by drawing an individual U uniformly at random from $[n]$, and then drawing a friend Z of U uniformly at random from the d_U friends of U and letting $X_n = d_U$ and $Y_n = d_Z$.

Theorem 1.2 (Your friends have more friends than you do) *Let $G = ([n], E)$ be a finite graph with degree sequence $\boldsymbol{d} = (d_v)_{v\in[n]}$. Assume that $d_v \ge 1$ for every $v \in [n]$. Let X_n be the degree of a vertex drawn uniformly at random from $[n]$, and Y_n be the degree of a uniformly drawn neighbor of this vertex. Then*

$$\mathbb{E}[Y_n] \ge \mathbb{E}[X_n], \tag{1.2.6}$$

the inequality being strict unless all degrees are equal.

When I view myself as the random individual, I see that Theorem 1.2 has the interpretation that, on average, a random friend of mine has more friends than I do! We now give a formal proof of this fact:

Proof We note that the joint law of (X_n, Y_n) is equal to

$$\mathbb{P}(X_n = k, Y_n = l) = \frac{1}{n} \sum_{(u,v)\in E'} \mathbb{1}_{\{d_u=k, d_v=l\}} \frac{1}{d_u}, \tag{1.2.7}$$

where the sum is over all *directed* edges E', i.e., we now consider (u, v) to be a different edge than (v, u) and we notice that, given that u is chosen as the uniform vertex, the probability that its neighbor v is chosen is $1/d_u$. Clearly, $|E'| = 2|E| = \sum_{i\in[n]} d_i$. Thus,

$$\mathbb{E}[X_n] = \frac{1}{n} \sum_{(u,v)\in E'} \sum_{k,l} k \mathbb{1}_{\{d_u=k, d_v=l\}} \frac{1}{d_u} = \frac{1}{n} \sum_{(u,v)\in E'} 1, \tag{1.2.8}$$

while

$$\mathbb{E}[Y_n] = \frac{1}{n} \sum_{(u,v)\in E'} \sum_{k,l} l \mathbb{1}_{\{d_u=k, d_v=l\}} \frac{1}{d_u} = \frac{1}{n} \sum_{(u,v)\in E'} \frac{d_v}{d_u}. \tag{1.2.9}$$

We will bound $\mathbb{E}[X_n]$ from above by $\mathbb{E}[Y_n]$. For this, we note that

$$1 \le \frac{1}{2}\left(\frac{x}{y} + \frac{y}{x}\right) \tag{1.2.10}$$

for every $x, y > 0$, to obtain that

$$\mathbb{E}[X_n] \le \frac{1}{n} \sum_{(u,v)\in E'} \frac{1}{2}\left(\frac{d_u}{d_v} + \frac{d_v}{d_u}\right) = \frac{1}{n} \sum_{(u,v)\in E'} \frac{d_u}{d_v} = \mathbb{E}[Y_n], \tag{1.2.11}$$

the penultimate equality following from the symmetry in (u, v). □

After the discussion of degrees in graphs, we continue with *graph distances*. For $u, v \in [n]$ and a graph $G = ([n], E)$, we let the graph distance $\mathrm{dist}_G(u, v)$ between u and v be equal to the minimal number of edges in a path linking u and v. When u and v are not in the same connected component, we set $\mathrm{dist}_G(u, v) = \infty$. We are interested in settings where G has a high amount of connectivity, so that many pairs of vertices are connected to one another by short paths. In order to describe how large distances between vertices typically are, we draw U_1 and U_2 uniformly at random from $[n]$ and we let

$$H_n = \mathrm{dist}_G(U_1, U_2). \tag{1.2.12}$$

Often, we will consider H_n conditionally on $H_n < \infty$. This means that we consider the typical number of edges between a uniformly chosen pair of *connected* vertices. As a result, H_n is sometimes referred to as the *typical distance*. Exercise 1.1 investigates the probability that $H_n < \infty$.

Just like the degree of a random vertex D_n, H_n is also a *random variable* even when the graph G is deterministic. The nice fact is that the distribution of H_n tells us something about

all distances in the graph. An alternative and frequently used measure of distances in a graph is the *diameter* $\mathrm{diam}(G)$, defined as

$$\mathrm{diam}(G) = \max_{u,v\in[n]} \mathrm{dist}_G(u, v). \tag{1.2.13}$$

However, the diameter has several disadvantages. First, in many instances, the diameter is algorithmically more difficult to compute than the typical distances (since one has to measure the distances between all pairs of vertices and maximize over them). Second, it is a number instead of the distribution of a random variable, and therefore contains far less information than the distribution of H_n. Finally, the diameter is highly sensitive to small changes of the graph. For example, adding a small string of connected vertices to a graph may change the diameter dramatically, while it hardly influences the typical distances. As a result, in this book as well as Volume II, we put more emphasis on the typical distances. For many real-world networks, we will give plots of the distribution of H_n.

1.3 Complex Networks: the Infamous Internet Example

Complex networks have received a tremendous amount of attention in the past decades. In this section we use the Internet as an example of a real-world network to illustrate some of their properties. For an artistic impression of the Internet, see Figure 1.1.

Measurements have shown that many real-world networks share two fundamental properties: the *small-world phenomenon* roughly stating that distances in real-world networks are quite small and the *scale-free phenomenon* roughly stating that the degrees in real-world networks show an enormous amount of variability. We next discuss these two phenomena in detail.

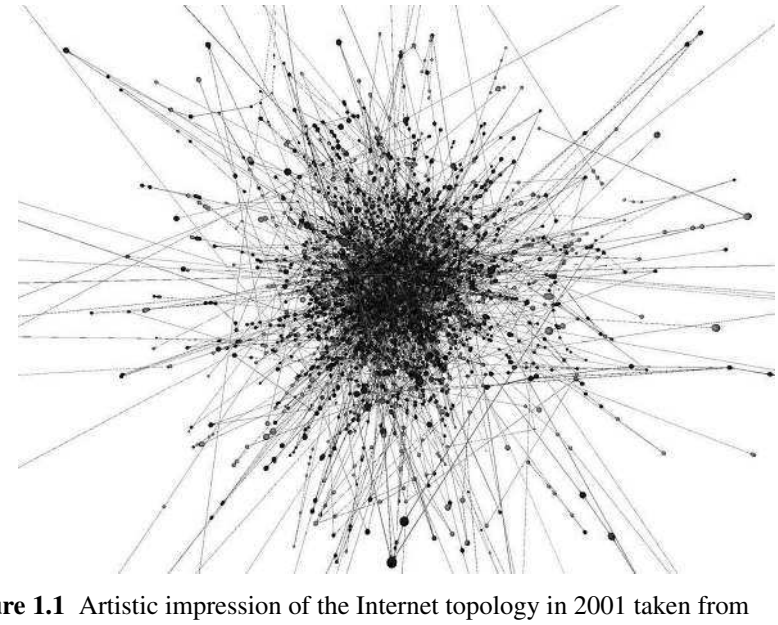

Figure 1.1 Artistic impression of the Internet topology in 2001 taken from `http://www.fractalus.com/steve/stuff/ipmap/`.

Small-World Phenomenon

The first fundamental network property is the fact that typical distances between vertices are small. This is called the 'small-world' phenomenon (see, e.g., the book by Watts (1999)). In particular, such networks are highly connected: their largest connected component contains a significant proportion of the vertices. Many networks, such as the Internet, even consist of *one* connected component, since otherwise e-mail messages could not be delivered. For example, in the Internet, IP-packets cannot use more than a threshold of physical links, and if distances in the Internet were larger than this threshold, then e-mail service would simply break down. Thus, the graph of the Internet has evolved in such a way that typical distances are relatively small, even though the Internet itself is rather large. For example, as seen in Figure 1.2(a), the number of Autonomous Systems (AS) traversed by an e-mail data set, sometimes referred to as the AS-count, is typically at most 7. In Figure 1.2(b), the proportion of routers traversed by an e-mail message between two uniformly chosen routers, referred to as the *hopcount*, is shown. It shows that the number of routers traversed is at most 27, while the distribution resembles a Poisson probability mass function.

Interestingly, various different data sets (focussing on different regional parts of the Internet) show roughly the same AS-counts. This shows that the AS-count is somewhat robust and it hints at the fact that the AS graph is relatively homogeneous. See Figure 1.3. For example, the AS-counts in North America and in Europe are quite close to the one in the entire AS graph. This implies that the dependence on geometry of the AS-count is rather weak, even though one would expect geometry to play a role. As a result, most of the models for the Internet, as well as for the AS graph, ignore geometry altogether.

Scale-Free Phenomenon

The second, maybe more surprising, fundamental property of many real-world networks is that the number of vertices with degree at least k decays slowly for large k. This implies that degrees are highly variable and that, even though the average degree is not so large, there exist vertices with extremely high degree. Often, the tail of the empirical degree distribution seems to fall off as an inverse power of k. This is called a 'power-law degree sequence', and resulting graphs often go under the name 'scale-free graphs'. It is visualized for the AS

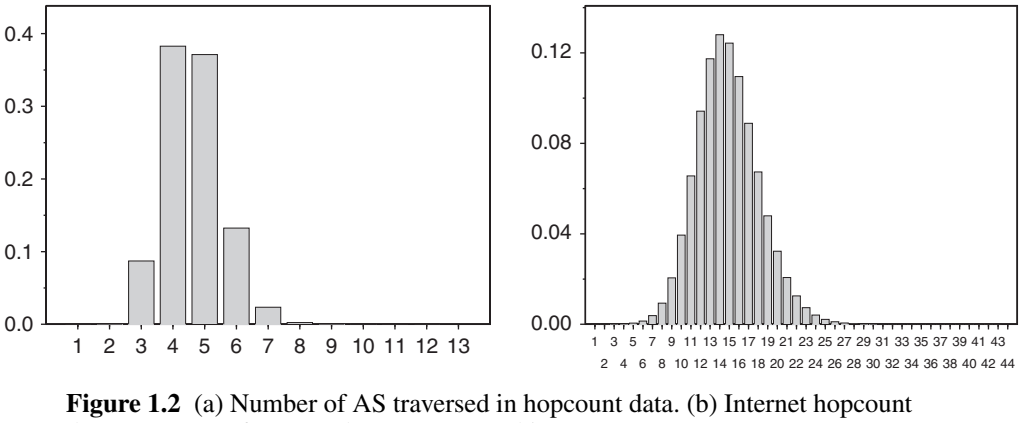

Figure 1.2 (a) Number of AS traversed in hopcount data. (b) Internet hopcount data. Courtesy of Hongsuda Tangmunarunkit.

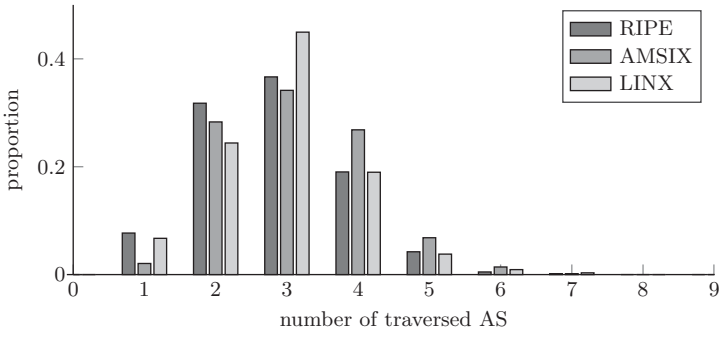

Figure 1.3 Number of AS traversed in various data sets. Data courtesy of Piet Van Mieghem.

graph in Figure 1.5, where the degree distribution of the AS graph is plotted on a log-log scale. Thus, we see a plot of $\log k \mapsto \log N_k$, where N_k is the number of vertices with degree k. When N_k is approximately proportional to an inverse power of k, i.e., when, for some normalizing constant c_n and some exponent τ,

$$N_k \approx c_n k^{-\tau}, \tag{1.3.1}$$

then

$$\log N_k \approx \log c_n - \tau \log k, \tag{1.3.2}$$

so that the plot of $\log k \mapsto \log N_k$ is close to a straight line. This is the reason why degree sequences in networks are often depicted in a log-log fashion, rather than in the more customary form of $k \mapsto N_k$. Here and in the remainder of this section, we write $\approx$ to denote an uncontrolled approximation. The power-law exponent τ can be estimated by the slope of the line in the log-log plot. Naturally, we must have that

$$\sum_k N_k = n < \infty, \tag{1.3.3}$$

so that it is reasonable to assume that $\tau > 1$.

For Internet, such log-log plots of the degrees first appeared in a paper by the Faloutsos brothers (1999) (see Figure 1.4 for the degree sequence in the Autonomous Systems graph). Here the power-law exponent is estimated as $\tau \approx 2.15 - 2.20$.

In recent years, many more Internet data sets have been collected. We particularly refer to the Center for Applied Internet Data Analysis (CAIDA) website for extensive measurements (see e.g., Krioukov et al. (2012) for a description of the data). See Figure 1.5, where the power-law exponent is now estimated as $\tau \approx 2.1$. See also Figure 1.7 for two examples of more recent measurements of the degrees of the Internet at the router or Internet Protocol (IP) level.

Measuring the Internet is quite challenging, particularly since the Internet is highly decentralized and distributed, so that a central authority is lacking. Huffaker et al. (2012) compare various data sets in terms of their coverage of the Internet and their accuracy.[1] The tool of the trade to obtain Internet data is called `traceroute`, an algorithm that allows you to send a

[1] See, in particular, `http://www.caida.org/research/topology/topo_comparison/`.

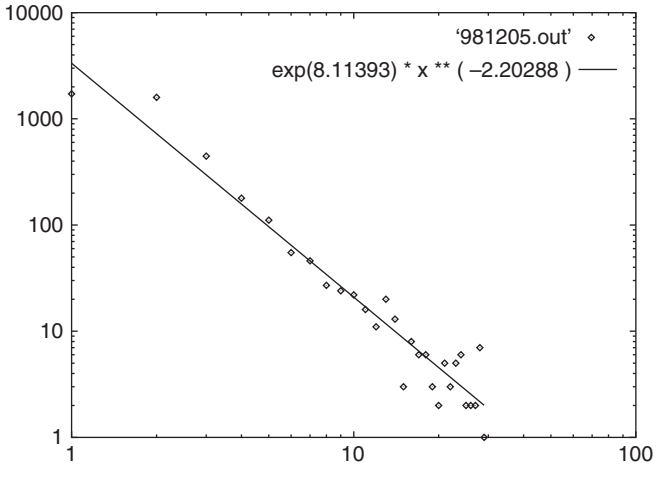

Figure 1.4 Degree sequence of Autonomous Systems (AS) on December 1998 on a log-log scale from Faloutsos, Faloutsos and Faloutsos (1999). These data suggest power-law degrees with exponent $\tau \approx 2.15 - 2.20$, the estimate on the basis of the data is 2.20288 with a multiplicative constant that is estimated as $e^{8.11393}$. This corresponds to c_n in (1.3.1).

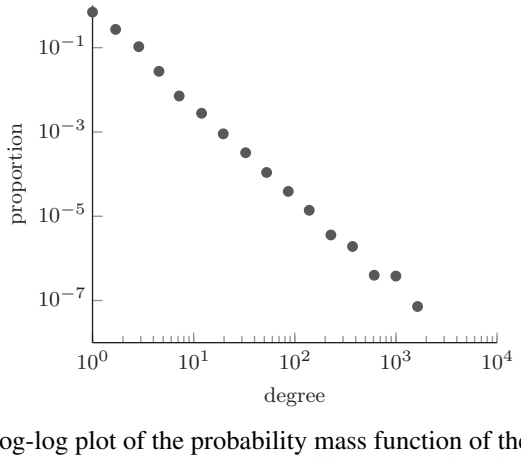

Figure 1.5 (a) Log-log plot of the probability mass function of the degree sequence of Autonomous Systems (AS) on April 2014 on a log-log scale from Krioukov et al. (2012) (data courtesy of Dmitri Krioukov). Due to the binning procedure that is applied, the figure looks smoother than many other log-log plots of degree sequences.

message between a source and a destination and to receive a list of the visited routers along the way. By piecing together many of such paths, one gets a picture of the Internet as a graph. This picture becomes more accurate when the number of sources and destinations increases, even though, as we describe in more detail below, it is not entirely understood how accurate these data sets are. In `traceroute`, also the direction of paths is obtained, and thus the graph reconstruction gives rise to a *directed* graph. The in- and out-degree sequences of this graph turn out to be quite different, as can be seen in Figure 1.6. It is highly interesting to explain such differences.

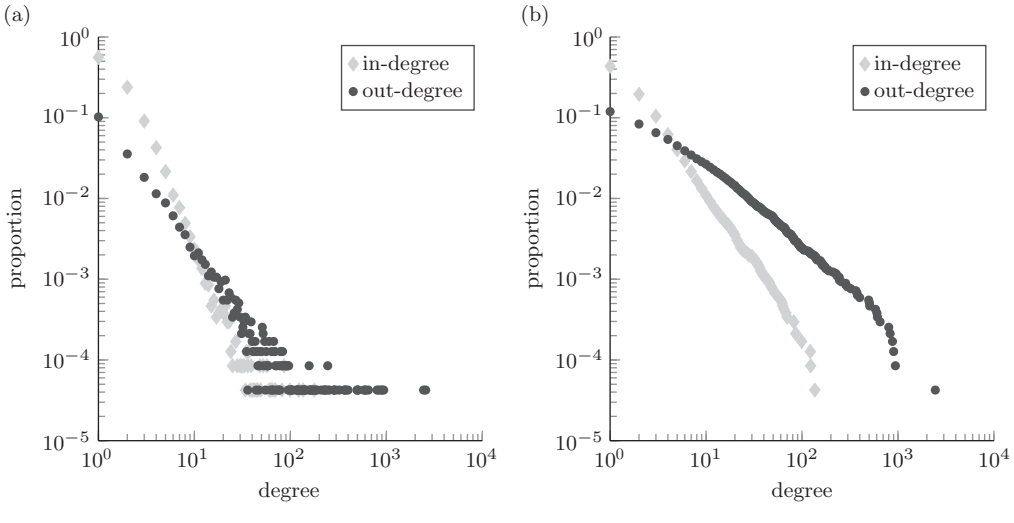

Figure 1.6 Log-log plots of in- and out-degree sequence of the Autonomous Systems graph. (a) Probability mass function. (b) Complementary cumulative distribution function.

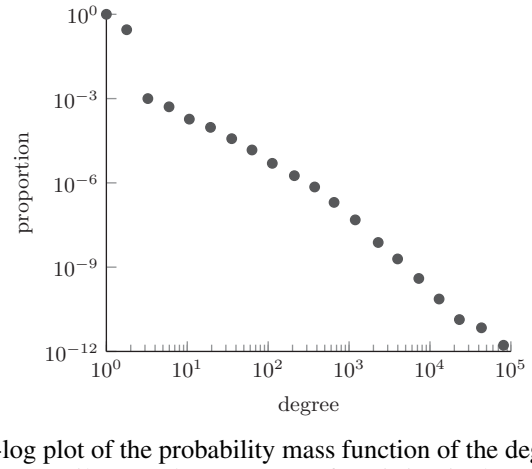

Figure 1.7 Log-log plot of the probability mass function of the degree distribution at router level from April 2014 (data courtesy of Dmitri Krioukov). The data set consists of 55,663,339 nodes and 59,685,901 edges. For a detailed explanation of how the data set was obtained, we refer to `http://www.caida.org/data/internet-topology-data-kit/`.

An interesting topic of research, receiving quite a bit of attention recently, is how the Internet behaves under malicious attacks or random breakdown (see, e.g., Albert et al. (2001) or Cohen et al. (2000, 2001)). The conclusion, based on various models for the Internet, is that the topology is critical for the vulnerability. When vertices with high degrees are taken out, the random graph models for the Internet cease to have the necessary connectivity properties. In particular, Albert et al. (2001) claim that when 2.5 percent of the Internet routers are *randomly* removed, the diameter of the Internet is unaffected, suggesting a remarkable tolerance to random attacks. Instead, when about 3 percent of the highest

degree routers are *deterministically* removed, the Internet breaks down completely. Such results would have great implications for the resilience of networks, both to random and deliberate attacks.

A critical look at the proposed models for the Internet, and particularly, the claim of power-law degree sequences and the suggestion that attachment of edges in the Internet has a preference towards high-degree vertices, was given by Willinger, Govindan, Paxson and Shenker (2002). The authors conclude that the Barabási–Albert model (as described in more detail in Chapter 8) does not model the growth of the AS or IP graph appropriately, particularly since the degrees of the receiving vertices in the AS graph are even larger than for the Barabási–Albert model.

This criticism was most vehemently argued by Willinger, Anderson and Doyle (2009), with the suggestive title 'Mathematics and the internet: A source of enormous confusion and great potential'. In this view, the problem comes from the quality of the data. Indeed, the data set on which Faloutsos et al. (1999) base their analysis, and which is used again by Albert et al. (2001) to investigate the resilience properties of the Internet, was collected by Pansiot and Grad (1998) in order to study the efficiency of multicast versus unicast, which are different ways to send packages. The data was collected using `traceroute`, a tool that was not designed to be used to reconstruct the Internet as a graph. Pansiot and Grad realized that their way of reconstructing the Internet graph had some problems, and they write, 'We mention some problems we found in tracing routes, and we discuss the realism of the graph we obtained.' However, the Faloutsos brothers (1999) simply used the Pansiot–Grad data set, and took it at face value. This was then repeated by Albert, Jeong and Barabási (2001), which puts their results in a somewhat different light.

We now give some details of the problems with the data sets. Readers who are eager to continue can skip this part and move to the next section. Let us follow Willinger et al. (2009) to discuss the difficulties in using `traceroute` data to reconstruct a graph, which are threefold:

IP Alias Resolution Problem. A fundamental problem is that `traceroute` reports so-called *input interfaces*. Internet routers, the nodes of the Internet graph, may consist of several input interfaces, and it is a non-trivial problem to map these interfaces to routers. When errors are made in this procedure, the data does not truthfully represent the connectivity structure of the Internet routers.

Opaque Layer-2 Clouds. The Internet consists of different layers that facilitate the interoperability between heterogeneous network topologies. Since `traceroute` acts on layer-3, it is sometimes unable to trace through layer-2 clouds. This means that the internal connectivity structure of a larger unit of routers in layer-2 could be invisible for `traceroute`, so that `traceroute` shows connections between many, or even all, of these routers, even though most of these connections actually do not exist. This causes routers to be wrongfully assigned a very high degree.

Measuring Biases. Due to the way `traceroute` data is collected, an incomplete picture of the Internet is obtained, since only connections between routers that are actually being used by the data are reported. When this data set would be unbiased, a truthful picture of the Internet could still be obtained. Unfortunately, routers with a high degree