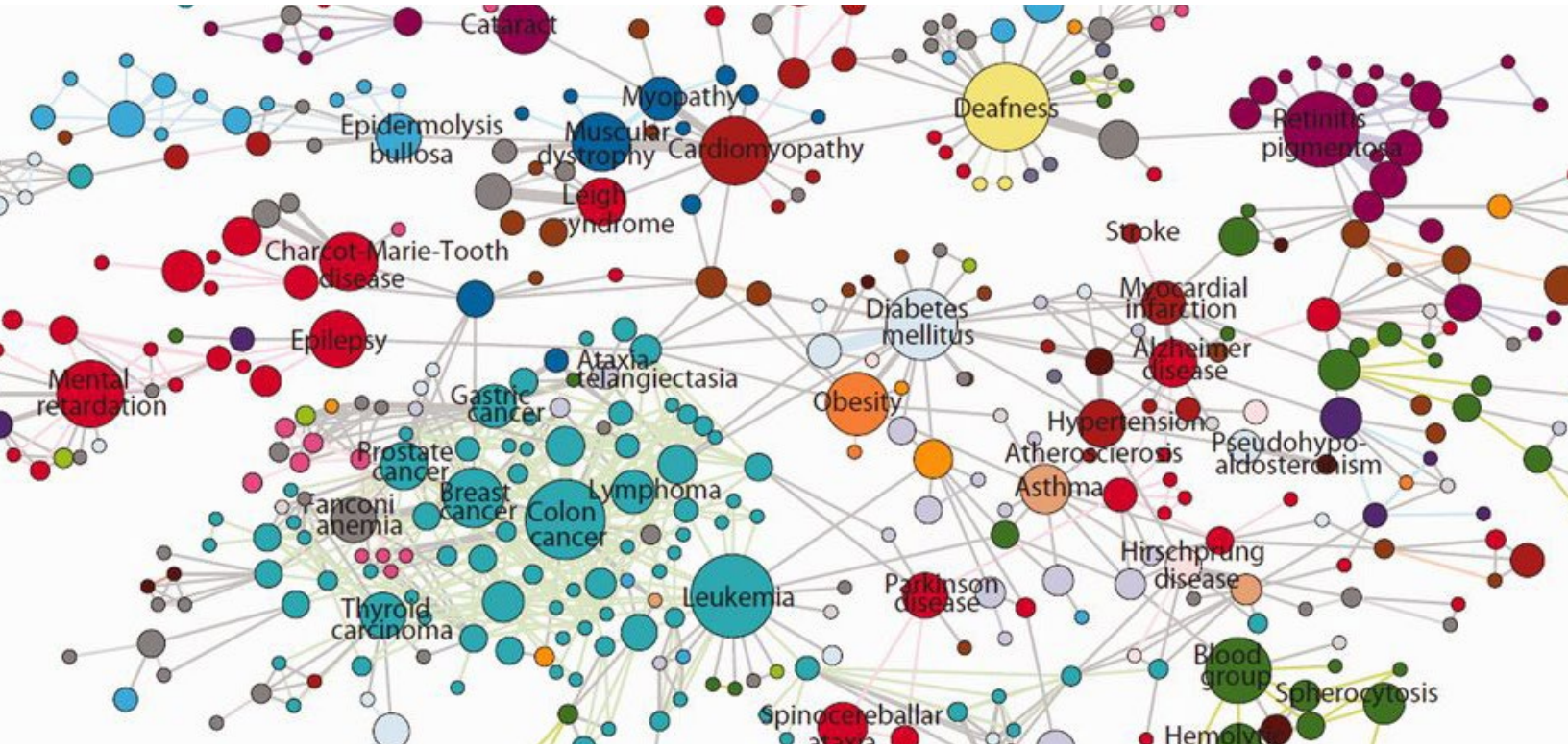
2

ALBERT-LÁSZLÓ BARABÁSI

# NETWORK SCIENCE
# GRAPH THEORY




**ACKNOWLEDGEMENTS**

MÁRTON PÓSFAI
GABRIELE MUSELLA
MAURO MARTINO
ROBERTA SINATRA
PHILIPP HOEVEL
SARAH MORRISON
AMAL HUSSEINI

**Figure 2.0 (front cover)**
**Human Disease Network**

The Human Disease Network, whose nodes are diseases connected if they have common genetic origin. Published as a supplement of the Proceedings of the National Academy of Sciences [1], the map was created to illustrate the genetic interconnectedness of apparently distinct diseases. With time it crossed disciplinary boundaries, taking up a life of its own. The New York Times created an interactive version of the map and the London-based Serpentine Gallery, one of the top contemporary art galleries in the world, have exhibited it part of their focus on networks and maps [2]. It is also featured in numerous books on design and maps [3, 4, 5].


**PDF V27, 05.09.2014**

# THE BRIDGES OF KÖNIGSBERG

Few research fields can trace their birth to a single moment and place in history. Graph theory, the mathematical scaffold behind network science, can. Its roots go back to 1735 in Königsberg, the capital of Eastern Prussia, a thriving merchant city of its time. The trade supported by its busy fleet of ships allowed city officials to build seven bridges across the river Pregel that surrounded the town. Five of these connected to the mainland the elegant island Kneiphof, caught between the two branches of the Pregel. The remaining two crossed the two branches of the river (Figure 2.1). This peculiar arrangement gave birth to a contemporary puzzle: Can one walk across all seven bridges and never cross the same one twice? Despite many attempts, no one could find such path. The problem remained unsolved until 1735, when Leonard Euler, a Swiss born mathematician, offered a rigorous mathematical proof that such path does not exist [6, 7].

Euler represented each of the four land areas separated by the river with letters A, B, C, and D (Figure 2.1). Next he connected with lines each piece of land that had a bridge between them. He thus built a graph, whose nodes were pieces of land and links were the bridges. Then Euler made a simple observation: if there is a path crossing all bridges, but never the same bridge twice, then nodes with odd number of links must be either the starting or the end point of this path. Indeed, if you arrive to a node with an odd number of links, you may find yourself having no unused link for you to leave it.

A walking path that goes through all bridges can have only one starting and one end point. Thus such a path cannot exist on a graph that has more than two nodes with an odd number of links. The Königsberg graph had four nodes with an odd number of links, A, B, C, and D, so no path could satisfy the problem.

Euler's proof was the first time someone solved a mathematical problem using a graph. For us the proof has two important messages: The first is that some problems become simpler and more tractable if they are represented as a graph. The second is that the existence of the path does not

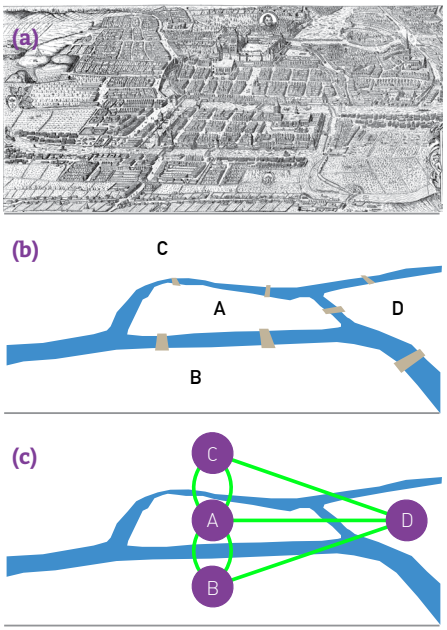



**Figure 2.1**
**The Bridges of Königsberg**

**(a)** A contemporary map of Königsberg (now Kaliningrad, Russia) during Euler's time.

**(b)** A schematic illustration of Königsberg's four land pieces and the seven bridges across them.

**(c)** Euler constructed a graph that has four nodes (A, B, C, D), each corresponding to a patch of land, and seven links, each corresponding to a bridge. He then showed that there is no continuous path that would cross the seven bridges while never crossing the same bridge twice. The people of Königsberg gave up their fruitless search and in 1875 built a new bridge between B and C, increasing the number of links of these two nodes to four. Now only one node was left with an odd number of links. Consequently we should be able to find the desired path. Can you find one yourself?

depend on our ingenuity to find it. Rather, it is a property of the graph. Indeed, given the structure of the Königsberg graph, no matter how smart we are, we will never find the desired path. In other words, networks have properties encoded in their structure that limit or enhance their behavior.

To understand the many ways networks can affect the properties of a system, we need to become familiar with graph theory, a branch of mathematics that grew out of Euler's proof. In this chapter we learn how to represent a network as a graph and introduce the elementary characteristics of networks, from degrees to degree distributions, from paths to distances and learn to distinguish weighted, directed and bipartite networks. We will introduce a graph-theoretic formalism and language that will be used throughout this book.


**Online Resource 2.1**

**The Bridges of Königsberg**

Watch a short video introducing the Königsberg problem and Euler's solution.

# NETWORKS AND GRAPHS

If we want to understand a complex system, we first need to know how its components interact with each other. In other words we need a map of its wiring diagram. A network is a catalog of a system's components often called *nodes* or *vertices* and the direct interactions between them, called *links* or *edges* (BOX 2.1). This network representation offers a common language to study systems that may differ greatly in nature, appearance, or scope. Indeed, as shown in Figure 2.2, three rather different systems have exactly the same network representation.

Figure 2.2 introduces two basic network parameters:

*Number of nodes*, or $N$, represents the number of components in the system. We will often call $N$ the *size of the network*. To distinguish the nodes, we label them with $i = 1, 2, \ldots, N$.

*Number of links*, which we denote with $L$, represents the total number of interactions between the nodes. Links are rarely labeled, as they can be identified through the nodes they connect. For example, the (2, 4) link connects nodes 2 and 4.

The networks shown in Figure 2.2 have $N = 4$ and $L = 4$.

The links of a network can be *directed* or *undirected*. Some systems have directed links, like the WWW, whose uniform resource locators (URL) point from one web document to the other, or phone calls, where one person calls the other. Other systems have undirected links, like romantic ties: if I date Janet, Janet also dates me, or like transmission lines on the power grid, on which the electric current can flow in both directions.

A network is called directed (or *digraph*) if all of its links are directed; it is called undirected if all of its links are undirected. Some networks simultaneously have directed and undirected links. For example in the metabolic network some reactions are reversible (i.e., bidirectional or undirected) and others are irreversible, taking place in only one direction (directed).

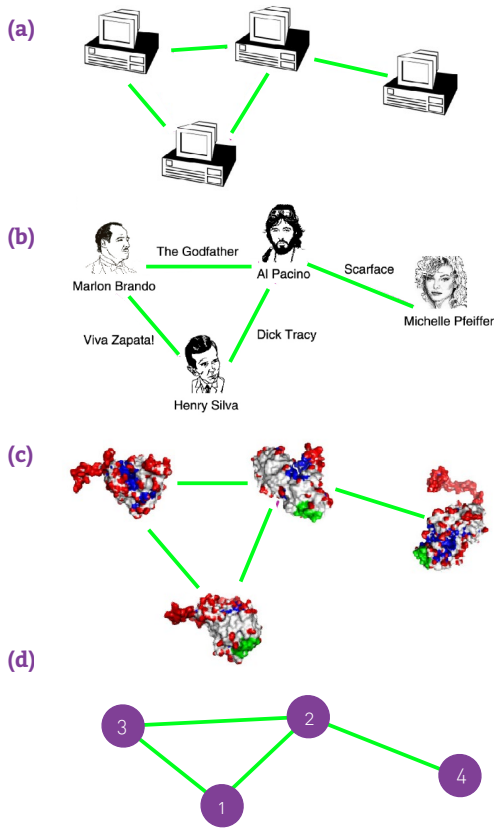



**Figure 2.2**
**Different Networks, Same Graph**

The figure shows a small subset of (a) the Internet, where routers (specialized computers) are connected to each other; (b) the Hollywood actor network, where two actors are connected if they played in the same movie; (c) a protein-protein interaction network, where two proteins are connected if there is experimental evidence that they can bind to each other in the cell. While the nature of the nodes and the links differs, these networks have the same graph representation, consisting of $N = 4$ nodes and $L = 4$ links, shown in (*d*).

The choices we make when we represent a system as a network will determine our ability to use network science successfully to solve a particular problem. For example, the way we define the links between two individuals dictates the nature of the questions we can explore:

**(a)** By connecting individuals that regularly interact with each other in the context of their work, we obtain the *organizational* or *professional network*, that plays a key role in the success of a company or an institution, and is of major interest to organizational research (Figure 1.7).

**(b)** By linking friends to each other, we obtain the *friendship network*, that plays an important role in the spread of ideas, products and habits and is of major interest to sociology, marketing and health sciences.

**(c)** By connecting individuals that have an intimate relationship, we obtain the *sexual network*, of key importance for the spread of sexually transmitted diseases, like AIDS, and of major interest for epidemiology.

**(d)** By using phone and email records to connect individuals that call or email each other, we obtain the *acquaintance network*, capturing a mixture of professional, friendship or intimate links, of importance to communications and marketing.

While many links in these four networks overlap (some coworkers may be friends or may have an intimate relationship), these networks have different uses and purposes.

We can also build networks that may be valid from a graph theoretic perspective, but may have little practical utility. For example, if we link all individuals with the same first name, Johns with Johns and Marys with Marys, we do obtain a well-defined graph, whose properties can be analyzed with the tools of network science. Its utility is questionable, however. Hence in order to apply network theory to a system, careful considerations must precede our choice of nodes and links, ensuring their significance to the problem we wish to explore.

Throughout this book we will use ten networks to illustrate the tools of network science. These *reference networks*, listed in Table 2.1, span social systems (mobile call graph or email network), collaboration and affiliation networks (science collaboration network, Hollywood actor network), information systems (WWW), technological and infrastructural systems (Internet and power grid), biological systems (protein interaction and metabolic network), and reference networks (citations). They differ widely in their sizes, from as few as $N$ =1,039 nodes in the E. coli metabolism, to almost half million nodes in the citation network. They cover several areas where networks are actively applied, representing 'canonical' datasets frequently

## BOX 2.1

### NETWORKS OR GRAPHS?

In the scientific literature the terms *network* and *graph* are used interchangeably:

| Network Science | Graph Theory |
|---|---|
| Network | Graph |
| Node | Vertex |
| Link | Edge |

Yet, there is a subtle distinction between the two terminologies: the {*network, node, link*} combination often refers to real systems: The WWW is a network of web documents linked by URLs; society is a network of individuals linked by family, friendship or professional ties; the metabolic network is the sum of all chemical reactions that take place in a cell. In contrast, we use the terms {*graph, vertex, edge*} when we discuss the mathematical representation of these networks: We talk about the web graph, the social graph (a term made popular by Facebook), or the metabolic graph. Yet, this distinction is rarely made, so these two terminologies are often synonyms of each other.

used by researchers to illustrate key network properties. As we indicate in Table 2.1, some of them are directed, others are undirected. In the coming chapters we will discuss in detail the nature and the characteristics of each of these datasets, turning them into the guinea pigs of our journey to understand complex networks.

| NETWORK | NODES | LINKS | DIRECTED UNDIRECTED | N | L | ⟨k⟩ |
|---|---|---|---|---|---|---|
| Internet | Routers | Internet connections | Undirected | 192,244 | 609,066 | 6.34 |
| WWW | Webpages | Links | Directed | 325,729 | 1,497,134 | 4.60 |
| Power Grid | Power plants, transformers | Cables | Undirected | 4,941 | 6,594 | 2.67 |
| Mobile Phone Calls | Subscribers | Calls | Directed | 36,595 | 91,826 | 2.51 |
| Email | Email addresses | Emails | Directed | 57,194 | 103,731 | 1.81 |
| Science Collaboration | Scientists | Co-authorship | Undirected | 23,133 | 93,439 | 8.08 |
| Actor Network | Actors | Co-acting | Undirected | 702,388 | 29,397,908 | 83.71 |
| Citation Network | Paper | Citations | Directed | 449,673 | 4,689,479 | 10.43 |
| E. Coli Metabolism | Metabolites | Chemical reactions | Directed | 1,039 | 5,802 | 5.58 |
| Protein Interactions | Proteins | Binding interactions | Undirected | 2,018 | 2,930 | 2.90 |

**Table 2.1**
**Canonical Network Maps**

The basic characteristics of ten networks used throughout this book to illustrate the tools of network science. The table lists the nature of their nodes and links, indicating if links are directed or undirected, the number of nodes ($N$) and links ($L$), and the average degree for each network. For directed networks the average degree shown is the average in- or out-degrees $\langle k \rangle = \langle k_{in} \rangle = \langle k_{out} \rangle$ (see Equation (2.5)).

# DEGREE, AVERAGE DEGREE, AND DEGREE DISTRIBUTION

A key property of each node is its *degree*, representing the number of links it has to other nodes. The degree can represent the number of mobile phone contacts an individual has in the call graph (i.e. the number of different individuals the person has talked to), or the number of citations a research paper gets in the citation network.

## DEGREE

We denote with $k_i$ the degree of the $i^{\text{th}}$ node in the network. For example, for the undirected networks shown in Figure 2.2 we have $k_1$=2, $k_2$=3, $k_3$=2, $k_4$=1. In an undirected network the *total number of links*, $L$, can be expressed as the sum of the node degrees:

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i \,. \tag{2.1}$$

Here the 1/2 factor corrects for the fact that in the sum (2.1) each link is counted twice. For example, the link connecting the nodes 2 and 4 in Figure 2.2 will be counted once in the degree of node 1 and once in the degree of node 4.

## AVERAGE DEGREE

An important property of a network is its *average degree* (BOX 2.2), which for an undirected network is

$$\langle k \rangle = \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N} \,. \tag{2.2}$$

In directed networks we distinguish between *incoming degree*, $k_i^{\text{in}}$, representing the number of links that point to node $i$, and *outgoing degree*, $k_i^{\text{out}}$, representing the number of links that point from node $i$ to other nodes. Finally, a node's *total degree*, $k_i$, is given by

$$k_i = k_i^{\text{in}} + k_i^{\text{out}}. \tag{2.3}$$

For example, on the WWW the number of pages a given document points to represents its outgoing degree, $k^{\text{out}}$, and the number of documents that point to it represents its incoming degree, $k^{\text{in}}$. The total number

## BOX 2.2

### BRIEF STATISTICS REVIEW

Four key quantities characterize a sample of $N$ values $x_1, \ldots, x_N$:

*Average (mean)*:

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

*The $n^{th}$ moment*:

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i^n$$

*Standard deviation*:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}$$

*Distribution of x*:

$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i} \,.$$

where $p_x$ follows

$$\sum_i p_x = 1 \left(\int p_x \, dx = 1\right)$$

of links in a directed network is

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} . \tag{2.4}$$

The 1/2 factor seen in (2.1) is now absent, as for directed networks the two sums in (2.4) separately count the outgoing and the incoming degrees. The average degree of a directed network is

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \frac{L}{N} \tag{2.5}$$

## DEGREE DISTRIBUTION

The *degree distribution*, $p_k$, provides the probability that a randomly selected node in the network has degree $k$. Since $p_k$ is a probability, it must be normalized, i.e.

$$\sum_{k=1}^{\infty} p_k = 1 . \tag{2.6}$$

For a network with $N$ nodes the degree distribution is the normalized histogram (Figure 2.3) is given by

$$p_k = \frac{N_k}{N} , \tag{2.7}$$

where $N_k$ is the number of degree-$k$ nodes. Hence the number of degree-$k$ nodes can be obtained from the degree distribution as $N_k = Np_k$.

The degree distribution has assumed a central role in network theory following the discovery of scale-free networks [8]. One reason is that the calculation of most network properties requires us to know $p_k$. For example, the average degree of a network can be written as

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k . \tag{2.8}$$

The other reason is that the precise functional form of $p_k$ determines many network phenomena, from network robustness to the spread of viruses.

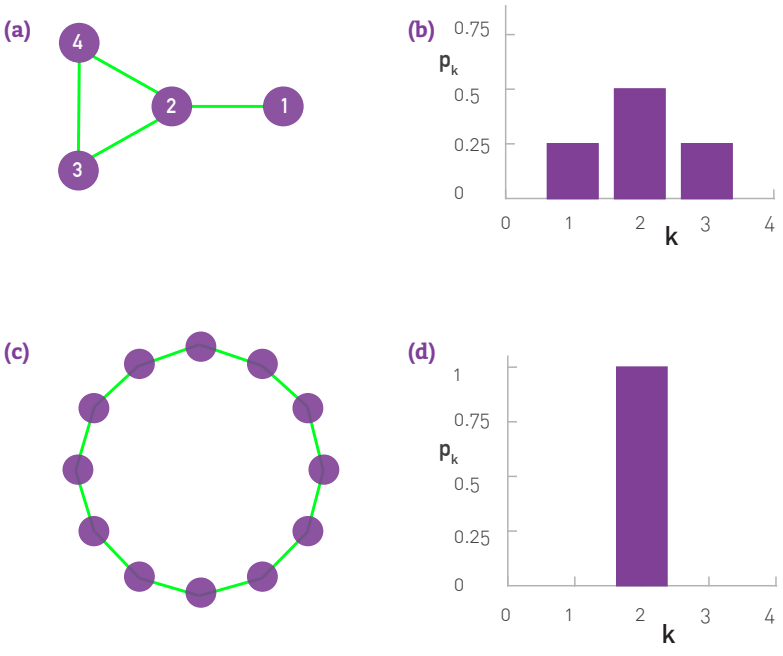



**Figure 2.3**
**Degree Distribution**

The degree distribution of a network is provided by the ratio (2.7).

**(a)** For the network in (a) with $N = 4$ the degree distribution is shown in (b).

**(b)** We have $p_1 = 1/4$ (one of the four nodes has degree $k_1 = 1$), $p_2 = 1/2$ (two nodes have $k_3 = k_4 = 2$), and $p_3 = 1/4$ (as $k_2 = 3$). As we lack nodes with degree $k > 3$, $p_k = 0$ for any $k > 3$.

**(c)** A one dimensional lattice for which each node has the same degree $k = 2$.

**(d)** The degree distribution of (c) is a Kronecker's delta function, $p_k = \delta(k - 2)$.

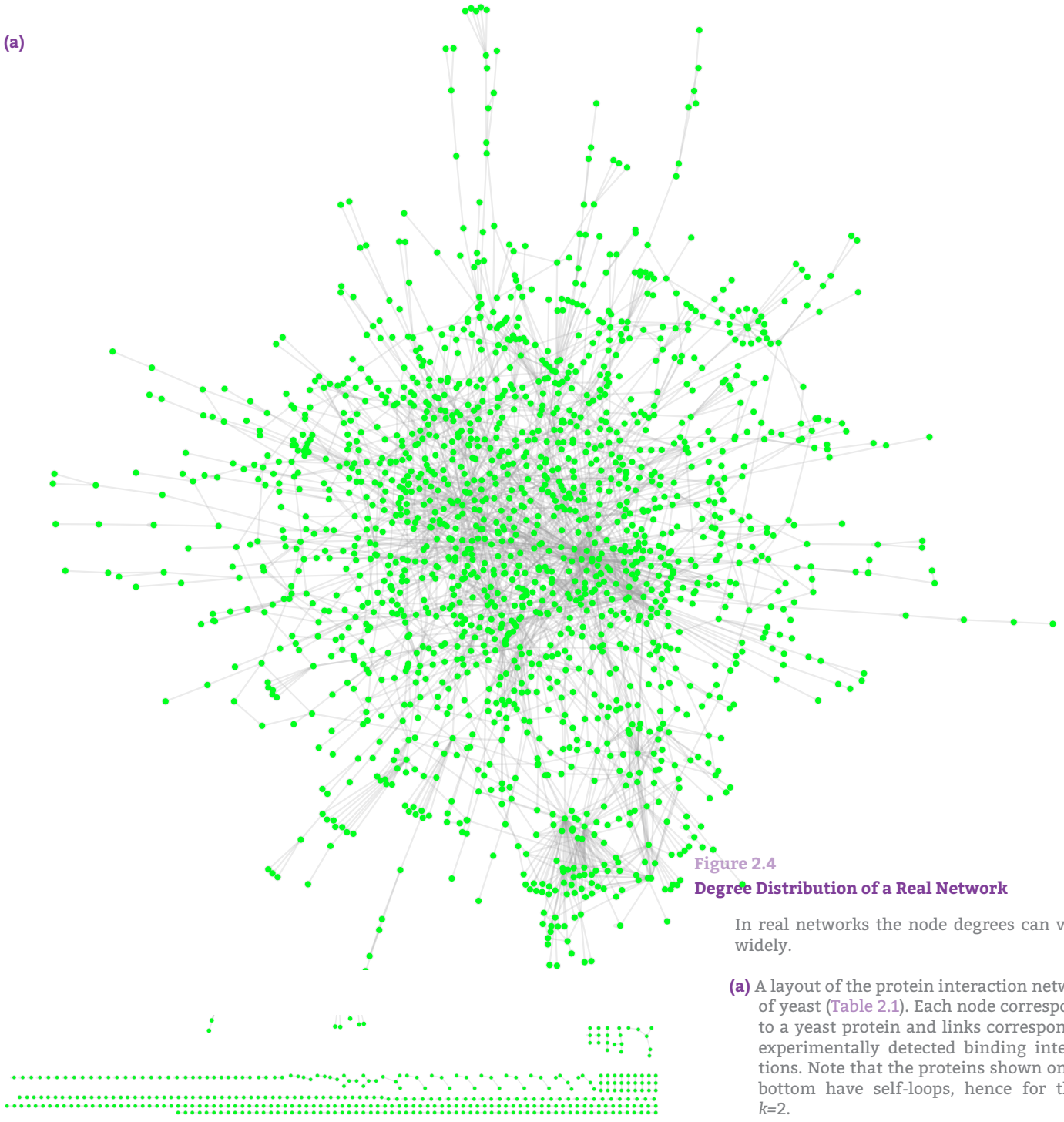



**Figure 2.4**
**Degree Distribution of a Real Network**

In real networks the node degrees can vary-widely.

**(a)** A layout of the protein interaction network of yeast (Table 2.1). Each node corresponds to a yeast protein and links correspond to experimentally detected binding interactions. Note that the proteins shown on the bottom have self-loops, hence for them $k$=2.

**(b)** The degree distribution of the protein interaction network shown in (a). The observed degrees vary between $k$=0 (isolated nodes) and $k$=92, which is the degree of the most connected node, called a *hub*. There are also wide differences in the number of nodes with different degrees: Almost half of the nodes have degree one (i.e. $p_1$=0.48), while we have only one copy of the biggest node (i.e. $p_{92} = 1/N$=0.0005).

**(c)** The degree distribution is often shown on a log-log plot, in which we either plot log $p_k$ in function of ln $k$, or, as we do in (c), or we use logarithmic axes. The advantages of this representation are discussed in Chapter 4.