Complex Graphs and Networks

Fan Chung
University of California at San Diego
La Jolla, California 92093
fan@ucsd.edu

Linyuan Lu
University of South Carolina,
Columbia, South Carolina 29208
lu@math.sc.edu

Contents

Preface

In many ways, working on graph theory problems over the years has always seemed like fun and games. Recently, through examples of large sparse graphs in realistic networks, research in graph theory has been forging ahead into an exciting new dimension. Graph theory has emerged as a primary tool for detecting numerous hidden structures in various information networks, including Internet graphs, social networks, biological networks, or more generally, any graph representing relations in massive data sets.

How will we explain from first principles the universal and ubiquitous coherence in the structure of these realistic but complex networks? In order to analyze these large sparse graphs we will need to use all the tools at our disposal, including combinatorial, probabilistic and spectral methods. Time and again, we have been pushed beyond the limit of the existing techniques and have had to create new and better tools to be able to analyze these networks. The examples of these networks have led us to focus on new, general and powerful ways to look at graph theory. In the other direction, we hope that these new perspectives on graph theory contribute to a sound scientific foundation for our understanding of the discrete networks that permeate this information age.

This book is based on ten lectures given at the *CBMS Workshop on the Combinatorics of Large Sparse Graphs* in June 2004 at the California State University at San Marcos. Various portions of the twelve chapters here are based on several papers coauthored with many collaborators. Indeed, to deal with the numerous leads in such an emerging area it is crucial to have partners to sound out the right approaches, to separate what can be rigorously proved and under what conditions from what cannot be proved, to face seemingly overwhelming obstacles and yet still gather enough energy to overcome one more challenge. Special thanks are due to our coauthors, including Bill Aiello, Reid Andersen, David Galas, Greg Dewey, Shirin Handjani, Doug Jungreis, and Van Vu.

We are particularly grateful to Ross Richardson and Reid Andersen for many beautiful illustrations in the book and to the students in Math261 spring 2004 at UCSD for taking valuable lecture notes. In the course of writing, we have greatly benefitted from discussions with Alan Frieze, Joe Buhler and Herb Wilf. Most of all, we are indebted to Steve Butler and Ron Graham for their thoughtful readings and invaluable comments without which this book would not have so swiftly converged.

Fan Chung and Lincoln Lu, May 2006

CHAPTER 1

Graph Theory in the Information Age

1.1. Introduction

Graph theory has a history dating back more than 250 years (starting with Leonhard Euler and his quest for a walk linking seven bridges in Königsberg [**17**]). Since then, graph theory, the study of networks in their most basic form as interconnections among objects, has evolved from its recreational roots into a rich and distinct subject. Of particular significance is its vital role in our understanding of the mathematics governing the discrete universe.

Throughout the years, graph theorists have been studying various types of graphs, such as planar graphs (drawn without edges crossing in the plane), interval graphs (arising in scheduling), symmetric graphs (hypercubes, platonic solids and those from group theory), routing networks (from communications) and computational graphs that are used in designing algorithms or simulations.

In 1999, at the dawn of the new Millennium, a most surprising type of graph was uncovered. Indeed, its universal importance has brought graph theory to the heart of a new paradigm of science in this information age. This family of graphs consists of a wide collection arising from diverse arenas but having completely unexpected coherence. Examples include the WWW-graphs, the phone graphs, the email graphs, the so-called "Hollywood" graphs of costars, the "collaboration" graph of coauthors, as well as legions of others from all branches of natural, social and life sciences. The prevailing characteristics of these graphs are the following:

- **Large** — The size of the network typically ranges from hundreds of thousands to billions of vertices. Brute force approaches are no longer feasible. Mathematical wizardry is in demand again — how can we use a relatively small number of parameters to capture the shape of the network?
- **Sparse** — The number of edges is *linear*, i.e., within a small multiple of the number of vertices. There might be *dense* graphs (having a quadratic number of edges in terms of vertices) out there but the large graphs that we encounter are mostly sparse.
- **The small world phenomenon** — This is used to refer to two distinct properties: *small distance* (two strangers are typically joined by a short chain of mutual acquaintances), and *the clustering effect* (two people who share a common neighbor are more likely to know each other)

- **Power law degree distribution** — The degree of a vertex is the number of vertices adjacent to it. The power law asserts that the number of vertices with degree k is proportional to $k^{-\beta}$ for some exponent $\beta \geq 1$.

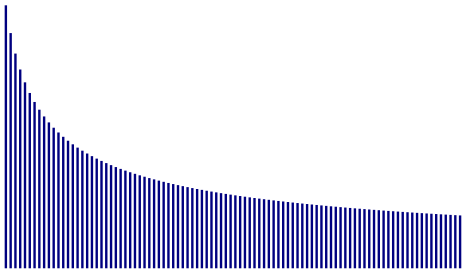

FIGURE 1. A power law distribution in the usual scale.

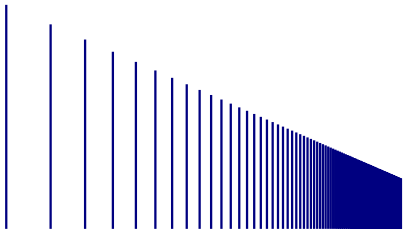

FIGURE 2. The same distribution in the log-log scale.

The first two characteristics (large and sparse) come naturally and the third (small world phenomenon) has long been within the mindset of the public consciousness. The most critical and striking fact is the power law. For example, why should the email graph and the collaboration graph have similar degree distributions? Why should the phone graphs have the same shape for different times of the day and different regions? Why should the biological networks constructed using the genome database have distributions similar to those of various social networks? Is Mother Nature finally revealing a glimpse of some first principles for the discrete world?

The power law allows us to use one single parameter (the exponent β) to describe the degree distribution of billions of nodes. With a short description of such a family of graphs, it is then possible to carry out a comprehensive analysis of these networks. On one hand, we can use various known methods and tools, combinatorial, probabilistic and spectral, to deal with problems on power law graphs. On the other hand, these realistic graphs (i.e., graphs derived from collection of data from real world applications) provide insight and suggest many new and exciting directions for research in graph theory. Indeed, in the pursuit of these large but attackable, sparse but complex graphs, we have to retool many methods from extremal and random graphs. Much is to be learned from this broad scope and new connections.

In fact, even at the end of the 19th century, the power law had been noted in various scenarios (more history will be mentioned in later sections). However, only in 1999 were the dots connected and a more complete picture emerged. The topic has spontaneously intrigued numerous researchers from diverse areas including physics, social science, computer science, telecommunications, biology and mathematics. A new area of network complexity has since been rapidly developing and is particularly enriched by the cross-fertilization of abundant disciplines. Mathematicians and especially graph theorists have much to contribute to building the scientific foundation of this area.

It is the goal of this monograph to cover some of the developments and mention what we believe are promising further directions. Since this is a fast moving field, there are already several books on this topic from the physics or heuristics points of view. The focus here is mainly on rigorous mathematical analysis via graph theory. The coverage is far from complete. There are perhaps too many models that have been introduced by various groups. Here we intend to give a consistent and simple (but not too simple!) picture rather than attempting to give an exhaustive survey. Instead, we direct the reader to several books [**10, 46, 118**] and related surveys [**3, 12, 102, 89, 106**].

REMARK 1.1. In some papers, power law graphs are referred to as "scale-free" graphs or networks. If the word "scale-free" is going to be used, the issue of "scale" should first be addressed. We will consider scale-free graphs (see Section 3.5) only after the notion of scale is clarified.

REMARK 1.2. In Figures 1 and 2, we illustrate a power law distribution in the usual scale and and in a log-log scale, respectively. Figures 3 and 4 contain the degree distribution of a call graph (with edges indicating telephone calls) and its power law approximation. In a way, the power law distribution is a straight line approximation for the log-log scale. Some might say that there are small "bumps" in the middle of the curves representing various degree distributions of realistic graphs. Indeed, the power law is a first-order estimate and an important basic case in our understanding of networks. We will interpret power law graphs in a broad sense including any graph that exhibits a power law degree distribution.

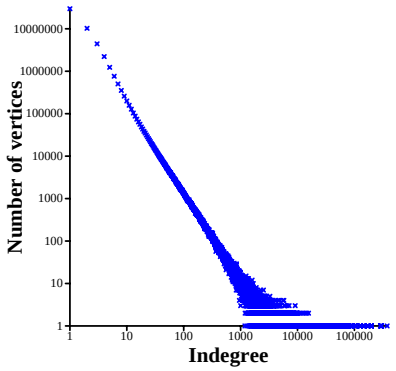

FIGURE 3. Degree distribution of a call graph in log-log scale.

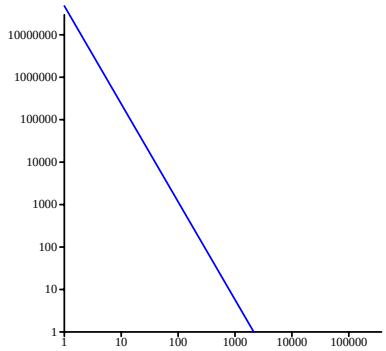

FIGURE 4. The power law approximation of Figure 3 in log-log scale.

1.2. Basic definitions

In the study of complex networks, there have been an increasingly large number of new and complicated definitions on various 'graph metrics'. Here we attempt to follow the advice of Einstein:

Everything should be made as simple as possible, but not simpler.
Albert Einstein

We intend to use as few new definitions as possible, as long as the ideas can be adequately addressed.

DEFINITION 1. *A* graph G *consists of a* vertex set $V(G)$ *and an* edge set $E(G)$*, where each edge is an unordered pair of vertices.*

For example, Figure 5 shows a graph $G = (V(G), E(G))$ defined as follows:

$$\begin{aligned} V(G) &= \{a, b, c, d\}, \\ E(G) &= \{\{a, b\}, \{a, c\}, \{b, c\}, \{b, d\}, \{c, d\}\}. \end{aligned}$$

The graph in Figure 5 is a *simple* graph since it does not contain *loops* or multiple edges. Figure 6 is a general graph with loops and multiple edges.

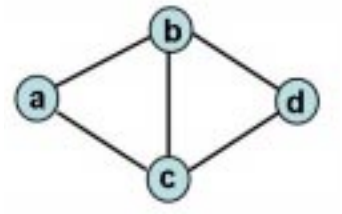

FIGURE 5. A simple graph G.

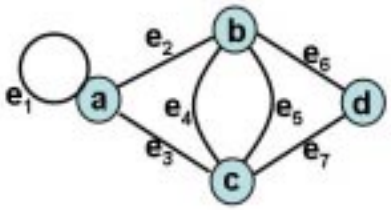

FIGURE 6. A multi-graph with a loop.

Figure 7 is a graph consisting of several mathematicians including the authors. Each edge denotes research collaboration that resulted in a mathematical paper reviewed by *Mathematical Reviews* of the American Mathematical Society.

Here are several equivalent ways to describe that an edge $\{u, v\}$ is in G:

- $\{u, v\} \in E(G)$.
- u and v are *adjacent*.
- u is a *neighbor* of v.
- The edge $\{u, v\}$ is incident to u (and also to v).

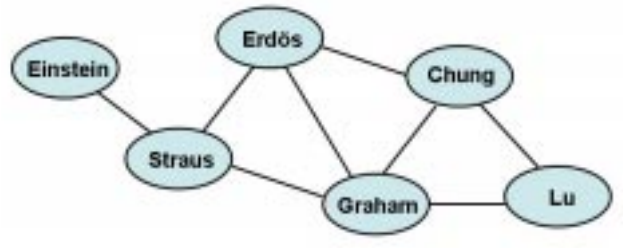

FIGURE 7. A small subgraph of the collaboration graph.

The degree of a vertex u is the number of edges incident to u. If a graph G has all its degrees equal to k, we say G is a k-regular graph.

DEFINITION 2. *A* path *from u to v of length t in G is an ordered sequence of distinct vertices $u = v_0, v_1, \ldots, v_t = v$ satisfying*

$$\{v_i, v_{i+1}\} \in E(G) \qquad \text{for } i = 0, 1, \ldots, t-1.$$

For example, in the graph of Figure 7, there is a path of length 4 from Einstein, Straus, Erdős, Chung and Lu.

DEFINITION 3. *A* walk *of a graph G of length t is an ordered sequence of vertices $v_0, v_1, \ldots, v_t$ satisfying*

$$\{v_i, v_{i+1}\} \in E(G) \qquad \text{for } i = 0, 1, \ldots, t-1.$$

We remark that vertices in a path are all distinct while a walk is allowed to have repeated vertices and edges.

DEFINITION 4. *For any two vertices $u, v \in V(G)$, the* distance *between u and v, denoted by $d(u, v)$, is the shortest length among all paths from u to v.*

For example, the distance between Einstein and Lu is 3, achieved by the path from Einstein, Straus, Graham, and Lu.

DEFINITION 5. *A graph is* connected *if for any two vertices u and v, there is a path from u to v.*

DEFINITION 6. *In a connected graph G, the* diameter *of G is the maximum distance over all pairs of vertices in G. If G is not connected, we use the convention that the diameter is defined to be the maximum diameter over the diameters of all connected components.*

DEFINITION 7. *The* average distance *of a connected graph G is the average taken over the distances of all pairs of vertices in G. If G is not connected, the average distance of G is the average taken over the distances of pairs of vertices with finite distance.*

DEFINITION 8. *A* directed graph *consists of the vertex set $V(G)$ and the edge set $E(G)$, where each edge is an ordered pair of vertices. We write $u \to v$ if an edge (u, v) is in $E(G)$. In this case, we say u is the tail and v is the head of the edge.*

Figure 8 is a directed graph associated with juggling patterns with period 3 and at most 2 balls. For an edge from a vertex labelled by (a_1, a_2) to a vertex

(a_2, a_3), the sequence (a_1, a_2, a_3) is a juggling pattern with period 3. Thus, a walk on this graph moves from one juggling pattern to another. It is of interest [**31**] to find as few cycles as possible to cover every edge once and only once. So, using this graph we can answer questions like these to pack all the juggling patterns with given period and a specified number of balls into sequences as short as possible.

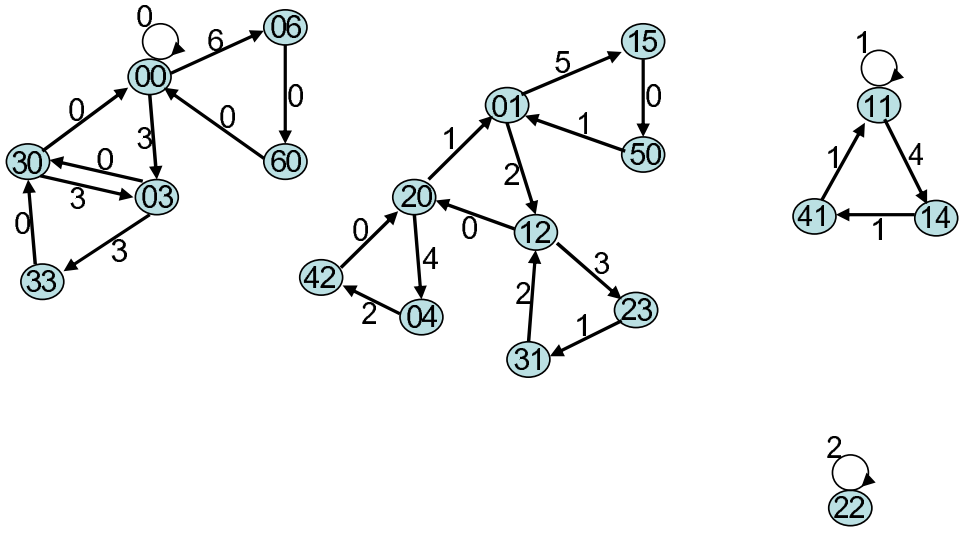

FIGURE 8. A directed graph associated with juggling patterns.

DEFINITION 9. *The indegree (or outdegree) of u is the number of edges with u as the head (or tail respectively).*

In this monograph, we are mainly concerned with finite graphs. Many real world graphs are huge but still finite. The Internet graph has a few billion nodes and keeps growing. The limit of the growth is perhaps infinite. Indeed, we dabble with infinity in several ways. We consider families of finite graphs on n vertices where n goes to infinity. In the enumeration of graphs satisfying various properties, we estimate the main order of magnitude or bound lower order terms by using the big "Oh" or little "oh" notation, namely, $O(\cdot)$ and $o(\cdot)$. The reader is referred to the book of Wilf [**121**] for a discussion of this notation.

1.3. Degree sequences and the power law

In a graph G, the collection of the degrees d_v for all vertices v can be viewed as a function defined on $V(G)$ or as a multi-set. There are several efficient ways to represent the degrees.

Typically, we can place the degrees in a list. If the vertex set consists of vertices $v_1, v_2, \ldots, v_n$, the degree sequence can be written as $d_{v_1}, d_{v_2}, \ldots, d_{v_n}$. For example, the graph in Figure 7 has a degree sequence

$$(1, 3, 4, 3, 3, 2).$$

Of course, the degree sequence depends on the choice of the order that we label the vertices. So, $(4, 3, 3, 3, 2, 1)$ is also a degree sequence for the graph in Figure 7.

For a given integer sequence $(d_1, d_2, \ldots, d_n)$, a natural question is if such a sequence is *graphical*, i.e., is a degree sequence of some graph. This question was answered by Erdős and Gallai in 1960. For a sequence to be graphical, it is necessary that the sum of all the degrees is even (as dictated by the Handshake Theorem). Another necessary condition is as follows: For each integer $r \leq n-1$,

$$\sum_{i=1}^{r} d_i \leq r(r-1) + \sum_{i=r+1}^{n} \min\{r, d_i\}. \tag{1.1}$$

Erdős and Gallai [**52**] showed that these two necessary conditions are in fact sufficient. In other words, an integer sequence $(d_1, d_2, \ldots, d_n)$ is graphical if $\sum_{i=1}^{n} d_i$ is even and (1.1) holds for all $r \leq n-1$.

Another characterization of graphical sequences was given by Havel [**74**] and Hakimi [**73**]. Namely, a sequence $(d_1, d_2, \ldots, d_n)$ with $d_i \geq d_{i-1}$, $n \geq 3$ and $d_1 \geq 1$ is graphical if and only if $(d_2 - 1, d_3 - 1, \ldots, d_{d_1+1} - 1, d_{d_1+2}, \ldots, d_n)$ is graphical.

An alternative way to present the collection of degrees is to consider the *frequencies* of the degrees. Let n_k denote the number of vertices of degree k. The *degree distribution* of G can be represented as $\langle n_0, n_1, n_2, \ldots, n_t \rangle$ where t denotes the maximum degree in G. For example, the degree distribution of the graph in Figure 7 is $\langle 0, 1, 2, 3, 1 \rangle$. We can also plot the degree distribution as shown in Figure 9.

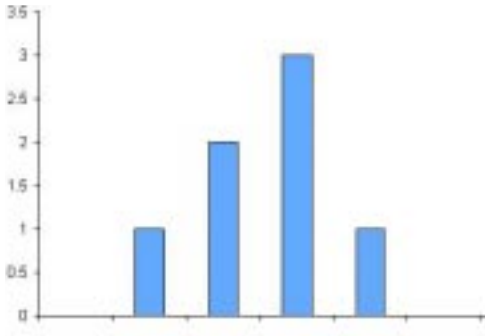

FIGURE 9. The degree distribution of the graph in Figure 7.

Suppose the degree distribution $\langle n_0, n_1, \ldots, n_t \rangle$ of a graph G satisfies the condition that n_k is proportional to $k^{-\beta}$ for some fixed $\beta > 1$, i.e.,

$$n_k \propto \frac{1}{k^{\beta}} \tag{1.2}$$

We would then say that G has a power law distribution with exponent β. We note that the expression in (1.2) is an asymptotic equation and is not exact. This is due to the fact that when dealing with a very large graph the precise numbers are either impossible to obtain or just unimportant. In such cases, what is important is to be able to control the error bounds. The asymptotic expression says the ratio of

the error bound and the main term goes to 0 as the number of vertices approaches infinity.

For a graph with a power law degree distribution, a good way to illustrate the degree distribution is by using a logarithmic scale. Namely, if we plot, for each k, the point $(\log x, \log y)$ with $x = k$ and $y = n_k$. The resulting curve should be approximately a straight line. If the power law has exponent β, the points satisfy the equation

$$\log y \approx \alpha - \beta \log x.$$

The negative slope of the line is β.

1.4. History of the power law

The earliest work on power laws can be traced back to 1896 in the lecture notes of the economist Wilfredo Pareto [**108**]. In these notes he argued that in all countries and times, the distribution of income and wealth follows a regular logarithmic pattern.

In 1926, Lotka [**90**] plotted the distribution of authors in the decennial index of Chemical Abstracts (1907-1916), and he found that the number of authors who published n papers is inversely proportional to the square of n (this is often called *Lotka's law*).

In 1932, Zipf [**126**] observed that the frequency of English words follows a power law function. That is, the word frequency that has rank i among all word frequencies is proportional to $1/i^a$ where a is close to 1. This is called *Zipf's law* or *Zipf's distribution*. Estoup [**54**] observed the same phenomenon for French in 1916. In fact, Zipf's law (which perhaps should be called Estoup's law) holds for other human languages, as well as for some artificial ones (e.g., programming languages) [**98**]. Similarly, Zipf [**127**] is often credited for noting that city sizes seem to follow a power law, although this idea can be traced back to Auerbach [**9**] in 1913.

In 1949, Yule [**125**] gave an explanation quite similar to preferential attachment for the distribution of species among genera of plants based on the empirical results of Willis [**123**]. The definition and analysis of the preferential attachment scheme will be given in Chapter 3.

In an influential paper of 1955, Simon [**111**] gave an argument of how the preferential attachment model leads to power law distributions and he listed five applications — the distribution of word frequencies in a document, the distribution of the number of papers published by scientists, the distribution of cities by population, the distribution of incomes, and the distribution of species among genera.

After Simon's article appeared, Mandelbrot raised vigorous objections to Simon's model and derivations based on preferential attachment. There was a series of heated exchanges between Simon and Mandelbrot in *Information and Control* [**95, 96, 97, 112, 113, 114**]. A scholarly report of this can be found in [**102**]. In the end, the economists seem to have sided with Simon and the preferential attachment model, as seen in the comprehensive survey by Gabaix [**63**].

In the study of random recursive trees, the parent is chosen from current vertices with probability proportional to the number of children of the node plus 1. This is a special case of preferential attachment. The degree distribution of such recursive trees was shown to obey a power law [**94**] (also see a 1993 survey [**115**]).

Then came the dawn of the new Millennium. The Internet and the vast amount of information flowing through it have touched every aspect of our lives as never before. Huge interconnection networks, physical as well as those derived from massive data, are becoming commonplace. It is then essential to understand the structure of these networks and their true nature. Around 1999, several research groups found power law distributions in numerous large networks. These include the Notre Dame group, the Santa Barbara group, the IBM group (and their consultants at the time), and the AT&T group (and their consultants including one of the authors) among others.

In 1999, Kumar et al. [**88**] from IBM reported that a web crawl of a pruned data set from 1997 containing about 40 million pages revealed that the indegree and outdegree distributions of the web followed a power law. At Notre Dame, Albert and Barabási [**11, 13**] independently reported the same phenomenon on the approximately 325 thousand node `nd.edu` subset of the web. Both reported an exponent of approximately 2.1 for the indegree power law and 2.7 for the outdegree (although the degree sequence for the outdegree deviates from the power law for small degrees). Later on, these figures were confirmed for a Web crawl of approximately 200 million nodes [**24**]. Thus, the power law fit of the degree distribution of the Web appears to be remarkably stable over time and scale.

Faloutsos et al. [**56**] have observed a power law for the degree distribution of the Internet network. They reported that the distribution of the outdegree for the interdomain routing tables fits a power law with an exponent of approximately 2.2 and that this exponent remained the same over several different snapshots of the network. At the router level the outdegree distribution for a single snapshot in 1995 followed a power law with an exponent of approximately 2.6. Their influential paper also includes data on various properties of the Internet graphs.

At AT&T, the researchers studied the graph derived from telephone calls during a period of time over one or more carriers' networks which is called a call graph. Using data collected by Abello et al. [**1**], Aiello et al. [**2**] observed that their call graphs are power law graphs. Both the indegrees and the outdegrees have an exponent of 2.1.

In addition to the Web graph and the call graph, many other massive graphs exhibit a power law for the degree distribution. The graphs derived from the U.S. power grid, the Hollywood graph of actors (where there is an edge between two actors if they have appeared together in a movie), the foodweb (links for ecological dynamics among diverse assemblages of species [**122**]), cellular and metabolic networks [**15**], and various social networks [**116**] all obey a power law. Thus, a power law fit for the degree distribution appears to be a generic and robust property for many massive real-world graphs.

Since 1999, several factors helped accelerate the progress on power law graphs — ample computing power for experimentalists, the usage of rigorous analysis from theoreticians and a conducive interdisciplinary nature of the area. There is room for all kinds of ideas and approaches, using modeling, analysis, optimization, algorithms, heuristics, biocomplexity and all their foundation in graph theory.

Time	Reference	Comments
1896	Pareto [**108**]	The distribution of income and wealth.
1926	Lotka [**90**]	Lotka's law for authors in Chemical Abstracts.
1932	Zipf [**126**]	Zipf's law for the frequency of English words.
1949	Yule [**125**]	The distribution of species among genera of plants.
1955	Simon [**111**]	Simon's model for various power law distributions.
1999	Faloutsos et al.[**56**] Kumar et al. [**88**] Barabási et al.[**11**]	The WWW graph is a power law graph.
1999	Abello et al. [**1**] Aiello et al. [**2**]	The call graphs are power law graphs.
1999	Bhalla et al.[**15**] Schilling [**110**]	Metabolic networks are power law graphs.
2000	Watts et al. [**119**]	Various social networks are power law graphs.

TABLE 1. A time table on the history of the power law.

1.5. Examples of power law graphs

1.5.1. Internet graphs. Here we mention several graphs that are related to the Internet.

(1) **AS-BGP networks:** An autonomous system (AS) is a network or a group of networks under a common administration with common routing policies, such as networks inside a university or a corporation. The Border

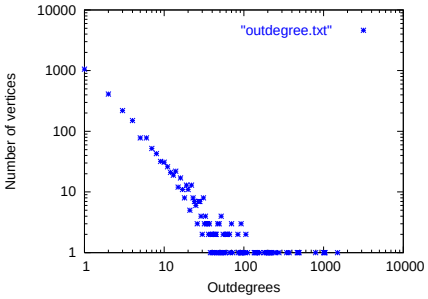

FIGURE 10. The number of vertices for each possible outdegree for an AS-BGP network.

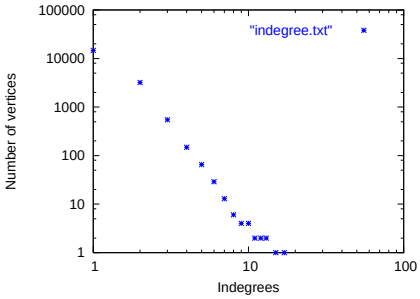

FIGURE 11. The number of vertices for each possible indegree for an AS-BGP network.

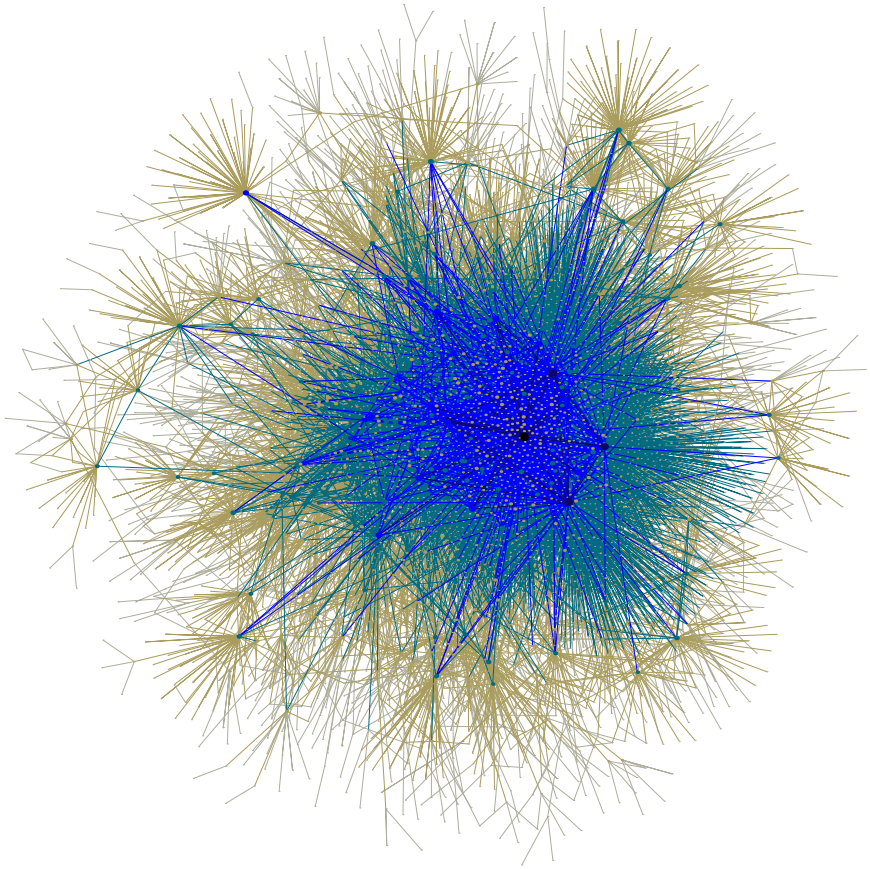

FIGURE 12. A subgraph of a BGP graph.

Gateway Protocol (BGP) is an inter-autonomous system routing protocol, for exchanging routing information between AS's or within an AS. For each destination, the router of an AS selects one AS path via BGP and records it to its BGP routing tables. The AS-BGP network is a graph with vertices consisting of AS's, and edges as AS pairs occurring in all AS paths [**25**]. Using the data collected by AS1221 (ASN-TELSTRA Telstra Pty Ltd), we examine a particular subgraph of the AS-BGP network, whose edge set is the union of AS paths recorded in AS1221's BGP routing table. The asymmetry of indegree distribution and outdegree distribution is apparent as seen in Figures 10 and 11. Figure 12 is a drawing of a subgraph of a BGP graph with about 6,400 vertices and 13,000 edges.

(2) The WWW-graphs are basically Internet topology maps. The vertices are URL's and the edges are those detected by traceroute-style path probes. For example, there are about 5 billion distinct web pages indexed by Google search engines. According to the *Internet Systems Consortium*, there are about 480,000 top level domain names as of July 2005.

(3) There are many large social networks based on various Internet communities such as the Instant Messaging networks of Yahoo, AOL and MSN.

One such examples, illustrated in Figure 13, is a subgraph of the Yahoo! Instant Messaging graph. It has about 29,000 vertices and 39,000 edges, courtesy of Kevin Lang for the data and Reid Andersen for the drawing.

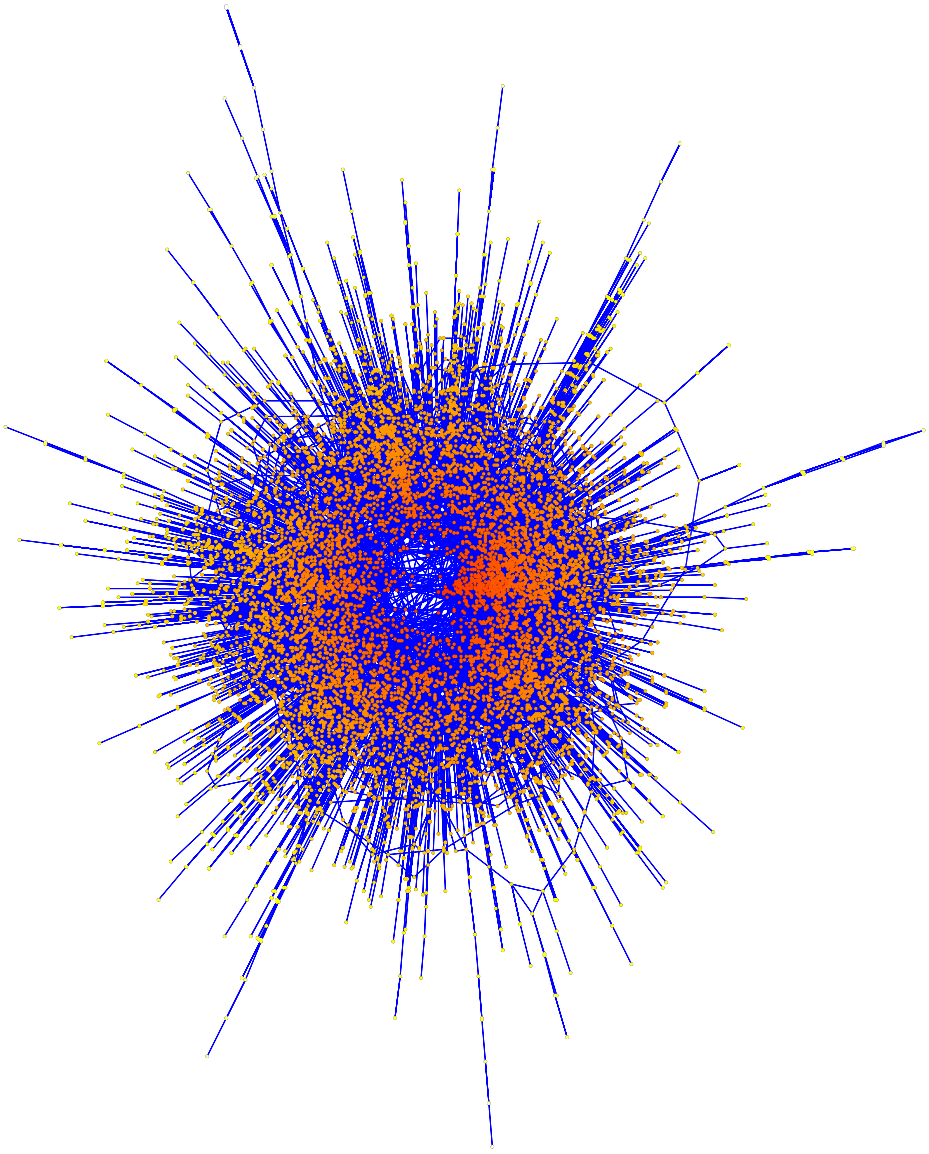

FIGURE 13. A subgraph of the Yahoo! Instant Messaging graph.

1.5.2. The call graph. The call graphs are generated by long distance telephone calls over different time intervals. For the sake of simplicity, we consider an example consisting of all the calls made in one day. A completed phone call is an edge in the graph. Every phone number which either originates or receives a call is

a node in the graph. When a node originates a call, the edge is directed out of the node and contributes to that node's outdegree. Similarly, when a node receives a call, the edge is directed into the node and contributes to that node's indegree.

In Figure 14, we plot the number of vertices versus the outdegree for the call graph of a particular day. A similar plot is shown in Figure 15 for the indegree. Plots of the number of vertices versus the indegree or outdegree for the call graphs for longer or shorter periods of time are extremely similar. For the call graph in Figures 14 and 15, we plot the number of connected components for each possible size in Figure 16.

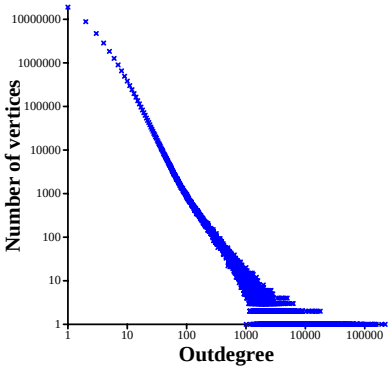

FIGURE 14. The number of vertices for each possible outdegree for a call graph (in log-log scale).

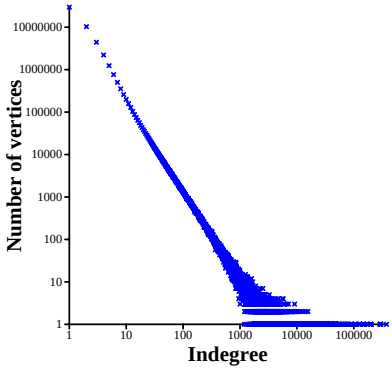

FIGURE 15. The number of vertices for each possible indegree for a call graph (in log-log scale).

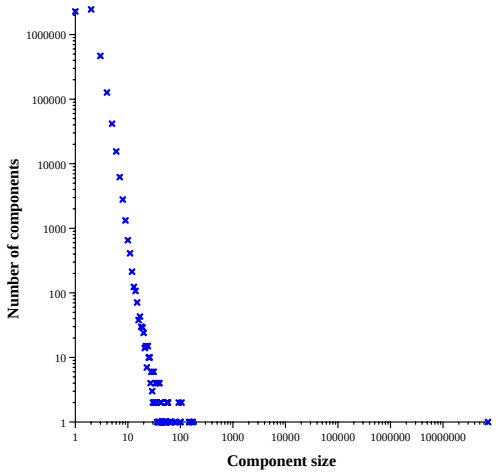

FIGURE 16. The number of connected components for each possible component size for a call graph (in log-log scale).

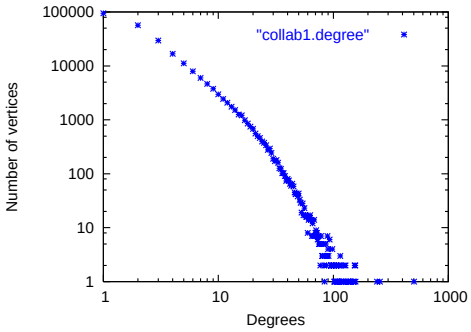

FIGURE 17. The number of vertices for each possible degree for the collaboration graph.

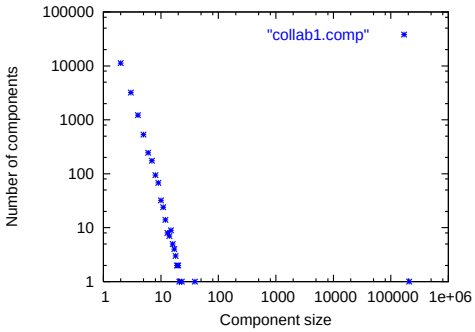

FIGURE 18. The number of components for each possible size for the collaboration graph.

1.5.3. Collaboration graphs. The collaboration graph is based on the database of Math Review of the American Mathematical Society. The database consists of 1.9 million authored items. There are several versions of the collaboration graph:

- *The collaboration graph* C has as its vertices roughly $401,000$ authors (as of July, 2004). Two authors are connected by an edge if and only if they have coauthored a paper. We remark that in this definition, a paper with five authors can introduce 10 edges. Also, C is a simple graph, not counting loops. The maximum degree of C is 1416, which of course is the number of coauthors of Paul Erdős, all of whom have Erdős number 1. Anyone who wrote a paper with someone with Erdős number 1 has Erdős number 2 and so on. The maximum Erdős number is 13. The collaboration graph has 84,000 isolated vertices, while the largest connected component of C has about 268,000 vertices and 676,000 edges. The reader is referred to the website of Grossman [**71**] for many interesting properties of C. For example, C is a power law graph with exponent 2.46. The collaboration graph C is sometimes called the *collaboration graph of the first kind*, in order to distinguish it from the other collaboration graphs below.
- *The collaboration graph of the second kind*, denoted by C', has the same vertex set as C. In contrast with C, only papers with two coauthors are considered. Two vertices in C' are joined by an edge if and only if the corresponding two authors have written a paper by themselves without other coauthors. Not surprisingly, C' has 84,000 isolated vertices. Among the remaining 235,000 vertices, there are 284,000 edges. The maximum degree of C' is 230, still due to Paul Erdős of course. The giant component of C' has 176,000 vertices. Additional properties on the giant component of C' can be found in Section 6.10.
- *The collaboration multigraph* allows multiple edges between two vertices. The number of edges between two authors are exactly the number of their

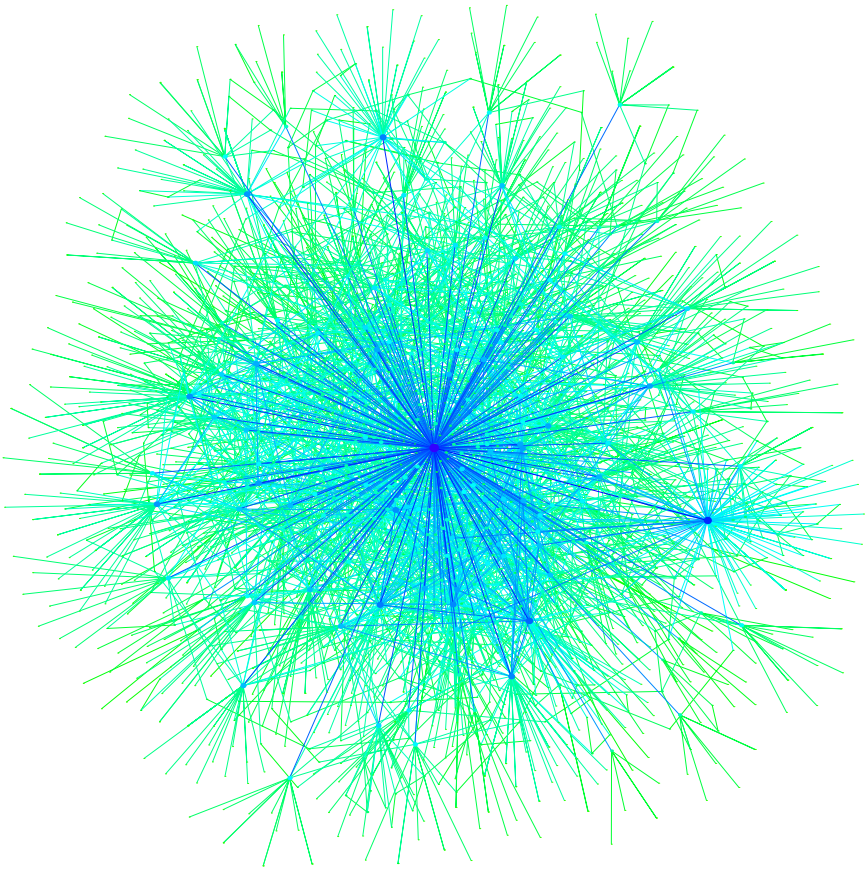

FIGURE 19. An induced subgraph of the collaboration graph.

joint papers. For example, András Sárkőzy has 62 joint papers with Erdős. Therefore there are 62 edges between the two vertices representing them. The collaboration multigraph has not been closely studied.

- *The fractional collaboration graph* has edge weights as inverses of the numbers of joint papers of two coauthors. For example, the edge between Sárkőzy and Erdős has weight 1/62. The edge between Chung and Erdős has weight 1/13. The edge weight has some geometrical interpretations. The smaller the weight is, the closer the coauthor relation is. The fractional collaboration graph also has not been closely examined.

The collaboration graph is growing rapidly. For example, the collaboration graph of the first kind as of May 2000 had about 333,000 vertices and 496,000 edges. We illustrate the degree distribution of such a collaboration graph in Figure 17. The distribution of connected component sizes is given in Figure 18. The drawing of an induced subgraph of the collaboration graph of the first kind (as of May 2000) is included in Figure 19.

1.5.4. Hollywood graph. The Hollywood graph is another version of a collaboration graph derived from movie databases. The vertices are 225,000 actors and an edge connects any two actors who have appeared in a feature film together. There are about 13 million edges. Barabási and Albert [**11**] found that the Hollywood graph satisfies the power law with exponent 2.3. Watts and Strogatz [**119**] have examined the Hollywood graph in their study of small world phenomenon. Similar to the Erdős number, the so-called *Kevin Bacon number* of an actor is the shortest distance to Kevin Bacon in the Hollywood graph. There are several websites dedicated to this topic as well a few variations of games. In Figure 20, an induced subgraph with about 10,000 vertices is illustrated.

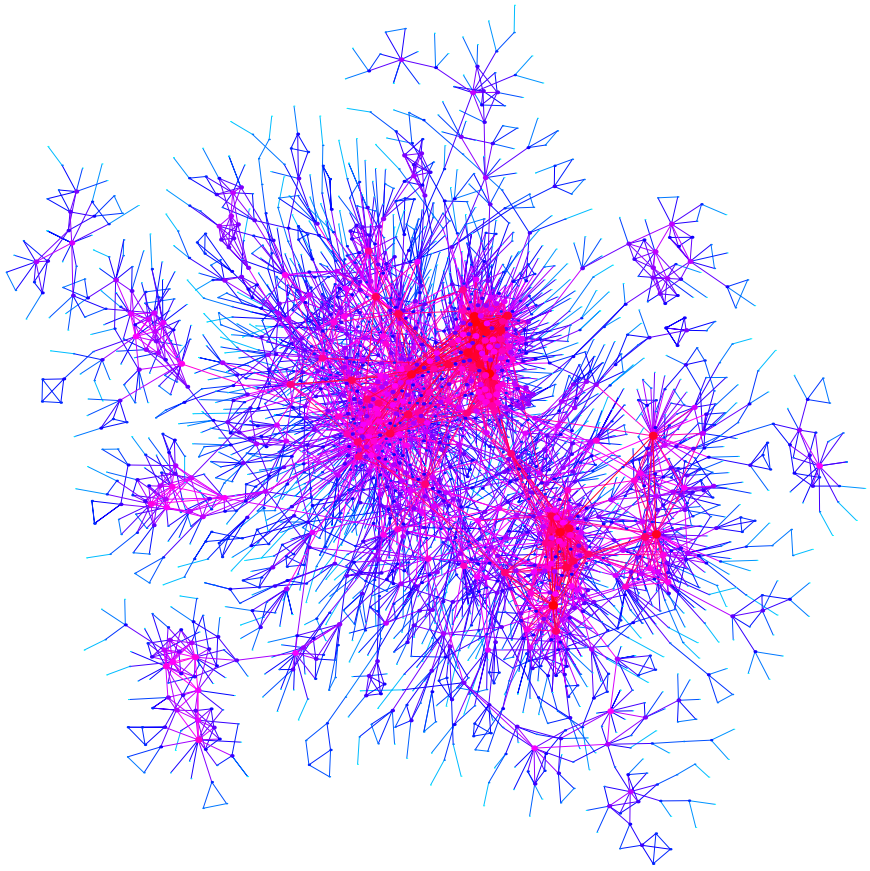

FIGURE 20. A subgraph of the Hollywood graph.

1.5.5. Biological networks. To exploit the huge amount of information from the genome data and the extensive bioreaction database, a major approach in the post-genome era is to understand the organizational principle of various genetic and metabolic networks. A great number of gene products are enzymes that catalyze cellular reactions forming a complex metabolic network. In fact, there are many kinds of biological networks with nodes corresponding to the metabolites and edges

representing reactions between the nodes. The adjacency can be defined using various reaction databases, including the enzyme-reaction database, chemical-reaction database, reversibility information of reactions, reaction-enzyme relation, enzyme-gene relations, and the evolving and updating of metabolic networks. Among the numerous biological networks, the yeast protein-protein networks are power law graphs with exponents about 1.6 (see [**47, 117**]). The E. coli metabolic networks are power law networks with exponents in the range of 1.7–2.2 (see [**2, 61**]). The yeast gene expression networks have exponents in the range of 1.4–1.7 (see [**47**]) and the gene functional interaction network has exponent 1.6 (see [**72**]). As can be seen, the range for the exponents of biological networks is somewhat different from the non-biological ones. This will be further discussed in Chapter 4.

1.6. An outline of the book

The main goal of this book is to study several random graph models and the tools required for analyzing these models.

When we say "a random graph", it means a probability space (consisting of some family $\mathcal{F}$ of graphs) together with a probability distribution (which assigns to each member of $\mathcal{F}$ a probability of being chosen).

All random graph models for power law graphs basically belong to the following two categories — the *off-line* model and the *on-line* model.

For the off-line model, the graph under consideration has a fixed number of vertices, say n vertices. For example, the probability space can be the set of all graphs on n vertices. The probability distribution of the random graph depends upon the choice of the model.

The on-line model is often called the generative model. At each tick of the clock, a decision is made for adding or deleting vertices or edges. The on-line model can be viewed as an infinite sequence of off-line models where the random graph model at time t may depend on all the earlier decisions.

The on-line models are of course much harder to analyze than the off-line models. Nevertheless, one might argue that the on-line models are closer to the way that realistic networks are generated. After the recent "rediscovery" of power law networks, the attention was first on the on-line models. In Chapter 3, we discuss the generative model coming from a preferential attachment scheme. In Chapter 4 we consider a duplication model, that is especially suited for studying networks that arise in biology.

Random graph theory has its roots in the early work of Erdős and Rényi. The classical model, that we call the Erdős-Rényi model, is an off-line model. There are two parameters — n, the number of vertices and p, the fixed probability for choosing edges. The probability space consists of all graphs with n vertices. Each pair of vertices $\{u, v\}$ is chosen to be an edge with probability p. Thus, the probability of choosing a specified graph on n vertices and e edges is $p^e(1-p)^{\binom{n}{2}-e}$.

There is a large literature and extensive research on random graphs of the Erdős-Rényi model which includes thousands of papers and dozens of books. There is a wealth of knowledge in classical random graph theory. Nevertheless, the Erdős-Rényi graphs have vertices which are almost regular and the expected degree is the same for every vertex. That is very different from realistic graphs that have uneven degree distributions as given by the power law. Furthermore, the study of classical random graphs mostly focuses on dense graphs and not as much on sparse graphs. (Here a sparse graph means a graph on n vertices with at most cn edges for some constant c.) The sparse random graphs in the Erdős-Rényi model do not have much local structure — locally the induced subgraphs are all like trees, while the power law graphs are sparse but with a great deal of local structures. In spite of these shortcomings, the classical random graph theory and in particular, the seminal work of Erdős and Rényi provide a solid foundation for our study of general random graphs. In Section 5.1, we review some of the significant results in classical random graphs.

In Chapter 5, we consider an off-line random graph model $G(\mathbf{w})$ for given degree distribution $\mathbf{w}$. Our model is a generalization of the Erdős-Rényi model. Each pair $\{u, v\}$ of vertices is independently chosen to be an edge with probability p_{uv}. Here p_{uv} is selected so that the expected degree at each vertex is as given. (For details, see Section 5.2.)

Because of the simplicity and elegance inherited from the Erdős-Rényi model, the random graph model $G(\mathbf{w})$ is quite amenable for probabilistic analysis. By sharpening the techniques in classical random theory (as seen in Chapter 2), we are able to examine a number of the major invariants of interest.

In Chapter 6, we analyze the sizes of the connected components and in particular the emergence of the giant component in a graph in $G(\mathbf{w})$. In Chapter 7, we study the diameter and average distance of a random graph in $G(\mathbf{w})$ and in particular the implications for power law graphs. In Chapter 8, we examine the eigenvalue distribution of the adjacency matrix of a random graph in $G(\mathbf{w})$. In Chapter 9, we analyze the spectra of the Laplacian of a random graph in $G(\mathbf{w})$ and particularly the semi-circle law.

In addition to the random graph $G(\mathbf{w})$ we also consider another off-line model called the configuration model. The original configuration model is a random graph model for k-regular graphs formed by combining k random matchings. The configuration model for a given degree sequence can be constructed by contracting random matchings appropriately (details in Section 11.1). In Chapter 11, we examine the evolution of random graphs in the configuration model and other related problems.

We consider two on-line random graphs — the generative model by preferential attachment schemes (in Chapter 3) and the duplication model that is particularly appropriate for biological networks (in Chapter 4). In addition, we also discuss the dynamic models that involve both addition and deletion of vertices/edges.

In Chapter 10, we analyze the on-line models using the knowledge that we have about the off-line models. We examine the comparisons of random graph models and the methods that are needed in this line of study.

Although random graph models are useful for analyzing realistic networks, there is no doubt that some aspects of realistic networks are not captured by random graphs. In Chapter 12, we look into a more general setting which uses random graphs to model the "global" aspects of networks while allowing further control of "local" aspects.

A flow chart in Figure 21 summarizes the interrelations of the chapters. Many chapters are mainly based on previous papers by the two authors and their collaborators. Chapter 1 is based on two papers with Bill Aiello [**2, 3**]. An earlier version of Chapter 2 has appeared as a survey paper [**36**] which contains additional examples. Chapter 3 is partly based on [**3, 40**] and Chapter 4 is based on [**41**]. Several sections of Chapter 5 contain material in [**33, 34, 38**]. Chapter 6 is mainly based on [**34, 38**] and Chapter 7 is based on [**35**]. Chapters 8 and 9 are based on two papers with Van Vu [**42, 43**]. Chapter 10 is partly in [**40**] and Chapter 11 is based on [**2**]. Chapter 12 has overlapped with [**37**] and the papers with Reid Andersen [**7, 8**]. Most of the illustrations for graph visualization in this chapter are due to the immense help of Ross Richarson and Reid Anderson.

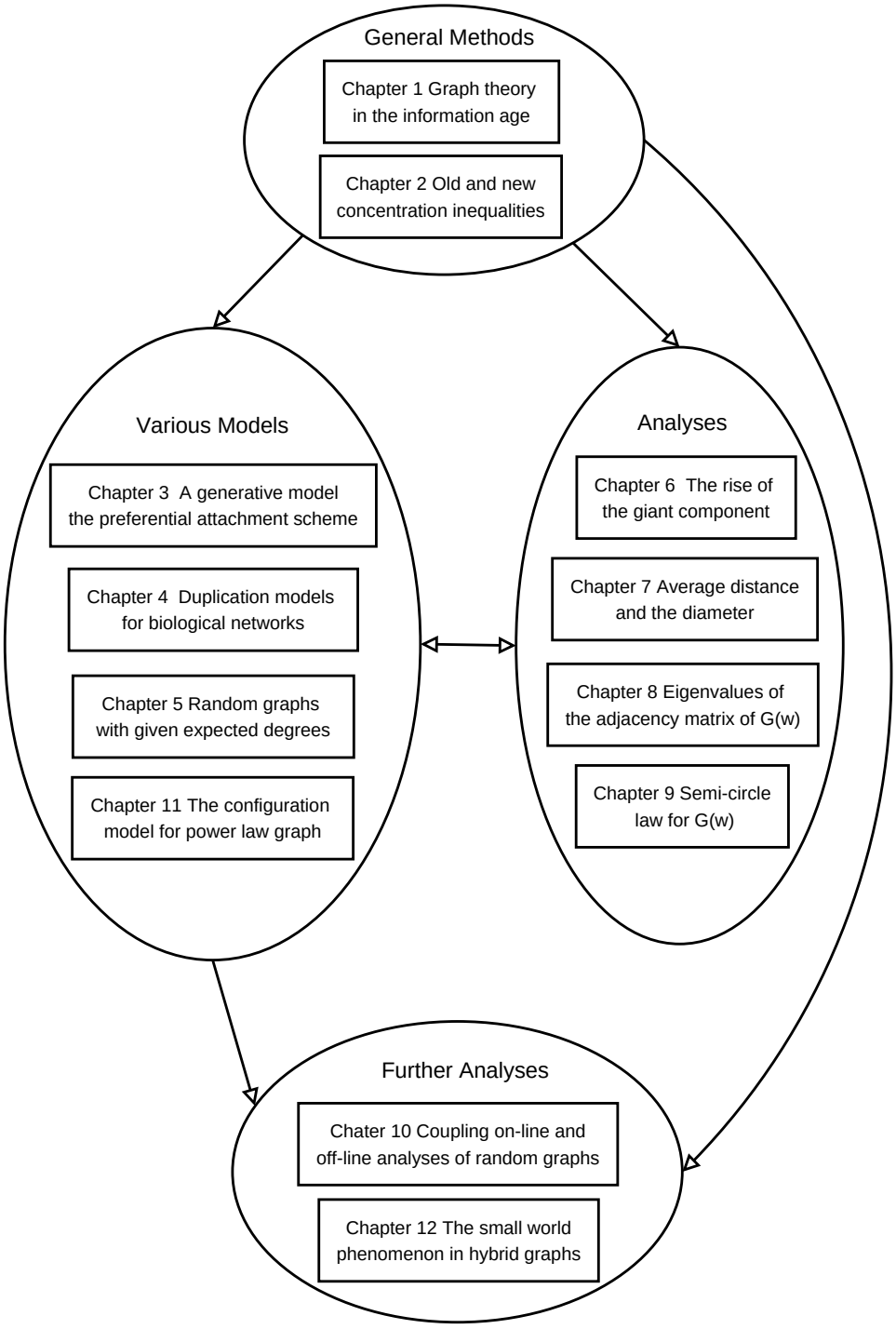

FIGURE 21. A flow chart of the chapters.