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

CHAPTER 1

Graph theory in the new millennium

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 [**18**]). 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 crossing in the plane), interval graphs (arising in scheduling), symmetric graphs (hypercubes, or 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 of graphs 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 from all branches of natural, social and the life sciences. The prevailing characteristics of these realistic 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. Perhaps there are many *dense* graphs (having quadratic number of edges) out there but the large graphs that we can hope to deal with are mostly sparse.
- **The Small world phenomenon** This is used to refer to two distinct properties: *small distance* and *the clustering effect.* Namely, two strangers are typically joined by a short chain of mutual acquaintances. and two people who share a common neighbor are more likely to know each other.

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

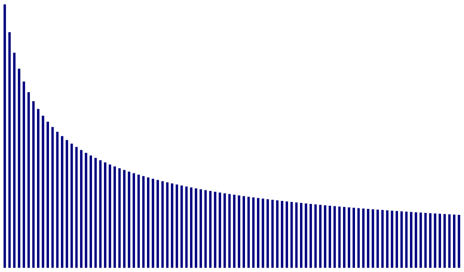

FIGURE 1. *A power law distribution in the usual scale.*

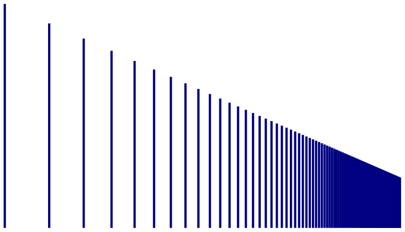

FIGURE 2. *The same distribution in the log-log scale.*

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, the realistic graphs 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 graphs 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 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 include references to several books [**13, 42, 113**] and related surveys [**3, 7, 97, 101**].

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 **??** 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.

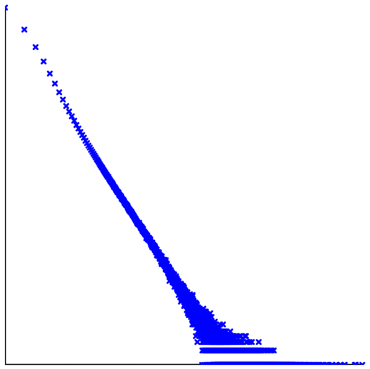

Figure 3. *Degree distribution of a call graph.*

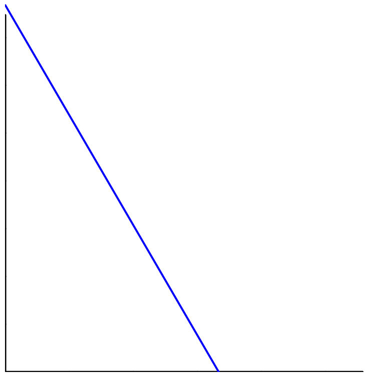

Figure 4. *The power law approximation of Figure 3.*

1.2. Basic definitions

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.

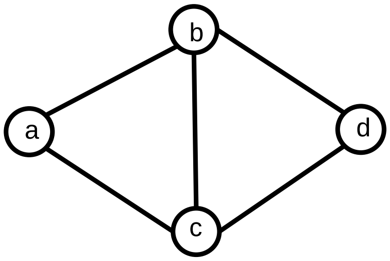

FIGURE 5. A simple graph G.

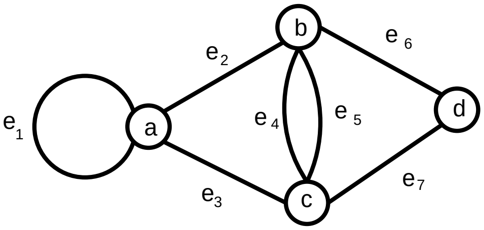

FIGURE 6. A multi-graph with a loop.

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 *Mathematicsl Reviews* of the American Mathematical Society.

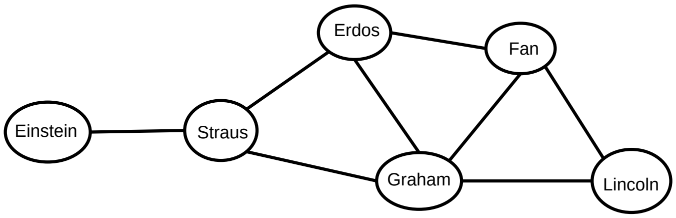

FIGURE 7. A small subgraph of the collaboration graph.

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).

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

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

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

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

Definition 3. *A* walk *of a graph G is an ordered sequence of vertices $v_0, v_1, \ldots, v_k$ satisfying*

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

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 Lincoln is 3, achieved by the path from Einstein, Straus, Graham, and Lincoln.

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 [**30**] 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.

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

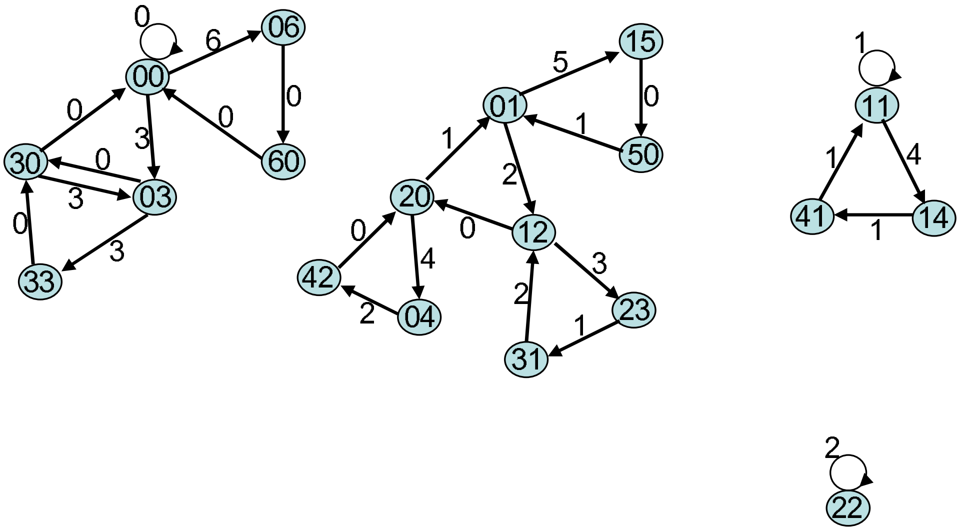

Figure 8. A directed graph associated with juggling patterns.

In this book, we are mainly concerned with finite graphs. Very many realistic graphs are huge but still finite. The Internet graph can has a few billion nodes and keeps growing. The limit of the growth is perhaps infinity. 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 [**116**] for a discussion of this notation.

1.3. Degree sequences and the power law

In a graph G, each vertex v has its degree, denoted by d_v, as the number of edges incident to v. The collection of the degrees d_v for all v can be viewed as a function defined on $V(G)$ or be considered as a multi-set. There are several efficient ways to represent the degrees.

Typically, we can place the degrees as 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 [**50**] 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 [**71**] and Hakimi [**70**]. 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 $(n_1, n_2, \ldots, n_t)$ where t denotes the maximum degree in G. For example, the degree distribution of the graph in Figure 7 is $\langle 1, 2, 3, 1 \rangle$. We can also plot the degree distribution as shown in Figure 9.

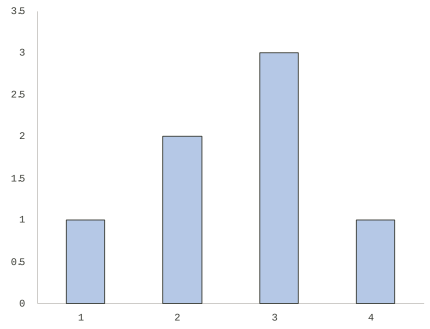

Figure 9. The degree distribution of the graph in Figure 7.

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 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 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 just β as indicated in Figure 4.

1.4. History of the power law

The earliest work on power laws can be traced back to the lecture notes of three volumes by the economist Wilfredo Pareto [**103**] in 1896 who argued that in all countries and times, the distribution of income and wealth follows a regular logarithmic pattern.

In 1926, Lotka [**85**] plotted the distribution of authors in the decennial index of Chemical Abstracts (1907-1916), and he found that the number of authors published n papers is inversely proportional to the square of n (which is often called *Lotka's law*).

In 1932, Zipf [**121**] 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 [**52**] 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) [**92**]. Similarly, Zipf [**122**] is often credited for noting that city sizes seem to follow a power law, although this idea can be traced back to Auerback [**12**] in 1913.

In 1949, Yule [**120**] gave an explanation quite similar to preferential attachment for the distribution of species among genera of plants based on the empirical results of Willis [**118**]. The definition and analysis of the preferential attachment scheme will be given later in Chapter 3.

In an influential paper of 1955, Simon [**106**] gave an argument of how the preferential attachment model leads to power law 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, distribution of income, 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* [**89, 90, 91, 107, 108, 109**]. A scholarly report of this can be found in [**97**]. In the end, the economists seem to have sided with Simon and the preferential attachment model, as seen in the comprehensive survey by Gabaix [**61**].

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 just a special case of preferential attachment. The degree distribution of such recursive trees was shown to obey a power law [**93**] (also see a 1993 survey [**110**]).

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 ubiquitous. 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. [**84**] from IBM reported that a web crawl of a pruned data set from 1997 containing about 40 million pages revealed that the in-degree and out-degree distributions of the web followed a power law. At Notre Dame, Albert and Barabási [**6, 14**] 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 in-degree power law and 2.7 for the out-degree (although the degree sequence for the out-degree deviates from the power law for small degree). Later on, these figures were confirmed for a Web crawl of approximately 200 million nodes [**27**]. Thus, the power law fit of the degree distribution of the Web appears to be remarkably stable over time and scale.

Faloutsos et al. [**54**] have observed a power law for the degree distribution of the Internet network. They reported that the distribution of the out-degree 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 out-degree distribution for a single snapshot in 1995 followed a power law with an exponent of approximately 2.6. Their influential paper [**54**] 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 in-degrees and the out-degrees 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 [**117**]), cellular and metabolic networks [**16**], and various social networks [**111**] all obey a power law. Thus, a power law fit for the degree distribution appears to be a ubiquitous 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 imagination, through modeling, analysis, optimization, algorithms, heuristics, biocomplexity and all their foundation in graph theory.

Time	Reference	Comments
1896	Pareto [**103**]	The distribution of income and wealth.
1926	Lotka [**85**]	Lotka's law for authors in Chemical Abstracts.
1932	Zipf [**121**]	Zipf's law for the frequency of English words.
1949	Yule [**120**]	The distribution of species among genera of plants.
1955	Simon [**106**]	Simon's model for various power law distributions.
1999	Faloutsos et al. [**54**] Kumar et al. [**84**] Barabási et al. [**6**]	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.[**16**] Schilling [**105**]	Cellular and metabolic networks are power law graphs.
2000	Watts, Strogatz [**114**]	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 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 Gateway Protocol (BGP) is an inter-autonomous system routing protocol, for exchanging routing information between ASes 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 ASes, and edges as AS pairs occurring in all AS paths. Using the data collected by AS1221 (ASN-TELSTRA Telstra Pty

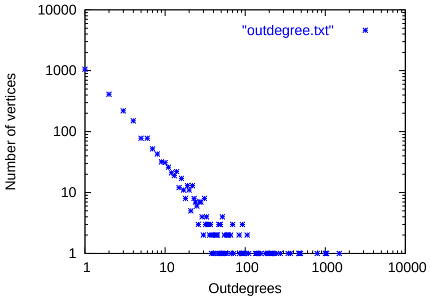

FIGURE 10. *The number of vertices for each possible outdegree for an AS-BGP network.*

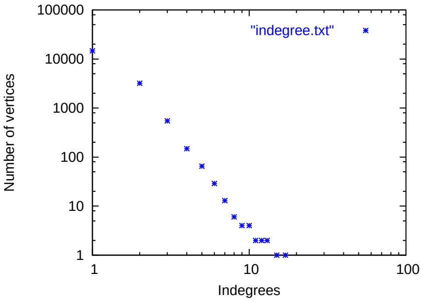

FIGURE 11. *The number of vertices for each possible indegree for an AS-BGP network.*

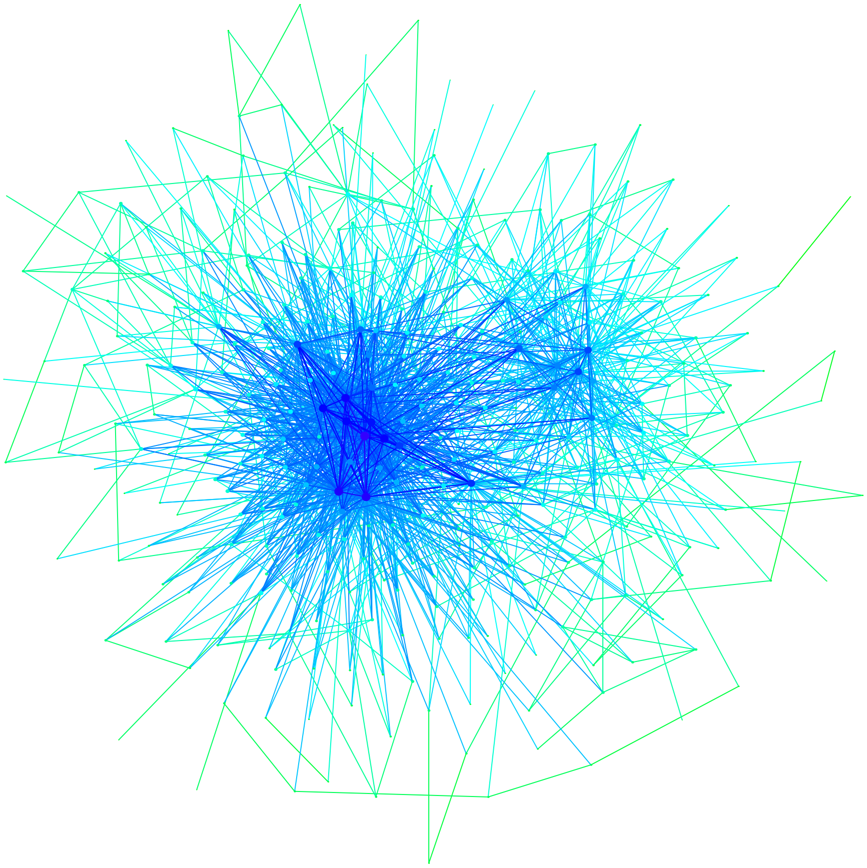

FIGURE 12. A subgraph of a BGP graph.

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 Figure 10 and 11.

(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. Figure 12 is a drawing of a subgraph of a BGP graph with about 6,400 vertices and 13,000 edges.

(3) There are many large social networks based on various Internet communities such as the Instant Messaging networks of Yahoo, AOL and MSN. One of such examples is illustrated in Figure **??**.

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 13, we plot the number of vertices versus the outdegree for the call graph of a particular day. A similar plot is shown in Figure 14 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 13 and 14, we plot the number of connected components for each possible size in Figure 15.

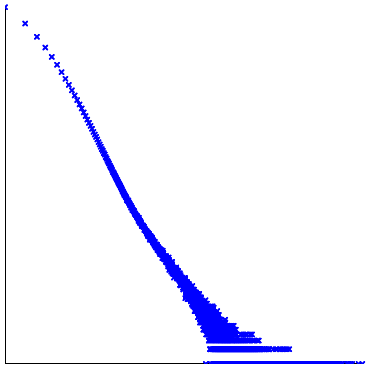

FIGURE 13. *The number of vertices for each possible outdegree for a call graph.*

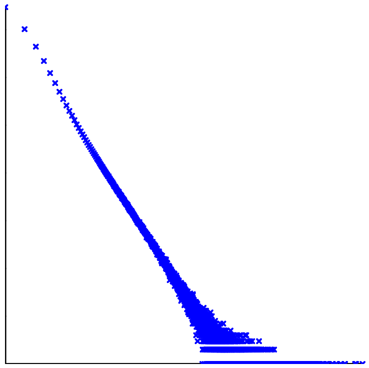

FIGURE 14. *The number of vertices for each possible indegree for a call graph.*

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 roughly $401,000$ authors as its vertices. 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, who 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. The largest connected component of C has about 268,000 vertices and 676,000 edges. The reader is referred to the website of Grossman [**68**] for many interesting properties of C. For example, C is a power law graph with exponent 2.46. The collaboration

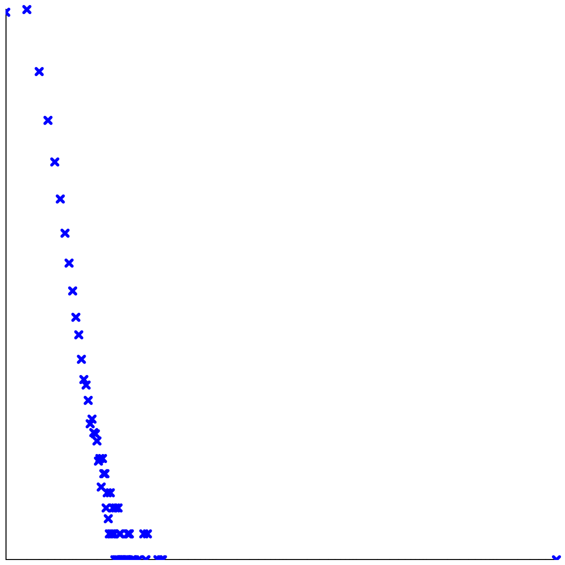

FIGURE 15. *The number of connected components for each possible component size for a call graph.*

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, of course still due to Paul Erdős. 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 joint papers. For example, Andras Sarkozy 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 Sarkozy 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. Here we illustrate the degree distribution of such a collaboration graph in Figure 16. The distribution of connected component sizes is given in Figure 17.

The drawing of the induced subgraph of the collaboration graph of the first kind (as of May 2000) is included in Figure 18.

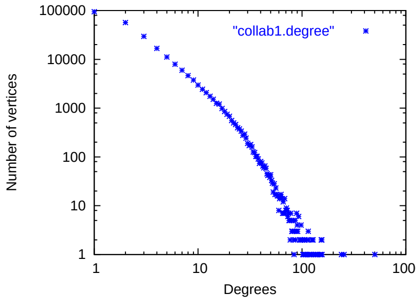

FIGURE 16. *The number of vertices for each possible degree for the collaboration graph.*

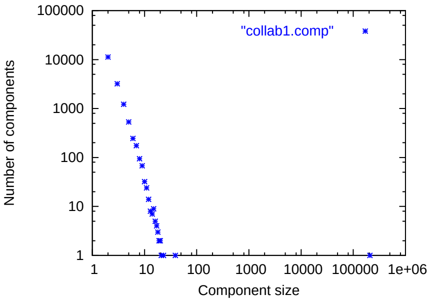

FIGURE 17. *The number of components for each possible size for the collaboration graph.*

1.5.4. Hollywood graph. The Hollywood graph is another version of a collaboration graph derived from the movies database. The vertices are about 225,000 actors and an edge connects any two actors who have appeared in a feature film together. There are about 13 million edges. In [**6**], Barabási and Albert found the Hollywood graph satisfies the power law with exponent 2.3. Watts and Strogatz [**114**] 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 delicated to this topic as well a few variations of games. In Figure 19, an induced subgraph with about 10,000 vertices is illustrated.

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 powerlaw graphs with exponents about 1.6 (see [**45, 112**]). The E. coli

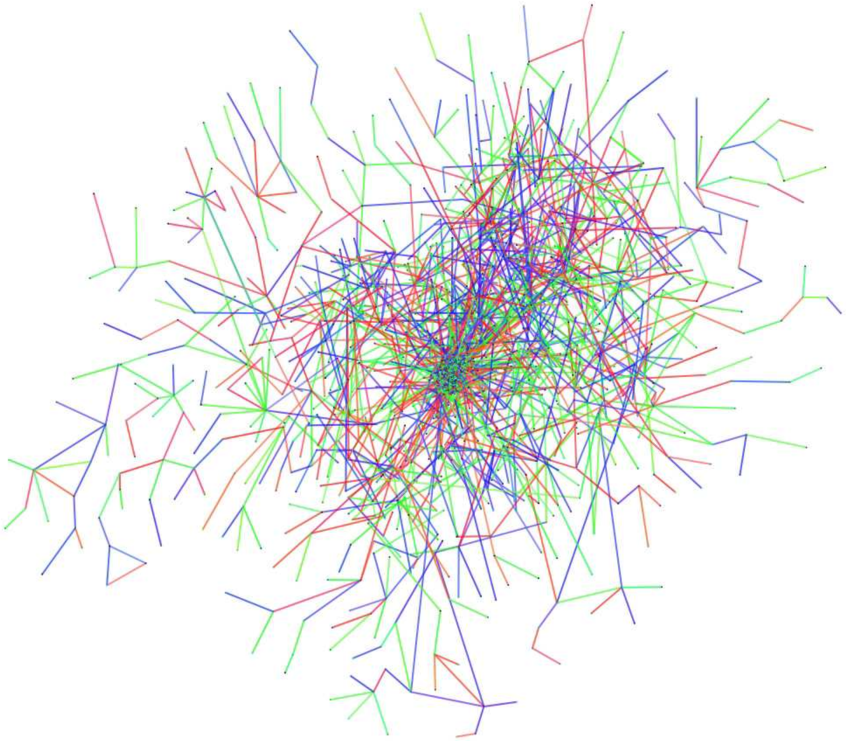

FIGURE 18. An induced subgraph of the collaboration graph.

metabolic networks are power law networks with exponents in the range of $1.7 - 2.2$ (see [**2, 59**]). The yeast gene expression networks have exponents $1.4 - 1.7$ (see [**45**]) and the gene functional interaction network has exponent 1.6 (see [**69**]). 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 *on-line* model.

For the off-line model, in the graph under consideration the number of vertices is fixed, say n vertices. For example, the probability space can be the set of all

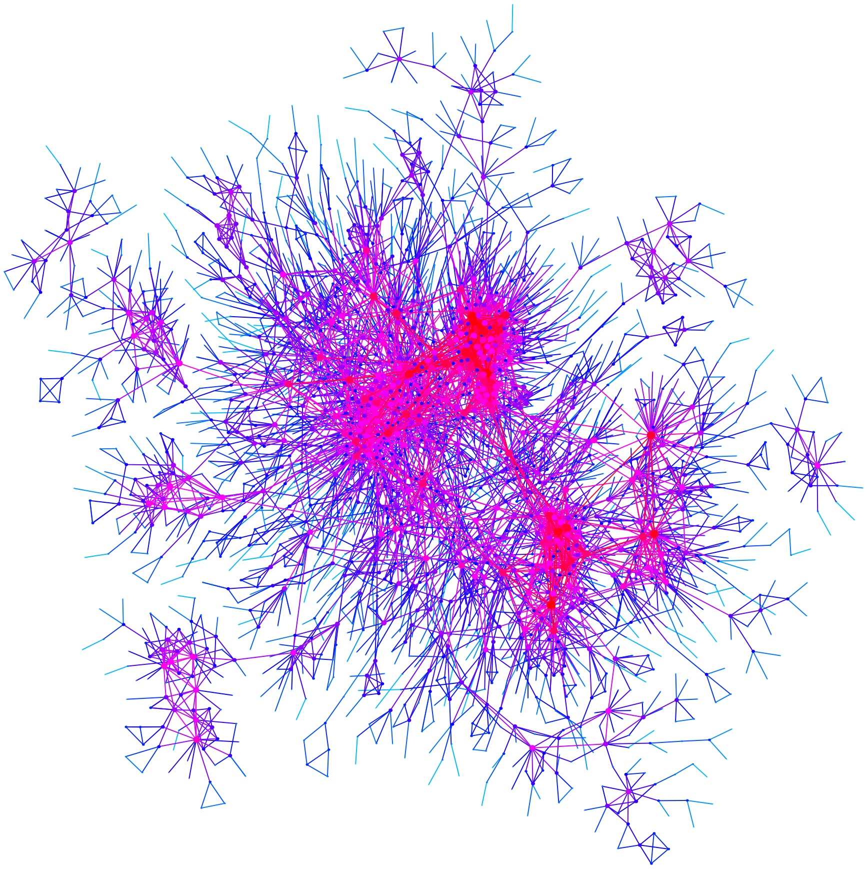

FIGURE 19. A subgraph of the Hollywood graph.

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/edges. The on-line model can be viewed as an infinite sequence of off-line models while 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. Soon 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 the duplication models, that are especially suitable 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 such as 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 $\mathcal{G}(\mathbf{w})$ 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 $\mathcal{G}(\mathbf{w})$ is quite amendable 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 $\mathcal{G}(\mathbf{w})$. In Chapter 7, we study the diameter and average distance of a random graph in $\mathcal{G}(\mathbf{w})$ 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 $\mathcal{G}(\mathbf{w})$. In Chapter 9, we analyze the spectra of the Laplacian of a random graph in $\mathcal{G}(\mathbf{w})$ and particularly the semi-circle law.

In addition to the random graph $\mathcal{G}(\mathbf{w})$ 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 20 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 [**?, ?**]. An earlier version of Chapter 2 has appeared as a survey paper [**35**] which contains additional examples. Chapter 3 is partly based on [**?, 40**] and Chapter 4 is based on [**41**]. Several sections of Chapter 5 contain material in [**32, 33, 37**]. Chapter 6 is mainly based on [**33, 37**] and Chapter 7 is based on [**34**]. Chapters 8 and 9 are based on two papers with Van Vu [**38, 39**]. Chapter 10 is partly in [**40**] and Chapter 11 is based on [**2**]. Chapter 12 has overlapped with [**36**] and the papers with Reid Andersen [**10, 11**].

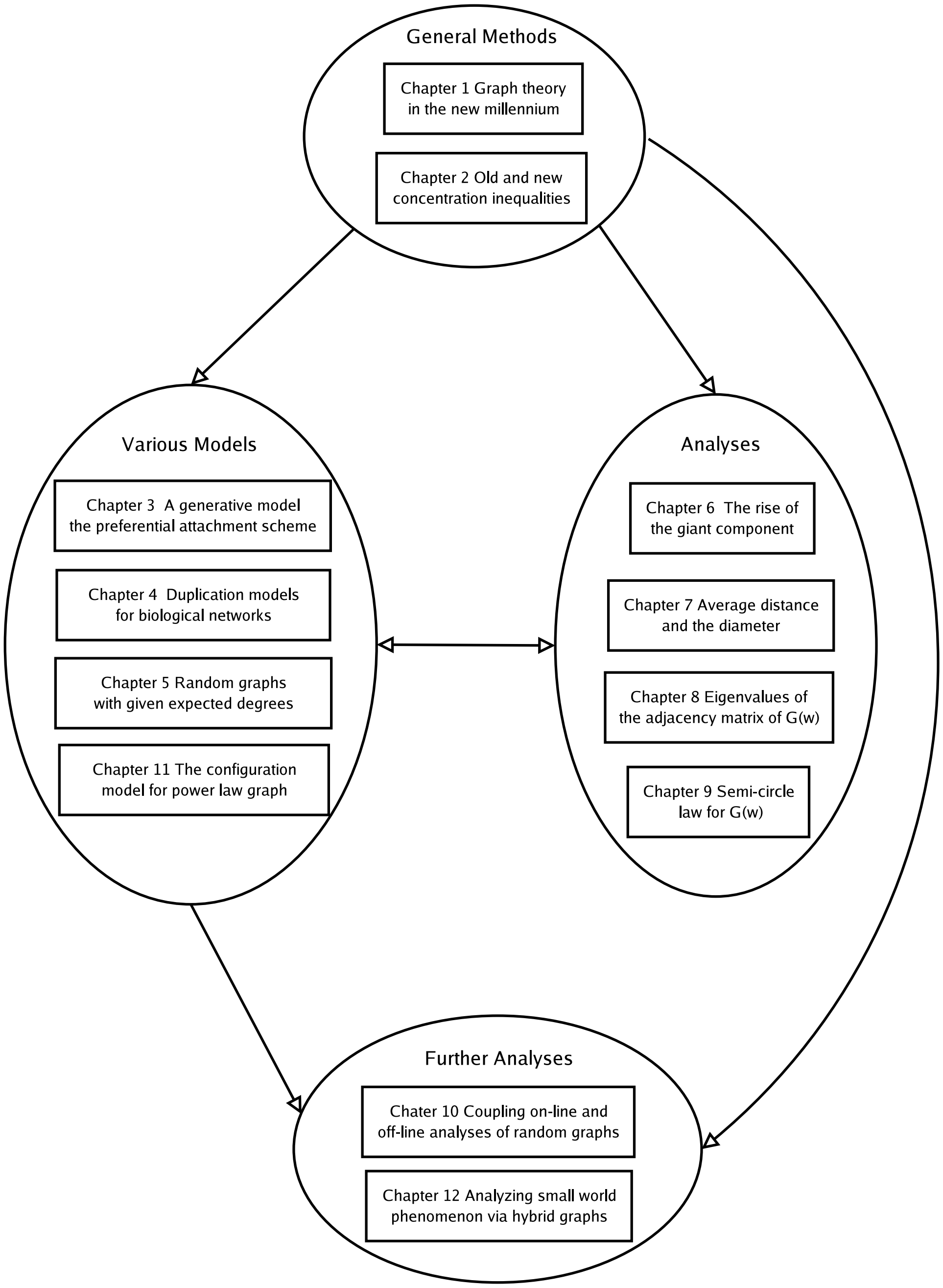
General Methods
Chapter 1 Graph theory in the new millennium
Chapter 2 Old and new concentration inequalities
Various Models
Chapter 3 A generative model the preferential attachment scheme
Chapter 4 Duplication models for biological networks
Chapter 5 Random graphs with given expected degrees
Chapter 11 The configuration model for power law graph
Analyses
Chapter 6 The rise of the giant component
Chapter 7 Average distance and the diameter
Chapter 8 Eigenvalues of the adjacency matrix of G(w)
Chapter 9 Semi-circle law for G(w)
Further Analyses
Chater 10 Coupling on-line and off-line analyses of random graphs
Chapter 12 Analyzing small world phenomenon via hybrid graphs

FIGURE 20. *A flow chart of the chapters*

CHAPTER 2

Old and new concentration inequalities

In the study of random graphs or any randomly chosen objects, the "tools of the trade" mainly concern various concentration inequalities and martingale inequalities.

To say this in layman's terms, suppose we wish to predict the outcome of a problem of interest. One reasonable guess is the expected value of the subject. However, how can we tell how good the expected value is, say, to the actual outcome of the event? Wouldn't it be nice if such a prediction can be accompanied by a guarantee of its accuracy (within a certain error estimate, for example)? This is exactly the role that the concentration inequalities play. In fact, the analysis can easily go astray without the rigorous control coming from the concentration inequalities.

In our study of random power law graphs, the usual concentration inequalities are simply not enough. The reasons are multi-fold: Due to uneven degree distribution, the error bound of those very large degrees offset the delicate analysis in the sparse part of the graph. Furthermore, our graph is dynamically evolving and therefore the probability space is changing at each tick of the time. The problems arising in the analysis of random power law provide impetus for improving our technical tools.

Indeed, in the course of our study of general random graphs, we need to use several strengthened versions of concentration inequalities and martingale inequalities. They are interesting in their own rights and may be useful for many other problems as well.

In the next several sections, we state and prove a number of variations and generalizations of concentration inequalities and martingale inequalities. Many of these will be used in later chapters. An earlier version of this chapter led to a survey paper in [**35**].

2.1. Binomial distribution and its asymptotic behavior

The Bernoulli trials, named after James Bernoulli, can be thought of as a sequence of coin-tossings. For some fixed value p, where $0 \leq p \leq 1$, the outcome of the coin-tossing has probability p of getting a "head". Let S_n denote the number of heads after n tosses. We can write S_n as a sum of independent random variables

X_i as follows:

$$S_n = X_1 + X_2 + \cdots + X_n$$

where, for each i, the random variable X satisfies

$$\begin{aligned} \Pr(X_i = 1) &= p, \\ \Pr(X_i = 0) &= 1 - p. \end{aligned} \tag{2.1}$$

A classical question is to determine the distribution of S_n. It is not too difficult to see that S_n has the *binomial distribution* $B(n, p)$:

$$\Pr(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad \text{for } k = 0, 1, 2, \ldots, n.$$

The expectation and variance of $B(n, p)$ are

$$\mathrm{E}(S_n) = np, \qquad \mathrm{Var}(S_n) = np(1-p).$$

To better understand the asymptotic behavior of the binomial distribution, we compare it with the normal distribution $N(a, \sigma)$, whose density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

where a denotes the expectation and σ^2 is the variance.

The case $N(0, 1)$ is called the *standard normal distribution* whose density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty.$$

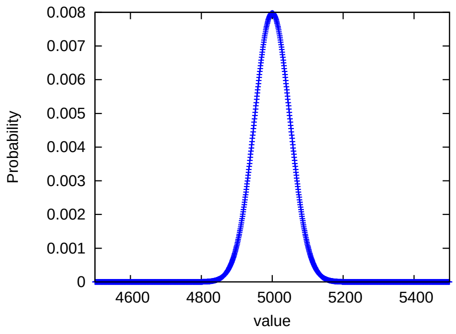

FIGURE 1. *The Binomial distribution* $B(10000, 0.5)$

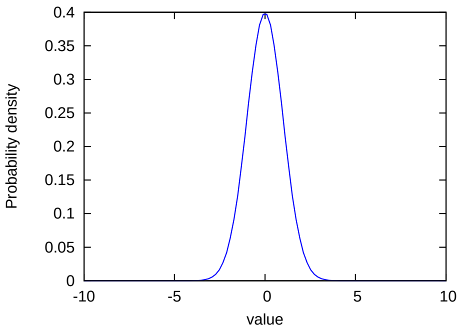

FIGURE 2. *The Standard normal distribution* $N(0, 1)$

When p is a constant, the limit of the binomial distribution, after scaling, is the standard normal distribution and can be viewed as a special case of the Central-Limit Theorem, sometimes called the DeMoivre-Laplace limit Theorem [**51**].

THEOREM 2.1. *The binomial distribution $B(n,p)$ for S_n, as defined in (2.1), satisfies, for two constants a and b,*

$$\lim_{n\to\infty} \Pr(a\sigma < S_n - np < b\sigma) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

where $\sigma = \sqrt{np(1-p)}$ provided $np(1-p) \to \infty$ as $n \to \infty$.

PROOF. We use the *Stirling formula* for $n!$ (see [**67**]).

$$\begin{aligned} n! &= (1+o(1))\sqrt{2\pi n}(\frac{n}{e})^n \\ \text{or, equivalently, } n! &\approx \sqrt{2\pi n}(\frac{n}{e})^n. \end{aligned}$$

For any constant a and b, we have

$$\begin{aligned} & \Pr(a\sigma < S_n - np < b\sigma) \\ = & \sum_{a\sigma<k-np<b\sigma} \binom{n}{k} p^k(1-p)^{n-k} \\ \approx & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}\frac{n^n}{k^k(n-k)^{n-k}}p^k(1-p)^{n-k} \\ = & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi np(1-p)}}(\frac{np}{k})^{k+1/2}(\frac{n(1-p)}{n-k})^{n-k+1/2} \\ = & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma}(1+\frac{k-np}{np})^{-k-1/2}(1-\frac{k-np}{n(1-p)})^{-n+k-1/2}. \end{aligned}$$

To approximate the above sum, we consider the following slightly simpler expression. Here, to estimate the lower ordered term, we use the fact that $k = np + O(\sigma)$ and $1 + x = e^{\ln(1+x)} = e^{x-x^2+O(x^3)}$, for $x = o(1)$. To proceed, we have

$$\begin{aligned} & \Pr(a\sigma < S_n - np < b\sigma) \\ \approx & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma}(1+\frac{k-np}{np})^{-k}(1-\frac{k-np}{n(1-p)})^{-n+k} \\ \approx & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{k(k-np)}{np}+\frac{(n-k)(k-np)}{n(1-p)}+\frac{k(k-np)^2}{n^2p^2}+\frac{(n-k)(k-np)^2}{n^2(1-p)^2}+O(\frac{1}{\sigma})} \\ = & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{k-np}{\sigma})^2+O(\frac{1}{\sigma})} \\ \approx & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{k-np}{\sigma})^2} \end{aligned}$$

Now, we set $x = x_k = \frac{k-np}{\sigma}$, and $dx = x_k - x_{k-1} = 1/\sigma$. Note that $a < x_1 < x_2 < \cdots < b$ form a $1/\sigma$-net for the interval (a,b). As n approaches the infinity, the limit exists. We have

$$\lim_{n\to\infty} \Pr(a\sigma < S_n - np < b\sigma) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Thus, the limit distribution of the normalized binomial distribution is the normal distribution. □

When np is upper bounded (by a constant), the above theorem is no longer true. For example, for $p = \frac{\lambda}{n}$, the limit distribution of $B(n,p)$ is the so-called *Poisson distribution* $P(\lambda)$.

$$\Pr(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}, \qquad \text{for } k = 0, 1, 2, \cdots .$$

The expectation and variance of the Poisson distribution $P(\lambda)$ is given by

$$\mathrm{E}(X) = \lambda, \qquad \text{and} \qquad \mathrm{Var}(X) = \lambda.$$

THEOREM 2.2. *For $p = \frac{\lambda}{n}$, where λ is a constant, the limit distribution of binomial distribution $B(n,p)$ is the Poisson distribution $P(\lambda)$.*

PROOF. We consider

$$\begin{aligned}
\lim_{n\to\infty} \Pr(S_n = k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{\lambda^k \prod_{i=0}^{k-1}(1 - \frac{i}{n})}{k!} e^{-p(n-k)} \\
&= \frac{\lambda^k}{k!}e^{-\lambda}.
\end{aligned}$$

□.

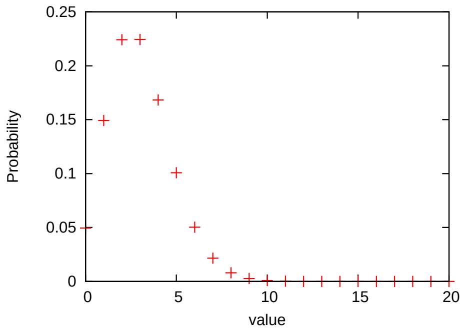

FIGURE 3. *The Binomial distribution* $B(1000, 0.003)$

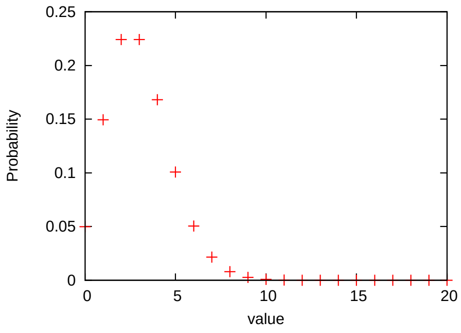

FIGURE 4. *The Poisson distribution* $P(3)$

As p decreases from $\Theta(1)$ to $\Theta(\frac{1}{n})$, the asymptotic behavior of the binomial distribution $B(n,p)$ changes from the normal distribution to the Poisson distribution. (Some examples are illustrated in Figure 5 and 6). Theorem 2.1 states that the asymptotic behavior of $B(n,p)$ within the interval $(np - C\sigma, np + C\sigma)$ (for any constant C) is close to the normal distribution. In some applications, we might need asymptotic estimates beyond this interval.

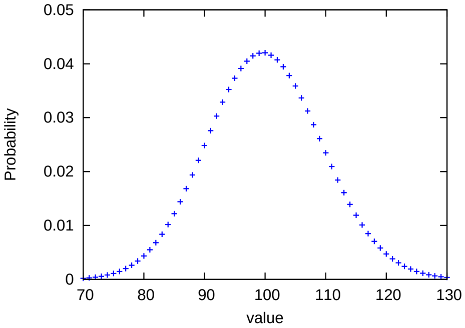

FIGURE 5. *The Binomial distribution* $B(1000, 0.1)$

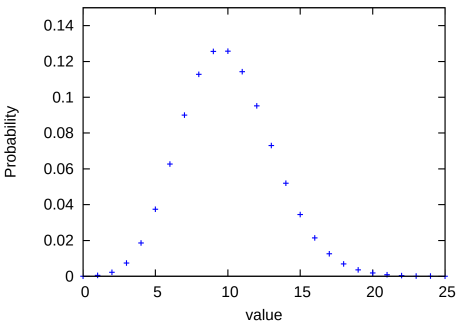

FIGURE 6. *The Binomial distribution* $B(1000, 0.01)$

2.2. General Chernoff inequalities

If the random variable under consideration can be expressed as a sum of independent variables, it is possible to derive good estimates. The binomial distribution is one such example where $S_n = \sum_{i=1}^n X_i$ and X_i's are independent and identical. In this section, we consider sums of independent variables that are not necessarily identical. To control the probability of how close a sum of random variables is to the expected value, various concentration inequalities are in play. A typical version of the Chernoff inequalities, attributed to Herman Chernoff, can be stated as follows:

THEOREM 2.3. [**28**] *Let $X_1, \ldots, X_n$ be independent random variables with $\mathrm{E}(X_i) = 0$ and $|X_i| \leq 1$ for all i. Let $X = \sum_{i=1}^n X_i$ and let σ^2 be the variance of X_i. Then*

$$\Pr(|X| \geq k\sigma) \leq 2e^{-k^2/4n},$$

for any $0 \leq k \leq 2\sigma$.

If the random variables X_i under consideration assume non-negative values, the following version of Chernoff inequalities is often useful.

THEOREM 2.4. [**28**] *Let $X_1, \ldots, X_n$ be independent random variables with*

$$Pr(X_i = 1) = p_i, \qquad Pr(X_i = 0) = 1 - p_i.$$

We consider the sum $X = \sum_{i=1}^n X_i$, with expectation $\mathrm{E}(X) = \sum_{i=1}^n p_i$. Then we have

$$\begin{array}{lrcl} \text{(Lower tail)} & Pr(X \leq E(X) - \lambda) & \leq & e^{-\lambda^2/2\mathrm{E}(X)}, \\ \text{(Upper tail)} & Pr(X \geq E(X) + \lambda) & \leq & e^{-\frac{\lambda^2}{2(\mathrm{E}(X)+\lambda/3)}}. \end{array}$$

We remark that the term $\lambda/3$ appearing in the exponent of the bound for the upper tail is significant. This covers the case when the limit distribution is Poisson distribution as well as the normal distribution.

There are many variations of the Chernoff inequalities. Due to the fundamental nature of these inequalities, we will state several versions and then prove the strongest version from which all the other inequalities can be deduced. (See Figure 7 for the flowchart of these theorems.) In this section, we will prove Theorem 2.8 and deduce Theorems 2.6 and 2.5. Theorems 2.10 and 2.11 will be stated and proved in the next section. Theorems 2.9, 2.7, 2.13, 2.14 on the lower tail can be deduced by reflecting X to $-X$.

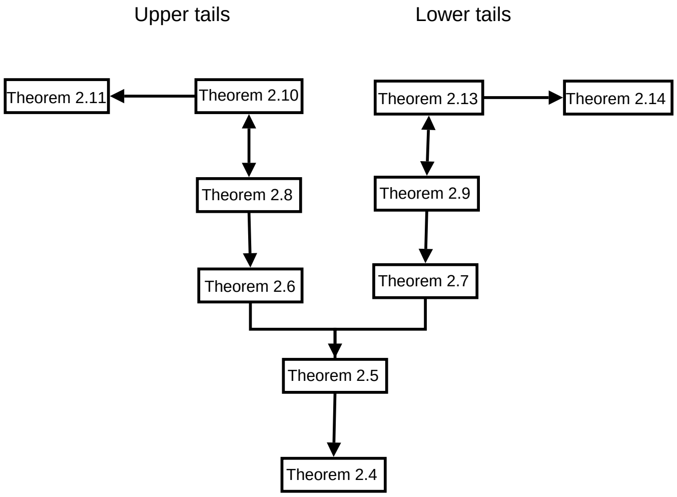

FIGURE 7. The flowchart for theorems on the sum of independent variables.

The following inequality is a generalization of the Chernoff inequalities for the binomial distribution:

THEOREM 2.5. [**33**] *Let $X_1, \ldots, X_n$ be independent random variables with*

$$Pr(X_i = 1) = p_i, \qquad Pr(X_i = 0) = 1 - p_i.$$

For $X = \sum_{i=1}^n a_i X_i$ with $a_i > 0$, we have $E(X) = \sum_{i=1}^n a_i p_i$ and we define $\nu = \sum_{i=1}^n a_i^2 p_i$. Then we have

$$Pr(X \leq E(X) - \lambda) \leq e^{-\lambda^2/2\nu} \tag{2.2}$$

$$Pr(X \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\nu + a\lambda/3)}} \tag{2.3}$$

where $a = \max\{a_1, a_2, \ldots, a_n\}$.

To compare inequalities (2.2) to (2.3), we consider an example in Figure 8. The cumulative distribution is the function $\Pr(X > x)$. The dotted curve in Figure 8 illustrates the cumulative distribution of the binomial distribution $B(1000, 0.1)$ with the value ranging from 0 to 1 as x goes from $-\infty$ to ∞. The solid curve at the lower-left corner is the bound $e^{-\lambda^2/2\nu}$ for the lower tail. The solid curve at the upper-right corner is the bound $1 - e^{-\frac{\lambda^2}{2(\nu + a\lambda/3)}}$ for the upper tail.

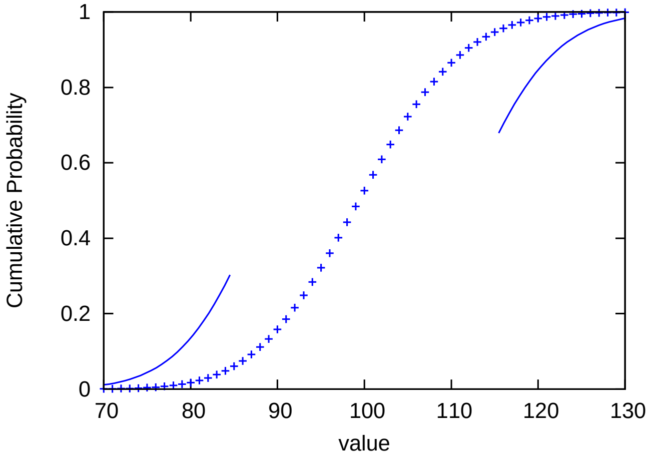

FIGURE 8. *Chernoff inequalities*

The inequality (2.3) in the above theorem is a corollary of the following general concentration inequality (also see Theorem 2.7 in the survey paper by McDiarmid [**94**]).

THEOREM 2.6. [**94**] *Let X_i ($1 \leq i \leq n$) be independent random variables satisfying $X_i \leq \mathrm{E}(X_i) + M$, for $1 \leq i \leq n$. We consider the sum $X = \sum_{i=1}^n X_i$ with expectation $\mathrm{E}(X) = \sum_{i=1}^n \mathrm{E}(X_i)$ and variance $\mathrm{Var}(X) = \sum_{i=1}^n \mathrm{Var}(X_i)$. Then we have*

$$\Pr(X \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X)+M\lambda/3)}}.$$

In the other direction, we have the following inequality.

THEOREM 2.7. *If $X_1, X_2, \ldots, X_n$ are non-negative independent random variables, we have the following bounds for the sum $X = \sum_{i=1}^n X_i$:*

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\frac{\lambda^2}{2\sum_{i=1}^n \mathrm{E}(X_i^2)}}.$$

A strengthened version of the above theorem is as follows:

THEOREM 2.8. *Suppose X_i are independent random variables satisfying $X_i \leq M$, for $1 \leq i \leq n$. Let $X = \sum_{i=1}^n X_i$ and $\|X\| = \sqrt{\sum_{i=1}^n \mathrm{E}(X_i^2)}$. Then we have*

$$\Pr(X \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\|X\|^2+M\lambda/3)}}.$$

Replacing X by $-X$ in the proof of Theorem 2.8, we have the following theorem for the lower tail.

THEOREM 2.9. *Let X_i be independent random variables satisfying $X_i \geq -M$, for $1 \leq i \leq n$. Let $X = \sum_{i=1}^n X_i$ and $\|X\| = \sqrt{\sum_{i=1}^n \mathrm{E}(X_i^2)}$. Then we have*

$$\Pr(X \leq E(X) - \lambda) \leq e^{-\frac{\lambda^2}{2(\|X\|^2+M\lambda/3)}}.$$

Before we give the proof of Theorems 2.8, we will first show the implications of Theorems 2.8 and 2.9. Namely, we will show that the other concentration inequalities can be derived from Theorems 2.8 and 2.9.

Fact: Theorem 2.8 $\Longrightarrow$ Theorem 2.6:

Proof. Let $X_i' = X_i - \mathrm{E}(X_i)$ and $X' = \sum_{i=1}^n X_i' = X - \mathrm{E}(X)$. We have

$$X_i' \le M \quad \text{for } 1 \le i \le n.$$

We also have

$$\begin{aligned} \|X'\|^2 &= \sum_{i=1}^n \mathrm{E}({X'}_i^2) \\ &= \sum_{i=1}^n \mathrm{E}((X_i - E(X_i))^2) \\ &= \sum_{i=1}^n \mathrm{Var}(X_i) \\ &= \mathrm{Var}(X). \end{aligned}$$

Applying Theorem 2.8, we get

$$\begin{aligned} \Pr(X \ge \mathrm{E}(X) + \lambda) &= \Pr(X' \ge \lambda) \\ &\le e^{-\frac{\lambda^2}{2(\|X'\|^2 + M\lambda/3)}} \\ &\le e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + M\lambda/3)}}. \end{aligned}$$

□

Fact: Theorem 2.9 $\Longrightarrow$ Theorem 2.7
The proof is straightforward by choosing $M = 0$.

Fact: Theorem 2.6 and 2.7 $\Longrightarrow$ Theorem 2.5

Proof. We define $Y_i = a_i X_i$. Note that

$$\|X\|^2 = \sum_{i=1}^n \mathrm{E}(Y_i^2) = \sum_{i=1}^n a_i^2 p_i = \nu.$$

Equation (2.2) follows from Theorem 2.7 since Y_i's are non-negatives.

For the other direction, we have

$$Y_i \le a_i \le a \le \mathrm{E}(Y_i) + a.$$

Equation (2.3) follows from Theorem 2.6. □

Fact: Theorem 2.8 and Theorem 2.9 $\Longrightarrow$ Theorem 2.3

The proof is by choosing $Y = X - E(X)$, $M = 1$ and applying Theorem 2.8 and 2.9 to Y.

Fact: Theorem 2.5 $\Longrightarrow$ Theorem 2.4

The proof is by choosing $a_1 = a_2 = \cdots = a_n = 1$.

Finally, we give the complete proof of Theorem 2.8 and thus finish the proofs for all the above theorems on Chernoff inequalities.

Proof of Theorem 2.8: We consider

$$E(e^{tX}) = E(e^{t\sum_i X_i}) = \prod_{i=1}^{n} E(e^{tX_i})$$

since the X_i's are independent.

We define $g(y) = 2\sum_{k=2}^{\infty} \frac{y^{k-2}}{k!} = \frac{2(e^y-1-y)}{y^2}$, and use the following facts about g:

- $g(0) = 1$.
- $g(y) \leq 1$, for $y < 0$.
- $g(y)$ is monotone increasing, for $y \geq 0$.
- For $y < 3$, we have

$$g(y) = 2\sum_{k=2}^{\infty} \frac{y^{k-2}}{k!} \leq \sum_{k=2}^{\infty} \frac{y^{k-2}}{3^{k-2}} = \frac{1}{1-y/3}$$

since $k! \geq 2 \cdot 3^{k-2}$. Then we have

$$\begin{aligned}
\mathrm{E}(e^{tX}) &= \prod_{i=1}^{n} \mathrm{E}(e^{tX_i}) \\
&= \prod_{i=1}^{n} \mathrm{E}(\sum_{k=0}^{\infty} \frac{t^k X_i^k}{k!}) \\
&= \prod_{i=1}^{n} \mathrm{E}(1 + t\mathrm{E}(X_i) + \frac{1}{2}t^2 X_i^2 g(tX_i)) \\
&\leq \prod_{i=1}^{n} (1 + t\mathrm{E}(X_i) + \frac{1}{2}t^2 \mathrm{E}(X_i^2) g(tM)) \\
&\leq \prod_{i=1}^{n} e^{t\mathrm{E}(X_i) + \frac{1}{2}t^2 \mathrm{E}(X_i^2) g(tM)} \\
&= e^{t\mathrm{E}(X) + \frac{1}{2}t^2 g(tM)\sum_{i=1}^{n}\mathrm{E}(X_i^2)} \\
&= e^{t\mathrm{E}(X) + \frac{1}{2}t^2 g(tM)\|X\|^2}.
\end{aligned}$$

Hence, for t satisfying $tM < 3$, we have

$$\begin{aligned}
\Pr(X \geq \mathrm{E}(X) + \lambda) &= \Pr(e^{tX} \geq e^{t\mathrm{E}(X) + t\lambda}) \\
&\leq e^{-t\mathrm{E}(X) - t\lambda}\mathrm{E}(e^{tX}) \\
&\leq e^{-t\lambda + \frac{1}{2}t^2 g(tM)\|X\|^2} \\
&\leq e^{-t\lambda + \frac{1}{2}t^2 \|X\|^2 \frac{1}{1-tM/3}}.
\end{aligned}$$

To minimize the above expression, we choose $t = \frac{\lambda}{\|X\|^2 + M\lambda/3}$. Therefore, $tM < 3$ and we have

$$\begin{aligned}
\Pr(X \geq \mathrm{E}(X) + \lambda) &\leq e^{-t\lambda + \frac{1}{2}t^2\|X\|^2 \frac{1}{1-tM/3}} \\
&= e^{-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}}.
\end{aligned}$$

The proof is complete. □

2.3. More concentration inequalities

Here we state several variations and extensions of the concentration inequalities as in Theorem 2.8. We first consider the upper tail.

Theorem 2.10. *Let X_i denote independent random variables satisfying $X_i \leq \mathrm{E}(X_i) + a_i + M$, for $1 \leq i \leq n$. For, $X = \sum_{i=1}^n X_i$, we have*

$$\Pr(X \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + \sum_{i=1}^n a_i^2 + M\lambda/3)}}.$$

Proof. Let $X_i' = X_i - \mathrm{E}(X_i) - a_i$ and $X' = \sum_{i=1}^n X_i'$. We have

$$X_i' \leq M \quad \text{for } 1 \leq i \leq n.$$

$$\begin{aligned}
X' - \mathrm{E}(X') &= \sum_{i=1}^n (X_i' - \mathrm{E}(X_i')) \\
&= \sum_{i=1}^n (X_i' + a_i) \\
&= \sum_{i=1}^n (X_i - \mathrm{E}(X_i)) \\
&= X - \mathrm{E}(X).
\end{aligned}$$

Thus,

$$\begin{aligned}
\|X'\|^2 &= \sum_{i=1}^n \mathrm{E}(X_i'^2) \\
&= \sum_{i=1}^n \mathrm{E}((X_i - \mathrm{E}(X_i) - a_i)^2) \\
&= \sum_{i=1}^n \mathrm{E}((X_i - \mathrm{E}(X_i))^2 + a_i^2 \\
&= \mathrm{Var}(X) + \sum_{i=1}^n a_i^2.
\end{aligned}$$

By applying Theorem 2.8, the proof is finished. □

Theorem 2.11. *Suppose X_i are independent random variables satisfying $X_i \leq \mathrm{E}(X_i) + M_i$, for $0 \leq i \leq n$. We order X_i's so that M_i are in increasing order. Let $X = \sum_{i=1}^n X_i$. Then for any $1 \leq k \leq n$, we have*

$$\Pr(X \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X)+\sum_{i=k}^n (M_i - M_k)^2 + M_k\lambda/3)}}.$$

Proof. For fixed k, we choose $M = M_k$ and

$$a_i = \begin{cases} 0 & \text{if } 1 \leq i \leq k \\ M_i - M_k & \text{if } k \leq i \leq n \end{cases}$$

We have

$$X_i - \mathrm{E}(X_i) \leq M_i \leq a_i + M_k. \quad \text{for } 1 \leq k \leq n.$$

$$\sum_{i=1}^n a_i^2 = \sum_{i=k}^n (M_i - M_k)^2.$$

Using Theorem 2.10, we have

$$\Pr(X_i \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X)+\sum_{i=k}^n (M_i - M_k)^2 + M_k\lambda/3)}}.$$

□

Example 2.12. *Let $X_1, X_2, \ldots, X_n$ be independent random variables. For $1 \leq i \leq n-1$, X_i follows the same distribution with*

$$\Pr(X_i = 0) = 1 - p \quad \textit{and} \quad \Pr(X_i = 1) = p.$$

X_n follows the distribution with

$$\Pr(X_n = 0) = 1 - p \quad \textit{and} \quad \Pr(X_n = \sqrt{n}) = p.$$

Consider the sum $X = \sum_{i=1}^n X_i$.

We have

$$\begin{aligned} \mathrm{E}(X) &= \sum_{i=1}^n \mathrm{E}(X_i) \\ &= (n-1)p + \sqrt{n}p. \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X) &= \sum_{i=1}^n \mathrm{Var}(X_i) \\ &= (n-1)p(1-p) + np(1-p) \\ &= (2n-1)p(1-p). \end{aligned}$$

Apply Theorem 2.6 with $M = (1-p)\sqrt{n}$. We have

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2((2n-1)p(1-p)+(1-p)\sqrt{n}\lambda/3)}}.$$

In particular, for constant $p \in (0,1)$ and $\lambda = \Theta(n^{\frac{1}{2}+\epsilon})$, we have

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\Theta(n^\epsilon)}.$$

Now we apply Theorem 2.11 with $M_1 = \ldots = M_{n-1} = (1-p)$ and $M_n = \sqrt{n}(1-p)$. We choose $k = n-1$, we have

$$\begin{aligned} \mathrm{Var}(X) + (M_n - M_{n-1})^2 &= (2n-1)p(1-p) + (1-p)^2(\sqrt{n}-1)^2 \\ &\leq (2n-1)p(1-p) + (1-p)^2 n \\ &\leq (1-p^2)n. \end{aligned}$$

Thus,

$$\Pr(X_i \geq E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2((1-p^2)n + (1-p)^2\lambda/3)}}.$$

For constant $p \in (0,1)$ and $\lambda = \Theta(n^{\frac{1}{2}+\epsilon})$, we have

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\Theta(n^{2\epsilon})}.$$

From the above examples, we note that Theorem 2.11 gives a significantly better bound than that in Theorem 2.6 if the random variables X_i have very different upper bounds.

For completeness, we also list the corresponding theorems for the lower tails. (These can be derived by replacing X by $-X$.)

THEOREM 2.13. *Let X_i denote independent random variables satisfying $X_i \geq \mathrm{E}(X_i) - a_i - M$, for $0 \leq i \leq n$. For $X = \sum_{i=1}^n X_i$, we have*

$$\Pr(X \leq E(X) - \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + \sum_{i=1}^n a_i^2 + M\lambda/3)}}.$$

THEOREM 2.14. *Let X_i denote independent random variables satisfying $X_i \geq \mathrm{E}(X_i) - M_i$, for $0 \leq i \leq n$. We order X_i's so that M_i are in increasing order. Let $X = \sum_{i=1}^n X_i$. Then for any $1 \leq k \leq n$, we have*

$$\Pr(X \leq E(X) - \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + \sum_{i=k}^n (M_i - M_k)^2 + M_k\lambda/3)}}.$$

Continuing the above example, we choose $M_1 = M_2 = \ldots = M_{n-1} = p$, and $M_n = \sqrt{n}p$. We choose $k = n-1$, so we have

$$\begin{aligned} \mathrm{Var}(X) + (M_n - M_{n-1})^2 &= (2n-1)p(1-p) + p^2(\sqrt{n}-1)^2 \\ &\leq (2n-1)p(1-p) + p^2 n \\ &\leq p(2-p)n. \end{aligned}$$

Using Theorem 2.14, we have

$$\Pr(X \leq E(X) - \lambda) \leq e^{-\frac{\lambda^2}{2(p(2-p)n + p^2\lambda/3)}}.$$

For a constant $p \in (0,1)$ and $\lambda = \Theta(n^{\frac{1}{2}+\epsilon})$, we have

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\Theta(n^{2\epsilon})}.$$

2.4. A concentration inequality with large error estimate

In the previous chapter, the Chernoff inequality gives very good probabilistic estimates when a random variable is close to its expected value. Suppose we allow the error bound to the expected value to be a positive fraction of the expected value. Then we can obtain even better bounds for the probability of the tails. The following two concentration inequalities can be found in [**100**].

Theorem 2.15. *Let X be a sum of independent random indicator variables. For any $\epsilon > 0$,*

$$\Pr(X \geq (1+\epsilon)\mathrm{E}(X)) \leq \left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{\mathrm{E}(X)}. \tag{2.4}$$

Theorem 2.16. *Let X be a sum of independent random indicator variables. For any $1 > \epsilon > 0$,*

$$\Pr(X \leq \epsilon\mathrm{E}(X)) \leq e^{-(1-\epsilon)^2\mathrm{E}(X)/2}. \tag{2.5}$$

The above inequalities, however, are still not enough for our applications in Chapter 7. We need the following somewhat stronger concentration inequality for the lower tail.

Theorem 2.17. *Let X be the sum of independent random indicator variables. For any $0 \leq \epsilon \leq e^{-1}$, we have*

$$\Pr(X \leq \epsilon\mathrm{E}(X)) \leq e^{-(1-2\epsilon(1-\ln\epsilon))\mathrm{E}(X)}. \tag{2.6}$$

Proof. Suppose that $X = \sum_{i=1}^{n} X_i$, where X_i's are independent random variables with

$$\Pr(X_i = 0) = 1 - p_i \text{ and } \Pr(X_i = 1) = p_i.$$

We have

$$\begin{aligned}
\Pr(X \leq \epsilon \mathrm{E}(X)) &= \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \Pr(X = k) \\
&= \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i \prod_{i \notin S} (1 - p_i) \\
&\leq \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i \prod_{i \notin S} e^{-p_i} \\
&= \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i e^{-\sum_{i \notin S} p_i} \\
&= \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i e^{-\sum_{i=1}^n p_i + \sum_{i \in S} p_i} \\
&\leq \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i e^{-\mathrm{E}(X) + k} \\
&\leq \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} e^{-\mathrm{E}(X) + k} \frac{(\sum_{i=1}^n p_i)^k}{k!} \\
&= e^{-\mathrm{E}(X)} \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \frac{(e\mathrm{E}(X))^k}{k!}.
\end{aligned}$$

When $\epsilon \mathrm{E}(X) < 1$, the statement is true since

$$\Pr(X \leq \epsilon \mathrm{E}(X)) \leq e^{-\mathrm{E}(X)} \leq e^{-(1 - 2\epsilon(1 - \ln \epsilon))\mathrm{E}(X)}.$$

Now we consider the case $\epsilon \mathrm{E}(X) \geq 1$.

Note that $g(k) = \frac{(e\mathrm{E}(X))^k}{k!}$ increases when $k < e\mathrm{E}(X)$. Let $k_0 = \lfloor \epsilon \mathrm{E}(X) \rfloor \leq \epsilon \mathrm{E}(X)$.

We have

$$\begin{aligned}
\Pr(X \leq \epsilon \mathrm{E}(X)) &\leq e^{-\mathrm{E}(X)} \sum_{k=0}^{k_0} \frac{(e\mathrm{E}(X))^k}{k!} \\
&\leq e^{-\mathrm{E}(X)} (k_0 + 1) \frac{(e\mathrm{E}(X))^{k_0}}{k_0!}.
\end{aligned}$$

By using the Stirling formula

$$n! \approx \sqrt{2\pi n}(\frac{n}{e})^n \geq (\frac{n}{e})^n,$$

we have

$$\begin{aligned}\Pr(X \leq \epsilon \mathrm{E}(X)) &\leq e^{-\mathrm{E}(X)}(k_0+1)\frac{(e\mathrm{E}(X))^{k_0}}{k_0!}\\ &\leq e^{-\mathrm{E}(X)}(k_0+1)(\frac{e^2\mathrm{E}(X)}{k_0})^{k_0}\\ &\leq e^{-\mathrm{E}(X)}(\epsilon\mathrm{E}(X)+1)(\frac{e^2}{\epsilon})^{\epsilon\mathrm{E}(X)}\\ &= (\epsilon\mathrm{E}(X)+1)e^{-(1-2\epsilon+\epsilon\ln\epsilon)\mathrm{E}(X)}.\end{aligned}$$

Here we replaced k_0 by $\epsilon\mathrm{E}(X)$ since the function $(x+1)(\frac{e^2\mathrm{E}(X)}{x})^x$ is increasing for $x < e\mathrm{E}(X)$.

To simplify the above expression, we have

$$\mathrm{E}(X) \geq \frac{1}{\epsilon} \geq \frac{1}{1-\epsilon}$$

since $\epsilon\mathrm{E}(X) \geq 1$ and $\epsilon \leq e^{-1} \leq 1-\epsilon$. Thus, $\epsilon\mathrm{E}(X)+1 \leq \mathrm{E}(X)$.

Also, we have $\mathrm{E}(X) \geq \frac{1}{\epsilon} \geq e$. The function $\frac{\ln x}{x}$ is decreasing for $x \geq e$. Thus,

$$\frac{\ln \mathrm{E}(X)}{\mathrm{E}(X)} \leq \frac{\ln\frac{1}{\epsilon}}{\frac{1}{\epsilon}} = -\epsilon\ln\epsilon.$$

We have

$$\begin{aligned}\Pr(X \leq \epsilon\mathrm{E}(X)) &\leq (\epsilon\mathrm{E}(X)+1)e^{-(1-2\epsilon+\epsilon\ln\epsilon)\mathrm{E}(X)}\\ &\leq \mathrm{E}(X)e^{-(1-2\epsilon)\mathrm{E}(X)}e^{-\epsilon\ln\epsilon\mathrm{E}(X)}\\ &\leq e^{-(1-2\epsilon)\mathrm{E}(X)}e^{-2\epsilon\ln\epsilon\mathrm{E}(X)}\\ &= e^{-(1-2\epsilon(1-\ln\epsilon))\mathrm{E}(X)}.\end{aligned}$$

The proof of Theorem 2.17 is complete. □

2.5. Martingales and Azuma's inequality

A martingale is a sequence of random variables $X_0, X_1, \ldots$ with finite means such that the conditional expectation of X_{n+1} given $X_0, X_1, \ldots, X_n$ is equal to X_n.

The above definition is given in the classical book of Feller [**51**], p. 210. However, the conditional expectation depends on the random variables under consideration and can be subtly difficult to deal with in various cases. In this book we will use the following definition which is concise and basically equivalent for the finite cases.

Suppose that Ω is a probability space with a probability distribution p. Let $\mathcal{F}$ denote a σ-field on Ω. (A σ-field on Ω is a collection of subsets of Ω which contains $\emptyset$ and Ω, and is closed under unions, intersections, and complementation.) In a σ-field $\mathcal{F}$ of Ω, the smallest set in $\mathcal{F}$ containing an element x is the intersection of all sets in $\mathcal{F}$ containing x. A function $f : \Omega \to \mathbb{R}$ is said to be $\mathcal{F}$-measurable if

$f(x) = f(y)$ for any y in the smallest set containing x. (For more terminology on martingales, the reader is referred to [**77**].)

If $f : \Omega \to \mathbb{R}$ is a function, we define the expectation $E(f) = E(f(x) \mid x \in \Omega)$ by

$$E(f) = E(f(x) \mid x \in \Omega) := \sum_{x \in \Omega} f(x)p(x).$$

If $\mathcal{F}$ is a σ-field on Ω, we define the conditional expectation $E(f \mid \mathcal{F}) : \Omega \to \mathbb{R}$ by the formula

$$E(f \mid \mathcal{F})(x) := \frac{1}{\sum_{y \in \mathcal{F}(x)} p(y)} \sum_{y \in \mathcal{F}(x)} f(y)p(y)$$

where $\mathcal{F}(x)$ is the smallest element of $\mathcal{F}$ which contains x.

A *filter* $\mathbf{F}$ is an increasing chain of σ-subfields

$$\{0, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}.$$

A martingale (obtained from) X is associated with a filter $\mathbf{F}$ and a sequence of random variables $X_0, X_1, \ldots, X_n$ satisfying $X_i = E(X \mid \mathcal{F}_i)$ and, in particular, $X_0 = E(X)$ and $X_n = X$.

EXAMPLE 2.18. *Given independent random variables $Y_1, Y_2, \ldots, Y_n$. We can define a martingale $X = Y_1 + Y_2 + \cdots + Y_n$ as follows. Let $\mathcal{F}_i$ be the σ-field generated by $Y_1, \ldots, Y_i$. (In other words, $\mathcal{F}_i$ is the minimum σ-field so that $Y_1, \ldots, Y_i$ are $\mathcal{F}_i$-measurable.) We have a natural filter $\mathbf{F}$:*

$$\{0, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}.$$

Let $X_i = \sum_{j=1}^{i} Y_j + \sum_{j=i+1}^{n} \mathrm{E}(Y_j)$. Then, $X_0, X_1, X_2, \ldots, X_n$ forms a martingale corresponding to the filter $\mathbf{F}$.

For $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ a vector with positive entries, the martingale X is said to be $\mathbf{c}$-Lipschitz if

$$|X_i - X_{i-1}| \le c_i \tag{2.7}$$

for $i = 1, 2, \ldots, n$. A powerful tool for controlling martingales is the following:

THEOREM 2.19 (Azuma's inequality). *If a martingale X is $\mathbf{c}$-Lipschitz, then*

$$\Pr(|X - E(X)| \ge \lambda) \le 2e^{-\frac{\lambda^2}{2\sum_{i=1}^{n} c_i^2}}, \tag{2.8}$$

where $\mathbf{c} = (c_1, \ldots, c_n)$.

THEOREM 2.20. *Let $X_1, X_2, \ldots, X_n$ be independent random variables satisfying*

$$|X_i - \mathrm{E}(X_i)| \le c_i \quad \textit{for } 1 \le i \le n.$$

Then we have the following bound for the sum $X = \sum_{i=1}^{n} X_i$.

$$\Pr(|X - \mathrm{E}(X)| \ge \lambda) \le 2e^{-\frac{\lambda^2}{2\sum_{i=1}^{n} c_i^2}}.$$

Proof of Azuma's inequality: For a fixed t, we consider the convex function $f(x) = e^{tx}$. For any $|x| \leq c$, $f(x)$ is below the line segment from $(-c, f(-c))$ to $(c, f(c))$. In other words, we have

$$e^{tx} \leq \frac{1}{2c}(e^{tc} - e^{-tc})x + \frac{1}{2}(e^{tc} + e^{-tc}).$$

Therefore, we can write

$$\begin{array}{rcl} \mathrm{E}(e^{t(X_i - X_{i-1})}|\mathcal{F}_{i-1}) & \leq & \mathrm{E}(\frac{1}{2c_i}(e^{tc_i} - e^{-tc_i})(X_i - X_{i-1}) + \frac{1}{2}(e^{tc_i} + e^{-tc_i})|\mathcal{F}_{i-1}) \\ & = & \frac{1}{2}(e^{tc_i} + e^{-tc_i}) \\ & \leq & e^{t^2c_i^2/2}. \end{array}$$

Here we apply the conditions $\mathrm{E}(X_i - X_{i-1}|\mathcal{F}_{i-1}) = 0$ and $|X_i - X_{i-1}| \leq c_i$.

Hence,

$$\mathrm{E}(e^{tX_i}|\mathcal{F}_{i-1}) \leq e^{t^2c_i^2/2}e^{tX_{i-1}}.$$

Inductively, we have

$$\begin{array}{rcl} \mathrm{E}(e^{tX}) & = & \mathrm{E}(\mathrm{E}(e^{tX_n}|\mathcal{F}_{n-1})) \\ & \leq & e^{t^2c_n^2/2}\mathrm{E}(e^{tX_{n-1}}) \\ & \leq & \cdots \\ & \leq & \prod_{i=1}^{n} e^{t^2c_i^2/2}\mathrm{E}(e^{tX_0}) \\ & = & e^{\frac{1}{2}t^2\sum_{i=1}^{n}c_i^2}e^{t\mathrm{E}(X)}. \end{array}$$

Therefore,

$$\begin{array}{rcl} \Pr(X \geq \mathrm{E}(X) + \lambda) & = & \Pr(e^{t(X - \mathrm{E}(X))} \geq e^{t\lambda}) \\ & \leq & e^{-t\lambda}\mathrm{E}(e^{t(X-\mathrm{E}(X))}) \\ & \leq & e^{-t\lambda}e^{\frac{1}{2}t^2\sum_{i=1}^{n}c_i^2} \\ & = & e^{-t\lambda + \frac{1}{2}t^2\sum_{i=1}^{n}c_i^2}. \end{array}$$

We choose $t = \frac{\lambda}{\sum_{i=1}^{n}c_i^2}$ (in order to minimize the above expression). We have

$$\begin{array}{rcl} \Pr(X \geq \mathrm{E}(X) + \lambda) & \leq & e^{-t\lambda + \frac{1}{2}t^2\sum_{i=1}^{n}c_i^2} \\ & = & e^{-\frac{\lambda^2}{2\sum_{i=1}^{n}c_i^2}}. \end{array}$$

To derive a similar lower bound, we consider $-X_i$ instead of X_i in the preceding proof. Then we obtain the following bound for the lower tail.

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\frac{\lambda^2}{2\sum_{i=1}^{n}c_i^2}}.$$

□

2.6. General martingale inequalities

Many problems which can be set up as a martingale do not satisfy the Lipschitz condition. It is desirable to be able to use tools similar to the Azuma inequality in such cases. In this section, we will first state and then prove several extensions of the Azuma inequality (see Figure 9).

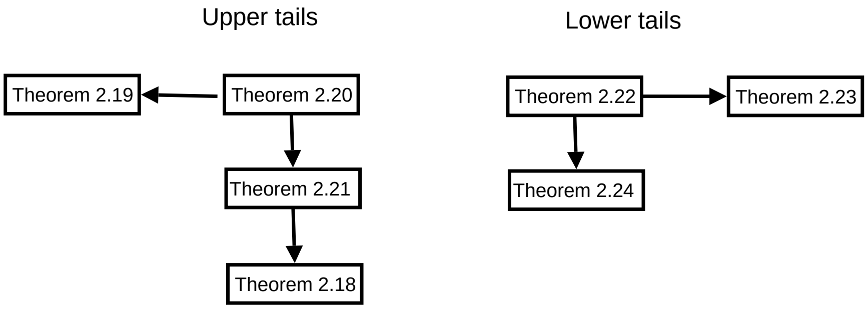

FIGURE 9. The flowchart for theorems on martingales.

Our starting point is the following well known concentration inequality (see [**94**]):

THEOREM 2.21. *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

(1) $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
(2) $|X_i - X_{i-1}| \leq M$, *for* $1 \leq i \leq n$.

Then, we have

$$\Pr(X - E(X) \geq \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n \sigma_i^2 + M\lambda/3)}}.$$

Since the sum of independent random variables can be viewed as a martingale (see Example 2.18), Theorem 2.21 implies Theorem 2.6. In a similar way, the following theorem is associated with Theorem 2.10.

THEOREM 2.22. *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

(1) $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
(2) $X_i - X_{i-1} \leq M_i$, *for* $1 \leq i \leq n$.

Then, we have

$$\Pr(X - E(X) \geq \lambda) \leq e^{-\frac{\lambda^2}{2\sum_{i=1}^n (\sigma_i^2 + M_i^2)}}.$$

The above theorem can be further generalized:

THEOREM 2.23. *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

(1) $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;

(2) $X_i - X_{i-1} \le a_i + M$, *for* $1 \le i \le n$.

Then, we have

$$\Pr(X - E(X) \ge \lambda) \le e^{-\frac{\lambda^2}{2(\sum_{i=1}^n (\sigma_i^2 + a_i^2) + M\lambda/3)}}.$$

Theorem 2.23 implies Theorem 2.21 by choosing $a_1 = a_2 = \cdots = a_n = 0$.

We also have the following theorem corresponding to Theorem 2.11.

THEOREM 2.24. *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

(1) $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \le \sigma_i^2$, *for* $1 \le i \le n$*;*
(2) $X_i - X_{i-1} \le M_i$, *for* $1 \le i \le n$.

Then, for any M, we have

$$\Pr(X - E(X) \ge \lambda) \le e^{-\frac{\lambda^2}{2(\sum_{i=1}^n \sigma_i^2 + \sum_{M_i > M} (M_i - M)^2 + M\lambda/3)}}.$$

Theorem 2.23 implies Theorem 2.24 by choosing

$$a_i = \begin{cases} 0 & \text{if } M_i \le M, \\ M_i - M & \text{if } M_i \ge M. \end{cases}$$

It suffices to prove Theorem 2.23 so that all the above stated theorems hold.

Proof of Theorem 2.23:

Recall that $g(y) = 2\sum_{k=2}^{\infty} \frac{y^{k-2}}{k!}$ satisfies the following properties:

- $g(y) \le 1$, for $y < 0$.
- $\lim_{y \to 0} g(y) = 1$.
- $g(y)$ is monotone increasing, for $y \ge 0$.
- When $b < 3$, we have $g(b) \le \frac{1}{1-b/3}$.

Since $\mathrm{E}(X_i|\mathcal{F}_{i-1}) = X_{i-1}$ and $X_i - X_{i-1} - a_i \le M$, we have

$$\begin{aligned}
\mathrm{E}(e^{t(X_i-X_{i-1}-a_i)}|\mathcal{F}_{i-1}) &= \mathrm{E}(\sum_{k=0}^{\infty} \frac{t^k}{k!}(X_i - X_{i-1} - a_i)^k|\mathcal{F}_{i-1}) \\
&= 1 - ta_i + \mathrm{E}(\sum_{k=2}^{\infty} \frac{t^k}{k!}(X_i - X_{i-1} - a_i)^k|\mathcal{F}_{i-1}) \\
&\le 1 - ta_i + \mathrm{E}(\frac{t^2}{2}(X_i - X_{i-1} - a_i)^2 g(tM)|\mathcal{F}_{i-1}) \\
&= 1 - ta_i + \frac{t^2}{2} g(tM)\mathrm{E}((X_i - X_{i-1} - a_i)^2|\mathcal{F}_{i-1}) \\
&= 1 - ta_i + \frac{t^2}{2} g(tM)(\mathrm{E}((X_i - X_{i-1})^2|\mathcal{F}_{i-1}) + a_i^2) \\
&\le 1 - ta_i + \frac{t^2}{2} g(tM)(\sigma_i^2 + a_i^2) \\
&\le e^{-ta_i + \frac{t^2}{2} g(tM)(\sigma_i^2+a_i^2)}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{E}(e^{tX_i}|\mathcal{F}_{i-1}) &= \mathrm{E}(e^{t(X_i-X_{i-1}-a_i)}|\mathcal{F}_{i-1})e^{tX_{i-1}+ta_i} \\
&\le e^{-ta_i + \frac{t^2}{2} g(tM)(\sigma_i^2+a_i^2)} e^{tX_{i-1}+ta_i} \\
&= e^{\frac{t^2}{2} g(tM)(\sigma_i^2+a_i^2)} e^{tX_{i-1}}.
\end{aligned}$$

Inductively, we have

$$\begin{aligned}
\mathrm{E}(e^{tX}) &= \mathrm{E}(\mathrm{E}(e^{tX_n}|\mathcal{F}_{n-1})) \\
&\le e^{\frac{t^2}{2} g(tM)(\sigma_n^2+a_n^2)} \mathrm{E}(e^{tX_{n-1}}) \\
&\le \cdots \\
&\le \prod_{i=1}^{n} e^{\frac{t^2}{2} g(tM)(\sigma_i^2+a_i^2)} \mathrm{E}(e^{tX_0}) \\
&= e^{\frac{1}{2}t^2 g(tM)\sum_{i=1}^n(\sigma_i^2+a_i^2)} e^{t\mathrm{E}(X)}.
\end{aligned}$$

Then for t satisfying $tM < 3$, we have

$$\begin{aligned}
\Pr(X \ge \mathrm{E}(X) + \lambda) &= \Pr(e^{tX} \ge e^{t\mathrm{E}(X)+t\lambda}) \\
&\le e^{-t\mathrm{E}(X)-t\lambda}\mathrm{E}(e^{tX}) \\
&\le e^{-t\lambda} e^{\frac{1}{2}t^2 g(tM)\sum_{i=1}^n(\sigma_i^2+a_i^2)} \\
&= e^{-t\lambda + \frac{1}{2}t^2 g(tM)\sum_{i=1}^n(\sigma_i^2+a_i^2)} \\
&\le e^{-t\lambda + \frac{1}{2}\frac{t^2}{1-tM/3}\sum_{i=1}^n(\sigma_i^2+a_i^2)}
\end{aligned}$$

We choose $t = \frac{\lambda}{\sum_{i=1}^n(\sigma_i^2+a_i^2)+M\lambda/3}$. Clearly $tM < 3$ and

$$\begin{aligned}
\Pr(X \ge \mathrm{E}(X) + \lambda) &\le e^{-t\lambda + \frac{1}{2}\frac{t^2}{1-tM/3}\sum_{i=1}^n(\sigma_i^2+c_i^2)} \\
&= e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+c_i^2)+M\lambda/3})}.
\end{aligned}$$

The proof of the theorem is complete. □

For completeness, we state the following theorems for the lower tails. The proofs are almost identical and will be omitted.

THEOREM 2.25. *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

(1) *$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \le \sigma_i^2$, for $1 \le i \le n$;*
(2) *$X_{i-1} - X_i \le a_i + M$, for $1 \le i \le n$.*

Then, we have

$$\Pr(X - E(X) \le -\lambda) \le e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+M\lambda/3)}}.$$

THEOREM 2.26. *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

(1) *$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \le \sigma_i^2$, for $1 \le i \le n$;*
(2) *$X_{i-1} - X_i \le M_i$, for $1 \le i \le n$.*

Then, we have

$$\Pr(X - E(X) \le -\lambda) \le e^{-\frac{\lambda^2}{2\sum_{i=1}^n(\sigma_i^2+M_i^2)}}.$$

THEOREM 2.27. *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

(1) *$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \le \sigma_i^2$, for $1 \le i \le n$;*
(2) *$X_{i-1} - X_i \le M_i$, for $1 \le i \le n$.*

Then, for any M, we have

$$\Pr(X - E(X) \le -\lambda) \le e^{-\frac{\lambda^2}{2(\sum_{i=1}^n \sigma_i^2+\sum_{M_i>M}(M_i-M)^2+M\lambda/3)}}.$$

2.7. Supermartingales and Submartingales

In this section, we consider further strengthened versions of the martingale inequalities that were mentioned so far. Instead of a fixed upper bound for the variance, we will assume that the variance $\mathrm{Var}(X_i|\mathcal{F}_{i-1})$ is upper bounded by a linear function of X_{i-1}. Here we assume this linear function is non-negative for all values that X_{i-1} takes. We first need some terminology.

For a filter $\mathbf{F}$:

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

a sequence of random variables $X_0, X_1, \ldots, X_n$ is called a *submartingale* if X_i is $\mathcal{F}_i$-measurable (i.e., $X_i(a) = X_i(b)$ if all elements of $\mathcal{F}_i$ containing a also contain b and vice versa) then $E(X_i \mid \mathcal{F}_{i-1}) \le X_{i-1}$, for $1 \le i \le n$.

A sequence of random variables $X_0, X_1, \ldots, X_n$ is said to be a *supermartingale* if X_i is $\mathcal{F}_i$-measurable and $E(X_i \mid \mathcal{F}_{i-1}) \ge X_{i-1}$, for $1 \le i \le n$.

To avoid repetition, we will first state a number of useful inequalities for for submartingales and supermartingales. Then we will give the proof for the general inequalities in Theorem 2.30 for submartingales and in Theorem 2.32) for supermartingales. Furthermore, we will show that all the stated theorems follow from

Theorems 2.30 and 2.32 (See Figure 10). Note that the inequalities for submartingales and supermartingales are not quite symmetric.

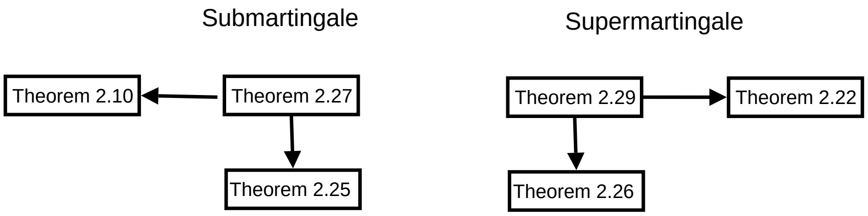

FIGURE 10. The flowchart for theorems on submartingales and supermartingales

THEOREM 2.28. *Suppose that a submartingale X, associated with a filter $\mathbf{F}$, satisfies*

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \phi_i X_{i-1}$$

and

$$X_i - E(X_i|\mathcal{F}_{i-1}) \leq M$$

for $1 \leq i \leq n$. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2((X_0+\lambda)(\sum_{i=1}^n \phi_i)+M\lambda/3)}}.$$

THEOREM 2.29. *Suppose that a supermartingale X, associated with a filter $\mathbf{F}$, satisfies, for $1 \leq i \leq n$,*

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \phi_i X_{i-1}$$

and

$$E(X_i|\mathcal{F}_{i-1}) - X_i \leq M.$$

Then we have

$$\Pr(X_n \leq X_0 - \lambda) \leq e^{-\frac{\lambda^2}{2(X_0(\sum_{i=1}^n \phi_i)+M\lambda/3)}},$$

for any $\lambda \leq X_0$.

THEOREM 2.30. *Suppose that a submartingale X, associated with a filter $\mathbf{F}$, satisfies*

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma^2 + \phi_i X_{i-1}$$

and

$$X_i - E(X_i|\mathcal{F}_{i-1}) \leq a_i + M$$

for $1 \leq i \leq n$. Here σ_i, a_i, ϕ_i and M are non-negative constants. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n (\sigma_i^2+a_i^2)+(X_0+\lambda)(\sum_{i=1}^n \phi_i)+M\lambda/3)}}.$$

REMARK 2.31. *Theorem 2.30 implies Theorem 2.28 by setting all σ_i's and a_i's to zero. Theorem 2.30 also implies Theorem 2.23 by choosing $\phi_1 = \cdots = \phi_n = 0$.*

The theorem for a supermartingale is slightly different due to the asymmetry of the condition on variance.

THEOREM 2.32. *Suppose a supermartingale X, associated with a filter $\mathbf{F}$, satisfies, for $1 \leq i \leq n$,*

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2 + \phi_i X_{i-1}$$

and

$$E(X_i|\mathcal{F}_{i-1}) - X_i \leq a_i + M,$$

where M, a_i's, σ_i's, and ϕ_i's are non-negative constants. Then we have

$$\Pr(X_n \leq X_0 - \lambda) \leq e^{-\frac{\lambda^2}{\sum_{i=1}^n (\sigma_i^2 + a_i^2) 2(X_0(\sum_{i=1}^n \phi_i) + M\lambda/3)}},$$

for any $\lambda \leq 2X_0 + \frac{\sum_{i=1}^n (\sigma_i^2 + a_i^2)}{\sum_{i=1}^n \phi}$.

REMARK 2.33. *Theorem 2.32 implies Theorem 2.29 by setting all σ_i's and a_i's to zero. Theorem 2.32 also implies Theorem 2.25 by choosing $\phi_1 = \cdots = \phi_n = 0$.*

Proof of Theorem 2.30:

For a positive t (to be chosen later), we consider

$$\begin{aligned}
E(e^{tX_i}|\mathcal{F}_{i-1}) &= e^{tE(X_i|\mathcal{F}_{i-1}) + ta_i} E(e^{t(X_i - E(X_i|\mathcal{F}_{i-1}) - a_i)}|\mathcal{F}_{i-1}) \\
&= e^{tE(X_i|\mathcal{F}_{i-1}) + ta_i} \sum_{k=0}^{\infty} \frac{t^k}{k!} E((X_i - E(X_i|\mathcal{F}_{i-1}) - a_i)^k|\mathcal{F}_{i-1}) \\
&\leq e^{tE(X_i|\mathcal{F}_{i-1}) + \sum_{k=2}^{\infty} \frac{t^k}{k!} E((X_i - E(X_i|\mathcal{F}_{i-1}) - a_i)^k|\mathcal{F}_{i-1})}
\end{aligned}$$

Recall that $g(y) = 2\sum_{k=2}^{\infty} \frac{y^{k-2}}{k!}$ satisfying

$$g(y) \leq g(b) < \frac{1}{1 - b/3}$$

for all $y \leq b$ and $0 \leq b \leq 3$.

Since $X_i - E(X_i|\mathcal{F}_{i-1}) - a_i \leq M$, we have

$$\begin{aligned}
\sum_{k=2}^{\infty} \frac{t^k}{k!} E((X_i - E(X_i|\mathcal{F}_{i-1}) - a_i)^k|\mathcal{F}_{i-1}) &\leq \frac{g(tM)}{2} t^2 \mathrm{E}((X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) - a_i)^2|\mathcal{F}_{i-1}) \\
&= \frac{g(tM)}{2} t^2 (\mathrm{Var}(X_i|\mathcal{F}_{i-1}) + a_i^2). \\
&\leq \frac{g(tM)}{2} t^2 (\sigma_i^2 + \phi_i X_{i-1} + a_i^2).
\end{aligned}$$

Since $E(X_i|\mathcal{F}_{i-1}) \leq X_{i-1}$, we have

$$\begin{aligned}
E(e^{tX_i}|\mathcal{F}_{i-1}) &\leq e^{tE(X_i|\mathcal{F}_{i-1}) + \sum_{k=2}^{\infty} \frac{t^k}{k!} E((X_i - E(X_i|\mathcal{F}_{i-1}-) - a_i)^k|\mathcal{F}_{i-1})} \\
&\leq e^{tX_{i-1} + \frac{g(tM)}{2} t^2 (\sigma_i^2 + \phi_i X_{i-1} + a_i^2)} \\
&= e^{(t + \frac{g(tM)}{2} \phi_i t^2) X_{i-1}} e^{\frac{t^2}{2} g(tM)(\sigma_i^2 + a_i^2)}.
\end{aligned}$$

We define $t_i \geq 0$ for $0 < i \leq n$, satisfying

$$t_{i-1} = t_i + \frac{g(t_0 M)}{2} \phi_i t_i^2,$$

while t_0 will be chosen later. Then

$$t_n \leq t_{n-1} \leq \cdots \leq t_0,$$

and

$$\begin{aligned} E(e^{t_i X_i}|\mathcal{F}_{i-1}) &\leq e^{(t_i + \frac{g(t_i M)}{2}\phi_i t_i^2)X_{i-1}} e^{\frac{t_i^2}{2} g(t_i M)(\sigma_i^2 + a_i^2)} \\ &\leq e^{(t_i + \frac{g(t_0 M)}{2} t_i^2 \phi_i)X_{i-1}} e^{\frac{t_i^2}{2} g(t_i M)(\sigma_i^2 + a_i^2)} \\ &= e^{t_{i-1} X_{i-1}} e^{\frac{t_i^2}{2} g(t_i M)(\sigma_i^2 + a_i^2)}. \end{aligned}$$

since $g(y)$ is increasing for $y > 0$.

By Markov's inequality, we have

$$\begin{aligned} \Pr(X_n \geq X_0 + \lambda) &\leq e^{-t_n(X_0+\lambda)} E(e^{t_n X_n}) \\ &= e^{-t_n(X_0+\lambda)} E(E(e^{t_n X_n}|\mathcal{F}_{n-1})) \\ &\leq e^{-t_n(X_0+\lambda)} E(e^{t_{n-1} X_{n-1}}) e^{\frac{t_i^2}{2} g(t_i M)(\sigma_i^2 + a_i^2)} \\ &\leq \cdots \\ &\leq e^{-t_n(X_0+\lambda)} E(e^{t_0 X_0}) e^{\sum_{i=1}^n \frac{t_i^2}{2} g(t_i M)(\sigma_i^2 + a_i^2)} \\ &\leq e^{-t_n(X_0+\lambda) + t_0 X_0 + \frac{t_0^2}{2} g(t_0 M) \sum_{i=1}^n (\sigma_i^2 + a_i^2)}. \end{aligned}$$

Note that

$$\begin{aligned} t_n &= t_0 - \sum_{i=1}^n (t_{i-1} - t_i) \\ &= t_0 - \sum_{i=1}^n \frac{g(t_0 M)}{2} \phi_i t_i^2 \\ &\geq t_0 - \frac{g(t_0 M)}{2} t_0^2 \sum_{i=1}^n \phi_i. \end{aligned}$$

Hence

$$\begin{aligned} \Pr(X_n \geq X_0 + \lambda) &\leq e^{-t_n(X_0+\lambda) + t_0 X_0 + \frac{t_0^2}{2} g(t_0 M) \sum_{i=1}^n (\sigma_i^2 + a_i^2)} \\ &\leq e^{-(t_0 - \frac{g(t_0 M)}{2} t_0^2 \sum_{i=1}^n \phi_i)(X_0+\lambda) + t_0 X_0 + \frac{t_0^2}{2} g(t_0 M) \sum_{i=1}^n (\sigma_i^2 + a_i^2)} \\ &= e^{-t_0 \lambda + \frac{g(t_0 M)}{2} t_0^2 (\sum_{i=1}^n (\sigma_i^2 + a_i^2) + (X_0+\lambda) \sum_{i=1}^n \phi_i)} \end{aligned}$$

Now we choose $t_0 = \frac{\lambda}{\sum_{i=1}^n (\sigma_i^2 + a_i^2) + (X_0+\lambda)(\sum_{i=1}^n \phi_i) + M\lambda/3}$. Using the fact that $t_0 M < 3$, we have

$$\begin{aligned} \Pr(X_n \geq X_0 + \lambda) &\leq e^{-t_0 \lambda + t_0^2 (\sum_{i=1}^n (\sigma_i^2 + a_i^2) + (X_0+\lambda) \sum_{i=1}^n \phi_i) \frac{1}{2(1 - t_0 M/3)}} \\ &= e^{-\frac{\lambda^2}{2(\sum_{i=1}^n (\sigma_i^2 + a_i^2) + (X_0+\lambda)(\sum_{i=1}^n \phi_i) + M\lambda/3)}}. \end{aligned}$$

The proof of the theorem is complete. □

Proof of Theorem 2.32:

The proof is quite similar to that of Theorem 2.30. The following inequality still holds.

$$\begin{aligned}
E(e^{-tX_i}|\mathcal{F}_{i-1}) &= e^{-tE(X_i|\mathcal{F}_{i-1})+ta_i}E(e^{-t(X_i-E(X_i|\mathcal{F}_{i-1})+a_i)}|\mathcal{F}_{i-1}) \\
&= e^{-tE(X_i|\mathcal{F}_{i-1})+ta_i}\sum_{k=0}^{\infty}\frac{t^k}{k!}E((E(X_i|\mathcal{F}_{i-1})-X_i-a_i)^k|\mathcal{F}_{i-1}) \\
&\leq e^{-tE(X_i|\mathcal{F}_{i-1})+\sum_{k=2}^{\infty}\frac{t^k}{k!}E((E(X_i|\mathcal{F}_{i-1})-X_i-a_i)^k|\mathcal{F}_{i-1})} \\
&\leq e^{-tE(X_i|\mathcal{F}_{i-1})+\frac{g(tM)}{2}t^2\mathrm{E}((X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1})-a_i)^2)} \\
&\leq e^{-tE(X_i|\mathcal{F}_{i-1})+\frac{g(tM)}{2}t^2(\mathrm{Var}(X_i|\mathcal{F}_{i-1})+a_i^2)} \\
&\leq e^{-(t-\frac{g(tM)}{2}t^2\phi_i)X_{i-1}}e^{\frac{g(tM)}{2}t^2(\sigma_i^2+a_i^2)}.
\end{aligned}$$

We now define $t_i \geq 0$, for $0 \leq i < n$ satisfying

$$t_{i-1} = t_i - \frac{g(t_nM)}{2}\phi_i t_i^2.$$

t_n will be defined later. Then we have

$$t_0 \leq t_1 \leq \cdots \leq t_n,$$

and

$$\begin{aligned}
E(e^{-t_iX_i}|\mathcal{F}_{i-1}) &\leq e^{-(t_i-\frac{g(t_iM)}{2}t_i^2\phi_i)X_{i-1}}e^{\frac{g(t_iM)}{2}t_i^2(\sigma_i^2+a_i^2)} \\
&\leq e^{-(t_i-\frac{g(t_nM)}{2}t_i^2\phi_i)X_{i-1}}e^{\frac{g(t_nM)}{2}t_i^2(\sigma_i^2+a_i^2)} \\
&= e^{-t_{i-1}X_{i-1}}e^{\frac{g(t_nM)}{2}t_i^2(\sigma_i^2+a_i^2)}.
\end{aligned}$$

By Markov's inequality, we have

$$\begin{aligned}
\Pr(X_n \leq X_0-\lambda) &= \Pr(-t_nX_n \geq -t_n(X_0-\lambda)) \\
&\leq e^{t_n(X_0-\lambda)}E(e^{-t_nX_n}) \\
&= e^{t_n(X_0-\lambda)}E(E(e^{-t_nX_n}|\mathcal{F}_{n-1})) \\
&\leq e^{t_n(X_0-\lambda)}E(e^{-t_{n-1}X_{n-1}})e^{\frac{g(t_nM)}{2}t_n^2(\sigma_n^2+a_n^2)} \\
&\leq \cdots \\
&\leq e^{t_n(X_0-\lambda)}E(e^{-t_0X_0})e^{\sum_{i=1}^{n}\frac{g(t_nM)}{2}t_i^2(\sigma_i^2+a_i^2)} \\
&\leq e^{t_n(X_0-\lambda)-t_0X_0+\frac{t_n^2}{2}g(t_nM)\sum_{i=1}^{n}(\sigma_i^2+a_i^2)}.
\end{aligned}$$

We note

$$\begin{aligned}
t_0 &= t_n + \sum_{i=1}^{n}(t_{i-1}-t_i) \\
&= t_n - \sum_{i=1}^{n}\frac{g(t_nM)}{2}\phi_i t_i^2 \\
&\geq t_n - \frac{g(t_nM)}{2}t_n^2\sum_{i=1}^{n}\phi_i.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\Pr(X_n \leq X_0 - \lambda) &\leq e^{t_n(X_0-\lambda)-t_0X_0+\frac{t_n^2}{2}g(t_nM)\sum_{i=1}^n(\sigma_i^2+a_i^2)} \\
&\leq e^{t_n(X_0-\lambda)-(t_n-\frac{g(t_nM)}{2}t_n^2)X_0+\frac{t_n^2}{2}g(t_nM)\sum_{i=1}^n(\sigma_i^2+a_i^2)} \\
&= e^{-t_n\lambda+\frac{g(t_nM)}{2}t_n^2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+(\sum_{i=1}^n\phi_i)X_0)}
\end{aligned}$$

We choose $t_n = \frac{\lambda}{\sum_{i=1}^n(\sigma_i^2+a_i^2)+(\sum_{i=1}^n\phi_i)X_0+M\lambda/3}$. We have $t_nM < 3$ and

$$\begin{aligned}
\Pr(X_n \leq X_0 - \lambda) &\leq e^{-t_n\lambda+t_n^2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+(\sum_{i=1}^n\phi_i)X_0)\frac{1}{2(1-t_nM/3)}} \\
&\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+X_0(\sum_{i=1}^n\phi_i)+M\lambda/3)}}.
\end{aligned}$$

It remains to verify that all t_i's are non-negative. Indeed,

$$\begin{aligned}
t_i &\geq t_0 \\
&\geq t_n - \frac{g(t_nM)}{2}t_n^2\sum_{i=1}^n\phi_i \\
&\geq t_n(1-\frac{1}{2(1-t_nM/3)}t_n\sum_{i=1}^n\phi_i) \\
&= t_n(1-\frac{\lambda}{2X_0+\frac{\sum_{i=1}^n(\sigma_i^2+a_i^2)}{\sum_{i=1}^n\phi_i}}) \\
&\geq 0.
\end{aligned}$$

The proof of the theorem is complete. □

2.8. The decision tree and relaxed concentration inequalities

In this section, we will extend and generalize previous theorems to a martingale which is not strictly Lipschitz but is *nearly* Lipschitz. Namely, the (Lipschitz-like) assumptions are allowed to fail for relatively small subsets of the probability space and we can still have similar but weaker concentration inequalities. Similar techniques have been introduced by Kim and Vu [**78**] in their important work on deriving concentration inequalities for multivariate polynomials. The basic setup for decision trees can be found in [**9**] and has been used in the work of Alon, Kim and spencer [**8**]. Wormald [**119**] considers martingales with a 'stopping time' that has a similar flavor. Here we use a rather general setting and we shall give a complete proof here.

We are only interested in finite probability spaces and we use the following computational model. The random variable X can be evaluated by a sequence of decisions $Y_1, Y_2, \ldots, Y_n$. Each decision has finitely many outputs. The probability that an output is chosen depends on the previous history. We can describe the process by a decision tree T, a complete rooted tree with depth n. Each edge uv of T is associated with a probability p_{uv} depending on the decision made from u to v. Note that for any node u, we have

$$\sum_v p_{u,v} = 1.$$

We allow p_{uv} to be zero and thus include the case of having fewer than r outputs for some fixed r. Let Ω_i denote the probability space obtained after the first i decisions. Suppose $\Omega = \Omega_n$ and X is the random variable on Ω. Let $\pi_i \colon \Omega \to \Omega_i$ be the projection mapping each point to the subset of points with the same first i decisions. Let $\mathcal{F}_i$ be the σ-field generated by $Y_1, Y_2, \ldots, Y_i$. (In fact, $\mathcal{F}_i = \pi_i^{-1}(2^{\Omega_i})$ is the full σ-field via the projection π_i.) The $\mathcal{F}_i$ form a natural filter:

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}.$$

The leaves of the decision tree are exactly the elements of Ω. Let $X_0, X_1, \ldots, X_n = X$ denote the sequence of decisions to evaluate X. Note that X_i is $\mathcal{F}_i$-measurable, and can be interpreted as a labeling on nodes at depth i.

There is one-to-one correspondence between the following:

- A sequence of random variables $X_0, X_1, \ldots, X_n$ satisfying X_i is $\mathcal{F}_i$-measurable, for $i = 0, 1, \ldots, n$.
- A vertex labeling of the decision tree T, $f \colon V(T) \to \mathbb{R}$.

In order to simplify and unify the proofs for various general types of martingales, here we introduce a definition for a function $f : V(T) \to \mathbb{R}$. We say f satisfies an *admissible* condition P if $P = \{P_v\}$ holds for every vertex v.

Examples of admissible conditions:

(1) **Supermartingale:** For $1 \leq i \leq n$, we have

$$\mathrm{E}(X_i | \mathcal{F}_{i-1}) \geq X_{i-1}.$$

Thus the admissible condition P_u holds if

$$f(u) \leq \sum_{v \in C(u)} p_{uv} f(v)$$

where C_u is the set of all children nodes of u and p_{uv} is the transition probability at the edge uv.

(2) **Subermartingale:** For $1 \leq i \leq n$, we have

$$\mathrm{E}(X_i | \mathcal{F}_{i-1}) \leq X_{i-1}.$$

In this case, the admissible condition of the submartingale is

$$f(u) \geq \sum_{v \in C(u)} p_{uv} f(v).$$

(3) **Martingale:** For $1 \leq i \leq n$, we have

$$\mathrm{E}(X_i | \mathcal{F}_{i-1}) = X_{i-1}.$$

The admissible condition of the martingale is then:

$$f(u) = \sum_{v \in C(u)} p_{uv} f(v).$$

(4) **c-Lipschitz:** For $1 \leq i \leq n$, we have

$$|X_i - X_{i-1}| \leq c_i.$$

The admissible condition of the **c**-Lipschitz property can be described as follows:

$$|f(u) - f(v)| \leq c_i, \quad \text{for any child } v \in C(u)$$

where the node u is at level i of the decision tree.

(5) **Bounded Variance:** For $1 \leq i \leq n$, we have

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$$

for some constants σ_i.

The admissible condition of the bounded variance property can be described as:

$$\sum_{v \in C(u)} p_{uv} f^2(v) - (\sum_{v \in C(u)} p_{uv} f(v))^2 \leq \sigma_i^2.$$

(6) **General Bounded Variance:** For $1 \leq i \leq n$, we have

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2 + \phi_i X_{i-1}$$

where σ_i, ϕ_i are non-negative constants, and $X_i \geq 0$. The admissible condition of the general bounded variance property can be described as follows:

$$\sum_{v \in C(u)} p_{uv} f^2(v) - (\sum_{v \in C(u)} p_{uv} f(v))^2 \leq \sigma_i^2 + \phi_i f(u), \quad \text{and } f(u) \geq 0$$

where i is the depth of the node u.

(7) **Upper-bound:** For $1 \leq i \leq n$, we have

$$X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) \leq a_i + M$$

where a_i's, and M are non-negative constants. The admissible condition of the upper bounded property can be described as follows:

$$f(v) - \sum_{v \in C(u)} p_{uv} f(v) \leq a_i + M, \quad \text{for any child } v \in C(u)$$

where i is the depth of the node u.

(8) **Lower-bound:** For $1 \leq i \leq n$, we have

$$\mathrm{E}(X_i|\mathcal{F}_{i-1}) - X_i \leq a_i + M$$

where a_i's, and M are non-negative constants. The admissible condition of the lower bounded property can be described as follows:

$$(\sum_{v \in C(u)} p_{uv} f(v)) - f(v) \leq a_i + M, \quad \text{for any child } v \in C(u)$$

where i is the depth of the node u.

For any labeling f on T and fixed vertex r, we can define a new labeling f_r as follows:

$$f_r(u) = \begin{cases} f(r) & \text{if } u \text{ is a descendant of } r. \\ f(u) & \text{otherwise.} \end{cases}$$

A property P is said to be *invariant* under subtree-unification if for any tree labeling f satisfying P, and a vertex r, f_r satisfies P.

We have the following theorem.

THEOREM 2.34. *The eight properties as stated in the preceding examples — supermartingale, submartingale, martingale,* **c**-*Lipschitz, bounded variance, general bounded variance, upper-bounded, and lower-bounded properties are all invariant under subtree-unification.*

PROOF. We note that these properties are all admissible conditions. Let P denote any one of these. For any node u, if u is not a descendant of r, then f_r and f have the same value on v and its children nodes. Hence, P_u holds for f_r since P_u does for f.

If u is a descendant of r, then $f_r(u)$ takes the same value as $f(r)$ as well as its children nodes. We verify P_u in each case. Assume that u is at level i of the decision tree T.

(1) For supermartingale, submartingale, and martingale properties, we have

$$\begin{aligned}\sum_{v\in C(u)} p_{uv} f_r(v) &= \sum_{v\in C(u)} p_{uv} f(r)\\ &= f(r)\sum_{v\in C(u)} p_{uv}\\ &= f(r)\\ &= f_r(u).\end{aligned}$$

Hence, P_u holds for f_r.

(2) For **c**-Lipschitz property, we have

$$|f_r(u) - f_r(v)| = 0 \le c_i, \qquad \text{for any child } v \in C(u).$$

Again, P_u holds for f_r.

(3) For the bounded variance property, we have

$$\begin{aligned}\sum_{v\in C(u)} p_{uv} f_r^2(v) - (\sum_{v\in C(u)} p_{uv} f_r(v))^2 &= \sum_{v\in C(u)} p_{uv} f^2(r) - (\sum_{v\in C(u)} p_{uv} f(r))^2\\ &= f^2(r) - f^2(r)\\ &= 0\\ &\le \sigma_i^2.\end{aligned}$$

(4) For the second bounded variance property, we have

$$f_r(u) = f(r) \ge 0.$$

$$\begin{aligned}\sum_{v\in C(u)} p_{uv} f_r^2(v) - (\sum_{v\in C(u)} p_{uv} f_r(v))^2 &= \sum_{v\in C(u)} p_{uv} f^2(r) - (\sum_{v\in C(u)} p_{uv} f(r))^2\\ &= f^2(r) - f^2(r)\\ &= 0\\ &\le \sigma_i^2 + \phi_i f_r(u).\end{aligned}$$

(5) For upper-bounded property, we have

$$\begin{aligned} f_r(v) - \sum_{v\in C(u)} p_{uv} f_r(v) &= f(r) - \sum_{v\in C(u)} p_{uv} f(r) \\ &= f(r) - f(r) \\ &= 0 \\ &\leq a_i + M. \end{aligned}$$

for any child v of u.

(6) For the lower-bounded property, we have

$$\begin{aligned} \sum_{v\in C(u)} p_{uv} f_r(v) - f_r(v) &= \sum_{v\in C(u)} p_{uv} f(r) - f(r) \\ &= f(r) - f(r) \\ &= 0 \\ &\leq a_i + M, \end{aligned}$$

for any child v of u.

Therefore, P_v holds for f_r and any vertex v. □.

For two admissible conditions P and Q, we define PQ to be the property, which is only true when both P and Q are true. If both admissible conditions P and Q are invariant under subtree-unification, then PQ is also invariant under subtree-unification.

For any vertex u of the tree T, an ancestor of u is a vertex lying on the unique path from the root to u. For an admissible condition P, the associated *bad* set B_i over X_i's is defined to be

$$B_i = \{v| \text{ the depth of } v \text{ is } i, \text{ and } P_u \text{ does not hold for some ancestor } u \text{ of } v\}.$$

LEMMA 2.35. *For a filter* $\mathbf{F}$

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose each random variable X_j *is* $\mathcal{F}_i$*-measurable, for* $0 \leq i \leq n$*. For any admissible condition* P*, let* B_i *be the associated bad set of* P *over* X_i*. There are random variables* $Y_0, \ldots, Y_n$ *satisfying:*

(1) Y_i *is* $\mathcal{F}_i$*-measurable.*
(2) $Y_0, \ldots, Y_n$ *satisfy condition* P*.*
(3) $\{x : Y_i(x) \neq X_i(x)\} \subset B_i$*, for* $0 \leq i \leq n$*.*

PROOF. We modify f and define f' on T as follows. For any vertex u,

$$f'(u) = \begin{cases} f(u) & \text{if } f \text{ satisfies } P_v \text{ holds for every ancestor } v \text{ of } u \text{ including } u \text{ itself.} \\ f(v) & v \text{ is the ancestor with smallest depth so that } f \text{ fails } P_v. \end{cases}$$

Let S be the set of vertices u satisfying

- f fails P_u,
- f satisfies P_v for every ancestor v of u.

It is clear that f' can be obtained from f by a sequence of subtree-unifications, where S is the set of the roots of subtrees. Furthermore, the order of subtree-unifications does not matter. Since P is invariant under subtree-unifications, the number of vertices that P fails decreases. Now we will show f' satisfies P.

Suppose to the contrary that f' fails P_u for some vertex u. Since P is invariant under subtree-unifications, f also fails P_u. By the definition, there is an ancestor v (of u) in S. After the subtree-unification on subtree rooted at v, P_u is satisfied. This is a contradiction.

Let $Y_0, Y_1, \ldots, Y_n$ be the random variables corresponding to the labeling f'. Y_i's satisfy the desired properties in (1)-(3). □

The following theorem generalizes Azuma's inequality. A similar but more restricted version can be found in [**78**].

Theorem 2.36. *For a filter* $\mathbf{F}$

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose the random variable X_i *is* $\mathcal{F}_i$*-measurable, for* $0 \leq i \leq n$. *Let* $B = B_n$ *denote the bad set associated with the following admissible condition:*

$$\begin{aligned} \mathrm{E}(X_i|\mathcal{F}_{i-1}) &= X_{i-1} \\ |X_i - X_{i-1}| &\leq c_i \end{aligned}$$

for $1 \leq i \leq n$ *where* $c_1, c_2, \ldots, c_n$ *are non-negative numbers. Then we have*

$$\Pr(|X_n - X_0| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}} + \Pr(B),$$

Proof. We use Lemma 2.35 which gives random variables $Y_0, Y_1, \ldots, Y_n$ satisfying properties (1)-(3) in the statement of Lemma 2.35. Then it satisfies

$$\begin{aligned} \mathrm{E}(Y_i|\mathcal{F}_{i-1}) &= Y_{i-1} \\ |Y_i - Y_{i-1}| &\leq c_i. \end{aligned}$$

In other words, $Y_0, \ldots, Y_n$ form a martingale which is $(c_1, \ldots, c_n)$-Lipschitz. By Azuma's inequality, we have

$$\Pr(|Y_n - Y_0| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}}.$$

Since $Y_0 = X_0$ and $\{x : Y_n(x) \neq X_n(x)\} \subset B_n = B$, we have

$$\begin{aligned} \Pr(|X_n - X_0| \geq \lambda) &\leq \Pr(|Y_n - Y_0| \geq \lambda) + \Pr(X_n \neq Y_n) \\ &\leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}} + \Pr(B). \end{aligned}$$

□

For $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ a vector with positive entries, a martingale is said to be near-$\mathbf{c}$-Lipschitz with an exceptional probability η if

$$\sum_i \Pr(|X_i - X_{i-1}| \geq c_i) \leq \eta. \tag{2.9}$$

Theorem 2.36 can be restated as follows:

Theorem 2.37. *For non-negative values, $c_1, c_2, \ldots, c_n$, a martingale X is near-$\mathbf{c}$-Lipschitz with an exceptional probability η. Then X satisfies*

$$\Pr(|X - E(X)| < a) \leq 2e^{-\frac{a^2}{2\sum_{i=1}^n c_i^2}} + \eta.$$

Now, we can use the same technique to relax all the theorems in the previous sections.

Here are the relaxed versions of Theorems 2.23, 2.28, and 2.30.

Theorem 2.38. *For a filter* $\mathbf{F}$

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a random variable X_j is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} E(X_i \mid \mathcal{F}_{i-1}) &\leq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \sigma_i^2 \\ X_i - E(X_i|\mathcal{F}_{i-1}) &\leq a_i + M \end{aligned}$$

where σ_i, a_i and M are non-negative constants. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+M\lambda/3)}} + \Pr(B).$$

Theorem 2.39. *For a filter* $\mathbf{F}$

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a non-negative random variable X_j is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} E(X_i \mid \mathcal{F}_{i-1}) &\leq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \phi_i X_{i-1} \\ X_i - E(X_i|\mathcal{F}_{i-1}) &\leq M \end{aligned}$$

where ϕ_i and M are non-negative constants. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2((X_0+\lambda)(\sum_{i=1}^n \phi_i)+M\lambda/3)}} + \Pr(B).$$

Theorem 2.40. *For a filter* $\mathbf{F}$

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a non-negative random variable X_j is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} E(X_i \mid \mathcal{F}_{i-1}) &\leq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \sigma_i^2 + \phi_i X_{i-1} \\ X_i - E(X_i|\mathcal{F}_{i-1}) &\leq a_i + M \end{aligned}$$

where σ_i, phi_i, a_i and M are non-negative constants. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+(X_0+\lambda)(\sum_{i=1}^n \phi_i)+M\lambda/3)}} + \Pr(B).$$

For supermartingales, we have the following relaxed versions of Theorem 2.25, 2.29, and 2.32.

Theorem 2.41. *For a filter* $\mathbf{F}$

$$\{\emptyset,\Omega\}=\mathcal{F}_0\subset\mathcal{F}_1\subset\cdots\subset\mathcal{F}_n=\mathcal{F},$$

suppose a random variable X_j *is* $\mathcal{F}_i$*-measurable, for* $0\leq i\leq n$. *Let* B *be the bad set associated with the following admissible conditions:*

$$\begin{array}{rcl} E(X_i\mid\mathcal{F}_{i-1}) & \geq & X_{i-1}\\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) & \leq & \sigma_i^2\\ E(X_i|\mathcal{F}_{i-1})-X_i & \leq & a_i+M \end{array}$$

where σ_i, a_i *and* M *are non-negative constants. Then we have*

$$\Pr(X_n\leq X_0-\lambda)\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+M\lambda/3)}}+\Pr(B).$$

Theorem 2.42. *For a filter* $\mathbf{F}$

$$\{\emptyset,\Omega\}=\mathcal{F}_0\subset\mathcal{F}_1\subset\cdots\subset\mathcal{F}_n=\mathcal{F},$$

suppose a random variable X_j *is* $\mathcal{F}_i$*-measurable, for* $0\leq i\leq n$. *Let* B *be the bad set associated with the following admissible conditions:*

$$\begin{array}{rcl} E(X_i\mid\mathcal{F}_{i-1}) & \geq & X_{i-1}\\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) & \leq & \phi_iX_{i-1}\\ E(X_i|\mathcal{F}_{i-1})-X_i & \leq & M \end{array}$$

where ϕ_i *and* M *are non-negative constants. Then we have*

$$\Pr(X_n\leq X_0-\lambda)\leq e^{-\frac{\lambda^2}{2(X_0(\sum_{i=1}^n\phi_i)+M\lambda/3)}}+\Pr(B).$$

for all $\lambda\leq X_0$.

Theorem 2.43. *For a filter* $\mathbf{F}$

$$\{\emptyset,\Omega\}=\mathcal{F}_0\subset\mathcal{F}_1\subset\cdots\subset\mathcal{F}_n=\mathcal{F},$$

suppose a non-negative random variable X_j *is* $\mathcal{F}_i$*-measurable, for* $0\leq i\leq n$. *Let* B *be the bad set associated with the following admissible conditions:*

$$\begin{array}{rcl} E(X_i\mid\mathcal{F}_{i-1}) & \geq & X_{i-1}\\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) & \leq & \sigma_i^2+\phi_iX_{i-1}\\ E(X_i|\mathcal{F}_{i-1})-X_i & \leq & a_i+M \end{array}$$

where σ_i,ϕ_i,a_i *and* M *are non-negative constants. Then we have*

$$\Pr(X_n\leq X_0-\lambda)\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+X_0(\sum_{i=1}^n\phi_i)+M\lambda/3)}}+\Pr(B),$$

for $\lambda<X_0$.

To see the powerful effect of the concentration and Martingale inequalities as stated in this chapter, the best way is to check out many interesting applications. Indeed, the inequalities here are especially useful for estimating the error bounds in the random graphs that we shall discuss in subsequent chapters. The applications for random graphs of the off-line models are easier than those for the on-line models. In fact, the concentration results in Chapter 3 (for the preferential attachment scheme) and Chapter 4 (for the duplication model) are all quite complicated. For a beginner, a good place to start is Chapter 5 on classical random graphs of the Erdős-Rényi model and the generalization of random graph models with given expected degrees. An earlier version of this chapter has appeared as a survey paper [**35**] and includes some applications there.

CHAPTER 3

A generative model - the preferential attachment scheme

The preferential attachment scheme is often attributed to Herbert Simon. In his paper [**106**] of 1955, he gave a model for word distribution using the preference attachment scheme and derived *Zipf's law*. Namely, the probability of a word having occurred exactly i times is proportional to $1/i$.

The basic setup for the preferential attachment scheme is a simple *local* growth rule, which however leads to a *global* consequence — a power law distribution. Since this local growth rule gives preferences to vertices with large degrees, the scheme is often described as "*the rich get richer*".

In this chapter, we shall give a clean and rigorous treatment of the preferential attachment scheme. Of interest is to determine the exponent of the power law from the parameters of the local growth rule.

3.1. Basic steps of the preferential attachment scheme

There are two parameters for the preferential attachment model:

- A probability p, where $0 \leq p \leq 1$.
- An initial graph G_0, that we have at time 0.

Usually, G_0 is taken to be the graph formed by one vertex having one loop. (We consider the degree of this vertex to be 1, and in general a loop adds 1 to the degree of a vertex.) Note, in this model multiple edges and loops are allowed.

We also have are two operations we do on a graph:

- *Vertex-step* — Add a new vertex v, and add an edge $\{u, v\}$ from v by randomly and independently choosing u in proportion to the degree of u in the current graph.
- *Edge-step* — Add a new edge $\{r, s\}$ by independently choosing vertices r and s with probability proportional to their degrees.

Note that for the edge-step, r and s could be the same vertex. Thus loops could be created. However, as the graph gets large, the probability of adding a loop can be well bounded and is quite small.

The random graph model $G(p, G_0)$ is assembled as follows:

```
Begin with the initial graph $G_0$.
For $t > 0$, at time $t$, the graph $G_t$ is formed by modifying $G_{t-1}$ as follows:
    with probability $p$, take a vertex-step,
    otherwise, take an edge-step.
```

When G_0 is the graph consisting of a single loop, we will simplify the notation and write $G(p) = G(p, G_0)$.

3.2. Analyzing the preferential attachment model

To analyze the graph generated by the preferential attachment model $G(p)$, we let n_t denote the number of vertices of $G(p)$ at time t and let e_t denote the number of edges of $G(p)$ at time t. We have

$$e_t = t + 1.$$

The number of vertices n_t, however, is a sum of t random indicator variables,

$$n_t = 1 + \sum_{i=1}^{t} s_t$$

where

$$\begin{aligned} Pr(s_j = 1) &= p, \\ Pr(s_j = 0) &= 1 - p. \end{aligned}$$

It follows that the expected value $E(n_t)$ satisfies

$$E(n_t) = 1 + pt.$$

To get a handle on the actual value of n_t, we use the binomial concentration inequality as described in Theorem 2.4. Namely,

$$Pr(|n_t - E(n_t)| > a) \leq e^{-a^2/(2pt+2a/3)}.$$

Thus, n_t is exponentially concentrated around $E(n_t)$.

The problem of interest is the degree distribution of a graph generated by $G(p)$. Let $m_{k,t}$ denote the number of vertices of degree k at time t. First we note that

$$m_{1,0} = 1, \text{ and } m_{0,k} = 0.$$

We wish to derive the recurrence for the expected value $E(m_{k,t})$. Note that a vertex of degree k at time t could have come from two cases, either it was a vertex of degree k at time $t - 1$ and had no edge added to it, or it was a vertex of degree $k - 1$ at time $t - 1$ and the new edge was put in adjacent to it. Let $\mathcal{F}_t$ be the σ-algebra associated with the probability space at time t. Thus, for $t > 0$ and $k > 1$, we have

$$\begin{aligned}\mathrm{E}(m_{k,t}|\mathcal{F}_{t-1}) &= m_{k,t-1}(1-\frac{kp}{2t}-\frac{(1-p)2k}{2t}) \\ &\quad +m_{k-1,t-1}(\frac{(k-1)p}{2t}+\frac{(1-p)2(k-1)}{2t}) \\ &= m_{k,t-1}(1-\frac{(2-p)k}{2t})+m_{k-1,t-1}(\frac{(2-p)(k-1)}{2t}). \qquad (3.1)\end{aligned}$$

If we take the expectation on both sides, we get the following recurrence formula.

$$\mathrm{E}(m_{k,t}) = \mathrm{E}(m_{k,t-1})(1-\frac{(2-p)k}{2t})+\mathrm{E}(m_{k-1,t-1})(\frac{(2-p)(k-1)}{2t}).$$

For $t > 0$ and $k = 1$, we have

$$E(m_{1,t}|\mathcal{F}_{t-1}) = m_{1,t-1}(1-\frac{(2-p)}{2t})+p. \qquad (3.2)$$

Thus,

$$\mathrm{E}(m_{1,t}) = \mathrm{E}(m_{1,t-1})(1-\frac{(2-p)}{2t})+p.$$

To solve this recurrence, some existing papers made the (unjustified) assumption $E(m_{k,t}) \approx a_k t$ where a_k is independent of k. The peril of such innocent-looking assumptions will be discussed later in this chapter.

Here we will give a rigorous proof that the expected values $E(m_{k,t})$ follow a power law when t goes to infinity. To do so, we invoke Lemma 3.1 (to be proved in the next section) which asserts that for a sequence $\{a_t\}$ satisfying the recursive relation $a_{t+1} = (1-\frac{b_t}{t})a_t + c_t$ the limit $\lim_{t\to\infty}\frac{a_t}{t}$ exists and

$$\lim_{t\to\infty}\frac{a_t}{t} = \frac{c}{1+b}$$

provided that $\lim_{t\to\infty} b_t = b > 0$ and $\lim_{t\to\infty} c_t = c$.

We proceed by induction on k to show that $\lim_{t\to\infty} E(m_{k,t})/t$ has a limit M_k for each k.

The first case is $k = 1$. In this case, we apply Lemma 3.1 with $b_t = b = (2-p)/2$ and $c_t = c = p$ to deduce that $\lim_{t\to\infty} E(m_{1,t})/t$ exists and

$$M_1 = \lim_{t\to\infty}\frac{E(m_{1,t})}{t} = \frac{2p}{4-p}.$$

Now we assume that $\lim_{t\to\infty} E(m_{k-1,t})/t$ exists and we apply the lemma again with $b_t = b = k(2-p)/2$ and $c_t = E(m_{k-1,t-1})(2-p)(k-1)/(2t)$, so $c = M_{k-1}(2-p)(k-1)/2$. Lemma 3.1 implies that the limit $\lim_{t\to\infty} E(m_{k,t})/t$ exists and is equal to

$$M_k = M_{k-1}\frac{(2-p)(k-1)}{2+k(2-p)} = M_{k-1}\frac{k-1}{k+\frac{2}{2-p}}. \qquad (3.3)$$

Thus we can write

$$M_k = \frac{2p}{4-p}\prod_{j=2}^{k}\frac{j-1}{j+\frac{2}{2-p}} = \frac{2p}{4-p}\frac{\Gamma(k)\Gamma(2+\frac{2}{2-p})}{\Gamma(k+1+\frac{2}{2-p})} \tag{3.4}$$

where $\Gamma(k)$ is the Gamma function.

We wish to show that the graph G generated by $G(p)$ is a power law graph with $M_k \propto k^{-\beta}$ (where $\propto$ means "is proportional to") for large k. If $M_k \propto k^{-\beta}$, then

$$\frac{M_k}{M_{k-1}} = \frac{k^{-\beta}}{(k-1)^{-\beta}} = (1-\frac{1}{k})^{\beta} = 1 - \frac{\beta}{k} + O(\frac{1}{k^2}).$$

From (3.3) we have

$$\frac{M_k}{M_{k-1}} = \frac{k-1}{k+\frac{2}{2-p}} = 1 - \frac{1+\frac{2}{2-p}}{k+\frac{2}{2-p}} = 1 - \frac{1+\frac{2}{2-p}}{k} + O(\frac{1}{k^2})$$

Thus we have an approximated power-law graph with

$$\beta = 1 + \frac{2}{2-p} = 2 + \frac{p}{2-p}.$$

Since p is between 0 and 1, the range for β is $2 \le \beta \le 3$ as illustrated in Figure 3.2.

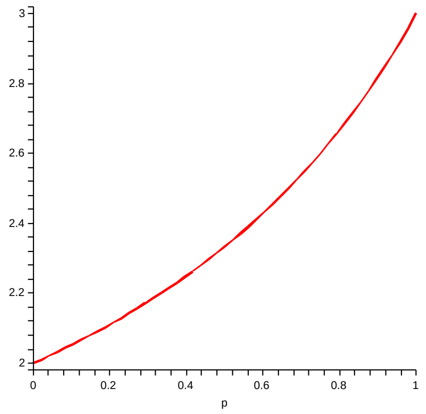

FIGURE 1. Exponent $\beta = 2 + \frac{p}{2-p}$ falls into the range $[2, 3]$.

The equation for M_k in (3.4) can be expressed by using the Beta function:

$$\begin{aligned} B(a,b) &= \int_0^1 x^{a-1}(1-x)^{b-1}dx \\ &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \end{aligned} \tag{3.5}$$

Therefore M_k satisfies

$$\begin{aligned}
M_k &= \frac{2p}{4-p}\frac{\Gamma(k)\Gamma(2+\frac{2}{2-p})}{\Gamma(k+1+\frac{2}{2-p})} \\
&= \frac{p(\beta-1)}{\beta}\frac{\Gamma(k)\Gamma(1+\beta)}{\Gamma(k+\beta)} \\
&= p(\beta-1)\frac{\Gamma(k)\Gamma(\beta)}{\Gamma(k+\beta)} \\
&= p(\beta-1)\int_0^1 x^{k-1}(1-x)^{\beta-1}dx \\
&= p(\beta-1)B(k,\beta)
\end{aligned}$$

Another consequence of the above derivation for M_k is the following nontrivial inequality:

$$\sum_{k=1}^{\infty}\frac{\Gamma(k)}{\Gamma(k+\beta)} = \frac{1}{\Gamma(\beta)(\beta-1)}. \tag{3.6}$$

One way to prove (3.6) is to use the fact that the expected number of vertices is $1+pt$. Since $\sum_{k=1}^{\infty} M_k = p$, the equation (3.6) immediately follows.

An alternative way to directly prove (3.6) is the following:

$$\begin{aligned}
\sum_{k=1}^{\infty}\frac{\Gamma(k)}{\Gamma(k+\beta)} &= \frac{1}{\Gamma(\beta)}\sum_{k=1}^{\infty}\frac{\Gamma(k)\Gamma(\beta)}{\Gamma(k+\beta)} \\
&= \frac{1}{\Gamma(\beta)}\sum_{k=1}^{\infty}B(k,\beta) \\
&= \frac{1}{\Gamma(\beta)}\sum_{k=1}^{\infty}\int_0^1 x^{k-1}(1-x)^{\beta-1}dx \\
&= \frac{1}{\Gamma(\beta)}\int_0^1\sum_{k=1}^{\infty}x^{k-1}(1-x)^{\beta-1}dx \\
&= \frac{1}{\Gamma(\beta)}\int_0^1(1-x)^{\beta-2}dx \\
&= \frac{1}{\Gamma(\beta)(\beta-1)}.
\end{aligned}$$

Equation 3.6 is proved.

3.3. A useful lemma for rigorous proofs

Lemma 3.1. *Suppose that a sequence $\{a_t\}$ satisfies the recursive relation*

$$a_{t+1} = (1-\frac{b_t}{t+t_1})a_t + c_t \textit{ for } t \geq t_0.$$

Furthermore, suppose $\lim_{t\to\infty} b_t = b > 0$ *and* $\lim_{t\to\infty} c_t = c$. *Then* $\lim_{t\to\infty} \frac{a_t}{t}$ *exists and*

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

Proof. Without loss of generality, we can assume $t_1 = 0$ after shifting t by t_1.

By rearranging the recurrence relation, we have

$$\begin{aligned}
\frac{a_{t+1}}{t+1} - \frac{c}{1+b} &= \frac{(1-\frac{b_t}{t})a_t + c_t}{t+1} - \frac{c}{1+b} \\
&= (\frac{a_t}{t} - \frac{c}{1+b})(\frac{t}{t+1})(1-\frac{b_t}{t}) + \frac{t}{t+1}(1-\frac{b_t}{t})(\frac{c}{1+b}) \\
&\quad + \frac{c_t}{t+1} - \frac{c}{1+b} \\
&= (\frac{a_t}{t} - \frac{c}{1+b})(1-\frac{1+b_t}{t+1}) + \frac{c_t}{t+1} - \frac{(1+b_t)c}{(t+1)(1+b)} \\
&= (\frac{a_t}{t} - \frac{c}{1+b})(1-\frac{1+b_t}{t+1}) + \frac{(1+b)c_t - (1+b_t)c}{(1+b)(1+t)}
\end{aligned}$$

Letting $s_t = |\frac{a_t}{t} - \frac{c}{1+b}|$, the triangle inequality now gives :

$$s_{t+1} \le s_t|1 - \frac{1+b_t}{t+1}| + |\frac{(1+b)c_t - (1+b_t)c}{(1+b)(1+t)}|.$$

Using the fact that $\lim_{t\to\infty} b_t = b$ and $\lim_{t\to\infty} c_t = c$, we have

$$|(1+b)c_t - (1+b_t)c| < \epsilon$$

for any fixed $\epsilon > 0$ provided t is sufficiently large. So, for some T, we have $b_t > b/2$ if $t \ge T$. Thus,

$$s_{t+1} - \epsilon < (s_t - \epsilon)(1 - \frac{1+b/2}{t})$$

Since $b > 0$, it is not difficult to show that $\prod(1-(1+b/2)/t)$ goes to 0 as $t \to \infty$. Repeated application of the above inequality gives $s_t < 2\epsilon$ for large t. Since ϵ can be arbitrarily chosen, we have $s_t \to 0$ as t goes to infinity as desired. Therefore we have proved that

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

□

3.4. The peril of heuristics via an example of balls-and-bins

Here we give an example of an incorrect deduction of the power law. This example of a balls-and-bins problem is a generalized version of Polya's urn problem and is quite interesting in its own.

The classical problem of Polya's urns has the following setup:

Start with a fixed number of bins each with one ball. At each tick of time, a new ball is placed in one of the bins with the probability of choosing the i-th bin proportional to the number of balls in the i-th bin.

Here we consider the ball-and-bin game when the number of bins is not fixed. We have two parameters, p, a probability between 0 and 1 and a real number r. We call this model Polya(p, r).

Imagine we have a stream of balls arriving one at a time.
    At the very beginning, we place the first ball in a bin.
    At time t, with probability p, we place the newly arrived ball in a new bin.
        Otherwise, we place the new ball in an existing bin
        and we choose a bin with probability proportional to the
        rth-power of the number of the balls in the bin.

We can modify Polya(p, r) into the following model, denoted by Polya$^*(p, r)$:

We have a stream of balls arriving two at a time.
  At the very beginning, we place the first set of two balls in a bin.
  At time t, with probability p, we place one new ball in a new bin and the other
        ball in an existing bin with probability proportional to the
        rth-power of the bin size.
        Otherwise, we place the each of the two new balls in an
        existing bin with probability proportional to the rth-power of
        the bin size.

For the case of $r = 1$, the model Polya$^*(p, 1)$ is just the preferential attachment model in Section 3.1 if we view the bins as vertices and edges are between the bins the two balls that arrive at the same time go into. The model Polya$^*(p, r)$ is regarded as a preferential attachment with *feedback*. When $r > 1$, it is preferential attachment with *positive* feedback. When $r < 1$, it is preferential attachment with *negative* feedback. This general form of preferential attachment has been examined in a number of papers [**31, 46, 47, 83, 102**]. For example, it was shown that for $r > 1$, a single bin dominates. In fact, for any $k > r/(r - 1)$, with high probability only finitely many bins ever reach size k.

In the remainder of this section, we will give a "proof" that for $r > 1$ in Polya(p, r), the bin sizes have a power law distribution. The exercise here is to find what is wrong in this "proof"!

Let $n_k(t)$ be the number of bins at time t with k balls. Note that

$$E[n_k(t+1)] = E[n_k(t)(p + (1-p)(1 - \frac{k^r}{w_t}))] + (1-p)E[\frac{n_{k-1}(t)(k-1)^r}{w_t}]$$

where w_t denotes $\sum_i n_i(t) i^r$. Let us assume that as t gets large $E(n_k(t))$ converge to fixed fractions of the total number of balls. In other words, $n_k(t) \approx a_k t$. *(A very dangerous assumption indeed!)* Furthermore, assume w_t converges to wt for some constant w. By plugging those assumptions in the above equation, we get

$$a_k(t+1) = a_k t(p + (1-p)(1 - \frac{k^r}{wt})) + (1-p)a_{k-1}t\frac{(k-1)^r}{wt}.$$

This implies

$$\begin{aligned}\frac{a_k}{a_{k-1}} &= \frac{(1-p)(k-1)^r}{w + (1-p)k^r} \\ &= \frac{(k-1)^r}{\frac{w}{1-p} + k^r} \\ &\approx \left(\frac{k-1}{k}\right)^r\end{aligned}$$

for k large. Thus, one might be inclined to conclude that the bin size distribution is a power law distribution with exponent r if $r > 1$!

However, the truth (see [**31**]) is that all but one of the a_i's are zero. A quick simulation will show that almost all balls go into one bin. In fact, it can be shown that all balls go into one bin with the exceptions of the balls in bins of size 1 and finitely many other balls. This model gives an explanation for the forming of a *monopoly*.

What went wrong in the above "proof"? The power law distribution is a consequence of an unfortunate ratio 0/0. That is exactly why rigorous mathematics is needed here.

3.5. Scale-free networks

Quite a few recent papers use the term "scale-free networks" to mean graphs with a power law degree distribution. However, power law and scale-free are very different concepts. In fact, the term "scale-free" has rarely been properly defined.

Here we intend to clarify the distinction of the two. To discuss "scale-free", first we have to answer the question concerning "scale". What is the appropriate scale or scales? How should "scale-free" be defined in a natural way?

Two types of scale come to mind — *space* and *time*. In fact, scales of space and time can coexist simultaneously. For example, the Call graphs have very similar shape (the same exponent in the power law distribution) while sampling at different geographical locations and at different sampling intervals. To simplify the issues, we separately discuss "scale-free in space" and "scale-free in time".

3.5.1. Scale-free in space. "Self-similarity" is one of the visible traits that exist in numerous networks. By comparing the web crawls of [**6, 14**] and [**27, 84**] we see that the same power law appears to govern various subgraphs of the web as well as the whole. However, while some subgraphs obey the same power law and appear to be self-similar, (i.e., similar to the entire graph), there exist subgraphs of the web which would not obey the power law (e.g., the subgraph defined by all nodes with out-degree 50). So, for what kind of subgraphs can "self-similarity" be considered or even formally defined?

As an example, for the family of recursive trees [**93**] as rooted trees, the definition comes naturally. The special subtrees consisting of all descendants of a vertex is similar to the whole tree.

For a general graph, additional information will be needed to help define the special subgraphs for which self-similarity will hold. One direction is to consider a geometric embedding of the graph into some specified metric space. Then we use the metric to define the special subgraphs. Another direction is to take the graph as given but to extract a so-called "local graph" from it. The graphic metric of the local graph provides the geometry of the graph. In Chapter 12, we will define the local graphs and discuss this idea further.

3.5.2. Scale-free in time. It is easier to define scale-free in terms of time than space perhaps because time is one-dimensional but space is multi-dimensional. The generative model is a process of growing graphs by adding nodes and edges one at a time. One way is to divide the time into almost equal units and combine all nodes born in the same unit time into one super-node. The bigger time unit one chooses, the fewer nodes the resulting graph has. We say a model is *scale-free* if it generates power law graphs with the same exponent regardless the choice of time scale. In other words, a generative model is invariant with respect to time in the sense that if we change the time scale by any given factor, then the original graph and the scaled graph should satisfy the power law with the same exponent for the degrees.

We can modify the previous model by adding an additional integer parameter m. Here are the two generalized steps:

- *Vertex-m-step* — Add a new vertex v, and m new edges $\{u_i, v\}$, $i = 1, \ldots, m$, by choosing u_i with probability proportional to the degree of u in the current graph.
- *Edge-m-step* — Add m new edges $\{r_i, s_i\}$, $i = 1, \ldots, m$, by choosing vertices r_i with probability proportional to the degree of r_i, and by choosing vertices s_i with probability proportional to the degree of s_i.

Now we assemble a graph $G(p, m, G_0)$:

```
Begin with the initial graph $G_0$.
For $t > 0$, at time $t$,
        with probability $p$, take a vertex-$m$-step,
        otherwise, take an edge-$m$-step.
```

If G_0 is taken to be the graph consisting of a vertex with m loops, we write $G(p, m) = G(p, m, G_0)$.

In this model every vertex has degree at least m. Let $m_{k,t}$ be the number of vertices with degree k at time t. At time t, G_t has exactly $e_0 + mt$ edges. We will denote this by e_t. Let $\mathcal{F}_t$ be the σ-algebra generated by the probability space at time t. Thus, for $t > 0$ and $k > m$, we have

$$\begin{aligned}
\mathrm{E}(m_{k,t}|\mathcal{F}_{t-1}) &= m_{k,t-1}(1 - \frac{kmp}{2e_{t-1}} - \frac{m(1-p)2k}{2e_{t-1}}) \\
&\quad + m_{k-1,t-1}(\frac{(k-1)mp}{2e_{t-1}} + \frac{(1-p)2m(k-1)}{2e_{t-1}}) + O(\frac{1}{t^2}) \\
&= m_{k,t-1}(1 - \frac{(2-p)mk}{2e_{t-1}}) + m_{k-1,t-1}(\frac{(2-p)m(k-1)}{2e_{t-1}}) + O(\frac{1}{t^2}). \qquad (3.7)
\end{aligned}$$

Note that the $O(1/t^2)$ term above makes it possible to absorb the error terms caused by loops or multiple edges. Now by taking the expectation on both sides, we get the following recurrence formula.

$$\mathrm{E}(m_{k,t}) = \mathrm{E}(m_{k,t-1})(1 - \frac{(2-p)mk}{2e_{t-1}}) + \mathrm{E}(m_{k-1,t-1})(\frac{(2-p)m(k-1)}{2e_{t-1}}) + O(\frac{1}{t^2}).$$

In the random graph model $G(p, m)$, we have $e_t = m(t+1)$. If we substitute e_t in the above inequality, all appearances of m are cancelled out. Indeed, we get exactly the same recurrence formula as we previously had for $G(p)$ in (3.1). Therefore, graphs generated by $G(p, m)$ has the same power law distribution as graphs generated by $G(p)$. So we see the exponent β is independent of the scale unit m.

If we compare the figures of the degree distributions of $G(p)$ and $G(p, m)$ in their logarithmic representation, the figures are almost identical in the sense that the shape of the curves are straight lines of the same slope. The only difference is that the line associated with $G(p, m)$ is a slight linear translation to the right. Mainly, the density of $G(p, m)$ differs from that of $G(p)$ by a factor of m. In the logarithmic representation, the difference is an additive term of $\log m$, which is rather small in comparison with n, the number of nodes. Nevertheless, the main characteristic of the power law is the exponent of the power law as seen from the same slope in both figures.

3.6. The sharp concentration of preference attachment scheme

In section 3.2 we considered the expected degrees for graphs generated by the preference attachment scheme and we derived the power law distribution for the expected degree sequence. However, the expected degree can be quite different from the actual degree of a random graph in hand. Can we give a (probabilistic) estimate of the difference? The goal of the section is to answer this question.

Since the preference attachment scheme is an on-line model, a concentration bound that we intend to give involves nontrivial arguments and is somewhat lengthy.

We will prove the following theorem.

THEOREM 3.2. *For the preferential attachment model $G(p)$, almost surely the number of vertices with degree k at time t is*

$$M_k t + O(2\sqrt{k^3 t \ln(t)}).$$

Recall $M_1 = \frac{2p}{4-p}$ *and* $M_k = \frac{2p}{4-p}\frac{\Gamma(k)\Gamma(1+\frac{2}{2-p})}{\Gamma(k+1+\frac{2}{2-p})} = O(k^{-(2+\frac{p}{2-p})})$*, for* $k \geq 2$*. In other words, almost surely the graphs generated by* $G(p)$ *have the power law degree distribution with the exponent* $\beta = 2 + \frac{p}{2-p}$*.*

Proof. We have shown that

$$\lim_{t\to\infty}\frac{\mathrm{E}(m_{k,t})}{t} = M_k,$$

where M_k is defined recursively in (3.3). It is sufficient to show $m_{k,t}$ concentrates on the expected value.

We shall prove the following claim.

Claim: For any fixed $k \geq 1$, for any $c > 0$, with probability at least $1 - 2(t+1)^{k-1}e^{-c^2}$, we have

$$|m_{k,t} - M_k(t+1)| \leq 2kc\sqrt{t}.$$

To see that the claim implies Theorem 3.2, we choose $c = \sqrt{k\ln t}$. Note that

$$2(t+1)^{k-1}e^{-c^2} = 2(t+1)^{k-1}t^{-k} = o(1).$$

From the Claim, with probability $1 - o(1)$, we have

$$|m_{k,t} - M_k(t+1)| \leq 2\sqrt{k^3 t\ln t},$$

as desired.

It remains to prove the claim.

Proof of Claim: We shall prove it by induction on k.

The base case of $k = 1$*:*

For $k = 1$, from equation 3.2, we have

$$\begin{aligned}
\mathrm{E}(m_{1,t} - M_1(t+1)|\mathcal{F}_{t-1}) &= \mathrm{E}(m_{1,t}|\mathcal{F}_{t-1}) - M_1(t+1) \\
&= m_{1,t-1}(1-\frac{2-p}{2t}) + p - M_1t - M_1 \\
&= (m_{1,t-1} - M_1t)(1-\frac{2-p}{2t}) + p - M_1\frac{2-p}{2} - M_1 \\
&= (m_{1,t-1} - M_1t)(1-\frac{2-p}{2t})
\end{aligned}$$

since $M_1 = \frac{2p}{4-p}$ and $p - M_1\frac{2-p}{2} - M_1 = 0$.

Let $X_{1,t} = \frac{m_{1,t} - M_1(t+1)}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}$. We consider the martingale formed by $1 = X_{1,0}, X_{1,1}, \cdots, X_{1,t}$.

We have

$$\begin{aligned} X_{1,t} - X_{1,t-1} &= \frac{m_{1,t} - M_1(t+1)}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})} - \frac{m_{1,t-1} - M_1 t}{\prod_{j=1}^{t-1}(1-\frac{2-p}{2j})} \\ &= \frac{1}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}[(m_{1,t} - M_1(t+1)) - (m_{1,t-1} - M_1 t)(1 - \frac{2-p}{2t})] \\ &= \frac{1}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}[(m_{1,t} - m_{1,t-1}) + \frac{2-p}{2t}(m_{1,t-1} - M_1 t) - M_1]. \end{aligned}$$

We note that $|m_{1,t} - m_{1,t-1}| \leq 2$, $m_{1,t-1} \leq t$, and $M_1 = \frac{2p}{4-p} < 1$. We have

$$|X_{1,t} - X_{1,t-1}| \leq \frac{4}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}. \tag{3.8}$$

Since $|m_{1,t} - m_{1,t-1}| \leq 2$, we have

$$\begin{aligned} \mathrm{Var}(m_{1,t}|\mathcal{F}_{t-1}) &\leq \mathrm{E}((m_{1,t} - m_{1,t-1})^2|\mathcal{F}_{t-1}) \\ &\leq 4. \end{aligned}$$

Therefore, we have the following upper bound for $\mathrm{Var}(X_{1,t}|\mathcal{F}_{t-1})$.

$$\begin{aligned} \mathrm{Var}(X_{1,t}|\mathcal{F}_{t-1}) &= \mathrm{Var}\big((m_{1,t} - M_1(t+1))\frac{1}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}\big|\mathcal{F}_{t-1}\big) \\ &= \frac{1}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})^2}\mathrm{Var}(m_{1,t} - M_1(t+1)|\mathcal{F}_{t-1}) \\ &= \frac{1}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})^2}\mathrm{Var}(m_{1,t}|\mathcal{F}_{t-1}) \\ &\leq \frac{4}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})^2}. \end{aligned} \tag{3.9}$$

We apply Theorem 2.22 on the martingale $\{X_{1,t}\}$ with $\sigma_i^2 = \frac{4}{\prod_{j=1}^{i}(1-\frac{2-p}{2j})^2}$, $M = \frac{4}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}$ and $a_i = 0$. We have

$$\Pr(X_{1,t} \geq \mathrm{E}(X_{1,t}) + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2 + M\lambda/3)}}.$$

Here $\mathrm{E}(X_{1,t}) = X_{1,0} = 1$. We will use the following approximation.

$$\begin{aligned} \prod_{j=1}^{i}(1 - \frac{2-p}{2j}) &= \frac{\Gamma(i+\frac{p}{2})}{\Gamma(i+1)\Gamma(\frac{p}{2})} \\ &= (\frac{1}{\Gamma(\frac{p}{2})} + O(\frac{1}{i}))i^{-1+p/2}. \end{aligned}$$

For any $c > 0$, we choose $\lambda = \frac{2c\sqrt{t}}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})} \approx 2\Gamma(\frac{p}{2})ct^{(3-p)/2}$. We have

$$\begin{aligned}\sum_{i=1}^{t}\sigma_i^2 &= \sum_{i=1}^{t}\frac{4}{\prod_{j=1}^{i}(1-\frac{2-p}{2j})^2}\\ &\approx \sum_{i=1}^{t}4\Gamma^2(\frac{p}{2})i^{2-p}\\ &\approx \frac{4\Gamma^2(\frac{p}{2})}{3-p}t^{3-p}\\ &< 2\Gamma^2(\frac{p}{2})t^{3-p}.\end{aligned}$$

We note that

$$M\lambda/3 \approx \frac{4}{3}\Gamma^2(\frac{p}{2})ct^{5/2-p} < 2\Gamma^2(\frac{p}{2})t^{3-p}$$

provided $c < \sqrt{t}$. We have

$$\begin{aligned}\Pr(X_{1,t} \geq 1+\lambda) &\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2+M\lambda/3)}}\\ &< e^{-\frac{4\Gamma^2(\frac{p}{2})c^2t^{3-p}}{(4+o(1))\Gamma^2(\frac{p}{2})t^{3-p}}}\\ &\approx e^{-c^2}.\end{aligned}$$

Since 1 is much smaller than λ, we can replace $1+\lambda$ by 1 without loss of generality. Thus, with probability at least $1-e^{-c^2}$, we have

$$X_{1,t} \leq \lambda.$$

Similarly, with probability at least $1-e^{-c^2}$, we have

$$m_{1,t} - M_1(t+1) \leq 2c\sqrt{t}. \tag{3.10}$$

We remark that the inequality 3.10 holds for any $c > 0$. In fact, it is trivial when $c > \sqrt{t}$ since $|m_{1,t} - M_1(t+1)| \leq 2t$ always holds.

Similarly, by applying Theorem 2.26 on the martingale, the following lower bound

$$m_{1,t} - M_1(t+1) \geq -2c\sqrt{t} \tag{3.11}$$

holds with probability at least $1-e^{-c^2}$.

We have proved the claim for $k = 1$.

The inductive step:

Suppose the claim holds for $k-1$. For k, we define

$$X_{k,t} = \frac{m_{k,t} - M_k(t+1) - 2(k-1)c\sqrt{t}}{\prod_{j=1}^{t}(1-\frac{(2-p)k}{2j})}.$$

we have

$$\begin{aligned}
&\mathrm{E}(m_{k,t} - M_k(t+1) - 2(k-1)c\sqrt{t}|\mathcal{F}_{t-1}) \\
&\quad = \mathrm{E}(m_{k,t}|\mathcal{F}_{t-1}) - M_k(t+1) - 2(k-1)c\sqrt{t} \\
&\quad = m_{k,t-1}(1 - \frac{(2-p)k}{2t}) + m_{k-1,t-1}(\frac{(2-p)(k-1)}{2t}) \\
&\qquad -M_k(t+1) - 2(k-1)c\sqrt{t}.
\end{aligned}$$

By the induction hypothesis, with probability at least $1 - 2t^{k-2}e^{-c^2}$, we have

$$|m_{k-1,t-1} - M_{k-1}t| \leq 2(k-1)c\sqrt{t-1}.$$

By using this estimate, with probability at least $1 - 2t^{k-2}e^{-c^2}$, we have

$$\mathrm{E}(m_{k,t} - M_k(t+1) - 2(k-1)c\sqrt{t}|\mathcal{F}_{t-1}) \leq (1 - \frac{(2-p)k}{2t})(m_{k,t-1} - M_k t - 2(k-1)c\sqrt{t-1})$$

by using the fact that $M_k \leq M_{k-1}$ as seen in (3.3).

Therefore, $0 = X_{k,0}, X_{k,1}, \cdots, X_{k,t}$ forms a submartingale with fail probability at most $2t^{k-2}e^{-c^2}$.

Similar to inequalities (3.8) and (3.9), it can be easily shown that

$$|X_{1,t} - X_{1,t-1}| \leq \frac{4}{\prod_{j=1}^{t}(1 - \frac{(2-p)k}{2j})} \tag{3.12}$$

and

$$\mathrm{Var}(X_{1,t}|\mathcal{F}_{t-1}) \quad \leq \quad \frac{4}{\prod_{j=1}^{t}(1 - \frac{(2-p)k}{2j})^2}.$$

We apply Theorem 2.39 on the submartingale with $\sigma_i^2 = \frac{4}{\prod_{j=1}^{i}(1 - \frac{(2-p)k}{2j})^2}$, $M = \frac{4}{\prod_{j=1}^{t}(1 - \frac{2-p}{2j})}$ and $a_i = 0$. We have

$$\Pr(X_{k,t} \geq \mathrm{E}(X_{k,t}) + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t} \sigma_i^2 + M\lambda/3)}} + \Pr(B),$$

where $\Pr(B) \leq t^{k-1}e^{-c^2}$ by induction hypothesis.

Here $\mathrm{E}(X_{k,t}) = X_{k,0} = 0$. We will use the following approximation.

$$\begin{aligned}
\prod_{j=1}^{i}(1 - \frac{(2-p)k}{2j}) &= \frac{\Gamma(i + 1 - \frac{(2-p)k}{2})}{\Gamma(i+1)\Gamma(1 - \frac{(2-p)k}{2})} \\
&= (\frac{1}{\Gamma(1 - \frac{(2-p)k}{2})} + O(\frac{1}{i}))i^{-k(2-p)/2}.
\end{aligned}$$

For any $c > 0$, we choose $\lambda = \frac{2c\sqrt{t}}{\prod_{j=1}^{t}(1-\frac{(2-p)k}{2j})} \approx 2\Gamma(1-\frac{(2-p)k}{2})ct^{1/2+k(2-p)/2}$. We have

$$\begin{aligned}
\sum_{i=1}^{t}\sigma_i^2 &\leq \sum_{i=1}^{t}\frac{4}{\prod_{j=1}^{i}(1-\frac{(2-p)k}{2j})^2}\\
&\approx \sum_{i=1}^{t}4\Gamma^2(1-\frac{(2-p)k}{2})i^{k(2-p)}\\
&\approx \frac{4\Gamma^2(1-\frac{(2-p)k}{2})}{1+(2-p)k}t^{1+k(2-p)}\\
&< 2\Gamma^2(1-\frac{(2-p)k}{2})t^{1+k(2-p)}.
\end{aligned}$$

We note that

$$M\lambda/3 \approx \frac{4}{3}\Gamma^2(1-\frac{(2-p)k}{2})ct^{\frac{1}{2}+(2-p)k} < 2\Gamma^2(1-\frac{(2-p)k}{2})t^{1+(2-p)k}$$

as long as $c < \sqrt{t}$. We have

$$\begin{aligned}
\Pr(X_{k,t} \geq \lambda) &\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2+M\lambda/3)}} + \Pr(B)\\
&< e^{-\frac{4\Gamma^2(1-\frac{(2-p)k}{2})c^2t^{1+(2-p)k}}{(4+o(1))\Gamma^2(1-\frac{(2-p)k}{2})t^{1+(2-p)k}}} + \Pr(B)\\
&< e^{-c^2} + t^{k-1}e^{-c^2}\\
&\leq (t+1)^{k-1}e^{-c^2}.
\end{aligned}$$

With probability at least $1-(t+1)^{k-1}e^{-c^2}$, we have

$$X_{k,t} \leq \lambda.$$

Equivalently, with probability at least $1-(t+1)^{k-1}e^{-c^2}$, we have

$$m_{k,t} - M_k(t+1) \leq 2kc\sqrt{t}. \tag{3.13}$$

We remark that the inequality (3.10) holds for any $c > 0$. In fact, it is trivial when $c > \sqrt{t}$ since $|m_{k,t} - M_k(t+1)| \leq 2kt$ always holds.

To obtain the lower bound, we consider

$$X'_{k,t} = \frac{m_{k,t} - M_k(t+1) + 2(k-1)c\sqrt{t}}{\prod_{j=1}^{t}(1-\frac{(2-p)k}{2j})}.$$

It can be easily shown that $X'_{k,t}$ is nearly a supermartingale. Similarly, if applying Theorem 2.42 to $X'_{k,t}$, the following lower bound

$$m_{k,t} - M_k(t+1) \geq -2kc\sqrt{t} \tag{3.14}$$

holds with probability at least $1-(t+1)^{k-1}e^{-c^2}$.

Together these prove the statement for k. The proof of Theorem 3.2 is complete. □

For completeness, we here state the corresponding theorem for $G(p, m)$.

THEOREM 3.3. *For the preferential attachment model $G(p, m, G_0)$, almost surely the number of vertices with degree k at time t is*

$$M_k t + m_{k,0} + O(2m\sqrt{(k+m-1)^3 t \ln(t)}).$$

Recall $M_m = \frac{2p}{4-p}$ and $M_k = \frac{2p}{4-p}\frac{\Gamma(k)\Gamma(1+\frac{2}{2-p})}{\Gamma(k+1+\frac{2}{2-p})} = O(k^{-(2+\frac{p}{2-p})})$, for $k \geq m+1$. In other words, almost surely the graphs generated by $G(p, m, G_0)$ have the power law degree distribution with the exponent $\beta = 2 + \frac{p}{2-p}$.

3.7. Models for directed graphs

Many real-world graphs are directed graphs. For example, the WWW-graph has edges each of which represents a link from a webpage to another. There are vertices with large in-degrees but relatively small out-degrees such as Yahoo, CNN or USA Today. Such vertices are often called *authorities* [**81**]. There are also vertices, called *hubs*, with large out-degrees but relatively small in-degrees. For directed graphs, we can have quite different distributions for in-degrees and out-degrees. For example, the in-degree sequence of the WWW graph follows the power law distribution with the exponent β about 2.1 while the out-degree sequence follows a different power law with exponent β about 2.7.

In this section, we will consider a preferential attachment model that can generate a directed graph with power-law in-degree distributon and power-law out-degree distribution. Furthermore, the exponents for the power law distributions are specified different values.

To generate such a directed graph, we have three parameters for the preferential attachment model:

- Two given probabilities p_1, p_2, satisfying $0 \leq p_1, p_2 \leq p_1 + p_2 \leq 1$.
- An initial graph G_0 at time 0.

We also have three operations:

- *Source-vertex-step* — Add a new vertex v, and add an directed edge $\{v, u\}$ from v by randomly and independently choosing u in proportion to the in-degree of u in the current graph.
- *Sink-vertex-step* — Add a new vertex v, and add an edge $\{u, v\}$ to v by randomly and independently choosing u in proportion to the out-degree of u in the current graph.
- *Edge-step* — Add a new edge $\{r, s\}$ by independently choosing vertices r and s with probability proportional to their in-degree (or out-degree), respectively.

The random graph model $D_0(p_1, p_2, G_0)$ is assembled as follows:

Begin with the initial graph G_0.
For $t > 0$, at time t, the graph G_t is formed by modifying G_{t-1} as follows:

```
        with probability $p_1$, take a source-vertex-step,
        with probability $p_2$, take a sink-vertex-step,
        otherwise, take an edge-step.
```

This simple model generates a power law graph with different exponents (as functions of p_1 and p_2) for in-degree and out-degree distributions. We remark that the vertices with in-degree zero (i.e., source vertices) will always have zero in-degree. Vice versa, the vertices with out-degree zero (i.e., sink vertices) will always have out-degree zero. Except for the vertices in G_0, the rest of vertices are partitioned into two groups — source vertices and sink vertices. This model might not be feasible for modeling most realistic networks.

We here consider a modified preferential attachment scheme with an additional parameter $\alpha \geq 0$, defined as follows:

α-preferential attachment scheme (or α-scheme, in short):
A vertex u is chosen for the tail (or head) of a new edge with probability proportional to its in-weight (or out-weight) where the in-weight of u is defined to be the sum of the in-degree of u and α. (The out-weight of u is the sum of the out-degree of u and α.)

The random graph model $D(p_1, p_2, \alpha, G_0)$ is assembled as follows:

```
Begin with the initial graph $G_0$.
For $t > 0$, at time $t$, the graph $G_t$ is formed by modifying $G_{t-1}$ as follows:
        with probability $p_1$, take a source-vertex-step using the $\alpha$-scheme,
        with probability $p_2$, take a sink-vertex-step using the $\alpha$-scheme,
        otherwise, take an edge-step.
```

We note that an alternative model is to add loops to a new vertex in each step. It is not hard to see that adding a loop is equivalent to the 1-preferential attachment scheme. In fact, the α-preferential attachment scheme can be viewed as adding α loops. When G_0 is the graph consisting of a single vertex, we simplify the notation and write $G(p_1, p_2, \alpha) = G(p_1, p_2, \alpha, G_0)$.

The number of edges of $G(p_1, p_2, \alpha)$ at time t is exactly t. The total weight at time t is just $t + \alpha n_t$. The number of vertices n_t at time t follows the binomial distribution. The expected value $E(n_t)$ satisfies

$$E(n_t) = 1 + (p_1 + p_2)t.$$

To deal with the actual value n_t, we use the binomial concentration inequality as described in Theorem 2.4. Namely,

$$Pr(|n_t - E(n_t)| > a) \leq e^{-a^2/(2pt+2a/3)}.$$

Thus, n_t is exponentially concentrated around $E(n_t)$.

Let $m^{in}_{k,t}$ denote the number of vertices of in-degree k at time t. We note that

$$m^{in}_{0,k} = 0.$$

We wish to derive a recurrence formula for the expected value $E(m^{in}_{k,t})$. A vertex of in-degree k at time t could have come from two cases, either it was a vertex of degree k at time $t-1$ and had no edge added to it, or it was a vertex of degree $k-1$ at time $t-1$ and the new edge was incident to it.

Let $\mathcal{F}_t$ denote the σ-algebra generated by the probability space at time t. For $t>0$ and $k>1$, we have

$$\begin{aligned}
\mathrm{E}(m^{in}_{k,t}|\mathcal{F}_{t-1}) &= m^{in}_{k,t-1}(1-\frac{(k+\alpha)p_1}{t-1+\alpha n_t}-\frac{(1-p_1-p_2)(k+\alpha)}{t-1+\alpha n_t}) \\
&\quad +m^{in}_{k-1,t-1}(\frac{(k-1+\alpha)p_1}{t-1+\alpha n_t}+\frac{(1-p_1-p_2)(k-1+\alpha)}{t-1+\alpha n_t}) \\
&= m^{in}_{k,t-1}(1-\frac{(1-p_2)(k+\alpha)}{t-1+\alpha n_t})+m^{in}_{k-1,t-1}(\frac{(1-p_2)(k-1+\alpha)}{t-1+\alpha n_t}). \qquad (3.15)
\end{aligned}$$

If we take the expectation on both sides and apply the estimation $n_t \approx (p_1+p_2)t$, we obtain the following recurrence formula.

$$\mathrm{E}(m^{in}_{k,t}) \approx \mathrm{E}(m^{in}_{k,t-1})(1-\frac{(1-p_2)(k+\alpha)}{t(1+(p_1+p_2)\alpha)})+\mathrm{E}(m^{in}_{k-1,t-1})(\frac{(1-p_2)(k-1+\alpha)}{t(1+(p_1+p_2)\alpha}).$$

For $t>0$ and $k=0,1$, we have

$$\begin{aligned}
E(m^{in}_{1,t}|\mathcal{F}_{t-1}) &= m^{in}_{1,t-1}(1-\frac{(1-p_2)(1+\alpha)}{t-1+\alpha n_t})+m^{in}_{0,t-1}(\frac{(1-p_2)\alpha}{t-1+\alpha n_t})+p_2 \\
E(m^{in}_{0,t}|\mathcal{F}_{t-1}) &= m^{in}_{0,t-1}(1-\frac{(1-p_2)\alpha}{t-1+\alpha n_t})+p_1.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{E}(m^{in}_{1,t}) &\approx \mathrm{E}(m^{in}_{1,t-1})(1-\frac{(1-p_2)(1+\alpha)}{t(1+(p_1+p_2)\alpha)})+\mathrm{E}(m^{in}_{0,t-1})\frac{(1-p_2)\alpha}{t(1+(p_1+p_2)\alpha)}+p_2. \\
\mathrm{E}(m^{in}_{0,t}) &\approx \mathrm{E}(m^{in}_{0,t-1})(1-\frac{(1-p_2)\alpha}{t(1+(p_1+p_2)\alpha)})+p_1.
\end{aligned}$$

Here these asymptotic equalities are by the fact that $n_t \approx (p_1+p_2)t$.

We proceed by induction on k to show that $\lim^{in}_{t\to\infty} E(m^{in}_{k,t})/t$ has a limit M^{in}_k for each k.

The first case is $k=0$. In this case, we apply Lemma 3.1 with $b_t=b=(1-p_2)\alpha/(1+(p_1+p_2)\alpha)$ and $c_t=c=p_2$ to deduce that $\lim_{t\to\infty} E(m^{in}_{0,t})/t = M^{in}_0$ exists. We have

$$\begin{aligned}
M^{in}_0 &= \frac{c}{1+b} \\
&= \frac{p_2}{1+\frac{(1-p_2)\alpha}{(1+(p_1+p_2)\alpha)}} \\
&= \frac{p_2(1+(p_1+p_2)\alpha)}{1+(1+p_1)\alpha}. \qquad (3.16)
\end{aligned}$$

For the case $k = 1$, we apply Lemma 3.1 with $b_t = b = (1-p_2)(1+\alpha)/(1+(p_1+p_2)\alpha))$ and $c_t = \mathrm{E}(m^{in}_{0,t-1})\frac{(1-p_2)\alpha}{t(1+(p_1+p_2)\alpha)} + p_1$. We have

$$c = \lim_{t\to\infty} c_t = M_0^{in}\frac{(1-p_2)\alpha}{1+(p_1+p_2)\alpha} + p_1.$$

It implies that $\lim_{t\to\infty} E(m^{in}_{0,t})/t = M_1^{in}$ exists. We have

$$\begin{aligned} M_1^{in} &= \frac{c}{1+b} \\ &= \frac{M_0^{in}\frac{(1-p_2)\alpha}{1+(1+p_1)\alpha} + p_1}{1 + \frac{(1-p_2)(1+\alpha)}{(1+(p_1+p_2)\alpha)}} \\ &= \frac{p_1 + (p_1+p_2+p_1^2-p_2^2)\alpha}{2-p_2+(1+p_1)\alpha}. \end{aligned} \tag{3.17}$$

For $k > 1$, we assume that $\lim_{t\to\infty} E(m^{in}_{k-1,t})/t = M^{in}_{k-1}$ exists and we apply the lemma again with $b_t = b = \frac{(1-p_2)(k+\alpha)}{(1+(p_1+p_2)\alpha)}$ and $c_t = E(m^{in}_{k-1,t-1})\frac{(1-p_2)(k-1+\alpha)}{t(1+(p_1+p_2)\alpha)}$, so $c = M^{in}_{k-1}\frac{(1-p_2)(k-1+\alpha)}{(1+(p_1+p_2)\alpha)}$. Lemma 3.1 implies that the limit $\lim^{in}_{t\to\infty} E(m^{in}_{k,t})/t = M^{in}_k$ exists and is equal to

$$\begin{aligned} M_k^{in} &= \frac{c}{1+b} \\ &= M^{in}_{k-1}\frac{\frac{(1-p_2)(k-1+\alpha)}{1+(1+p_1)\alpha}}{1+\frac{(1-p_2)(k+\alpha)}{(1+(p_1+p_2)\alpha)}} \\ &= M^{in}_{k-1}\frac{k-1+\alpha}{k+\alpha+\frac{1+(p_1+p_2)\alpha}{1-p_2}}. \end{aligned} \tag{3.18}$$

Thus we can write

$$\begin{aligned} m_k^{in} &= m_k^{in}\prod_{j=2}^{k}\frac{j-1+\alpha}{j+\alpha+\frac{1+(p_1+p_2)\alpha}{1-p_2}} \\ &= m_1^{in}\frac{\Gamma(k+\alpha)\Gamma(2+\alpha+\frac{1+(p_1+p_2)\alpha}{1-p_2})}{\Gamma(1+\alpha)\Gamma(k+1+\alpha+\frac{1+(p_1+p_2)\alpha}{1-p_2})} \\ &\approx M_1^{in}\frac{\Gamma(2+\alpha+\frac{1+(p_1+p_2)\alpha}{1-p_2})}{\Gamma(1+\alpha)}k^{1+\frac{1+(p_1+p_2)\alpha}{1-p_2}} \end{aligned}$$

where $\Gamma(k)$ is the Gamma function.

Thus we have a power-law graph for the in-degree sequence with

$$\beta^{in} = 1 + \frac{1+(p_1+p_2)\alpha}{1-p_2} = 2 + \frac{p_2+(p_1+p_2)\alpha}{1-p_2}.$$

Let $m^{out}(t,k)$ be the number of vertices with out-degree at time t. Similarly we can show $\lim_{t\to\infty}\frac{\mathrm{E}(m^{out}_{t,k})}{t}$ exists. We denote it by M^{out}_k. We have

$$M_0^{out} = \frac{p_2(1+(p_1+p_2)\alpha)}{1+(1+p_2)\alpha} \tag{3.19}$$

$$M_1^{out} = \frac{p_2+(p_1+p_2+p_2^2-p_1^2)\alpha}{2-p_1+(1+p_2)\alpha} \tag{3.20}$$

(3.21)

For $k > 1$, we have

$$M_k^{out} = M_1^{out}\frac{\Gamma(k+\alpha)\Gamma(2+\alpha+\frac{1+(p_1+p_2)\alpha}{1-p_1})}{\Gamma(1+\alpha)\Gamma(k+1+\alpha+\frac{1+(p_1+p_2)\alpha}{1-p_1})} \tag{3.22}$$

$$\approx M_1^{out}\frac{\Gamma(2+\alpha+\frac{1+(p_1+p_2)\alpha}{1-p_1})}{\Gamma(1+\alpha)}k^{1+\frac{1+(p_1+p_2)\alpha}{1-p_1}}. \tag{3.23}$$

The exponent β^{out} for the out-degree distribution is

$$\beta^{in} = 1 + \frac{1+(p_1+p_2)\alpha}{1-p_1} = 2 + \frac{p_1+(p_1+p_2)\alpha}{1-p_1}.$$

Similar to section 3.6, we can prove the sharp concentration result for the in-degree and out-degree distributions. For completeness, we state the following theorem for the directed preferential attachment model.

THEOREM 3.4. *For the preferential attachment model $G(p_1, p_2, \alpha)$, we have*

(1) *Almost surely the number of vertices with in-degree k at time t is*

$$M_k^{in}t + O(2\sqrt{k^3 t\ln(t)}),$$

where M_k^{in} is defined in equation (3.16), (3.17), and (3.19).

(2) *Almost surely the number of vertices with out-degree k at time t is*

$$M_k^{out}t + O(2\sqrt{k^3 t\ln(t)}),$$

where M_k^{out} is defined in equation (3.19), (3.20), and (3.22).

(3) *Almost surely it is a power law directed graph with the exponent $\beta^{in} = 2 + \frac{p_2+(p_1+p_2)\alpha}{1-p_2}$ for the in-degree distribution and the exponent $\beta^{out} = 2 + \frac{p_1+(p_1+p_2)\alpha}{1-p_1}$ for the out-degree distribution.*

The exponents β^{in} and β^{out} have special meanings. It is not difficult to see that both values are greater than 2. It can be observed that $p_2 + (p_1+p_2)\alpha$ is the expected increment for the in-degree of the new vertex while $1 - p_2$ is the expected increment for the in-degrees of the current graphs. Hence, $\beta^{in} - 2$ is the ratio of the increment of edges to the new vertex and the increment of edges to the current graph. There is a similar interpretation for $\beta^{out} - 2$ as well.