0 Preface

This book comes in two parts. The first is an introduction to the asymptotic theory of combinatorial structures that can be decomposed into component elements. The second is a detailed study of such structures, in the style of a research monograph. The reason for this split is that the main ideas are rather straightforward, and can be relatively simply explained. However, using these ideas and a fair amount of technical application, there are many sharp results that can be derived as consequences. We present some of these, to illustrate that the method is not only simple but also powerful.

We are specifically concerned with the component frequency spectrum – that is, with the numbers and sizes of the component elements – of a 'typical' structure of given (large) size n. A classic example of a decomposable combinatorial structure is that of permutations of n elements, with cycles as the component elements; here, the component spectrum is just the cycle type. Our approach is to take 'typical' to mean 'with high probability', when a structure is chosen at random according to some given probability distribution from the set of all structures of size n; most commonly, but not necessarily, according to the uniform distribution. This enables us to introduce ideas from probability theory, such as conditioning, Stein's method and distributional approximation, as tools in our investigation.

We gain our understanding of the component spectrum by comparison with simpler random objects. Sometimes these objects are discrete; indeed, our fundamental comparisons are with sequences of independent random variables and with the Ewens Sampling Formula. However, we also use continuous approximations, such as Brownian motion, the scale invariant Poisson process and the Poisson–Dirichlet process. Our comparisons are formulated not only as limit theorems as $n \to \infty$, but also as approximations with concrete error bounds, valid for any fixed n. In the first eight chapters, we introduce our approach, prove some of the basic approximations, and outline the more detailed results and their consequences. From Chapter 9 onwards, the treatise becomes (unashamedly) technical.

In a decomposable structure of size n, the component spectrum consists of the numbers $C_1^{(n)}$ counting components of size 1, $C_2^{(n)}$ counting components of size 2, ..., $C_n^{(n)}$ counting components of size n, where the $C_i^{(n)}$ have to satisfy the equation

$$C_1^{(n)} + 2C_2^{(n)} + \cdots + nC_n^{(n)} = n, \tag{0.1}$$

because the structure has total size n. A quantity of traditional interest and frequent study is then

$$K_{0n} = C_1^{(n)} + C_2^{(n)} + \cdots + C_n^{(n)},$$

the number of components in the structure. The vector of component counts $(C_1^{(n)}, C_2^{(n)}, \ldots, C_n^{(n)})$ can be viewed as a stochastic process, if the structure is chosen at random from among the $p(n)$ structures of size n. A 'typical' property then corresponds to an event, defined in terms of the stochastic process, which has 'high' probability; for the uniform distribution over all possible structures of size n, this is equivalent to a property of the structure which is true of a 'high' proportion of all such structures.

We are thus concerned with the behavior of the discrete dependent nonnegative integer-valued random processes

$$C^{(n)} = (C_1^{(n)}, C_2^{(n)}, \ldots, C_n^{(n)}), \quad n = 1, 2, \ldots$$

arising from randomly chosen combinatorial structures. These processes have to satisfy (0.1), of course, but all the classical examples have much more in common. A key common feature is that, for each $n \geq 1$, the joint distribution $\mathcal{L}(C_1^{(n)}, \ldots, C_n^{(n)})$ satisfies the **Conditioning Relation**

$$\mathcal{L}(C_1^{(n)}, \ldots, C_n^{(n)}) = \mathcal{L}(Z_1, Z_2, \ldots, Z_n | T_{0n} = n), \tag{0.2}$$

for a fixed sequence of independent random variables $Z_1, Z_2, \ldots$ taking values in $\mathbb{Z}_+$, where

$$T_{0n} = T_{0n}(Z) = Z_1 + 2Z_2 + \cdots + nZ_n.$$

For the classical combinatorial structures, the random variables Z_i have either Poisson, negative binomial or binomial distributions. For example, random permutations, discussed in detail in Chapter 1.1, satisfy the Conditioning Relation for random variables Z_i that have Poisson distributions with means $\mathbb{E}Z_i = 1/i$.

Our unifying approach is developed in a context motivated by a large sub-class of classical combinatorial structures that share, in addition to the Conditioning Relation, the following common feature. We assume that the random variables $(Z_i,\ i \geq 1)$ are such as to satisfy the **Logarithmic Condition**

$$i\mathbb{P}[Z_i = 1] \to \theta, \quad i\mathbb{E}Z_i \to \theta \text{ as } i \to \infty \tag{0.3}$$

for some $\theta > 0$. In our probabilistic setting, there is no need to be more specific about the distributions of the Z_i, so that we are free to move away from the classical Poisson, binomial and negative binomial families; this added flexibility has its uses, for example when investigating random characteristic polynomials over a finite field. And even within the classical families, we can choose θ to take a value different from that normally associated with the uniform distribution over a well-known set of combinatorial objects. The simplest example of this arises when the Z_j have Poisson distributions with mean $\mathbb{E}Z_j = \theta/j$, for any $\theta > 0$; the special case $\theta = 1$ corresponds to the uniform distribution. In the general case, the distribution of $C^{(n)}$ is called the Ewens Sampling Formula. This distribution, discussed in detail in Chapter 5, plays a central rôle in our work.

Our main theme is that the Conditioning Relation and the Logarithmic Condition are together enough to ensure that the component spectrum of a large decomposable combinatorial structure has a prescribed, universal form; the numbers of small components of different sizes are almost independent, with distributions approximated by those of the Z_i, and the sizes of the large components are jointly distributed almost as those of the Ewens Sampling Formula. We complement this broad picture with many detailed refinements.

We note that the Conditioning Relation by itself, even without the Logarithmic Condition, is a powerful tool, though not the subject of this book; a general treatment is given in Arratia and Tavaré (1994). Perhaps the simplest example is that of set partitions, in which (0.2) is satisfied for Poisson distributed random variables Z_i with means $\mathbb{E}Z_i = 1/i!$. These Z_i do *not* satisfy (0.3), and the distribution of the number $C_i^{(n)}$ of components (in this case blocks) of size i is not well approximated by that of Z_i. However, as noted in (2.7), the Poisson random variables Z_i in the Conditioning Relation (0.2) may also be taken with $\mathbb{E}Z_i = x^i/i!$, for any choice of $x \in (0, \infty)$. No fixed choice of x works any better than $x = 1$, but by choosing x to vary with n, in particular taking $x = x(n)$ to be the solution of $xe^x = n$, a very good approximation for the joint distribution of the component spectrum of a random set partition may be achieved: see Pittel (1997b). In this book, we only need to use (0.2) with a *fixed* choice of the random variables Z_i, although there are some questions, even for logarithmic combinatorial structures satisfying (0.3), for which it may be useful to allow the Z_i to vary with n; see Section 5.2 of Arratia and Tavaré (1994), and also Stark (1997a).

History

The comparison of the component spectrum of a combinatorial structure to an independent process, with or without further conditioning, has a long history. Perhaps the best-known example is the representation of the multinomial distribution with parameters n and $p_1, \ldots, p_k$ as the joint law of independent Poisson random variables with means $\lambda p_1, \ldots, \lambda p_k$, conditional on their sum being equal to n.

Holst (1979a) provides an approach to urn models that unifies multinomial, hypergeometric and Pólya sampling. The joint laws of the dependent counts of the different types sampled are represented, respectively, as the joint distribution of independent Poisson, negative binomial, and binomial random variables, conditioned on their sum. See also Holst (1979b, 1981). The quality of such approximations is assessed using metrics, including the total variation distance, by Stam (1978) and Diaconis and Freedman (1980).

The Conditioning Relation also appears in the context of certain reversible Markovian models of migration and clustering. In that setting, n individuals are classified as belonging to different groups, with the number of groups of size j being denoted by $C_j^{(n)}$. At stationarity, the distribution of $C^{(n)}$ satisfies the Conditioning Relation for independent random variables $Z_1, Z_2, \ldots$, whose distributions under the natural 'mass action' mixing hypothesis are Poisson. The models can also be used as descriptions of coagulation, fragmentation, aggregation and polymerization. See Whittle (1965, 1967, 1986) and Kelly (1979) for further details.

The books by Kolchin, Sevast'yanov and Chistyakov (1978) and Kolchin (1986, 1999) use the representation of the component spectrum of combinatorial structures, including random permutations and random mappings, in terms of independently distributed random variables, conditioned on the value of their sum. However, Kolchin's technique uses independent random variables that are identically distributed, and the number of components $C_i^{(n)}$ of size i is the number of random variables which take on the value i.

Conditioning was exploited by Shepp and Lloyd (1966) in their seminal paper on the asymptotics of random permutations, and also used by Watterson (1974) in a study of the Ewens Sampling Formula. The unpublished lecture notes of Diaconis and Pitman (1986) also emphasize the rôle of conditioning and probabilistic methods. Hansen (1989, 1990) uses conditioning to study the Ewens Sampling Formula and random mappings. Fristedt (1992, 1993) exploits conditioning to study random partitions of a set and random partitions of an integer; the sharpest results for random partitions of sets and integers are given in Pittel (1997a,b). Hansen (1994) has a systematic treatment of the behavior of the large components of logarithmic combinatorial structures.

Logarithmic combinatorial structures are usually studied without appeal to the conditioning relation, but using generating function methods instead. General discussions focussing on probabilistic properties include Knopfmacher (1979), Flajolet and Soria (1990), Flajolet and Odlyzko (1990a), Odlyzko (1995), Hwang (1994, 1998a,b), Zhang (1996a,b, 1998), Gourdon (1998), Panario and Richmond (2001) and Flajolet and Sedgewick (1999). For further treatment of the algebraic aspects of decomposable structures, the reader is referred to Foata (1974) and Joyal (1981) and to the books by Goulden and Jackson (1983), Stanley (1986) and Bergeron, Labelle and Leroux (1998).

Organization of the book

We begin in Chapter 1 with a survey of the main features of the joint behavior of the numbers of cycles of different sizes in a random permutation of n elements, to give a concrete and simple illustration of phenomena which occur throughout the class of logarithmic combinatorial structures. Even though the joint distribution of the cycle counts is specified precisely by Cauchy's formula, it is surprisingly difficult to derive useful information from it by a direct approach, when n is large; hence, even in this simple case, our methods have much to offer. Then, for the sake of historical perspective, we outline the analogous results for the prime factorization of a random integer, even though this is *not* an example of the class of combinatorial structures studied in our book.

Chapter 2 gives the *combinatorial* description of decomposable combinatorial structures, both logarithmic and non-logarithmic, first by way of specific examples such as mappings, partitions, and trees, and then in terms of general classes: assemblies, multisets, and selections. Next we give the *probabilistic* description of these classic combinatorial objects, focusing first on the **Conditioning Relation** (0.2), which is an algebraic condition; and then on the **Logarithmic Condition** (0.3), an analytic condition which characterizes the Logarithmic Class. We provide a combinatorial perspective on refining and coloring, including for example wreath products, and we discuss *tilting*, which may be considered as a probabilistic extension of coloring.

Chapter 3 begins the discussion of Logarithmic Combinatorial structures in the full generality of an arbitrary sequence of independent nonnegative integer-valued random variables Z_i satisfying the **Logarithmic Condition** (0.3), and made into a dependent process by the **Conditioning Relation** (0.2), so that the classical combinatorial examples are included as special cases. We discuss the probability metrics – total variation distance and various Wasserstein distances – used to assess the accuracy of our probabilistic approximations. We then give a brief survey of the results that we are able to derive, and conclude with an introduction to Stein's method, a technique that is essential for many of our proofs.

A central family of discrete distributions, the Ewens Sampling Formula, is the subject of Chapters 4 and 5. We investigate the family itself, as well as certain infinitely divisible random variables which are closely related to it. We also discuss the tools and limiting processes that are used in describing its properties: size biasing, the scale invariant Poisson process, the GEM distribution, and the Poisson–Dirichlet distribution. The result is a rather extensive asymptotic description of the distribution of the main features of the component spectrum, when the full distribution is specified by the Ewens Sampling Formula.

The same tools are used in Chapter 6 to extend the asymptotic description to more general logarithmic combinatorial structures. A single, relatively simple technical condition, the local limit theorem (LLT) of (6.6), is shown to imply the naive limit laws (3.4) for small components and (3.5) for large components. We then show that, for combinatorial structures such as assemblies, multisets and selections, the mild **Logarithmic Condition** (0.3) is already enough to imply (LLT).

For logarithmic combinatorial structures more general than assemblies, multisets and selections, this simple approach fails, and more sophisticated tools are needed. Chapter 7 sets the scene. We use the Conditioning Relation to show that the joint distribution of the large components of a logarithmic combinatorial structure is close to that of the large components of the Ewens Sampling Formula, provided that, for large i, the distribution of Z_i is close to that of Z_i^*, which has the Poisson distribution with mean θ/i, and that the distribution of $T_{0n}(Z)$ is close to that of $T_{0n}(Z^*)$. We then discuss how to measure the difference between $\mathcal{L}(Z_i)$ and $\mathrm{Po}(\theta/i)$, and establish working conditions under which the influence of these differences can be controlled. Under these conditions, Stein's method can be used to show the closeness of $\mathcal{L}(T_{0n}(Z))$ to $\mathcal{L}(T_{0n}(Z^*))$, and it turns out that this also enables one to show the closeness of the joint distributions $\mathcal{L}(C_1^{(n)}, \ldots, C_b^{(n)})$ and $\mathcal{L}(Z_1, \ldots, Z_b)$ for $b = o(n)$, thus treating the small components as well. We illustrate the conditions as applied to some of the basic examples, such as random mappings and random polynomials. Then we present the statements of our main approximation theorems – refining the naive limit theorems such as (3.4) for small components and (3.5) for large components by giving error bounds. We state both local and global approximations. The proofs themselves are presented in Chapters 9 through 13, which constitute the technical core of this monograph.

Chapter 8 gives a number of consequences of the approximation theorems of the preceding chapter, illustrating the power inherent in discrete functional limit theorems and approximations. Each is based on earlier limiting results, improving upon them in two ways. First, the context is broadened from an often quite restrictive setting to that of a very general logarithmic combinatorial structure. Secondly, the limit theorems are supplied with error bounds. The first setting is that of the usual “functional (Brownian motion) central limit theorem” for the number of components in various size ranges. Then we give several metrized comparison results relating to the Poisson–Dirichlet limit for the sizes of large components. For the very simplest functional of the component counting process, the total number of components, we investigate the accuracy of Poisson and related approximations to its distribution. Another famous theorem that we consider is the Erdős–Turán law for the order of a random permutation. Finally, we extend the theory of additive functions on additive arithmetic semigroups to general logarithmic structures.

The number theory connection

Our fascination with the component spectrum of logarithmic combinatorial structures is based partly on similarities to the prime factorization of a random integer selected uniformly from $\{1, 2, \ldots, n\}$, as observed in Knuth and Trabb Pardo (1976). The similarities include: having an independent process limit for small component counts; having Poisson–Dirichlet and GEM process limits for large components, as in Billingsley (1972, 1974, 1999), Bach (1985), Vershik (1987) and Donnelly and Grimmett (1993); and having a conditioning relation, here a related bias relation, to construct the dependent system from the independent system. The celebrated Dickman and Buchstab functions familiar to number theorists (cf. Tenenbaum (1995)) also arise in the combinatorial setting, described in Chapter 2. A further similarity involves the “Fundamental Lemma of Kubilius” in number theory; see Kubilius (1964), and Elliott (1979, 1980). This lemma corresponds to Theorem 7.7 for logarithmic combinatorial structures, stating that the total variation distance between the law of $(C_1^{(n)}, \ldots, C_b^{(n)})$ and the law of the approximating process $(Z_1, \ldots, Z_b)$ tends to zero when $b/n \to 0$, and giving an explicit upper bound for the error.

To see these similarities, one must view an integer as a multiset of primes. The most basic difference then lies in the sizes allowed for components: for the combinatorial structures considered in this monograph, the sizes allowed are $1, 2, 3, \ldots$, while, for prime factorizations, the sizes allowed are $\log p$ for primes p. For example, the integer $1848 = 2^3 \cdot 3 \cdot 7 \cdot 11$ is the instance having three components of size $\log 2$, one component each of sizes $\log 3, \log 7$, and $\log 11$, and no other components. This brief description suffices for a preface; for a somewhat longer discussion of the connections, see Chapter 1, or Arratia, Barbour and Tavaré (1997a) and Arratia (2002).

Notation

We end the preface with a brief description of our notation. A more extensive list may be found in the corresponding index. We write $\mathbb{N}$ for the natural numbers $\{1, 2, 3, \ldots\}$, $\mathbb{Z}_+$ for the nonnegative integers $\{0, 1, 2, \ldots\}$, and for the set of the first n natural numbers we write either $[n]$ or $\{1, 2, \ldots, n\}$. We write $A \subset B$ for the relation that A is a subset of B, allowing $A = B$. We denote the falling factorial by $x^{[r]} = x(x-1)\cdots(x-r+1)$ and rising factorial with $x^{(r)} = x(x+1)\cdots(x+r-1)$; in both cases, the value is 1 if $r = 0$. For the harmonic numbers we use $h(n+1) = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$. The first order asymptotic relation is written $a_n \sim b_n$, meaning $\lim a_n/b_n = 1$. We write $a \doteq b$ to denote a deliberately vague approximation, for heuristics or crude numerical values. We use the standard big-oh and little-oh notation: $a_n = O(b_n)$ means that $\limsup_n |a_n/b_n| < \infty$, and $a_n = o(b_n)$ means that $\lim_n a_n/b_n = 0$. We write $a_n \asymp b_n$ for the symmetric relation that both $a_n = O(b_n)$ and $b_n = O(a_n)$. We use $=$ to show that alternative notation may be used for a single object, for example $C^{(n)} = (C_1^{(n)}, \ldots, C_n^{(n)})$.

We write $\mathcal{L}(X)$ to denote the *law* (probability distribution) of a random object X, so that $\mathcal{L}(X) = \mathcal{L}(Y)$ means that X and Y have the same distribution; here, we also write $X =_d Y$. We use the notation $X_n \to_d X$ to indicate that X_n converges in distribution to X. We use $\sim$ when specifying the distribution of a random element; for example, $Z_i \sim \mathrm{Po}(1/i)$ states that Z_i has the Poisson distribution with mean $1/i$.