An Invitation to Analytic Combinatorics

διὸ δὴ συμμειγνύμενα αὐτά τε πρὸς αὐτὰ καὶ πρὸς ἄλληλα τὴν
ποικιλίαν ἐστὶν ἄπειρα· ἧς δὴ δεῖ θεωροὺς γίγνεσθαι τοὺς
μέλλοντας περὶ φύσεως εἰκότι λόγῳ

— PLATO, The Timaeus[1]

ANALYTIC COMBINATORICS is primarily a book about *combinatorics*, that is, the study of finite structures built according to a finite set of rules. *Analytic* in the title means that we concern ourselves with methods from mathematical analysis, in particular complex and asymptotic analysis. The two fields, combinatorial enumeration and complex analysis, are organized into a coherent set of methods for the first time in this book. Our broad objective is to discover how the continuous may help us to understand the discrete and to *quantify* its properties.

COMBINATORICS is, as told by its name, the science of combinations. Given basic rules for assembling simple components, what are the properties of the resulting objects? Here, our goal is to develop methods dedicated to *quantitative* properties of combinatorial structures. In other words, we want to measure things. Say that we have n different items like cards or balls of different colours. In how many ways can we lay them on a table, all in one row? You certainly recognize this counting problem—finding the number of *permutations* of n elements. The answer is of course the factorial number

$$n\,! = 1 \cdot 2 \cdot \ldots \cdot n.$$

This is a good start, and, equipped with patience or a calculator, we soon determine that if $n = 31$, say, then the number of permutations is the rather large quantity

$$31\,! = 8222838654177922817725562880000000, .$$

an integer with 34 decimal digits. The factorials solve an enumeration problem, one that took mankind some time to sort out, because the sense of the "$\cdots$" in the formula for $n!$ is not that easily grasped. In his book *The Art of Computer Programming*

[1] *"So their combinations with themselves and with each other give rise to endless complexities, which anyone who is to give a likely account of reality must survey."* Plato speaks of Platonic solids viewed as idealized primary constituents of the physical universe.

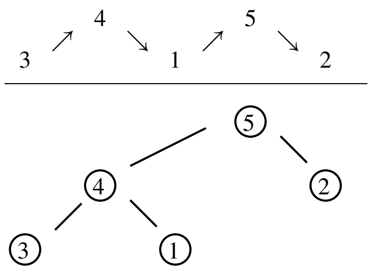

Figure 0.1. An example of the correspondence between an alternating permutation (top) and a decreasing binary tree (bottom): each binary node has two descendants, which bear smaller labels. Such *constructions*, which give access to *generating functions* and eventually provide solutions to counting problems, are the main subject of Part A.

(vol III, p. 23), Donald Knuth traces the discovery to the Hebrew *Book of Creation* (c. AD 400) and the Indian classic *Anuyogadvāra-sutra* (c. AD 500).

Here is another more subtle problem. Assume that you are interested in permutations such that the first element is smaller than the second, the second is larger than the third, itself smaller than the fourth, and so on. The permutations go up and down and they are diversely known as up-and-down or zigzag permutations, the more dignified name being *alternating* permutations. Say that $n = 2m + 1$ is odd. An example is for $n = 9$:

$$4 \nearrow 8 \searrow 6 \nearrow 7 \searrow 5 \nearrow 9 \searrow 1 \nearrow 3 \searrow 2$$

The number of alternating permutations for $n = 1, 3, 5, \ldots, 15$ turns out to be

$$1, 2, 16, 272, 7936, 353792, 22368256, 1903757312.$$

What are these numbers and how do they relate to the total number of permutations of corresponding size? A glance at the corresponding figures, that is, $1!, 3!, 5!, \ldots, 15!$, or

$$1, 6, 120, 5040, 362880, 39916800, 6227020800, 1307674368000,$$

suggests that the factorials grow somewhat faster—just compare the lengths of the last two displayed lines. But how and by how much? This is the prototypical question we are addressing in this book.

Let us now examine the counting of alternating permutations. In 1881, the French mathematician Désiré André made a startling discovery. Look at the first terms of the Taylor expansion of the trigonometric function $\tan z$:

$$\tan z = 1\frac{z}{1!} + 2\frac{z^3}{3!} + 16\frac{z^5}{5!} + 272\frac{z^7}{7!} + 7936\frac{z^9}{9!} + 353792\frac{z^{11}}{11!} + \cdots.$$

The counting sequence for alternating permutations, $1, 2, 16, \ldots$, curiously surfaces. We say that the function on the left is a *generating function* for the numerical sequence (precisely, a generating function of the *exponential* type, due to the presence of factorials in the denominators).

André's derivation may nowadays be viewed very simply as reflecting the construction of permutations by means of certain labelled binary trees (Figure 0.1 and p. 143): given a permutation σ a tree can be obtained once σ has been decomposed as a triple $\langle \sigma_L, \max, \sigma_R \rangle$, by taking the maximum element as the root, and appending, as left and right subtrees, the trees recursively constructed from σ_L and σ_R. Part A of this book develops at length *symbolic methods* by which the construction of the class $\mathcal{T}$ of all such trees,

$$\mathcal{T} \quad = \quad ① \quad \cup \quad (\mathcal{T}, \max, \mathcal{T}),$$

translates into an equation relating generating functions,

$$T(z) \quad = \quad z \quad + \quad \int_0^z T(w)^2 \, dw.$$

In this equation, $T(z) := \sum_n T_n z^n / n!$ is the exponential generating function of the sequence (T_n), where T_n is the number of alternating permutations of (odd) length n. There is a compelling formal analogy between the combinatorial *specification* and its generating function: Unions ($\cup$) give rise to sums ($+$), max-placement gives an integral ($\int$), forming a pair of trees corresponds to taking a square ($[\cdot]^2$).

At this stage, we know that $T(z)$ must solve the differential equation

$$\frac{d}{dz} T(z) = 1 + T(z)^2, \qquad T(0) = 0,$$

which, by classical manipulations[2], yields the explicit form

$$T(z) = \tan z.$$

The generating function then provides a simple *algorithm* to compute the coefficients recurrently. Indeed, the formula,

$$\tan z = \frac{\sin z}{\cos z} = \frac{z - \frac{z^3}{3!} + \frac{z^5}{5!} - \cdots}{1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \cdots},$$

implies, for n odd, the relation (extract the coefficient of z^n in $T(z)\cos z = \sin z$)

$$T_n - \binom{n}{2} T_{n-2} + \binom{n}{4} T_{n-4} - \cdots = (-1)^{(n-1)/2}, \quad \text{where} \quad \binom{a}{b} = \frac{a!}{b!(a-b)!}$$

is the conventional notation for binomial coefficients. Now, the exact enumeration problem may be regarded as solved since a very simple algorithm is available for determining the counting sequence, while the generating function admits an explicit expression in terms of well-known mathematical objects.

ANALYSIS, by which we mean mathematical analysis, is often described as the art and science of *approximation*. How fast do the factorial and the tangent number sequences grow? What about *comparing* their growths? These are typical problems of analysis.

[2] We have $T'/(1+T^2) = 1$, hence $\arctan(T) = z$ and $T = \tan z$.

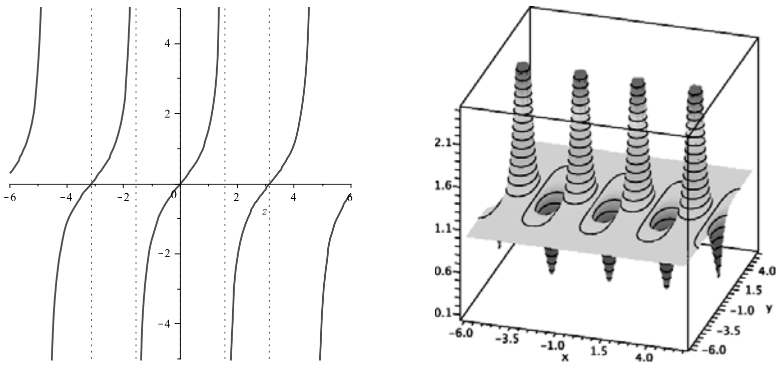

Figure 0.2. Two views of the function $z \mapsto \tan z$. Left: a plot for real values of $z \in [-6, 6]$. Right: the modulus $|\tan z|$ when $z = x + iy$ (with $i = \sqrt{-1}$) is assigned complex values in the square $\pm 6 \pm 6i$. As developed at length in Part B, it is the nature of *singularities* in the *complex domain* that matters.

First, consider the number of permutations, $n!$. Quantifying its growth, as n gets large, takes us to the realm of *asymptotic analysis*. The way to express factorial numbers in terms of elementary functions is known as Stirling's formula[3]

$$n! \sim n^n e^{-n} \sqrt{2\pi n},$$

where the $\sim$ sign means "approximately equal" (in the precise sense that the ratio of both terms tends to 1 as n gets large). This beautiful formula, associated with the name of the Scottish mathematician James Stirling (1692–1770), curiously involves both the basis e of natural logarithms and the perimeter 2π of the circle. Certainly, you cannot get such a thing without analysis. As a first step, there is an estimate

$$\log n! = \sum_{j=1}^{n} \log j \sim \int_1^n \log x \, dx \sim n \log \left(\frac{n}{e}\right),$$

explaining at least the $n^n e^{-n}$ term, but already requiring a certain amount of elementary calculus. (Stirling's formula precisely came a few decades after the fundamental bases of calculus had been laid by Newton and Leibniz.) Note the utility of Stirling's formula: it tells us almost instantly that 100! has 158 digits, while 1000! borders the astronomical 10^{2568}.

We are now left with estimating the growth of the sequence of tangent numbers, T_n. The analysis leading to the derivation of the generating function $\tan(z)$ has been so far essentially algebraic or "formal". Well, we can plot the graph of the tangent function, for real values of its argument and see that the function becomes infinite at the points $\pm\frac{\pi}{2}$, $\pm 3\frac{\pi}{2}$, and so on (Figure 0.2). Such points where a function ceases to be

[3]In this book, we shall encounter five different proofs of Stirling's formula, each of interest for its own sake: (i) by singularity analysis of the Cayley tree function (p. 407); (ii) by singularity analysis of polylogarithms (p. 410); (iii) by the saddle-point method (p. 555); (iv) by Laplace's method (p. 760); (v) by the Mellin transform method applied to the logarithm of the Gamma function (p. 766).

"smooth" (differentiable) are called *singularities*. By methods amply developed in this book, it is the local nature of a generating function at its "dominant" singularities (i.e., the ones closest to the origin) that determines the asymptotic growth of the sequence of coefficients. From this perspective, the basic fact that $\tan z$ has dominant singularities at $\pm\frac{\pi}{2}$ enables us to reason as follows: first approximate the generating function $\tan z$ near its two dominant singularities, namely,

$$\tan(z) \underset{z\to\pm\pi/2}{\sim} \frac{8z}{\pi^2-4z^2};$$

then extract coefficients of this approximation; finally, get in this way a valid approximation of coefficients:

$$\frac{T_n}{n!} \underset{n\to\infty}{\sim} 2\cdot\left(\frac{2}{\pi}\right)^{n+1} \qquad (n \text{ odd}).$$

With present day technology, we also have available *symbolic manipulation* systems (also called "computer algebra" systems) and it is not difficult to verify the accuracy of our estimates. Here is a small pyramid for $n = 3, 5, \ldots, 21$,

2	***1***
16	1***5***
272	27***1***
7936	793***5***
353792	35379***1***
22368256	2236825***1***
1903757312	1903757***267***
209865342976	20986534***2434***
29088885112832	290888851***04489***
4951498053124096	495149805***2966307***
(T_n)	$(T_n^\star)$

comparing the exact values of T_n against the approximations $T_n^\star$, where (n odd)

$$T_n^\star := \left\lfloor 2\cdot n!\left(\frac{2}{\pi}\right)^{n+1}\right\rfloor,$$

and discrepant digits of the approximation are displayed in bold. For $n = 21$, the error is only of the order of one in a billion. Asymptotic analysis (p. 269) is in this case wonderfully accurate.

In the foregoing discussion, we have played down a fact—one that is important. When investigating generating functions from an analytic standpoint, one should generally assign *complex* values to arguments not just real ones. It is singularities in the complex plane that matter and complex analysis is needed in drawing conclusions regarding the asymptotic form of coefficients of a generating function. Thus, a large portion of this book relies on a *complex analysis* technology, which starts to be developed in Part B dedicated to *Complex asymptotics*. This approach to combinatorial enumeration parallels what happened in the nineteenth century, when Riemann first recognized the deep relation between complex analytic properties of the *zeta* function, $\zeta(s) := \sum 1/n^s$, and the distribution of primes, eventually leading to the long-sought proof of the Prime Number Theorem by Hadamard and de la Vallée-Poussin in 1896. Fortunately, relatively elementary complex analysis suffices for our purposes, and we

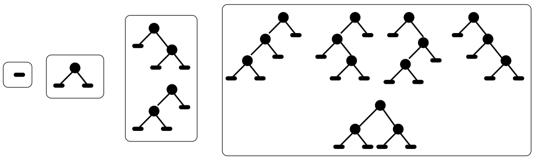

Figure 0.3. The collection of binary trees with $n = 0, 1, 2, 3$ binary nodes, with respective cardinalities 1, 1, 2, 5.

can include in this book a complete treatment of the fragment of the theory needed to develop the fundamentals of analytic combinatorics.

Here is yet another example illustrating the close interplay between combinatorics and analysis. When discussing alternating permutations, we have enumerated binary trees bearing distinct integer labels that satisfy a constraint—to decrease along branches. What about the simpler problem of determining the number of possible *shapes* of binary trees? Let C_n be the number of binary trees that have n binary branching nodes, hence $n+1$ "external nodes". It is not hard to come up with an exhaustive listing for small values of n (Figure 0.3), from which we determine that

$$C_0 = 1, \quad C_1 = 1, \quad C_2 = 2, \quad C_3 = 5, \quad C_4 = 14, \quad C_5 = 42.$$

These numbers are probably the most famous ones of combinatorics. They have come to be known as the *Catalan numbers* as a tribute to the Franco-Belgian mathematician Eugène Charles Catalan (1814–1894), but they already appear in the works of Euler and Segner in the second half of the eighteenth century (see p. 20). In his reference treatise *Enumerative Combinatorics*, Stanley, over 20 pages, lists a collection of some 66 different types of combinatorial structures that are enumerated by the Catalan numbers.

First, one can write a combinatorial equation, very much in the style of what has been done earlier, but without labels:

$$\mathcal{C} \quad = \quad \square \quad \cup \quad (\mathcal{C}, \bullet, \mathcal{C}) \,.$$

(Here, the $\square$–symbol represents an external node.) With symbolic methods, it is easy to see that the *ordinary generating function* of the Catalan numbers, defined as

$$C(z) := \sum_{n \geq 0} C_n z^n,$$

satisfies an equation that is a direct reflection of the combinatorial definition, namely,

$$C(z) \quad = \quad 1 \quad + \quad z\, C(z)^2.$$

This is a quadratic equation whose solution is

$$C(z) = \frac{1 - \sqrt{1-4z}}{2z}.$$

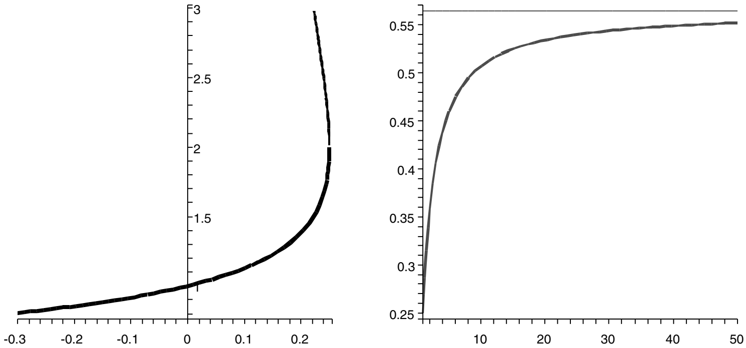

Figure 0.4. Left: the real values of the Catalan generating function, which has a square-root singularity at $z = \frac{1}{4}$. Right: the ratio $C_n/(4^n n^{-3/2})$ plotted together with its asymptote at $1/\sqrt{\pi} \doteq 0.56418$. The correspondence between *singularities* and *asymptotic* forms of *coefficients* is the central theme of Part B.

Then, by means of Newton's theorem relative to the expansion of $(1+x)^\alpha$, one finds easily ($x = -4z$, $\alpha = \frac{1}{2}$) the *closed form* expression

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

Stirling's asymptotic formula now comes to the rescue: it implies

$$C_n \sim C_n^\star \qquad \text{where} \qquad C_n^\star := \frac{4^n}{\sqrt{\pi n^3}}.$$

This last approximation is quite usable[4]: it gives $C_1^\star \doteq 2.25$ (whereas $C_1 = 1$), which is off by a factor of 2, but the error drops to 10% already for $n = 10$, and it appears to be less than 1% for any $n \geq 100$.

A plot of the generating function $C(z)$ in Figure 0.4 illustrates the fact that $C(z)$ has a *singularity* at $z = \frac{1}{4}$ as it ceases to be differentiable (its derivative becomes infinite). That singularity is quite different from a pole and for natural reasons it is known as a square-root singularity. As we shall see repeatedly, under suitable conditions in the complex plane, a square root singularity for a function at a point ρ invariably entails an asymptotic form $\rho^{-n} n^{-3/2}$ for its coefficients. More generally, it suffices to estimate a generating function near a singularity in order to deduce an asymptotic approximation of its coefficients. This correspondence is a major theme of the book, one that motivates the five central chapters (Chapters IV to VIII).

A consequence of the complex analytic vision of combinatorics is the detection of *universality phenomena* in large random structures. (The term is originally borrowed from statistical physics and is nowadays finding increasing use in areas of mathematics such as probability theory.) By universality is meant here that many quantitative

[4]We use $\alpha \doteq \mathsf{d}$ to represent a numerical approximation of the real α by the decimal d, with the last digit of d being at most ± 1 from its actual value.

properties of combinatorial structures only depend on a few global features of their definitions, not on details. For instance a growth in the counting sequence of the form

$$K \cdot A^n n^{-3/2},$$

arising from a square-root singularity, will be shown to be universal across *all* varieties of trees determined by a finite set of allowed node degrees—this includes unary–binary trees, ternary trees, 0–11–13 trees, as well as many variations such as non-plane trees and labelled trees. Even though generating functions may become arbitrarily complicated—as in an algebraic function of a very high degree or even the solution to an infinite functional equation—it is still possible to extract with relative ease *global asymptotic laws* governing *counting sequences*.

RANDOMNESS is another ingredient in our story. How useful is it to determine, exactly or approximately, counts that may be so large as to require hundreds if not thousands of digits in order to be written down? Take again the example of alternating permutations. When estimating their number, we have indeed quantified the proportion of these among all permutations. In other words, we have been predicting the *probability* that a random permutation of some size n is alternating. Results of this sort are of interest in all branches of science. For instance, biologists routinely deal with genomic sequences of length 10^5, and the interpretation of data requires developing enumerative or probabilistic models where the number of possibilities is of the order of 4^{10^5}. The language of probability theory then proves of great convenience when discussing characteristic parameters of discrete structures, since we can interpret exact or asymptotic enumeration results as saying something concrete about the likelihood of values that such parameters assume. Equally important of course are results from several areas of probability theory: as demonstrated in the last chapter of this book, such results merge extremely well with the analytic–combinatorial framework.

Say we are now interested in runs in permutations. These are the longest fragments of a permutation that already appear in (increasing) sorted order. Here is a permutation with 4 runs, separated by vertical bars:

$$2\,5\,8 \mid 3\,9 \mid 1\,4\,7 \mid 6\,.$$

Runs naturally present in a permutation are for instance exploited by a sorting algorithm called “natural list mergesort”, which builds longer and longer runs, starting from the original ones and merging them until the permutation is eventually sorted. For our understanding of this algorithm, it is then of obvious interest to quantify how many runs a permutation is likely to have.

Let $P_{n,k}$ be the number of permutations of size n having k runs. Then, the problem is once more best approached by generating functions and one finds that the coefficient of $u^k z^n$ inside the *bivariate* generating function,

$$P(z,u) \equiv \frac{1-u}{1-ue^{z(1-u)}} = 1 + zu + \frac{z^2}{2!}u(u+1) + \frac{z^3}{3!}u(u^2+4u+1) + \cdots,$$

gives the desired numbers $P_{n,k}/n!$. (A simple way of establishing the last formula bases itself on the tree decomposition of permutations and on the symbolic method; the numbers $P_{n,k}$, whose importance seems to have been first recognized by Euler,

Cambridge University Press
978-0-521-89806-5 - Analytic Combinatorics
Philippe Flajolet and Robert Sedgewick
Excerpt
More information

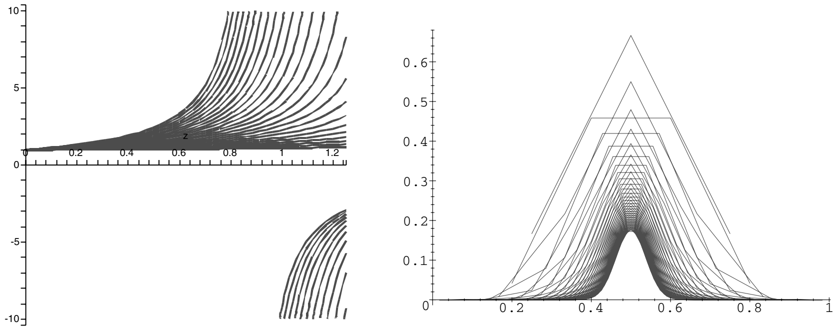

Figure 0.5. Left: A partial plot of the real values of the Eulerian generating function $z \mapsto P(z, u)$ for $z \in [0, \frac{5}{4}]$, illustrates the presence of a movable pole for A as u varies between 0 and $\frac{5}{4}$. Right: A suitable superposition of the histograms of the distribution of the number of runs, for $n = 2, \ldots, 60$, reveals the convergence to a Gaussian distribution (p. 695). Part C relates systematically the analysis of such a collection of singular behaviours to *limit distributions*.

are related to the *Eulerian numbers*, p. 210.) From here, we can easily determine effectively the mean, variance, and even the higher moments of the number of runs that a random permutation has: it suffices to expand blindly, or even better with the help of a computer, the bivariate generating function above as $u \to 1$:

$$\frac{1}{1-z} + \frac{1}{2}\frac{z\,(2-z)}{(1-z)^2}\,(u-1) + \frac{1}{2}\frac{z^2\left(6-4\,z+z^2\right)}{(1-z)^3}\,(u-1)^2 + \cdots .$$

When $u = 1$, we just enumerate all permutations: this is the constant term $1/(1-z)$ equal to the exponential generating function of all permutations. The coefficient of the term $u - 1$ gives the generating function of the *mean* number of runs, the next one provides the second moment, and so on. In this way, we discover the expectation and standard deviation of the number of runs in a permutation of size n:

$$\mu_n = \frac{n+1}{2}, \qquad \sigma_n = \sqrt{\frac{n+1}{12}}.$$

Then, by easy analytic–probabilistic inequalities (Chebyshev inequalities) that otherwise form the basis of what is known as the second moment method, we learn that the distribution of the number of runs is concentrated around its mean: in all likelihood, if one takes a random permutation, the number of its runs is going to be very close to its mean. The effects of such quantitative laws are quite tangible. It suffices to draw a *sample of one element* for $n = 30$ to get, for instance:

13, 22, 29|12, 15, 23|8, 28|18|6, 26|4, 10, 16|1, 5, 27|3, 14, 17, 20|2, 21, 30|25|11, 19|9|7, 24.

For $n = 30$, the mean is $15\frac{1}{2}$, and this sample comes rather close as it has 13 runs. We shall furthermore see in Chapter IX that even for moderately large permutations of size 10 000 and beyond, the probability for the number of observed runs to deviate

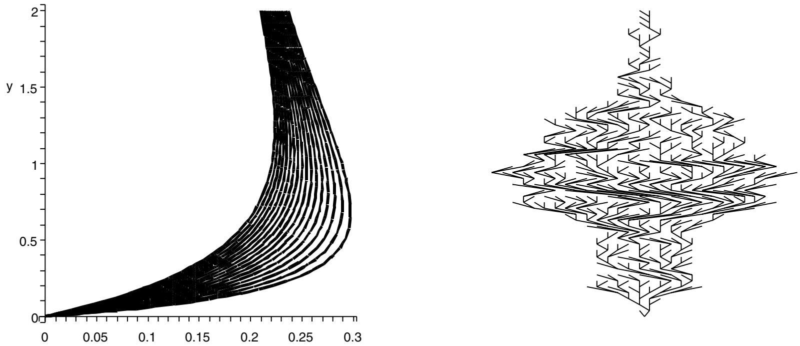

Figure 0.6. Left: The bivariate generating function $z \mapsto C(z, u)$ enumerating binary trees by size and number of leaves exhibits consistently a square-root singularity, for several values of u. Right: a binary tree of size 300 drawn uniformly at random has 69 leaves. As shown in Part C, *singularity perturbation* properties are at the origin of many randomness properties of combinatorial structures.

by more than 10% from the mean is less than 10^{-65}. As witnessed by this example, much regularity accompanies properties of large combinatorial structures.

More refined methods combine the observation of singularities with analytic results from probability theory (e.g., continuity theorems for characteristic functions). In the case of runs in permutations, the quantity $P(z, u)$ viewed as a function of z when u is fixed appears to have a pole: this fact is suggested by Figure 0.5 [left]. Then we are confronted with a fairly regular *deformation* of the generating function of all permutations. A parameterized version (with parameter u) of singularity analysis then gives access to a description of the asymptotic behaviour of the Eulerian numbers $P_{n,k}$. This enables us to describe very precisely what goes on: in a random permutation of large size n, once it has been centred by its mean and scaled by its standard deviation, *the distribution of the number of runs is asymptotically Gaussian*; see Figure 0.5 [right].

A somewhat similar type of situation prevails for binary trees. Say we are interested in leaves (also sometimes figuratively known as "cherries") in trees: these are binary nodes that are attached to two external nodes (□). Let $C_{n,k}$ be the number of trees of size n having k leaves. The bivariate generating function $C(z, u) := \sum_{n,k} C_{n,k} z^n u^k$ encodes all the information relative to leaf statistics in random binary trees. A modification of previously seen symbolic arguments shows that $C(z, u)$ still satisfies a quadratic equation resulting in the explicit form,

$$C(z, u) = \frac{1 - \sqrt{1 - 4z + 4z^2(1 - u)}}{2z}.$$

This reduces to $C(z)$ for $u = 1$, as it should, and the bivariate generating function $C(z, u)$ is a deformation of $C(z)$ as u varies. In fact, the network of curves of Figure 0.6 for several fixed values of u illustrates the presence of a smoothly varying square-root singularity (the aspect of each curve is similar to that of Figure 0.4). It is possible to analyse the *perturbation* induced by varying values of u, to the effect that