Largest Component in Random Combinatorial Structures

Xavier Gourdon

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Largest Component in Random Combinatorial Structures

Xavier GOURDON

N° 2548
Mai 1995

PROGRAMME 2

Largest Component in Random Combinatorial Structures

Xavier Gourdon
Algorithms Project, INRIA Rocquencourt

Abstract

We present a unified analytic framework dedicated to the estimation of the size of the largest component in random combinatorial structures.

Plus grande composante dans les structures combinatoires aléatoires

Résumé

Nous présentons un cadre analytique général destiné à l'estimation de la taille de la plus grande composante dans une structure combinatoire aléatoire.

To appear in *Proceedings of the 7th Conference on Formal Power Series and Algebraic Combinatorics.*

LARGEST COMPONENT IN RANDOM COMBINATORIAL STRUCTURES

XAVIER GOURDON

ABSTRACT. We present a unified analytic framework dedicated to the estimation of the size of the largest component in random combinatorial structures.

RÉSUMÉ. Nous présentons un cadre analytique général destiné à l'estimation de la taille de la plus grande composante dans une structure combinatoire aléatoire.

1. INTRODUCTION

The problem of predicting the occurrence of large components in random combinatorial structures is of interest in many branches of combinatorial modelling.

Given a window of 500 bases in a DNA sequence (with an alphabet $\{A, G, C, T\}$ of size 4) how "significant" is it to observe a sequence of 5 or 6 consecutive identical bases? In a similar spirit, Révész [13] describes the way an examination of "runs" in coin-flipping sequences (*i. e.* contiguous runs of identical outcomes) may be used to distinguish efficiently random sequences from man-made pseudo-random sequences.

In another context, that of cryptography, Quisquater and Delescaille [12] have conducted extensive computations to determine the behaviour of the standard DES (Data Enscription Standard) cryptosystem under iteration. They detect the usual occurrence of a "giant component" to which are attached a few "giant trees" in DES graphs. It is then of obvious interest to compare such observations against the random functional graph model since any significant deviation from randomness there may indicate the presence of a "hidden" structure that could be exploited by cryptanalysts.

In this paper, we present a unified analytic framework dedicated to the analysis of largest components in composite structures. This framework is based on an essential subdivision into three cases (non-critical, critical, super-critical) that depends on simple analytic conditions on intervening generating functions. A similar subdivision is also essential in characterising the distribution of the number of components, as shown by Soria and Flajolet [7]. Though they don't cover all cases, our conditions do lead to explicit distribution estimates in the asymptotic limit that are applicable to a large number of classical combinatorial structures. Three prototypical applications illustrating the fundamental trichotomy are size distribution of

(1) the largest root subtree in a random unlabelled rooted tree (the Catalan statistics, $\frac{1}{n+1}\binom{2n}{n}$);
(2) largest tree in a random mapping (the n^n statistics);
(3) largest summand in a random integer composition (the 2^{n-1} statistics).

Our work generalises several specific studies on largest components while placing the problem within a general theory of combinatorial schemas of [3, 7, 15]. Part of it extends results of Stepanov [16] relative to random mappings and of Knuth [10] relative to carry propagation in binary adders. Our results do not address however problems like largest cycles in random permutations, a

problem treated by Shepp and Lloyd [14] by means of a Tauberian argument (this schema should be discussed in a companion paper.)

2. Generating functions of largest components

2.1. Algebraic framework. The relation between generating functions (GF's)

$$C(z) = F(P(z))$$

is fundamental. It corresponds to combinatorial *substitution*

$$\mathcal{C} \approx \Phi(\mathcal{P}), \tag{1}$$

with $C(z)$, $F(w)$, $P(z)$ the GF's associated to $\mathcal{C}$, Φ, $\mathcal{P}$ respectively.

In the labelled case, this operation is the usual labelled substitution described by Joyal in [9]. In the unlabelled case, it is a form of "marked" substitution. Roughly, the meaning is that $\mathcal{C}$ is formed by substituting objects of $\mathcal{P}$ inside "atoms" of Φ. For example, the GF's of the generic set and generic sequence are

$$F(w) = e^w \quad \text{and} \quad F(w) = \frac{1}{1-w}.$$

This paper aims at studying the limit distribution of the random variable L_n representing the size of the largest $\mathcal{P}$-component in a random structure of size n in $\mathcal{C}$. In our context, the generating function of $\Phi(\mathcal{P})$-structures whose $\mathcal{P}$ components all have size less than or equal to m is $F(s_m(z))$, where

$$s_m(z) = \sum_{k=0}^{m} P_k\, z^k$$

denotes the truncation of the series $P(z)$ to order m. Thus,

$$\Pr(L_n \le m) = \frac{[z^n]F(s_m(z))}{[z^n]F(P(z))}, \tag{2}$$

and the problem is reduced to evaluate asymptotically $[z^n]F(s_m(z))$.

2.2. Analytic framework. It is known from classical analysis and analytic number theory that the asymptotic growth of coefficients of a series is determined by its analytic properties, especially its singular behaviour. Many functions used will be of the so-called *algebraic-logarithmic* (AL) type.

Definition 1. A complex function $f(z)$ is said to be Algebraic-Logarithmic (AL) at $\rho > 0$ if

- with the sole exception of $z = \rho$, $f(z)$ is analytic in an indented domain

  $$\Delta = \{z \in \mathbb{C},\ |z| \le \rho(1+\eta), |\arg(z-1)| \ge \phi\} \tag{3}$$

  for some $\eta > 0$ and $0 < \phi < \pi/2$;
- as z tends to ρ in Δ,

  $$f(z) = c + \left(1 - \frac{z}{\rho}\right)^{\alpha} \left(\log \frac{1}{1 - z/\rho}\right)^{\beta} (d + o(1)), \tag{4}$$

  with c, d, α and β complex numbers.

For reasons explained in [3], many elementary combinatorial structures have GF's of this type. Examples include simple families of trees in the sense of Meir and Moon [11], families of random mappings as considered by Arney and Bender [1], many classes of permutations defined by cycle constraints, etc. A basic theorem of Flajolet and Odlyzko [6] states that under the analytic continuation (3), the asymptotic condition (4) transfers to coefficients as

$$[z^n]f(z) = \frac{1}{\rho^n}\frac{1}{\Gamma(-\alpha)}\frac{(\log n)^\beta}{n^{1+\alpha}}(d + o(1))$$

whenever $\alpha \notin \{0, 1, 2, \dots\}$.

3. Non-critical occurrence

We discuss here the substitution schema $C(z) = F(P(z))$ when the dominant singularity of C is induced by the dominant singularity of P and P itself is AL. In that case, the analytic character of the outer function F is "non-critical". This situation covers for instance the size of the largest root subtree in a random Catalan tree (Example 1).

Theorem 1 (Non-critical case). *Assume*

(i) *the series $F(w)$ has a non zero radius of convergence R;*

(ii) *the function $P(z)$ is AL at $z = \rho > 0$. It satisfies, as z tends to ρ in its domain of analyticity*

$$P(z) = c - d\left(1 - \frac{z}{\rho}\right)^\alpha \left(\log\frac{1}{1 - z/\rho}\right)^\beta (1 + o(1)),$$

with c, d and α positive constants, α not an integer, and β a real number;

(iii) *the function $P(z)$ becomes singular before reaching the singularity of $F(w)$, that is $c = P(\rho) < R$.*

Then the distribution of the random variable $n - L_n$ (where L_n represents the size of the largest $\mathcal{P}$-component in a random $\Phi(\mathcal{P})$ structure of size n) tends to a discrete law: for any fixed non negative integer k, we have

$$\lim_{n\to\infty} \Pr(L_n = n - k) = \frac{b_k \rho^k}{F'(c)} \quad \text{where} \quad b_k = [z^k]F'(P(z)). \tag{5}$$

When $k \to \infty$, the value of (5) is asymptotically

$$\frac{d\,F''(c)}{\Gamma(-\alpha)F'(c)}\frac{(\log k)^\beta}{k^{\alpha+1}}.$$

Proof. Following (2), we have

$$\Pr(L_n = n - k) = \frac{[z^n]F(s_{n-k}(z)) - F(s_{n-k-1}(z))}{[z^n]F(P(z))}. \tag{6}$$

Denoting by $r_m(z) = P(z) - s_m(z)$ the rest of order m of $P(z)$, we have

$$F(s_m(z)) = F(P(z) - r_m(z)) = \sum_{\ell\ge 0}\frac{F^{(\ell)}(P(z))}{\ell!}(-1)^\ell r_m(z)^\ell.$$

When $n \geq 2m$, we have $[z^n]r_m(z)^\ell = 0$ for $\ell \geq 2$, thus $[z^n]F(s_m(z))$ is the coefficient of z^n in $F(P(z)) - r_m(z)F'(P(z))$. Thus, when $n \geq 2k$, (6) leads to

$$\Pr(L_n = n-k) = \frac{[z^n]\,(r_{n-k-1}(z) - r_{n-k}(z))\,F'(P(z))}{[z^n]F(P(z))} = \frac{[z^{n-k}]P(z)}{[z^n]F(P(z))}\,[z^k]F'(P(z)).$$

To conclude, we look at the behaviour of the two AL-functions $F(P(z))$ and $P(z)$ near their singularity and we use the transfer lemma of [6] on these functions. □

Comparing the distribution of L_n with that of the number of components is of interest. In the non-critical case, it is known (see [15]) that the distribution of the number of components in the non-critical case tends to a discrete law, which is consistent with our result: when the number of components is small, the largest component is large.

Distribution tail. The convergence to the discrete law (5) is not uniform enough to make a precise evaluation of the mean and variance directly (the only information we can derive is $E(L_n) \sim n$ and $V(L_n) = o(n^2)$, which is not satisfying). Thus we study the distribution tail, that is $\Pr(L_n \leq n-k)$ when k is large. This problem is much more difficult than the simple evaluation of the discrete law limit in the previous Theorem. For this purpose, we make use of singularity analysis.

Behaviour of the rests near the singularity. First, we study the behaviour of the rests $r_m(z) = \sum_{k>m} P_k\, z^k$ near the singularity $z = \rho$. For a fixed t, the generating function of $r_m(t)$ has a closed form in terms of $P(z)$

$$\sum_{m=0}^{+\infty} \frac{r_m(t)}{t^{m+1}}\, z^m = \frac{P(z) - P(t)}{z - t}. \tag{7}$$

Performing singularity analysis on this function from Cauchy's formulae (in the same vein as in [6]) gives the behaviour of $r_m(t)$ as $m \to \infty$.

Lemma 1 (Behaviour of the rests). *Let $P(z)$ be an AL-function at $z = \rho$, analytic (with the exception of $z = \rho$) in the indented domain (3) and satisfying (4) as $z \to \rho$ in Δ. We denote by $\mathcal{H}_0$ the Hankel contour that is the union of the two semi-axis*

$$\rho \mapsto \rho e^{i\phi} \quad (\rho \geq 0) \qquad \text{and} \qquad \rho \mapsto \rho e^{-i\phi} \quad (\rho \geq 0).$$

For all $a > 0$, we denote by $\mathcal{H}_a$ the Hankel contour clockwise oriented rounding $\mathcal{H}_0$ at a distance a at the left, and we denote by R_a the set of complex numbers at the left of $\mathcal{H}_a$. Then the rest $r_m(z)$ of $P(z)$ satisfies, as $m \to \infty$

$$r_m\left[\rho\left(1 + \frac{u}{m}\right)\right] = \rho^{m+1}\left(1 + \frac{u}{m}\right)^{m+1} \frac{(\log m)^\beta}{m^\alpha}\,(d\psi_\alpha(u) + o(1)), \tag{8}$$

where the $o(1)$ is uniform for $u \in R_1$, the function $\psi_\alpha(u)$ being defined by

$$\psi_\alpha(u) = \frac{1}{2i\pi}\int_{\mathcal{H}_{1/2}} \frac{(-p)^\alpha}{p-u}\,\frac{dp}{ep}.$$

This form of $\psi_\alpha(u)$ is not very explicit, but as we shall see further, it can be nicely written in terms of a Laplace transform.

Theorem 2 (Tail estimates in the non-critical case). *Under the assumptions of Theorem* 1, *we have the estimate, for all integer* $\ell \geq 2$

$$\Pr(L_n \leq m) = \frac{\Gamma(-\alpha)\, F^{(\ell)}(c)}{\ell!\, F'(c)} \left(\frac{(\log n)^\beta}{n^\alpha}\right)^{\ell-1} \lambda^{1+\ell\alpha} \left(d\, K_{\alpha,\ell}(\lambda) + o(1)\right), \tag{9}$$

uniformly as $n \to \infty$ *for* $\lambda = n/m$ *in any closed subinterval of* $]\ell - 1, \ell]$, *where*

$$K_{\alpha,\ell}(\gamma) = \frac{1}{2i\pi} \int_{\mathcal{H}} \left([(-u)^\alpha - e^u \psi_\alpha(u)]^\ell - (-1)^\ell e^{\ell u}\, \psi_\alpha(u)^\ell\right) \frac{du}{e^{u\gamma}}, \tag{10}$$

with $\mathcal{H}$ *a Hankel contour clockwise oriented rounding the positive real semi-axis.*

Proof. Suppose $\ell - 1 < n/m \leq \ell$. Let $\widehat{s}_m(z) = s_m(z) - c$. When $j < \ell$, we have $[z^n]\widehat{s}_m(z)^j = 0$, thus

$$[z^n]F(s_m(z)) = [z^n]\left(\sum_{j \geq \ell} \frac{F^{(j)}(c)}{j!}\, \widehat{s}_m(z)^j\right). \tag{11}$$

Let $g(z) = P(z) - c$. Replacing $\widehat{s}_m(z)$ by $g(z) - r_m(z)$ in (11), expanding $(g(z) - r_m(z))^j$ and using the fact that $[z^n]r_m(z)^j = 0$ for $j \geq \ell$, we find that $[z^n]F(s_m(z))$ is equal to the coefficient of z^n in the function

$$\frac{F^{(\ell)}(c)}{\ell!}\, B(z) + C(z),$$

with

$$B(z) = (g(z) - r_m(z))^\ell - (-1)^\ell r_m(z)^\ell \quad \text{and} \quad C(z) = \sum_{j=0}^{\ell-1} r_m(z)^j\, g(z)^{\ell+1-j}\, C_j(z),$$

where the $C_j(z)$ are AL at $z = \rho$ with $C_j(\rho)$ finite. Lemma 1 makes it possible to perform singularity analysis on this function, leading to the estimate

$$[z^n]F(s_m(z)) = \frac{F^{(\ell)}(c)}{\ell!} \frac{(\log m)^{\ell\beta}}{m^{1+\ell\alpha}} \left(d\, K_{\alpha,\ell}\left(\frac{n}{m}\right) + o(1)\right)$$

uniformly for $\lambda = n/m$ in any closed subinterval of $]\ell - 1, \ell]$. Dividing by the asymptotic value of $[z^n]F(P(z))$ finally gives the result. □

The function $K_{\alpha,\ell}(\lambda)$ is expressed in the next paragraph as an integral convolution.

The tail estimates make it possible to get a quite precise evaluation of the mean and variance.

Corollary 1 (Mean and variance in the non-critical case). *The mean and variance of the random variable* L_n *satisfy asymptotically, as* $n \to \infty$,

$$E(L_n) = n - \kappa_1\, n^{1-\alpha} (\log n)^\beta (1 + o(1)), \qquad \mathrm{Var}(L_n) \sim \kappa_2\, n^{2-\alpha} (\log n)^\beta,$$

where κ_1 *and* κ_2 *are constants defined by*

$$\kappa_1 = \frac{d\,\Gamma(-\alpha)\, F''(c)}{2F'(c)} \int_1^2 t^{2\alpha-1}\, K_{\alpha,2}(t)\, dt \quad \text{and} \quad \kappa_2 = \frac{d\,\Gamma(-\alpha)\, F''(c)}{F'(c)} \int_1^2 (t-1) t^{2\alpha-2} K_{\alpha,2}(t)\, dt.$$

An expression of $K_{\alpha,2}$ as a simple integral is given in the next paragraph, making possible to express the constants κ_1 and κ_2 as double integrals. When $\alpha = 1/2$ (which is the most common case encountered in the practice), their value can be computed explicitly:

$$\kappa_1 = \frac{d\,F''(c)}{\sqrt{\pi}F'(c)} \quad \text{and} \quad \kappa_2 = \frac{d\,F''(c)}{\sqrt{\pi}F'(c)}\left(1 - \frac{\pi}{4}\right). \tag{12}$$

Computation of $K_{\alpha,\ell}(\lambda)$. We make explicit the expression (10) of the function $K_{\alpha,\ell}(\lambda)$.

Lemma 2. *Denote by $H_\alpha(t)$ the function defined over $\mathbb{R}$ by*

$$H_\alpha(t) = \frac{1}{t^{\alpha+1}} \quad \text{if} \quad 0 < t \leq 1, \qquad H_\alpha(t) = 0 \quad \text{otherwise.}$$

The value of $K_{\alpha,\ell}(\lambda)$ for $\ell - 1 < \lambda \leq \ell$ is equal to $H_\alpha^{(\ell)}(\lambda)/\Gamma(-\alpha)^\ell$, where $H_\alpha^{(\ell)}$ is the ℓ-fold convolution of H_α with itself, that is

$$K_{\alpha,\ell}(\lambda) = \frac{1}{\Gamma^\ell(-\alpha)} \int_{t_1+t_2+\cdots+t_\ell=\lambda} H_\alpha(t_1)\cdots H_\alpha(t_\ell)\,dt_1\cdots dt_\ell.$$

Proof. The function $\psi_\alpha(u)$ looks like a Stieljes transform taken at the value $-u$ with a different contour. A Stieljes transform is a Laplace transform iterated twice. A similar property is true for $\psi_\alpha(u)$: suppose $\Re(u) < 0$ and u is at the left of the Hankel contour $\mathcal{H}$. We write

$$\psi_\alpha(u) = \frac{1}{2i\pi}\int_{\mathcal{H}} \frac{e^{-p}}{p^{1+\alpha}}\left(\int_0^\infty e^{x(p-u)}\,dx\right)dp = \frac{1}{2i\pi}\int_0^{+\infty}\left(\int_{\mathcal{H}} \frac{e^{-(x+1)p}}{p^{1+\alpha}}\,dp\right)e^{ux}\,dx.$$

The Hankel representation of the Γ-function gives an explicit value of the last integral $\int_{\mathcal{H}}$ from which we derive

$$\psi_\alpha(u) = \frac{e^{-u}}{\Gamma(-\alpha)}\mathcal{L}[G_\alpha(x)](-u), \qquad G_\alpha(x) = \begin{cases} x^{-\alpha-1} & \text{if } x > 1 \\ 0 & \text{if } 0 < x \leq 1 \end{cases}$$

where $\mathcal{L}$ denotes the Laplace transform.

Now suppose α is a complex number such that $\Re(\alpha) < 0$. Within the integral defining $\psi_\alpha(\lambda)$, we can shift the contour $\mathcal{H}$ to the line $\Re(c) = -1$. Since

$$(-u)^\alpha = \frac{1}{\Gamma(-\alpha)}\mathcal{L}\left[\frac{1}{x^{\alpha+1}}\right](-u),$$

$K_{\alpha,\ell}(\lambda)$ finally writes as

$$\frac{1}{\Gamma(-\alpha)^\ell}\left(\frac{1}{2i\pi}\int_{-1-i\infty}^{-1+i\infty}\left(\mathcal{L}[H_\alpha(-u)](p)^\ell - \mathcal{L}[G_\alpha(-u)](p)^\ell\right)e^{-p\lambda}\,du\right).$$

The last term is the inverse Laplace transform of the function $\mathcal{L}[H_\alpha]^\ell - \mathcal{L}[G_\alpha]^\ell$ taken at the point λ, thus it is equal to $H_\alpha^{(\ell)}(\lambda) - G_\alpha^{(\ell)}(\lambda)$. Since $\lambda < \ell$, we have $G_\alpha^{(\ell)}(\lambda) = 0$, which yields the result when $\Re(\alpha) < 0$. When λ is fixed, our functions are analytic in α, thus the result is true for all α. □

Example 1 (Largest subtree of a Catalan tree). A Catalan tree is a planar rooted unlabelled tree. Such a tree can be described recursively as a node followed by a sequence of subtrees which are of Catalan type. Thus, with our notations, the distribution of the size of the largest subtree of a Catalan tree corresponds to

$$F(w) = \frac{1}{1-w} \quad \text{and} \quad P(z) = \frac{1-\sqrt{1-4z}}{2} \quad \text{(Catalan tree generating function)}.$$

By Theorem 1, the random variable L_n counting the size of the largest subtree of a random Catalan tree satisfies

$$\lim_{n\to\infty} \Pr(L_n = n-k) = c_k, \qquad c_k = \frac{1}{4}, \frac{1}{8}, \frac{5}{64}, \frac{7}{128}, \frac{21}{512}, \ldots \quad \text{for} \quad k = 1, 2, \ldots$$

As for the mean and variance, Corollary 1 with (12) give

$$E(L_n) = n - \frac{2\sqrt{n}}{\sqrt{\pi}}(1+o(1)) \quad \text{and} \quad \operatorname{Var}(L_n) \sim \frac{2}{\sqrt{\pi}}\left(1-\frac{\pi}{4}\right) n^{3/2}.$$

4. Critical occurrence

In this case, the dominant singularity of $F(P(z))$ arises simultaneously from $P(z)$ and $F(w)$, a situation occurring for instance when studying the size of the largest tree in a random mapping (Example 2).

Theorem 3. *Assume*

(i) *the series $F(w)$ is AL at $w = R > 0$, and as $z \to R$ in its domain of analyticity*

$$F(w) = C + D\left(1-\frac{w}{R}\right)^{\gamma}\left(\log\frac{1}{1-w/R}\right)^{\delta}(1+o(1)),$$

where C, D, γ and δ are some constants with $D \neq 0$ and $\gamma < 1$, $\gamma \neq 0$;

(ii) *the function $P(z)$ is AL at $z = \rho > 0$ and $P(\rho) = R$. It satisfies, as $z \to \rho$ in its domain of analyticity*

$$P(z) = R - d\left(1-\frac{z}{\rho}\right)^{\alpha}\left(\log\frac{1}{1-z/\rho}\right)^{\beta}(1+o(1)),$$

with d and α positive constants, $\alpha < 1$, and β a real number.

Then the random variable L_n satisfies $\lim_{n\to\infty} \Pr(L_n \le n/\lambda) = f_{\alpha,\gamma}(\lambda)$ for $\lambda \ge 1$, where

$$f_{\alpha,\gamma}(\lambda) = \frac{\Gamma(-\alpha\gamma)\lambda^{\alpha\gamma+1}}{2i\pi}\int_{1-i\infty}^{1+i\infty}\left[p^{\alpha} - \frac{\mathcal{L}[G_{\alpha}(u)](p)}{\Gamma(-\alpha)}\right]^{\gamma} e^{\lambda p}\,dp. \tag{13}$$

The integral in (13) is an analytic function of γ for $\Re(\gamma) < 0$ which can be analytically continued for all complex number γ; for $\gamma > 0$ its value is defined by analytic continuation.

Proof. We can restrict to the case where $\rho = 1$. The function $C(z) = F(P(z))$ is AL at $z = 1$ and its asymptotic form near $z = 1$ is obtained from those of $F(w)$ and $P(z)$ near $w = R$ and $z = 1$. Using singularity analysis, we deduce

$$C_n = [z^n]F(P(z)) \sim \frac{K_n}{\Gamma(-\alpha\gamma)}, \qquad K_n = K\frac{(\log n)^{\beta\gamma+\delta}}{n^{\alpha\gamma+1}}, \quad K = D\,d^{\gamma}\,\alpha^{\delta}. \tag{14}$$

Now, following formulae (2), we study the behaviour of $[z^n]F(s_m(z))$. Let $\lambda = n/m$ and $\ell = \lfloor \lambda \rfloor$. From Taylor formula, $[z^n]F(s_m(z))$ is equal to the coefficient of z^n in the function

$$\sum_{j=0}^{\ell} (-1)^j \frac{F^{(j)}(P(z))}{j!} r_m(z)^j. \tag{15}$$

A classical result from analytic function theory states that the derivatives of $F^{(j)}(w)$ behave like the derivatives of the behaviour of $F(w)$ near $w = R$. Plugging this information into (15) together with the result of lemma 1 and performing singularity analysis finally gives for $[z^n]F(s_m(z))$ the asymptotic value

$$K_m\, g_{\alpha,\gamma}(\lambda), \qquad g_{\alpha,\gamma}(\lambda) = \frac{1}{2i\pi} \int_{\mathcal{H}} (-u)^{\alpha\gamma} \left[\sum_{j=0}^{\ell} \frac{(-1)^j (\gamma)_j}{j!} \left(\frac{c^u \psi_\alpha(u)}{(-u)^\alpha} \right)^j \right] e^{-u\lambda}\, du. \tag{16}$$

When α and λ are fixed, this integral is analytic in γ. When $\gamma < 0$, we can move the contour $\mathcal{H}$ to the vertical line $\Re(z) = -1$, and a change of variable leads to an expression in terms of an inverse Laplace transform

$$g_{\alpha,\gamma}(\lambda) = \frac{1}{2i\pi} \int_{1-i\infty}^{1+i\infty} p^{\alpha\gamma} \left[\sum_{j=0}^{\ell} \frac{(-1)^j (\gamma)_j}{j!} \left(\frac{\mathcal{L}[G_\alpha(u)](p)}{\Gamma(-\alpha) p^\alpha} \right)^j \right] e^{p\lambda}\, dp. \tag{17}$$

The function $G_\alpha(u)$ vanishes when $u < 1$, thus $G_\alpha^{(j)}(\lambda) = 0$ for $j > \ell$, and since a product transforms (by Laplace) into a convolution, the sum in the last integral can be extended until infinity and corresponds exactly to the expansion of $\left(1 - \frac{\mathcal{L}[G_\alpha(u)](p)}{\Gamma(-\alpha)p^\alpha}\right)^\gamma$. The result follows easily. □

Corollary 2 (Mean and variance in the critical occurrence). *Under the assumptions of the previous Theorem, the mean and variance of the random variable L_n satisfy asymptotically, as $n \to \infty$,*

$$E(L_n) \sim c_1\, n, \qquad \mathrm{Var}(L_n) \sim c_2\, n^2,$$

where the constants c_1 and c_2 are defined by

$$c_1 = \frac{1}{\alpha\gamma} \int_0^\infty \left[\left(1 - \frac{1}{\Gamma(-\alpha)} \int_x^\infty \frac{e^{-t}}{t^{\alpha+1}}\, dt \right)^\gamma - 1 \right] dx \tag{18}$$

and

$$c_2 = \frac{2}{\alpha\gamma(1-\alpha\gamma)} \int_0^\infty \left[\left(1 - \frac{1}{\Gamma(-\alpha)} \int_x^\infty \frac{e^{-t}}{t^{\alpha+1}}\, dt \right)^\gamma - 1 \right] x\, dx - c_1^2.$$

Proof. Thanks to the previous Theorem, we have

$$E(L_n) = \sum_{m \le n} [1 - \Pr(L_n \le m)] \sim \sum_{m \le n} \left[1 - f_{\alpha,\gamma}\left(\frac{n}{m} \right) \right] \sim c_1\, n \tag{19}$$

where $c_1 = \int_0^1 [1 - f_{\alpha,\gamma}(1/t)]\, dt$. This integral expression for the constant c_1 is not satisfying; we give another way of expressing c_1. Let

$$H(z) = \sum_{m \ge 0} [F(P(z)) - F(s_m(z))], \tag{20}$$

so that $E(L_n) = \frac{1}{c_n}[z^n]H(z)$. Formula (19) with (14) give an asymptotic value of $[z^n]H(z)$ from which it is easy to deduce the behaviour (we can restrict to $\rho = 1$)

$$H(z) \sim K(\alpha\gamma c_1)(1-z)^{\alpha\gamma-1}\left(\log\frac{1}{1-z}\right)^{\beta\gamma+\delta}, \tag{21}$$

valid as $z \to 1^-$, z being a real number. The behaviour of $H(z)$ near 1^- can also be computed directly from (20). Approaching sums by integrals (a technique used in [5, proof of Theorem 8] for example) gives for the behaviour of $H(z)$ near 1^- a formula like (21), but with $\alpha\gamma c_1$ replaced by an integral. By identification, this gives the expression (18) for c_1.

Starting with the formula $E(L_n^2) = \sum_{m\le n}(2m+1)\Pr(L_n > m)$, the same technique gives the result for the variance. □

Example 2 (Largest tree in random mappings). A random mapping is a set of cycles of labelled rooted general trees $\mathcal{T}$, which can be also interpreted as a sequence of trees $\mathcal{T}$. Thus it corresponds to the case

$$F(w) = \frac{1}{1-w} \quad \text{and} \quad P(z) = T(z), \qquad \text{where} \quad T(z) = ze^{T(z)}$$

(Cayley tree-function). The tree-function is AL near $z = 1/e$ and satisfies near this singularity $T(z) = 1 - \sqrt{2(1-ez)} + O(1-ez)$. Theorem 3 gives the limit distribution of the random variable L_n representing the size of the largest tree in a random mapping of size n in the form

$$\lim_{n\to\infty}\Pr(L_n \le n/\lambda) = f_{1/2,-1}(\lambda), \qquad \lambda \ge 1$$

with $f_{1/2,-1}(\lambda)$ given in (13). The value of $f_{1/2,-1}(\lambda)$ can be computed explicitly for $1 \le \lambda < 2$; instead of (13), we use the expression (17) which gives

$$f_{1/2,-1}(\lambda) = \frac{\Gamma(1/2)\lambda^{1/2}}{2i\pi}\int_{1-i\infty}^{1+i\infty}\left(p^{-1/2} + \frac{\mathcal{L}[G_{1/2}(u)](p)}{\Gamma(-1/2)p}\right)e^{\lambda p}\,dp.$$

We rewrite this in the form

$$f_{1/2,-1}(\lambda) = \frac{\Gamma(1/2)\lambda^{1/2}}{2i\pi}\int_{1-i\infty}^{1+i\infty}\left(\frac{\mathcal{L}[u^{-1/2}](p)}{\Gamma(1/2)} + \frac{\mathcal{L}[G_{1/2}(u)](p)\,\mathcal{L}[1](p)}{\Gamma(-1/2)}\right)e^{\lambda p}\,dp.$$

Since the integral is the inverse Laplace transform, this rewrites as

$$f_{1/2,-1}(\lambda) = \Gamma(1/2)\lambda^{1/2}\left(\frac{\lambda^{-1/2}}{\Gamma(1/2)} + \frac{(G_{1/2}(u)\star 1)(\lambda)}{\Gamma(-1/2)}\right) = 1 - \frac{\lambda^{1/2}}{2}\int_0^\lambda G_{1/2}(u)\,du = 2 - \lambda^{1/2}.$$

This is true for $1 \le \lambda \le 2$. The same technique can be used to compute $f_{1/2,-1}(\lambda)$ for $k \le \lambda < k+1$ when k is any positive integer, leading to much more complicated expressions when $k \ge 2$.

As for the mean and variance, Corollary 2 combined with numerical computations give $E(L_n) \sim c_1 n$ and $\mathrm{Var}(L_n) \sim c_2 n^2$ where $c_1 = 0.4834983\ldots$ and $c_2 = 0.0494698\ldots$ (the mean value of the size of the largest tree in a random mapping of size n was determined by Flajolet and Odlyzko in [5, Theorem 8] who gave the same expression for c_1).

5. Super-critical occurrence

In this last case, the singularity of $F(P(z))$ is dictated only by $F(w)$. The technique is a generalisation of the one Knuth used in [10] while analysing the average time for carry propagation. It consists essentially in studying the way the dominant singularity of $F(s_m(z))$ is modified as m increases.

Theorem 4 (Super-critical case). *Assume*

(i) *the series $F(w)$ is AL at $w = R > 0$;*

(ii) *the function $P(z)$ is AL at $z = \rho > 0$ and $P(\rho) > R$. It satisfies, as z tends to ρ in its domain of analyticity*

$$P(z) = R + d\left(1 - \frac{z}{\rho}\right)^{\alpha} \left(\log \frac{1}{1 - z/\rho}\right)^{\beta} (1 + o(1)),$$

with d, α and β real numbers, $\alpha \notin \{0, 1, 2, \ldots\}$ and $d \neq 0$.

Then the distribution of the random variable L_n satisfies

$$\Pr(L_n \leq m) = \exp(-n\, J_m)\,(1 + O(m J_m)) \quad \text{with} \quad J_m \sim \frac{d}{\Gamma(-\alpha) P'(a)(\rho - a)} \frac{(\log m)^{\beta}}{m^{\alpha+1}} \left(\frac{a}{\rho}\right)^{m}$$

where a is the unique number in $(0, \rho)$ such that $P(a) = R$.

Proof. We use formulae (2). The function $F(P(z))$ has only one dominant singularity at $z = a$; as for $F(s_m(z))$, it becomes singular at $z = a_m$ where a_m is the unique positive number such that $s_m(a_m) = R$. Since $a < \rho$ and $s_m(a) \to P(a)$, we have $a_m \to a$ as $m \to \infty$. These considerations permits to derive an estimation of $r_m(a_m)$ from the first terms of its expansion, leading to

$$r_m(a_m) \sim \left(\frac{a_m}{\rho}\right)^{m+1} \cdot \frac{1}{1 - a/\rho} \frac{d}{\Gamma(-\alpha)} \frac{(\log m)^{\beta}}{m^{\alpha+1}}.$$

Then from $r_m(a_m) = P(a_m) - P(a) \sim (a_m - a)P'(a)$, our estimate of $r_m(a_m)$ easily leads to

$$\frac{a_m}{a} = e^{J_m}, \qquad J_m \sim \left(\frac{a}{\rho}\right)^{m} \frac{d}{\Gamma(-\alpha)\, P'(a)(\rho - a)} \frac{(\log m)^{\beta}}{m^{\alpha+1}}.$$

Now it remains to perform singularity analysis on the two functions $F(P(z))$ and $F(s_m(z))$ near $z = a$ and $z = a_m$ which finally yields the result. □

Corollary 3 (Mean and variance in the super-critical occurrence). *Under the assumptions of the previous Theorem, the mean and variance of the random variable L_n satisfy asymptotically, as $n \to \infty$,*

$$E(L_n) = \log_T n - (\alpha + 1) \log_T \log n + \beta \log_T \log \log n + O(1), \qquad \mathrm{Var}(L_n) = O(1),$$

where $T = \rho/a$.

Proof. Since (J_m) tends geometrically to zero, we easily prove

$$E(L_n) = \sum_{m} \left[1 - e^{-nJ_m}\right] + O(1) = \sum_{m : nJ_m \leq 1} 1 + O(1), \tag{22}$$

the result for the mean follows then from the inverse asymptotic

$$nJ_m \leq 1 \quad \text{iff} \quad m \leq \log_T n - (\alpha + 1) \log_T \log n + \beta \log_T \log \log n + O(1).$$

The variance $\mathrm{Var}(L_n) = \sum_m \Pr(L_n = m)[m - E(L_n)]^2$ is $O(1)$ as can be proved by cutting the sum at $m = \lfloor E(L_n) \rfloor$ and using easy inequalities on each term of the two parts. □

The estimate of *harmonic sum* in (22) is generally treated thanks to Mellin transform technique [4], as done also in [10] or in [8] in the case $\alpha = -1$ and $\beta = 0$.

Example 3 (Largest summands in compositions). A composition is a sequence of positive integers, called *summands*, the size of a composition being the sum of its summands. The distribution of the largest summand L_n in a random composition of size n correspond to the case where

$$F(w) = \frac{1}{1-w} \quad \text{and} \quad P(z) = \frac{z}{1-z}.$$

The function $F(P(z))$ becomes singular at $z = 1/2$. At this point, $P(z)$ is regular so that we are in the super-critical case. Theorem 4 gives

$$\Pr(L_n \le m) = \exp(-nJ_m)(1 + O(nJ_m)) \quad \text{with} \quad J_m \sim \frac{1}{2}\frac{1}{2^m},$$

and from the Corollary

$$E(L_n) = \log_2 n + O(1), \qquad \mathrm{Var}(L_n) = O(1).$$

A similar analysis applies to longest runs in random strings [2, 10].

Example 4 (Longest sequence of unary nodes in unary-binary trees). We work with rooted plain trees. The family of unary-binary trees $\mathcal{C}$ can be obtained from the family of binary trees $\mathcal{B}$ by substituting each node with a non-empty sequence of unary nodes. In other terms, $\mathcal{C} = \mathcal{B}(\mathcal{S})$ where $\mathcal{S}$ is the family of non-empty sequence of unary nodes. We wish to study the random variable L_n counting the size of the longest $\mathcal{S}$ component in a random unary-binary tree of size n. This corresponds to the case where

$$P(z) = \frac{z}{1-z} \quad \text{and} \quad F(w) = B(w), \qquad \text{where} \quad B(w) = \frac{1 - \sqrt{1-4w^2}}{2w}$$

(binary trees generating function). The function $F(P(z))$ becomes singular when $z = 1/3$, near which $P(z)$ is regular: we are in the super-critical case. Theorem 4 gives

$$\Pr(L_n \le m) = \exp(-nJ_m)(1 + O(nJ_m)) \quad \text{with} \quad J_m \sim \frac{3}{2}\frac{1}{3^m}$$

and from the corollary

$$E(L_n) = \log_3 n + O(1), \qquad \mathrm{Var}(L_n) = O(1).$$

6. Conclusion

Methods of this paper also permit us to determine more complete asymptotic expansions while giving access to local limit theorem. Possible extensions of this work are

- largest components in product schemas of the form $\mathcal{C} = \mathcal{A} \times \Phi(\mathcal{P})$;
  distribution of the r-th largest component;
- distribution of the smallest components;
- problems like counting the largest cycle in a random mapping.

As done in [15, 7], it would be also of interest to study systematically the framework where generating functions are of the *exp-log* type.

Acknowledgment. This work was partly supported by the ESPRIT Basic Research Action of the E. C. No. 7141 (ALCOM II).

REFERENCES

1. J. Arney and E. D. Bender. Random mappings with constraints on coalescence and number of origins. *Pacific Journal of Mathematics*, 103:269–294, 1982.
2. W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. John Wiley, 1971.
3. P. Flajolet, B. Salvy, and P. Zimmermann. Automatic average–case analysis of algorithms. *Theoretical Computer Science, Series A*, 79(1):37–109, February 1991.
4. Philippe Flajolet, Xavier Gourdon, and Philippe Dumas. Mellin transforms and asymptotics: harmonic sums. *Theoretical Computer Science*, special issue on Analysis of Algorithms, 1995. to appear.
5. Philippe Flajolet and Andrew M. Odlyzko. Random mapping statistics. In J-J. Quisquater and J. Vandewalle, editors, *Advances in Cryptology*, volume 434 of *Lecture Notes in Computer Science*, pages 329–354. Springer Verlag, 1990. Proceedings of EUROCRYPT'89, Houtalen, Belgium, April 1989.
6. Philippe Flajolet and Andrew M. Odlyzko. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, 3(2):216–240, 1990.
7. Philippe Flajolet and Michèle Soria. General combinatorial schemas: Gaussian limit distributions and exponential tails. *Discrete Mathematics*, (114):159–180, 1993.
8. Xavier Gourdon and Helmut Prodinger. Random subgraphs of the n-cycle. *Discrete Mathematics*, submitted as short communication, 1994.
9. André Joyal. Une théorie combinatoire des séries formelles. *Advances in Mathematics*, 42(1):1–82, 1981.
10. D. E. Knuth. The average time for carry propagation. *Indagationes Mathematicae*, 40:238–242, 1978.
11. A. Meir and J. W. Moon. On the altitude of nodes in random trees. *Canadian Journal of Mathematics*, 30:997–1015, 1978.
12. J.-J. Quisquater and J.-P. Delescaille. How easy is collision search? Application to DES. In *Proceedings of EUROCRYPT'89*, volume 434 of *Lecture Notes in Computer Science*, pages 429–433. Springer-Verlag, 1989.
13. P. Révész. Strong theorems in coin tossing. In *Proc. 1978 Int. Congress of Mathematicians*, Helsinki, 1980.
14. L. A. Shepp and S. P. Lloyd. Ordered cycle lengths in a random permutation. *Transactions of the American Mathematical Society*, 121:340–357, 1966.
15. Michèle Soria-Cousineau. *Méthodes d'analyse pour les constructions combinatoires et les algorithmes*. Doctorat d'état, Université de Paris–Sud, Orsay, July 1990.
16. V. E. Stepanov. Limit distributions of certain characteristics of random mappings. *Theory of Probability and Applications*, 14:612–626, 1969.

ALGORITHMS PROJECT, INRIA ROCQUENCOURT, BP 105 - 78153 LE CHESNAY CEDEX (FRANCE)
E-mail address: Xavier.Gourdon@inria.fr

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine - Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes - IRISA, Campus universitaire de Beaulieu 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 Grenoble Cedex 1 (France)
Unité de recherche INRIA Sophia Antipolis - 2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

ISSN 0249 - 6399