HAL
open science

Generalized digital trees and their difference-differential equations

Philippe Flajolet, Bruce Richmond

To cite this version:

Philippe Flajolet, Bruce Richmond. Generalized digital trees and their difference-differential equations. [Research Report] RR-1423, INRIA. 1991. inria-00075137

HAL Id: inria-00075137

https://inria.hal.science/inria-00075137

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

N° 1423

Programme 2
Calcul Symbolique, Programmation
et Génie logiciel

GENERALIZED DIGITAL TREES AND THEIR DIFFERENCE-DIFFERENTIAL EQUATIONS

Philippe FLAJOLET
Bruce RICHMOND

Avril 1991

RR_1423

Generalized Digital Trees and their Difference–differential Equations

Philippe FLAJOLET *and* Bruce RICHMOND

Abstract. We consider a tree partitioning process in which n elements are split into b at the root of a tree (b a design parameter), the rest going recursively into two subtrees with a binomial probability distribution. This extends some familiar tree structures of computer science like the digital trie and the digital search tree. The exponential generating function for the expected size of the tree satisfies a difference–differential equation of order b.

The solution involves going to ordinary (rather than exponential) generating functions, analyzing singularities by means of Mellin transforms and contour integration. The method is of some general interest since a large number of related problems on digital structures can be treated in this way via singularity analysis of ordinary generating functions.

Arbres digitaux généralisés et equations différentielles et aux différences associées

Philippe FLAJOLET et Bruce RICHMOND

Résumé. On considère un processus de partitionnement arborescent dans lequel n éléments se scindent en b qui vont à la racine de l'arbre (b un paramètre du problème), le reste se partageant récursivement entre deux sous–arbres, selon une loi de probabilité binomiale. Ce processus généralise diverses structures usuelles de l'informatique telles l'arbre digital (dit "treille") et l'arbre de recherche digital. La série génératrice de la taille moyenne de l'arbre vérifie une équation différentielle et aux différence d'ordre b.

La solution met en jeu le passage aux séries génératrices ordinaires (plutôt qu'exponentielles), ainsi qu'une analyse de singularité par transformée de Mellin et intégration de contour. La méthode utilisée est de quelque intérêt car elle permet de traiter de la sorte par l'analyse de singularités de séries ordinaires bon nombre de problèmes sur des structures digitales.

Generalized Digital Trees and their Difference–differential Equations

Philippe Flajolet*
Algorithms Project
INRIA Rocquencourt
F-78153 Le Chesnay, France

Bruce Richmond†
Dept of Combinatorics and Optimization
University of Waterloo, Waterloo
Ontario, N2L 3G1, Canada

April 13, 1991

Abstract. We consider a tree partitioning process in which n elements are split into b at the root of a tree (b a design parameter), the rest going recursively into two subtrees with a binomial probability distribution. This extends some familiar tree structures of computer science like the digital trie and the digital search tree. The exponential generating function for the expected size of the tree satisfies a difference–differential equation of order b,

$$\frac{d^b}{dz^b} f(z) = e^z + 2e^{z/2} f(\frac{z}{2}).$$

The solution involves going to ordinary (rather than exponential) generating functions, analyzing singularities by means of Mellin transforms and contour integration. The method is of some general interest since a large number of related problems on digital structures can be treated in this way via singularity analysis of *ordinary* generating functions.

*Work of this author was supported in part by the Basic Research Action of the E.C. under contract No. 3075 (Project ALCOM).

†The research of this author was done while visiting INRIA and was also supported in part by NSERC under grant A–4067.

1 Introduction

We deal with a recursive (tree) partitioning process that depends on some integer parameter b. Given n items, with $n > b$, the process puts b of them aside in the root of a binary tree. The remaining $n - b$ elements separate into two subgroups (subtrees), each of them flipping an unbiased coin. The probability that the first subgroup (left subtree) has size k is thus the binomial probability,

$$\pi_{n,k} = \frac{1}{2^{n-b}} \binom{n-b}{k}. \tag{1}$$

The subgroups again split recursively by the same process. If a group has a cardinality n such that $n \leq b$, then its recursive splitting stops. A realization of this process is clearly attached to a particular binary tree in which internal nodes contain b items, while external nodes contain between 0 and b nodes. Nodes corresponding to groups of cardinality 0 are called empty nodes.

When $b = 1$, we obtain in this way the classical *digital search tree* structure invented by Coffman and Eve in 1970, see [13, p. 489] or [19, p. 245]. For $b = 0$, the process defined above determines an infinite tree; however, if we retain only a suitable "finite" part of the tree, we obtain the classical *search "trie"*, see [13, p. 481] or [19, p. 248]. For general b, the corresponding tree structure seems to have been the basis of a folk algorithm in the late 1970's for maintaining paged hashing tables, the idea being to replace the usual chaining technique by a faster dichotomic access based on bits of hashed records.

Let f_n be the expected number of non-empty nodes of such a tree. (In computer implementations empty nodes need not be represented effectively thanks of the use of "null" pointers.) We propose to analyze f_n as a function of n, when the parameter b is kept fixed (we shall suppress the dependence upon b except where it is important). The basic recurrence is

$$\begin{cases} f_0 &= 0 \\ f_n &= 1 & \text{if } 1 \leq n \leq b \\ f_n &= 1 + \sum_{k=0}^{n-b} \pi_{n,k}(f_k + f_{n-k}) & \text{if } n > b \end{cases} \tag{2}$$

with $\pi_{n,k}$ the Bernoulli probability of Eq. (1).

We introduce the ordinary generating function (OGF) and the exponential generating function (EGF),

$$F(z) = \sum_{n \geq 0} f_n z^n \qquad f(z) = \sum_{n \geq 0} f_n \frac{z^n}{n!}.$$

Proposition 1 *The exponential generating function $f(z)$ satisfies the difference-differential equation*

$$\frac{d^b}{dz^b} f(z) = e^z + 2e^{z/2} f(\frac{z}{2}), \tag{3}$$

with initial conditions

$$f(z)\Big|_{z=0} = 0, \qquad \frac{d^j}{dz^j} f(z)\Big|_{z=0} = 1 \quad \textit{for } j = 1, 2, \ldots, b.$$

This model includes as subcases the usual model of tries ($b = 0$) and the usual model of binary search trees ($b = 1$). When $b = 0, 1$, there is a route by now classical (see [8, 13]) to such equations; it consists in an asymptotic analysis with several stages: (i) explicit solution of the functional equation; (ii) a Taylor expansion providing the coefficients; (iii) analysis of the coefficients via either Mellin transforms or contour integrals of the type used in the calculus of finite differences. We refer to [8] for a partial survey of these techniques.

The itinerary we follow here is different. It starts with the observation that corresponding ordinary generating functions satisfy *functional equations* of a simpler form that can be solved by iteration. *Mellin transforms* are then used to determine the behaviour of these OGF's near the dominant singularity, $z = 1$. This singularity is an accumulation point of simple poles. The behaviour of the OGF $F(z)$ is determined by Mellin transforms, through an analysis that requires extended asymptotics in the complex plane, see e.g., [7] for other examples.

Mellin transforms are computed here via standard techniques of contour integrals. They involve higher order basic hypergeometric functions, and in the particular case of $b = 2$, they lead to q–Bessel functions.

We find that near 1, $F(z)$ satisfies

$$F(z) = \frac{1}{(1-z)^2}[q_0 + \widehat{S}(\log_2(1-z)^{-1})] + O((1-z)^{-3/2}), \tag{4}$$

where $\widehat{S}(u)$ is a periodic function of u with mean value 0. We can then apply the technique of *singularity analysis* (for a comparable situation, see Odlyzko's counting of 2-3 trees [18], and [6] for general theorems). They enable us to derive an asymptotic form of f_n from the singular expansion (4),

$$f_n = n\,[q_0 + S(\log n)] + O(n^{1/2}), \tag{5}$$

where S is also a periodic function with mean value 0, see Theorem 1 below.

The path taken here is thus a *two stage analysis* based on (complex) Mellin asymptotics combined with singularity analysis like [6]. A similar two level procedure where (complex) Mellin asymptotics is combined with the saddle point method constitutes the method of Meinardus in the asymptotic theory of integer partitions [2].

Here are values of q_0 corresponding to a few values of b, as computed by numerical integration.

$$\begin{aligned} q_0(2) &= 0.5747 \\ q_0(3) &= 0.4069 \\ q_0(4) &= 0.3159 \\ q_0(5) &= 0.2585 \\ q_0(10) &= 0.1360 \end{aligned} \tag{6}$$

As a byproduct of the analysis, we obtain the estimate $q_0(b) \approx (b \log 2)^{-1}$, for large b. In the context of paging in computer applications, this means that nodes tend to be about 69% full ($\log 2 = 0.69$). Interestingly enough, the same filling ratio is achieved by paged tries and various dynamic hashing strategies [5, 15].

2 A Streamlined Analysis

In this section, we present the major steps of the analysis in all detail, except for some more technical Mellin transform computations that are relegated to the next section in order not to obscure the proof structure too much.

Our purpose here is to prove the following result

Theorem 1 *The expected number of non-empty nodes in a random tree built from n elements satisfies*

$$f_n = n\,[q_0 + S(\log n)] + O(\sqrt{n})$$

where

$$q_0 = \frac{1}{\log 2}\int_0^\infty \frac{(1+t)^{b-1}}{[(1+t)(1+t/2)(1+t/4)\cdots]^b}\,dt,$$

and $S(u)$ is a periodic function with mean value 0 and Fourier expansion

$$S(u) = \sum_{k\in\mathbb{Z}\setminus 0} q_k e^{2ik\pi u}$$

with

$$q_k = \frac{1}{(\log 2)\cdot\Gamma(2+2ik\pi/\log 2)}\int_0^\infty \frac{(1+t)^{b-1}}{[(1+t)(1+t/2)(1+t/4)\cdots]^b} t^{2ik\pi/\log 2}\,dt.$$

2.1 Basic Equations

We obtain first an explicit form of the OGF, $F(z)$.

Lemma 2 *The ordinary generating function $F(z)$ is given by*

$$F(z) = \frac{1}{1-z}G(\frac{z}{1-z}), \tag{7}$$

where

$$G(z) = \sum_{j=0}^{\infty} \frac{2^{-bj(j-1)/2}\,2^j z^{bj} P(z/2^j)}{[(1+z)(1+z/2)\cdots(1+z/2^j)]^b}, \tag{8}$$

and $P(z) = z(1+z)^{b-1}$.

PROOF. Like in several other analysis on digital trees, we start by setting

$$g(z) = e^{-z}f(z), \qquad g(z) = \sum_{n\geq 0} g_n \frac{z^n}{n!} \qquad \text{so that } f_n = \sum_{j=0}^{n}\binom{n}{j} g_j. \tag{9}$$

The induced differential equation of $g(z)$ is

$$\sum_{j=0}^{b}\binom{b}{j}\frac{d^j}{dz^j}g(z) = 1 + 2g(\frac{z}{2}),$$

or equivalently for coefficients ($n > b$)

$$\sum_{j=0}^{b} \binom{b}{j} g_{n+j} = 2^{1-n} g_n.$$

It is now natural to solve this recurrence by the use of *ordinary* generating functions. Let $G(z) = \sum_{n\geq 0} g_n z^n$. The relation between f_n and g_n first gives rise to the functional relation $F(z) = (1-z)^{-1}G(z/(1-z))$. The recurrence relation on the g_n leads to the simple functional equation (multiply by z^{n+b} and sum),

$$G(z)(1+z)^b = 2z^b G(\frac{z}{2}) + P(z), \tag{10}$$

where $P(z)$ is obtained by adjusting the terms of order 0 to b, using the relations

$$f(z) = e^z - 1 + O(z^{b+1}), \quad g(z) = 1 - e^{-z} + O(z^{b+1}), \quad G(z) = \frac{z}{1+z} + O(z^{b+1}).$$

Finally, the solution of the functional equation (10) is obtained by iteration:

$$\begin{aligned} G(z) &= \frac{P(z)}{(1+z)^b} + \frac{2z^b}{(1+z)^b} G(z/2) \\ &= \frac{P(z)}{(1+z)^b} + \frac{2z^b \cdot P(z/2)}{(1+z)^b \cdot (1+z/2)^b} + \frac{2z^b \cdot 2(z/2)^b}{(1+z)^b \cdot (1+z/2)^b} G(z/4), \end{aligned}$$

and so on. ■

A closely related form of $G(z)$ is also especially useful:

$$G(t^{-1}) = \sum_{j=0}^{\infty} \frac{2^j \bar{P}(2^j t)}{[(1+t)(1+2t)\cdots(1+2^j t)]^b},$$

with $\bar{P}(z)$ a reciprocal function of $P(z)$,

$$\bar{P}(z) = z^b P(1/z) = (1+z)^{b-1}.$$

Thus generating functions admit expansions as sums of rational functions. These rational functions can in turn be expanded into partial fractions. From there, explicit expressions could be derived for the coefficients f_n: these expressions appear to be somewhat impractical when b exceeds 2, so that we do not attempt to make them explicit; for $b = 0, 1$, they coincide with the forms obtained for tries and standard digital trees by classical methods, and we thus have a new way of carrying out these analyses.

The singularities of $F(z)$ and $G(z)$ are also apparent from these expressions: $G(z)$ has singularities at $z = -2^j$ with accumulation point at $-\infty$; accordingly, $F(z)$ is singular at $z = (1-2^{-k})^{-1}$, for $k = 1, 2, \ldots$, and at their accumulation point $z = 1$.

2.2 Mellin transforms

The idea is to estimate $G(z)$ as z tends to ∞ directly, for z in a neighbourhood of $+\infty$. This asymptotic expansion provides via the mapping $z \mapsto z/(1-z)$ the behaviour of $F(z)$ near its singularity $z = 1$. The analysis relies on Mellin transforms.

The convenient form is to set $z = 1/t$ and we need to consider $t \to 0$. Define first

$$Q(u) = \prod_{j=0}^{\infty}(1 + \frac{u}{2^j}).$$

Then, from Proposition 1, we have

$$\frac{G(t^{-1})}{Q(t/2)^b} = \sum_{k=0}^{\infty} \frac{2^k (2^k t)^b P(\frac{1}{2^k t})}{(Q(2^k t))^b}.$$

By introducing the infinite product, we have thus reduced the analysis of $G(t^{-1})$ to that of the sum,

$$H(t) = \sum_{k=0}^{\infty} 2^k \frac{\bar{P}(2^k t)}{Q^b(2^k t)},$$

a particular case of a "harmonic sum" which is naturally treated by Mellin transforms.

We recall that the Mellin transform of a function $\psi(t)$ is the function denoted $\psi^*(s)$ such that

$$\psi^*(s) = \int_0^{\infty} \psi(t) t^{s-1}\, dt.$$

Lemma 3 *The Mellin transform of $H(t) = G(t^{-1})/Q^b(t/2)$ is defined for $\Re(s) > 1$ and it satisfies*

$$H^*(s) = \frac{h^*(s)}{1 - 2^{1-s}} \qquad \textit{where} \qquad h^*(s) = \int_0^{\infty} \frac{(1+t)^{b-1}}{Q^b(t)} t^{s-1}\, dt.$$

PROOF. Let us write $h(t) = \bar{P}(t)/Q(t)$. Since $h(at)$ transforms into $a^{-s}h^*(s)$, by linearity, the Mellin transform of $H(t)$ is found to be

$$H^*(s) = h^*(s) \cdot \sum_{k=0}^{\infty} 2^k 2^{-ks} = \frac{h^*(s)}{1 - 2^{1-s}}.$$

Sufficient conditions for the validity are: (i) the absolute convergence of the sum, which means $\Re(s) > 1$; (ii) the absolute convergence of $h^*(s)$ which necessitates $\Re(s) > 0$. ∎

2.3 Mellin Analysis

It is not too hard to use Mellin analysis in order to establish estimates for $G(z)$ when z tends to $+\infty$ along the *real* line, and accordingly, for $F(z)$ when z tends to 1 along the *real* ray $[0, 1]$. Some of the difficulty of our problems comes from the fact that we need a *continuation into the complex plane* of these asymptotic expansions.

As is usual in a Mellin analysis, we apply the inversion theorem in order to recover $H(t)$ from $H^*(s)$. We have [4]

$$\frac{G(t^{-1})}{Q^b(t)} = \frac{1}{2i\pi}\int_{d-i\infty}^{d+i\infty} H^*(s)t^{-s}\,ds, \qquad d > 1. \tag{11}$$

Apart from the explicit form of $H^*(s)$, we also require in passing some growth estimates for the Mellin transform $h^*(s)$

$$h^*(c+ix) = O\left(|x|^{b-1}e^{-\pi|x|}\right) \qquad \text{as } x \to \pm\infty, \text{ for } c > 0 \tag{12}$$

whose proof will be deferred till the next section (see Lemma 5).

Lemma 4 *The function $F(z)$ satisfies locally around $z = 1$, $|\mathrm{Arg}(z-1)| > \pi/4$,*

$$F(z) = \frac{1}{(1-z)^2}[q_0 + \widehat{S}(\log_2(1-z)^{-1})] + O((1-z)^{-3/2}),$$

where $\widehat{S}(u)$ is defined by the Fourier expansion

$$\widehat{S}(u) = \frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus 0} h^*(\xi_k)e^{2ik\pi u} \text{ with } \xi_k = 1 + \frac{2ik\pi}{\log 2}.$$

PROOF. We apply the standard technique of Mellin analysis and in the inverse Mellin integral (11) we shift d to the left. If $t = re^{i\phi}$,

$$|t^{d-ix}| = r^{-d}e^{x\phi}.$$

Thus from the fundamental growth property (12), we can choose any $d > 0$ in order to ensure the convergence of (11), provided $|\mathrm{Arg}(t)| < \pi$ and we avoid the line of poles of $(1-2^{1-s})^{-1}$, i.e., $d = 1$.

Recall that the standard technique is to pick a contour consisting of two vertical lines with real parts d and d_1, and two horizontal lines with $|\Im(s)| \to +\infty$ that pass in between the poles of $H^*(s)$, at $|\Im(s)| = 2i\pi(k+1/2)/\log 2$. The smallness properties implies that the integrals along horizontal lines are negligible. By Cauchy's residue theorem, we thus get

$$\frac{G(t^{-1})}{Q^b(t/2)} = \sum \mathrm{Res}(H^*(s)t^{-s})_{\Re(s)>1/2} + O(t^{-1/2}), \tag{13}$$

provided we take, as we may, $d_1 = 1/2$. Now the singularities of $H^*(s)$ are at the points

$$\xi_k = 1 + \frac{2ik\pi}{\log 2},$$

k an integer. These are simple poles with residue $h^*(\xi_k)/\log 2$. Thus, from (13), and using the fact that $Q(t/2) = 1 + O(t)$, we find our main estimate

$$G(t^{-1}) = \frac{1}{t\log 2}[h^*(1) + \sum_{k\neq 0} h^*(\xi_k)t^{2ik\pi/\log 2}] + O(t^{-1/2}).$$

The residues at the non-real poles contribute a Fourier series in $\log_2 t$,

$$\widehat{S}(\log_2 t) = \frac{1}{\log 2} \sum_{k \in \mathbb{Z} \setminus 0} h^*(\xi_k) e^{2ik\pi \log_2 t}.$$

By the growth conditions on $h^*(s)$, this series has exponentially fast convergence.

From these developments, the asymptotic form of $F(z)$ finally follows via the basic relation $F(z) = (1-z)^{-1} G(z/(1-z))$. The validity condition $|\mathrm{Arg}(t)| < \pi$ for Mellin inversion is amply fullfilled since we have assumed $|\mathrm{Arg}(z-1)| > \pi/4$. We find $q_0 = h^*(1)/\log 2$, and periodic fluctuations in the form of the Fourier series $\widehat{S}$. ∎

2.4 Singularity analysis.

We now need to translate term by term an infinite number of periodic fluctuations. The situation is identical to the case of 2-3 trees of Odlyzko [18] and also to that of register allocation in [7].

The basic method is the one called singularity analysis and detailed in [6]. Let $[z^n]a(z)$ denote the coefficient of z^n in the Taylor expansion of $a(z)$. We have the following "transfers" from function to coefficients,

$$\begin{cases} [z^n](1-z)^{-2} & = & n+1 \\ [z^n](1-z)^{-1-\xi_k} & = & \dfrac{n^{\xi_k}}{\Gamma(1+\xi_k)}(1+O(\dfrac{1}{n})) \\ [z^n]O((1-z)^{-3/2}) & = & O(n^{1/2}). \end{cases}$$

The last relation is applicable in the context of our problem since the singular expansion holds in an extended area of the complex plane, as guaranteed by Lemma 4.

From these principles, we derive

$$f_n = n[q_0 + \frac{1}{\log 2} \sum_{k \in \mathbb{Z} \setminus 0} \frac{h^*(\xi_k)}{\Gamma(1+\xi_k)} n^{2ik\pi/\log 2}] + O(n^{1/2}),$$

which, apart from notational details, coincides with our main estimate in the statement of Theorem 1.

3 Mellin Integrals and q–analogs

We now establish some of the properties of the Mellin transform $h^*(s)$ that are needed in order to complete the proof of Theorem 1. The computations developed here also provide alternative series forms for the integral representation of q_0 and reveal some connections with basic hypergeometric functions that are especially useable for low values of b (we work out the case $b = 2$). Finally, they show that the periodic fluctuations in the form of the function $S(u)$ are from a standard set of functions encountered everywhere in this range of problems.

It will prove convenient to operate here with the simple Mellin integral

$$I^*(s) = \int_0^\infty \frac{1}{Q^b(t)} t^{s-1}\, dt,$$

on whose properties we now concentrate. Properties of $I^*(s)$ easily carry over to $h^*(s)$ because of the relation

$$h^*(s) = \sum_{j=0}^{b-1} \binom{b-1}{j} I^*(s+j).$$

Lemma 5 *The function $I^*(s)$ admits the representation*

$$I^*(s) = \frac{\pi}{\sin \pi s}[A_0(2^s) + (s-1)A_1(2^s) + \cdots + (s-1)(s-2)\cdots(s-b+1)A_{b-1}(s)],$$

where the $A_k(x)$'s are entire functions.

We observe right away that the $A_k(2^s)$ are themselves entire functions of s with the complex period $2i\pi/\log 2$, and thus they stay bounded as $s = c + ix$ when $x \to \pm\infty$. Therefore,

$$I^*(c+ix) = O\left(|x|^{b-1}e^{-\pi|x|}\right) \qquad \text{as } |x| \to \infty, \text{ for } c > 0. \tag{14}$$

This is sufficient to justify the growth estimates used in the previous section.

PROOF. We find it useful to write the Mellin transform as a loop integral, using a Hankel contour. Let

$$J(s) = \int_{\mathcal{H}} \frac{1}{Q^b(t)}(-t)^{s-1}\, dt,$$

where $\mathcal{H}$ denote a contour that goes from $+\infty - i0$, circles around 0 clockwise and returns to $+\infty + i0$.

Then by a standard argument (see [21, p. 244]), comparing the determinations in the upper and lower half planes, we find

$$J(s) = 2i\sin(\pi s)I^*(s).$$

By another standard argument, see [21] again, we can evaluate the loop integral by residues,

$$J(s) = 2i\pi \sum_{j=0}^{\infty} \operatorname{Res}\left(\frac{(-t)^{s-1}}{(Q(t))^b}\right)_{t=-2^j}.$$

Each residue provides one term in the power series expansion of the $A_k(x)$.

We now describe a procedure to calculate a residue for any fixed b. The problem is to expand the integrand till order $b-1$ around $t = -2^j$, j an integer.

First, we observe that

$$\frac{(-t)^{s-1}}{Q^b(t)} = 2^{jb}\frac{1}{(t+2^j)^b}[(-t)^{s-1}\prod_{\ell\neq j}(1+t2^{-\ell})^{-b}]. \tag{15}$$

Thus the problem reduces to evaluating the $(b-1)$st coefficient in the expansion of the quantity inside square brackets at $t = -2^j$. Let $X(t)$ represent the product

$$X(t) = \prod_{\ell \neq j} (1 + t2^{-\ell})^{-b}.$$

From Eq. (15), we expand separately around $t = -2^j$ the two quantities, $X(t)$ and $(-t)^{s-1}$. We set $t = -2^j + w$. First we have

$$(-t)^{s-1} = 2^{j(s-1)}[1 + \frac{(1-s)}{1!}\frac{w}{2^j} + \frac{(1-s)(2-s)}{2!}(\frac{w}{2^j})^2 + \cdots]. \tag{16}$$

We next proceed to expand $X(-2^j + w)$ around $w = 0$. We have

$$X(-2^j + w) = \prod_{\ell \neq j}(1 - 2^{-\ell+j})^{-b} \prod_{\ell \neq j}(1 + \frac{w}{2^\ell - 2^j})^{-b}. \tag{17}$$

Now,

$$\log(1 + \frac{w}{2^\ell - 2^j}) = \sum_{\alpha \geq 1} \frac{(-1)^{\alpha-1}}{\alpha} \frac{w^\alpha}{(2^\ell - 2^j)^\alpha}. \tag{18}$$

Set

$$Q_m = \prod_{\ell=1}^{m}(1 - 2^{-\ell}) \qquad \text{and} \qquad Q_\infty = \prod_{\ell=1}^{\infty}(1 - 2^{-\ell}).$$

From the major equations (15,16,17,18), we get at last (!)

$$\operatorname{Res}\left(\frac{(-t)^{s-1}}{(Q(t))^b}\right)_{t=-2^j} = (-1)^{jb}2^{-bj(j+1)/2}\frac{2^{j(s-1)}2^{jb}}{Q_j^b Q_\infty^b} \cdot \sum_{k=0}^{b-1} \frac{(1-s)(2-s)\cdots(k-s)}{k!2^{jk}} Y_{b-1-k}(j), \tag{19}$$

where the $Y_\beta(j)$ are defined by

$$\exp\left(b \sum_{\alpha \geq 1} \frac{(-1)^\alpha w^\alpha}{\alpha} (\sum_{\ell \neq j} \frac{1}{(2^\ell - 2^j)^\alpha})\right) = \sum_{\beta \geq 0} Y_\beta(j) w^\beta. \tag{20}$$

Finally

$$A_k(x) = \frac{(-1)^k}{k!} \cdot \frac{1}{Q_\infty^b} \sum_{j=0}^{\infty} (-1)^{jb} \frac{2^{-bj(j+1)/2}}{Q_j^b} Y_{b-1-k}(j)(x2^{b-1-k})^j, \tag{21}$$

which concludes the proof of the lemma. ∎

The technique of using loop contours for Mellin transforms gives rise to the famous Hankel representation of the Gamma function [21, p. 244]. Supplemented by a residue calculation as done here, it is one of the ways that are used to derive the functional equation for the Zeta function, following Riemann. It is also the approach used for the computation of Mellin transforms of rational functions [20].

Furthermore, the forms (21) in Lemma 5 relate to the vast domain of q-analogues of special functions—here, basic hypergeometric series with $q = 1/2$; for this, the reader can consult [9], see especially Chapter 4, on integral representations. When $b = 1$, we have $q_0 = 1$ for obvious combinatorial reasons, and we are brought to q-exponentials (or q-gamma functions [3]) and to a classical identity going back at least to Ramanujan [10, Chap. XI],

$$\begin{aligned}\int_0^\infty \frac{1}{(1+t)(1+qt)(1+q^2t)\cdots} t^{s-1}\,dt &= \frac{\pi}{\sin \pi s} \prod_{m=1}^{\infty} \frac{1}{1-q^m} \\ &\quad \cdot \sum_{j=0}^{\infty} (-1)^j \frac{q^{j(j-1)/2} q^{-j(s-1)}}{(1-q)(1-q^2)\cdots(1-q^j)} \\ &= \frac{\pi}{\sin \pi s} \prod_{m=1}^{\infty} \frac{1-q^{m-s}}{1-q^m}.\end{aligned}$$

The forms obtained become intricate as b gets large. We cite here:

Corollary 2 *When $b = 2$, we have*

$$q_0(2) = \frac{1}{\log 2}[2\log 2(2A_0'(4) + 2A_1'(4) - A_0'(2)) + A_1(4) - A_1(2)],$$

where A_0 and A_1 are q-Bessel functions,

$$\begin{cases} A_0(x) &= -\dfrac{2}{Q_\infty^2} \displaystyle\sum_{j=0}^{\infty} \frac{2^{-j(j+1)}}{Q_j^2} [\alpha - \zeta_j(1)] x^j \\ A_1(x) &= -\dfrac{1}{Q_\infty^2} \displaystyle\sum_{j=0}^{\infty} \frac{2^{-j(j+1)}}{Q_j^2} x^j, \end{cases} \tag{22}$$

with

$$\alpha = \frac{1}{1} + \frac{1}{3} + \frac{1}{7} + \frac{1}{15} + \cdots \quad \textit{and} \quad \zeta_j(1) = \frac{1}{1-2^{-1}} + \frac{1}{1-2^{-2}} + \cdots + \frac{1}{1-2^{-j}},$$

the analogues of Euler's constant and of the j-th harmonic number respectively.

From this, we find in a matter of seconds $q_0(2) = 0.57470\ 90927\ 57031\ 98404$.

PROOF. At s an integer, the representation provided by Lemma 5 has the indeterminate form $\frac{0}{0}$. We thus apply de L'Hôpital's rule, by which at an integer point $s = 1, 2, \ldots$, we get

$$I^*(s) = (-1)^s [2^s \log 2(A_0'(2^s) + (s-1)A_1'(2^s) + A_1(2^s)].$$

The proof then relies on the computation of the A_k's that was detailed in Eqns. (19,20,21). ∎

The expression of $q_0(2)$ is thus a combination of q-analogues of the Bessel functions J_0 and Y_0. For any given $b \geq 2$, the expressions obtained involve q-analogues of ${}_0F_{b-1}$ hypergeometrics.

From Lemma 5 and the periodicity of the Mellin transforms $h^*(s)$ and $I^*(s)$ that are implied, we also get:

Corollary 3 *The periodic fluctuation $S(u)$ in the coefficient f_n is expressible as a finite linear combination of the standard functions*

$$S^{\langle r\rangle}(u) = \sum_{k\in\mathbb{Z}\setminus 0} \Gamma(r-\xi_k)e^{2ik\pi u}.$$

PROOF. This follows from the form of the Fourier coefficients of $S(u)$, namely

$$\frac{1}{\log 2}\cdot\frac{h^*(\xi_k)}{\Gamma(1+\xi_k)},$$

the decomposition of $I^*(s)$ (with its factor of $\pi/\sin\pi s$ and the periodicities of 2^s) and the complement formula for the Gamma function. ∎

We conclude this section by investigating the dependency of $q_0(b)$ on b. In computer applications, b normally represents a page ("bucket") capacity, measured in the number of records that a disk page can contain. The value of b usually varies from a few tens to a few hundreds.

We prove that for large b, $q_0(b) \approx 1/(b\log 2)$. Therefore, neglecting periodic fluctuations, the number of nodes needed to store a file of size n is about

$$\frac{n}{b\log 2}.$$

Since $1/\log 2 = 1.44$, this represents a loss of about 44%, compared to a perfect packing that would require n/b pages. Alternatively, the global storage occupation behaves as though nodes were 69% full ($\log 2 = 0.69$).

Theorem 4 *As b tends to ∞, we have*

$$q_0(b) \sim \frac{1}{b\log 2}.$$

PROOF. Take the integral representation of q_0,

$$q_0 = \frac{1}{\log 2}\int_0^\infty \left(\frac{1+t}{Q(t)}\right)^b \frac{dt}{1+t},$$

and use Laplace's method to evaluate the integral for large b. Everything rests on the local expansion at 0, $(1+t)/Q(t) = \exp(-t+O(t^2))$, from which there follows that the integral is $\sim 1/b$. ∎

4 Other analyses

It should be clear from the way that our discussion was organized that any suitable additive tree parameter will also yield to these techniques. We can solve recurrences of the general form

$$f_n = c_n + \sum_{k=0}^{n-b} \pi_{n,k}(f_k + f_{n-k})$$

where c_n is the sum of a polynomial in n and of terms of the type $\sum_{j=0}^{b} \lambda_j \delta_{n,j}$ that may be adjusted to any specific initial conditions.

The resulting equation is the one corresponding to Lemma 2, with $P(z)$ being in general a rational function of the form $N(z)/(1+z)^B$ that fully characterizes the problem. This opens a new perspective on the analysis of tries by means of ordinary generating functions since we can proceed in an almost automatic fashion from a problem specification to its asymptotic estimates.

For instance, if we want to analyze the number of internal–external nodes in standard digital trees ($b = 1$), see [8], we find immediately from the problem specification that

$$P(z) = z/(1+z) \qquad \text{and} \qquad \bar{P}(t) = tP(1/t) = t/(1+t).$$

Thus (!), the proportion of internal–external nodes is of order $O(n)$ and its principal linear term is of the form $(q_0^* \cdot n)$, with

$$q_0^* = \frac{1}{\log 2} \int_0^\infty \frac{t(1+t)^{-1}}{(1+t)(1+t/2)(1+t/4)\cdots} dt.$$

This is a new integral representation for the old constant 0.37204 that was found in [8, Thm 2] and derived there under a sum form.

Similarly, if we want to count the total number of nodes in a randomly grown tree, *including* empty nodes, we get $P(z) = (1+z)^b$, and we find the number of these nodes to be approximately $(q_0^{**} \cdot n)$ on average, where

$$q_0^{**} = \frac{1}{\log 2} \int_0^\infty \frac{(1+t)^b}{Q^b(t)} dt.$$

Finally, still in the same vein, the trees under consideration are fairly well balanced since their path length, computed by these methods, is found to be $n \log_2 n + O(n)$.

Another tree model (based on order statistics rather than random bits) is discussed by Mahmoud and Pittel in [17]; their model also has internal nodes that contain b elements and external nodes that can contain from 0 to b elements. They show that, under certain conditions, the model leads to an asymptotically Gaussian distribution for the size of the tree. A similar limit law holds for the number of nodes in random tries as shown by Jacquet and Régnier [11]. In view of these facts, we expect the number of nodes in a generalized digital tree of size n to be asymptotically normally distributed for large n.

Distributions of other parameters on digital search trees are discussed by Louchard in [16], who obtained the expected "profile" (*i.e.*, the proportion of nodes at each level in the tree) of standard digital trees. Louchard's result (see also [14]) can be extended to generalized digital trees via our approach, and the profile is again non Gaussian. Precise height estimates might yield to the probabilistic techniques of Aldous and Shields [1]. Finally, it could be of some interest to investigate the variance of node levels, after Kirschenhofer and Prodinger revealed surprising connections between some of the corresponding asymptotic analysis and certain identities that belong to the elementary theory of modular forms [12].

Acknowledgement. We are indebted to Pr. Hosam M. Mahmoud for rekindling our interest in this problem.

References

[1] ALDOUS, D., AND SHIELDS, P. A diffusion limit for a class of randomly growing binary trees. *Probability Theory and Related Fields 79*, 4 (1988), 509–542.

[2] ANDREWS, G. E. *The Theory of Partitions*, vol. 2 of *Encyclopedia of Mathematics and its Applications.* Addison–Wesley, 1976.

[3] ASKEY, R. Ramanujan's extensions of the Gamma and Beta functions. *American Mathematical Monthly 87* (1980), 346–359.

[4] DOETSCH, G. *Handbuch der Laplace Transformation,* Vol. 1–3. Birkhäuser Verlag, Basel, 1955.

[5] FAGIN, R., NIEVERGELT, J., PIPPENGER, N., AND STRONG, R. Extendible hashing: A fast access method for dynamic files. *A.C.M. Transactions on Database Systems 4* (1979), 315–344.

[6] FLAJOLET, P., AND ODLYZKO, A. M. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics 3*, 2 (1990), 216–240.

[7] FLAJOLET, P., AND PRODINGER, H. Register allocation for unary–binary trees. *SIAM Journal on Computing 15*, 3 (August 1986), 629–640.

[8] FLAJOLET, P., AND SEDGEWICK, R. Digital search trees revisited. *SIAM Journal on Computing 15*, 3 (August 1986), 748–767.

[9] GASPER, G., AND RAHMAN, M. *Basic Hypergeometric Series*, vol. 35 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, 1990.

[10] HARDY, G. H. *Ramanujan: Twelve Lectures on Subjects Suggested by his Life and Work*, third ed. Chelsea Publishing Company, New-York, 1978. Reprinted and Corrected from the First Edition, Cambridge, 1940.

[11] JACQUET, P., AND RÉGNIER, M. Trie partitioning process: Limiting distributions. In *CAAP'86* (1986), P. Franchi-Zanetacchi, Ed., vol. 214 of *Lecture Notes in Computer Science*, pp. 196–210. Proceedings of the 11th Colloquium on Trees in Algebra and Programming, Nice France, March 1986.

[12] KIRSCHENHOFER, P., AND PRODINGER, H. On some applications of formulæ of Ramanujan in the analysis of algorithms, 1988. Preprint.

[13] KNUTH, D. E. *The Art of Computer Programming*, vol. 3: Sorting and Searching. Addison-Wesley, 1973.

[14] KONHEIM, A. G., AND NEWMAN, D. J. A note on growing binary trees. *Discrete Mathematics*, 4 (1973), 57–63.

[15] LARSON, P. A. Dynamic hashing. *BIT 18* (1978), 184–201.

[16] LOUCHARD, G. Exact and asymptotic distributions in digital and binary search trees. *RAIRO Theoretical Informatics and Applications 21*, 4 (1987), 479–495.

[17] MAHMOUD, H. M., AND PITTEL, B. Analysis of the space of search trees under the random insertion algorithm. *Journal of Algorithms 10* (1989), 52–75.

[18] ODLYZKO, A. M. Periodic oscillations of coefficients of power series that satisfy functional equations. *Advances in Mathematics 44* (1982), 180–205.

[19] SEDGEWICK, R. *Algorithms*, second ed. Addison–Wesley, Reading, Mass., 1988.

[20] SNEDDON, I. N. *The Use of Integral Transforms.* McGraw–Hill Inc., 1972.

[21] WHITTAKER, E. T., AND WATSON, G. N. *A Course of Modern Analysis*, fourth ed. Cambridge University Press, 1927. Reprinted 1973.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique.

ISSN 0249-6399