Normal limiting distribution for the size and the external path length of tries

Philippe Jacquet, Mireille Regnier

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France

Tél. (1) 39 63 55 11

Rapports de Recherche

N° 827

NORMAL LIMITING DISTRIBUTION FOR THE SIZE AND THE EXTERNAL PATH LENGTH OF TRIES

Philippe JACQUET
Mireille REGNIER

AVRIL 1988

RR 827

NORMAL LIMITING DISTRIBUTION FOR THE SIZE AND THE EXTERNAL PATH LENGTH OF TRIES

DISTRIBUTION LIMITE GAUSSIENNE DE LA TAILLE ET DE LA LONGUEUR DE CHEMINEMENT EXTERNE DES TRIES

Philippe Jacquet et Mireille Régnier

Abstract:

This paper studies the limiting distribution of the size and of the external path length of random tries. We consider Bernoulli and Poisson models, for uniform or biased data. We prove the convergence to the Gaussian distribution, as well as the convergence of the moments of any order.

Résumé: *Dans ce papier, nous étudions la distribution limite de la taille et de la longueur de cheminement externe des tries. Nous considérons les modèles de Bernoulli et de Poisson, pour des distributions de données uniformes ou biaisées. Pour ces deux paramètres, nous prouvons la convergence en loi vers une distribution normale (*i.e. *gaussienne), ainsi que la convergence des moments d'ordre quelconque.*

NORMAL LIMITING DISTRIBUTION FOR THE SIZE AND THE EXTERNAL PATH LENGTH OF TRIES

Philippe Jacquet and Mireille Régnier
INRIA-Rocquencourt
78 153 Le Chesnay-FRANCE

Abstract

This paper studies the limiting distribution of the size and of the external path length of random tries. We consider Bernoulli and Poisson models, for uniform or biased data. We prove the convergence to the Gaussian distribution, as well as the convergence of the moments of any order.

Keywords : performance analysis, tries, distributions, complex analysis.

I INTRODUCTION

This paper is devoted to the study of tries. Tries are a basic tree structure associated to a recursive partitioning process which appears in quite a large number of computing problems such as: fast retrieval of digital data (Dynamic Hashing Algorithms: [2], [3], [6], [12], [13], [15], [19], [22]), communication protocols (the Tree Protocol of Capetanakis and Tsybakov-Mikhailov: [1], [4], [16], [17], [18]), polynomial factorization ([10], [14]), data compression ([25]).

The trie structure appeared initially as a device for storing a collection of digital data. A binary trie is a binary tree in which the records are stored in the leaves. There is a maximum number b of records that can be stored in a single leaf. Each record is assumed to be an infinite sequence of 0 and 1. By reading this key, where 0 means "go left" and 1 means "go right", we get a path which starts from the root of the tree and reaches the leaf where the record is effectively stored. The following picture gives an example of a trie where $b = 1$.

A trie can be dynamically built by successive insertions of the items. Note that the relative order of insertions is irrelevant.

Our purpose is to determine the behaviour of the size of the trie, *i.e.* the number of internal nodes, and of the external path length, when the number of stored records goes to infinity. These

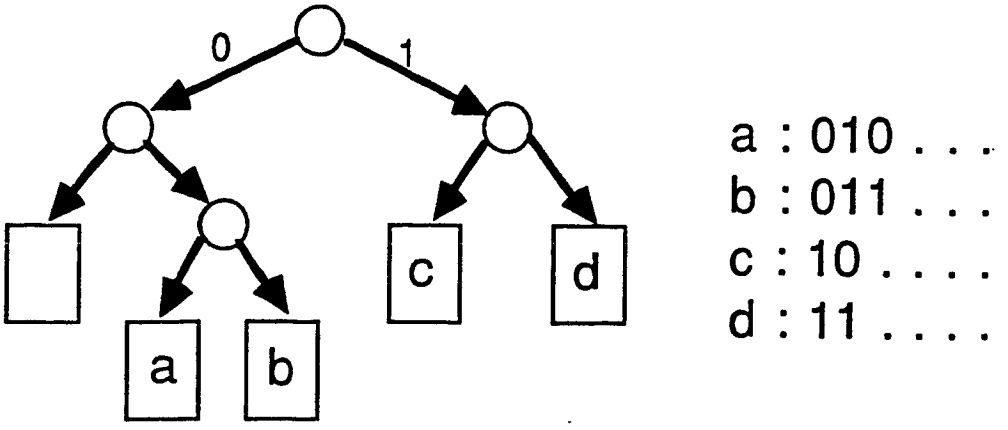

Figure 1

parameters describe the performance of the underlying probabilistic partitioning system. In the case of trie data structure of hashing schemes, the size characterizes the *storage occupation*, and the external path length the *processing time* in central unit. In the context of the tree protocol, the size of the associated trie is exactly the *length of the Collision Resolution Interval* (CRI) that separates colliding stations. The external path length is used to analyze the *waiting time* of the customers.

We assume that the records constitute random bit streams; thus the size S_n of the trie (or its external path length) is a random variable conditioned by the number n of keys. As n increases, those parameters have intricate (exact) distributions that, as we will prove, tend to limiting distributions of simple form.

The process which generates the random keys is such that

$$\begin{cases} \Pr(\text{bit} = 0) = p \\ \Pr(\text{bit} = 1) = q \end{cases}$$

with $p + q = 1$. This probabilistic model is a classical one ([11], [12], [22]). The biased case, when $p \neq q$, is relevant in numerous problems; for example, versions of the tree protocol are optimized using biased random bit streams ([4], [17]).

In this paper, we will prove the two following theorems:

THEOREM 1A: *Let S_n be the random variable representing the size of a trie which contains n random keys. As $n \to \infty$, the distribution of the random variable S_n, once centered and normalized ([5]), tends to a limiting Gaussian distribution. Moreover, the moments of any order converge to the corresponding moments. These results hold true for biased and uniform bit distributions.*

THEOREM 1B: *Let L_n be the random variable representing the external path length of a trie which contains n random keys. As $n \to \infty$, the distribution of the random variable L_n, once*

centered and normalized ([5]), tends to a limiting Gaussian distribution. These results hold true for biased and uniform bit distributions.

Let P_n^k be the probability that the size (resp. the external path length) of the trie S_n be equal to k, and introduce the bivariate generating functions

$$\begin{cases} \mathcal{P}(z,u) = \sum_{n,k} P_n^k u^k \frac{z^n}{n!} . P(z,u) \\ = \sum_{n,k} P_n^k u^k \frac{z^n}{n!} e^{-z} = \mathcal{P}(z,u) e^{-z} . \end{cases}$$

Function P is also a Poisson generating function when z is fixed: it represents the probability generating function (pgf) of S_N when N is a Poisson random variable with parameter z. Our proof will proceed in stages, as explained below for the size parameter.

1 *Recursion and functional equations:* We use the recursive nature of the tree process to set up a *non linear difference equation* satisfied by $P(z,u)$, namely (when $b = 1$)

$$P(z,u) = uP(pz,u)P(qz,u) + (1-u)(1+z)e^{-z} .$$

2 *Exponential lower and upper bounds:* The problem is to obtain good lower and upper bounds for $|P(z,u)|$ when u is fixed and z is large. The bounds will be of the form:

$$|P(z,u)| > |ce^{-\alpha z}|$$

and

$$|P(z,u)| < ce^{\alpha |z|} ,$$

with $0 < \alpha < 1$, c nonnegative and the respective domains of u and z to be made precise later.

3 *Asymptotic evaluation:* The problem is now to obtain a good asymptotic approximation for $P(z,u)$ for fixed u and large z. Setting $L(z,u) = \log P(z,u)$, L satisfies a *quasi linear difference equation.* From there we prove the estimate $L(z,u) = O(z)$ and we determine the growth of the moments $P_u(z,1)$ and $P_{uu}(z,1)$.

4 *Central Limit Theorem for the Poisson generating function:* The characteristic function $P(z, e^{-it})$ - after normalisation using mean and variance estimates from Point 3 - converges as $z \to \infty$ to the characteristic function of a normally distributed variable, namely $e^{-t^2/2}$.

5 *Limiting distribution under the Bernoulli model:* There now remains to translate the previous limit result under a Poisson model with parameter z to the case where n is fixed but large (the latter is the so-called Bernoulli model). We make use of the Cauchy formula:

$$P_n(u) = \frac{n!}{2i\pi} \oint P(z,u) e^z \frac{dz}{z^{n+1}} .$$

The integration will be done on the circle of center 0 and radius n. By this device and previous estimates, results translate from the Poisson to the Bernoulli case.

For the external path length, the scheme of the proof is similar, as we will show in Sections VII, VIII and IX.

II RECURSION AND FUNCTIONAL EQUATION

The purpose of this section is to establish the basic equations which are satisfied by the pgf of S_n. Let P_n^k be the probability that the size of the trie S_n be equal to k, and introduce the probability generating function (pgf) of the distribution of S_n

$$P_n(u) = \sum_k P_n^k u^k$$

where u is a complex variable *a priori* such that $|u| \leq 1$. We introduce also the bivariate Poisson generating function

$$P(z,u) = \sum_n P_n(u) \frac{z^n}{n!} e^{-z}$$

where z is an arbitrary complex variable. Note that when z is fixed as a real positive number, then function P represents the pgf of S_N when N is itself Poisson with parameter z.

We will show that we can enlarge the region of the complex plane where $P(z,u)$ is defined and analytic with respect to the variable u.

Proposition 2: *Let $P(z,u)$ be the generating function for the size of random tries. It is defined and analytical for every z and u, such that $|u| < 1/(p^{b+1} + q^{b+1})$ and it satisfies the functional equation:*

$$P(z,u) = uP(pz,u)P(qz,u) + (1-u)e_b(z)e^{-z} \ , \quad (1)$$

with the notation

$$e_b(z) = 1 + z + .. + \frac{z^b}{b!} \ .$$

Proof: According to the recursive definition of tries, the number I of internal nodes of a trie $\mathcal{T}$ with left subtrie $\mathcal{T}_0$ and right subtrie $\mathcal{T}_1$ satisfies:

$$I(\mathcal{T}) = I(\mathcal{T}_0) + I(\mathcal{T}_1) + 1 - \chi(|\mathcal{T}| \leq b).$$

Here, $\chi(P)$ is the characteristic function of property P (i.e. $\chi(P) = 1$ if P is true, $\chi(P) = 0$ otherwise). Applying to this equation the algebraic methods defined in [7] , one may derive "directly" (1). And from (1), one gets the recurrence:

$$P_0(u) = \cdots = P_b(u) = 1$$

$$P_n(u)\,(1 - u(p^n + q^n)) \;=\; u \sum_{0<j<n} \binom{n}{j}\, p^j P_j(u) q^{n-j} P_{n-j}(u),$$

which shows that the $P_n(u)$ are rational fractions and insures the analyticity of $P_n(u)$ for $|u| < 1/(p^{b+1} + q^{b+1})$. Moreover, chosing $|u| < 1/(p^{b+1} + q^{b+1})$ and noting $\beta = \max(|u/(1 - u(p^{b+1} + q^{b+1}))\,, 1)$, we get $|P_n(u)| < \beta^{n-1}$ by induction. The analyticity of P follows.

III EXPONENTIAL LOWER AND UPPER BOUNDS

We said previously that z and u were complex variables. It will prove convenient in order to treat the Bernoulli case, to have z vary in cones with vertex 0. Thus, we note:

$$C_\theta = \{z;\ |\arg(z)| < \theta\}$$

with θ between 0 and $\frac{\pi}{2}$. All over this paper, the notation $V(a)$ represents a neighbourhood of a.

Proposition 3: *For every $\theta \in [0,\ \frac{\pi}{2}[$ there exist $\alpha \in]0,\ 1[$, a neighbourhood $V(1)$ and positive constants c_1, c_2, such that, for $u \in V(1)$:*

$$z \in C_\theta \;\Rightarrow\; |P(z,u)| > c_1 |e^{-\alpha z}|\,,$$

$$z \notin C_\theta \;\Rightarrow\; |P(z,u)e^z| < c_2 e^{\alpha|z|}\,.$$

Proof: We arbitrarily fix α in $]0,1[$ (we shall give constraints on α only for the second part of the proposition). We proceed by induction. To start the recurrence, we choose A such that:

$$z \in C_\theta \text{ and } \mathrm{Re}(z) \geq A \;\Rightarrow\; |e_b(z)e^{-(1-\alpha)z}| < 1 \text{ and } |e^{-\alpha z}| < \frac{1}{2}\,.$$

Since $P(z,1)$ is 1 and P is uniformly continuous on compact sets, there exists $V(1)$ such that, in the domain $\mathcal{D}_0 = \{z \in C_\theta,\ A \leq \mathrm{Re}(z) \leq \frac{A}{p}\} \times V(1)$, P satisfies:

$$|P(z,u)| \geq 2|e^{-\alpha z}|.$$

Assuming $p < q$, let the $\mathcal{D}_m$ be the domains $\{z \in C_\theta,\ A \leq \mathrm{Re}(z) \leq \frac{A}{pq^m}\} \times V(1)$, m being an integer. They satisfy:

$$(z,u) \in \mathcal{D}_{m+1} - \mathcal{D}_0 \implies (pz,u) \in \mathcal{D}_m \text{ and } (qz,u) \in \mathcal{D}_m \ .$$

For (z,u) in $\mathcal{D}_0$, $|P(z,u)e^{\alpha z}| > 2$ holds. Let us suppose that the inequality holds for every (z,u) in $\mathcal{D}_m$. Assuming $|u-1| < 1$ and $|u| > \frac{3}{4}$, one gets, from (1), for (z,u) in $\mathcal{D}_{m+1} - \mathcal{D}_0$:

$$|P(z,u)| \geq \frac{3}{4}.4|e^{-\alpha(p+q)z}| - |e^{-\alpha z}| = 2|e^{-\alpha z}| \ .$$

And the first assertion follows by induction over m.

$$z \in C_\theta \ \text{ and} |z| \geq A \ \Rightarrow\ |e_b(z)e^{-(1-\alpha)|z|}| < 1 \text{ and } e^{-\alpha|z|} < \frac{1}{2} \ .$$

The proof of the second assertion proceeds similarily. It is convenient to note $S(z,u) = P(z,u)e^z$. We have the following functional equation, which is derived from (1),

$$S(z,u) = uS(pz,u)S(qz,u) + (1-u)e_b(z) \ .$$

At first, we choose α such that for every z outside C_θ

$$|e^z| \ < \ e^{\alpha|z|} \ ,$$

thus $\cos\theta \ < \ \alpha \ < \ 1$. Now, one can chose A such that:

Since $S(z,1)$ is e^z and S is uniformly continuous on compact sets, there exists $V(1)$ such that, in any domain of the form $\mathcal{F} = \{z \notin C_\theta,\ A \leq |z| \leq \frac{A}{p}\} \times V(1)$, S satisfies:

$$|S(z,u)| \leq \frac{1}{2}e^{\alpha|z|} \ .$$

We complete the proof with the same type of induction as for the first part of the proposition.

IV ASYMPTOTIC ESTIMATES FOR THE POISSON MODEL

This section provides various estimates on $P(z,u)$ and its logarithms, and concludes by a determination of the mean and variance of the size of tries under the Poisson model.

Proposition 4: *For any cone C_θ, $0 < \theta < \frac{\pi}{2}$, there exists $V(1)$ such that $L(z,u) = \log P(z,u)$ be defined and analytic when $(z,u) \in C_\theta \times V(1)$. Also, in that domain, there exists a constant $B > 0$ such that*

$$|L(z,u)| < B.|z|,$$

uniformly in u.

Proof: The existence and analyticity follow from Proposition 3, as P is analytic and non zero. Let us set, for $(z,u) \in \mathcal{C}_\theta \times V(1)$:

$$g(z,u) = \log\Big(1 - \frac{(1-u)e_b(z)e^{-z}}{P(z,u)}\Big) \ , \quad \varphi(z,u) = \frac{L(z,u) - \log(u)}{z} \ .$$

We get a "linear" form for (1):

$$\varphi(z,u) = p\varphi(pz,u) + q\varphi(qz,u) - \frac{g(z,u)}{z} \ , \quad (2)$$

with $|g(z,u)/z| < C|e^{\alpha z}e^{-z}|$. Let now :

$$\mathcal{D}_n = \ \{(z,u) \ ; \ u \in V(1), \ z \in \mathcal{C}_\theta, \ \mathrm{Re}(z) < \frac{1}{q^n}\}$$

be an increasing sequence of truncated cones, and B_n be:

$$B_n = \sup_{z \in \mathcal{D}_n} |\varphi(z,u)|.$$

As the maximum on $\mathcal{D}_n$ is obtained either for z in $\mathcal{D}_{n-1}$ or for z satisfying the equation (2), with pz and qz in $\mathcal{D}_{n-1}$, we get:

$$B_n \leq B_{n-1} + C \ e^{-(1-\alpha)q^{-n}} \ \leq B_0 + C \sum_n e^{-(1-\alpha)q^{-n}} \ \leq B.$$

When z is real positive, $P(z,u)$ is the pgf of the size of trie when the number of inserted records is itself Poisson of parameter z (this is the so-called Poisson model). The mean $X(z)$ and variance $v(z)$ of this pgf are defined as:

$$\begin{cases} X(z) = P_u(z,1) \\ v(z) = P_{uu}(z,1) + P_u(z,1) - \Big(\dot{P}_u(z,1)\Big)^2 \ . \end{cases}$$

By extension, we will name these last expressions "mean" and "variance", even when z is not a real positive number.

Corollary 5: *For z varying in a cone $\mathcal{C}_\theta$ with $\theta \in]0, \frac{\pi}{2}[$, the mean and variance are $O(z)$ when $z \to \infty$, and*

$$L(z,e^{it}) = iX(z)t - \frac{v(z)}{2}t^2 + O(zt^3) \ ,$$

for $z \in \mathcal{C}_\theta$ and t in a neighbourhood of 0.

Proof: L is analytic with respect to $u = e^{it}$ and thus with respect to t, when t is in a neighbourhood of 0. Computing its first derivatives, we get, for u in $V(1)$ or t in $v0$:

$$L(z,u) = iX(z)t - \frac{v(z)}{2}t^2 + g_z(t).t^3,$$

where $g_z(t)$ is analytic. Applying the Cauchy formula on a contour C included in $V(0)$ and encircling 0:

$$X(z) = \frac{1}{2i\pi}\oint_C \frac{L(z,e^{i\omega})}{\omega^2}d\omega$$

$$\frac{v(z)}{2} = \frac{1}{2i\pi}\oint_C \frac{L(z,e^{i\omega})}{\omega^3}d\omega$$

$$g_z(t) = \frac{1}{2i\pi}\oint_C \frac{L(z,e^{i\omega})}{\omega^3(\omega - t)}d\omega\ ,$$

where $\frac{1}{\omega - t}$ is upper bounded on the contour (we can redefine $V(0)$ as interior to C). Thus $X(z)$, $v(z)$ are functions that are $O(z)$. Moreover, $g_z(t)$ is also uniformly $O(z)$ within t.

Remark: We proved, in passing, that the cumulants (the coefficients of the expansion of $L(z,e^{it})$ with respect to t) are all $O(z)$.

To prove the convergence to the normal distribution, we need more precise estimate on the growth of $X(z)$ and $v(z)$.

Lemma 6: *The mean and variance $X(z)$ and $v(z)$ under the Poisson model satisfy the equations:*

$$\begin{cases} X(z) = X(pz) + X(qz) + 1 - e_b(z)e^{-z} \\ v(z) = v(pz) + v(qz) + (2X(z) - 1 + e_b(z)e^{-z})e_b(z)e^{-z}\ . \end{cases}$$

Asymptotically, we have:

$$\begin{cases} X(z) \sim z\ Q_1(z) \\ v(z) \sim z\ Q_2(z) \end{cases}$$

where Q_1 and Q_2 are upper and lower bounded by strictly positive constants. When $p = q = \frac{1}{2}$, Q_1 and Q_2 are also periodic in $\log_2(z)$.

Proof: The equations are derived from the definitions and (1). The asymptotic values are derived by the methods (Mellin transform [8]) developed in [8, 12, 20, 22]. An extensive analysis, with numerical computations of the variance, is given in [23] .

V CENTRAL LIMIT THEOREM FOR THE POISSON MODEL

Theorem 7: *The distribution of the size of the tries in the Poisson model, once centered and normalized, converges to the normal distribution. moreover, the moments of any order of the centered and normalized distribution converge to the corresponding moments of the normal distribution.*

Proof: Let $\sigma(z)$ be the standard deviation $\sqrt{v(z)}$. According to the previous propositions, we have the estimate

$$P(z, e^{it/\sigma(z)})e^{-itX(z)} = exp\{-\frac{t^2}{2} + O\Big(\frac{z.t^3}{\sigma^3(z)}\Big)\} .$$

Since $|v(z)| > A.|z|$, we have: $\frac{z}{\sigma(z)^3} < \frac{B}{\sqrt{|z|}}$. Thus:

$$P(z, e^{it/\sigma(z)})e^{-itX(z)} \to e^{-\frac{t^2}{2}} ,$$

when $z \to \infty$, uniformly in any neighbourhood of $t = 0$ (since asymptotically we have $e^{it/\sigma(z)} \in V(1)$).

VI. BACK TO THE BERNOULLI CASE

Now, we are able to finish the proof of our main Theorem 1 about the normal limiting distribution of the size of tries under the Bernoulli model. We make use of the Cauchy formula:

$$P_n(u) = \frac{n!}{2i\pi} \oint P(z,u)e^z \frac{dz}{z^{n+1}} ,$$

the integration being done along the circle $|z| = n$ as illustrated by figure number 2. We first evaluate the integral, using the asymptotic development of $P(z,u)$ derived in the preceding sections. The Bernoulli mean and variance follow from this computation, and their asymptotic estimates are given in Lemma 8. Hence, the limiting distribution result will follow.

We first proceed with the evaluation of:

$$e^{-\frac{itX(n)}{\sqrt{V(n)}}} P_n(e^{\frac{it}{\sqrt{V(n)}}}) .$$

Using the Stirling formula and noting: $z = ne^{i\theta}$, this is: $\sqrt{\frac{n}{2\pi}}\ I_n\ (1 + O(\frac{1}{n}))$ with:

$$I_n = e^{-n}\ e^{-\frac{itX(n)}{\sqrt{V(n)}}} \int_{-\pi}^{\pi} P(ne^{i\theta}, e^{\frac{-it}{\sqrt{V(n)}}})e^{ne^{i\theta}} e^{-in\theta} d\theta .$$

This is an integration on the contour in Figure 2.

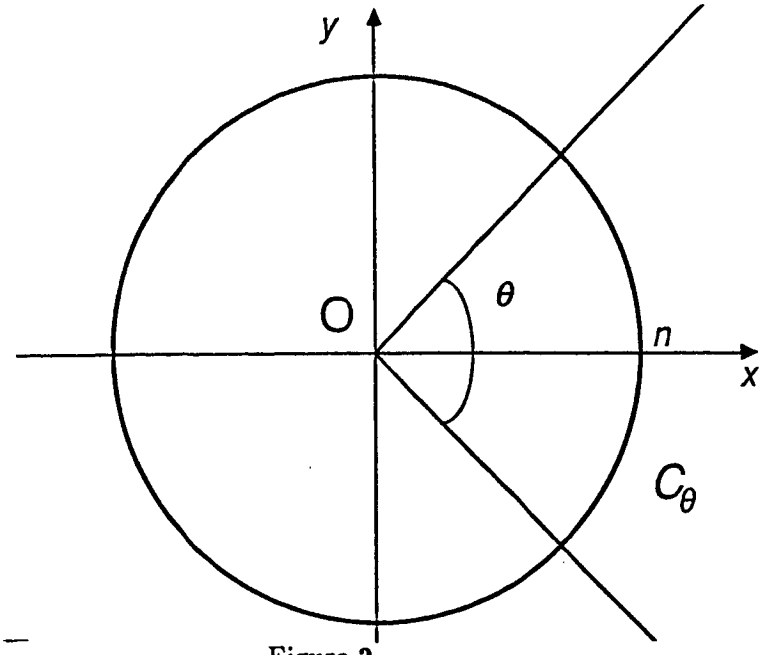

Figure 2

We split this integral into two parts, inside and outside the cone C_Θ. First, for $0 < |\theta| < \theta_{max}$, we use the developments of $e^{i\theta}, X(ne^{i\theta})$, and $V(ne^{i\theta})$, valid on the cone $C_{\theta_{max}}$, namely:

$$\begin{cases} X(ne^{i\theta}) = X(n) + nX'(n)i\theta + O(n\theta^2) \\ V(ne^{i\theta}) = V(n) + O(n\theta) . \end{cases}$$

The change of variable $y = \sqrt{n}\theta$ yields the integrand:

$$e^{-\frac{1}{2}(y-\frac{\sqrt{n}X'(n)t}{sqrtV(n)})^2} e^{-\frac{t^2}{2}\frac{V(n)-nX'^2(n)}{V(n)} + O(\frac{y+y^2}{\sqrt{n}})} .$$

When $n \to \infty$, the bounds of integration also tend to infinity, and the dominated convergence theorem readily gives:

$$\int_{|\theta|<\theta_{max}} = e^{-\frac{t^2}{2}\frac{V(n)-nX'^2(n)}{V(n)}} \sqrt{\frac{2\pi}{n}} (1 + O(\frac{1}{\sqrt{n}})).$$

Second, we prove that the contribution for $|\theta| > \theta_{max}$ is negligible. From Proposition 3, we have:

$$|P(z,u)e^z| < e^{n\alpha} = O(\frac{e^n}{n}) .$$

Hence:

$$e^{-\frac{itX(n)}{\sqrt{V(n)}}} P_n(e^{\frac{it}{\sqrt{V(n)}}}) = e^{-\frac{t^2}{2}\frac{V(n)-nX'^2(n)}{V(n)}} (1 + O(\frac{1}{\sqrt{n}})) .$$

This computation gives in passing the asymptotic expressions of the Bernoulli mean and variance, for which we can state asymptotic estimates:

Lemma 8: *The mean $X_B(n)$ and the variance $V_B(n)$ under the Bernoulli model satisfy:*

$$\begin{cases} X_B(n) \sim X(n) \\ V_B(n) \sim V(n) - nX'^2(n) \\ \alpha n \leq V_B(n) \leq \beta n \ . \end{cases}$$

for some positive reals α and β.

Proof: The upper bound follows from Lemmas 5 and 6. To get the lower bound, one derives $(E1)$ and gets:

$$Q(z) = V(z) - zX'^2(z) = Q(pz) + Q(qz) + \phi(z)$$

where ϕ can be shown to be strictly positive. Thus $\phi^*(0)$ is strictly positive, and it follows from basic properties of Mellin transform that Q is at least linear.

We can turn to the main Theorem 1A. As a straightforward consequence of Lemma 8, $\frac{|t|}{\sqrt{V_B(n)}}$ is bounded iff $|t|$ is bounded. Thus, we can rewrite our estimate of $P_n(u)$ as:

$$e^{\frac{-itX(n)}{\sqrt{V_B(n)}}} P_n(e^{\frac{-it}{\sqrt{V_B(n)}}}) = e^{-\frac{t^2}{2}} (1 + O(\frac{1}{\sqrt{n}})) \ .$$

Applying the Levy's Continuity Theorem yields Theorem 1.

VII. THE EXTERNAL PATH LENGTH: THE GENERATING FUNCTION:

We turn now, in this section and the following one, to the study of the external path length. The proof scheme is roughly the same as before.

Proposition 9: *Let $P(z,u)$ be the generating function for the external path length of random tries, and let $\mathcal{P}(z,u) = P(z,u)e^z$. $\mathcal{P}(z,u)$ satisfies the functional equation:*

$$\mathcal{P}(z,u) = \mathcal{P}(puz,u)\mathcal{P}(quz,u) + e_b(z) - e_b(uz).$$

Moreover, $\mathcal{P}(z,u)$ and $P(z,u)$ are defined and analytical for every z and for $|u| < (p^{b+1} + q^{b+1})^{1/b+1}$, ∎

Remark: This appearance of a new connection between u and z, in the terms $\mathcal{P}(puz,u)$ and $\mathcal{P}(quz,u)$ implies the differences between the following proof and the preceding one. Unfortunately, the cone C_θ is not stable by the transformations: $z \to puz$ and $z \to quz$, when u is complex. Thus, to be able to exploit the functional equation, we restrict ourself to u ranging in a real interval. Then, $P_n(u)$ is the *Laplace transform* of the distribution.

Proof: With the notations of II, $\mathcal{L}$ being the external path length of a trie $\mathcal{T}$, one has:

$$\mathcal{L}(\mathcal{T}) = \mathcal{L}(\mathcal{T}_0) + \mathcal{L}(\mathcal{T}_1) + |\mathcal{T}| - \sum_{j=1}^{b} j\chi(|\mathcal{T}| = j).$$

Again, one gets "directly" (1') with the algebraic methods defined in [FRS85b]. To get the analyticity, one makes use of the following majoration.

Majoration Lemma: *For any* $R < \frac{1}{(p^{b+1}+q^{b+1})^{1/b+1}}$, *there exist two constants* C *and* $\beta > 1$, *such that:*

$$|\frac{P_n(u)}{n!}| < C.\beta^{-n\log n} \ , |u| < R.$$

Proof: A is chosen such that: $\frac{q}{(p^{b+1}+q^{b+1})^{1/b+1}} < A < 1$. From the condition:

$$|u| < R < \frac{1}{(p^{b+1}+q^{b+1})^{1/b+1}} < 1/q$$

and from the inequality: $(p^n+q^n)^{1/n} < (p^{b+1}+q^{b+1})^{1/b+1} < 1$, one gets:

$$\exists n_0 : n > n_0 \Longrightarrow \frac{(uq)^n}{1-(u(p^n+q^n)^{1/n})^n} < \frac{A^{n+1}}{n}.$$

Now, we fix $\gamma > 1$ and one can choose $C > 1$ such that: $\sup_{\substack{n\le n_0 \\ |u|<R}} |\frac{P_n(u)\gamma^{n\log n}}{n!}| < C$. At last, we choose $1 < \beta < \gamma$ such that: $AC(A\beta^{\log 2})^n < 1$. We have then: $|\frac{P_n(u)}{n!}| < C\gamma^{-n\log n} < C\beta^{-n\log n}$. Then, we can prove the result by induction, for $n > n_0$:

$$|\frac{P_n(u)}{n!}| < \frac{(uq)^n}{1-u^n(p^n+q^n)} \sum_{n1+n2<n} \frac{P_{n1}(u)}{n1!}\frac{P_{n2}(u)}{n2!}$$

As:$n1\log n1 + n2\log n2 > n(\log n - \log 2)$, one has:

$$\frac{|P_n(u)|}{n!} < \frac{A^{n+1}}{n}.n.C^2.\beta^{-n(\log n-\log 2)} < C.\beta^{n\log n}.$$

Proposition 10: *For every* $\theta \in [0, \frac{\pi}{2}[$, *there exists a constant:* $0 \le m_1 \le 1$, *a real neighbourhood of* 1, $V(1)$, *and positive constants* c_1, c_2, *such that, for* $u \in V(1)$*:*

$$\begin{cases} z \in C_\theta \Rightarrow |\mathcal{P}(z,u)| > c_1 e^{|z|^{m_1}} \\ z \notin C_\theta \Rightarrow |\mathcal{P}(z,u)| < c_2 e^{|z|^{1+\sqrt{|1-u|}}}. \end{cases}$$

Proof: m_1 is arbitrarily fixed, $m_1 < 1$, and we proceed by induction. We start the recurrence with A such that:

$$z \in C_\theta \text{ and } \Re(z) \ge A \Rightarrow \begin{cases} e^{|z|cos\theta} > 2e^{|z|^{m_1}} \\ e_b(|z|) < e^{|z|^{m_1}} \end{cases}$$

With $\mathcal{D}_0$ as before, noticing $|\mathcal{P}(z,1)| = |e^z| = e^{|z|cos\theta}$, a similar argument of compacity yields:

$$|\mathcal{P}(z,u)| > 2e^{|z|^{m_1}} \ .$$

By induction, this equality holds on the increasing domains $\mathcal{D}_m$, as:

$$u \in V(1) \Rightarrow \begin{cases} (pu)^{m_1} + (qu)^{m_1} > 1 \ \ , u \in V(1) \\ 0 < |e_b(z) - e_b(uz)| < 2e^{|z|^{m_1}}. \end{cases}$$

Finally, for z in the compact set $\{|z| < A, Arg(z) < \theta\}$, the continuity of these functions ensures the existence of c_1 such that:

$$|\mathcal{P}(z,u)| > c_1 e^{|z|^{m_1}}.$$

To derive of the upper bound, we note $v = |1-u|$ and we choose A and $V(1)$ such that:

$$\begin{cases} 1/2e^{\alpha|z|^{1+\sqrt{v}}} > 1/2e^{\alpha|z|} > 1 & ,|z| > A, u \in V(1) \\ |e_b(z) - e_b(uz)| < 1/4e^{\alpha|z|} < 1/4e^{\alpha|z|^{1+\sqrt{v}}} & ,|z| > A, u \in V(1) \\ |\mathcal{P}(z,u)| < 1/2e^{\alpha|z|^{1+\sqrt{v}}} & , z \in \mathcal{D}_0 \end{cases}$$

The majoration is proved by induction in the increasing domains $\mathcal{D}_m$ as:

$$\begin{cases} |\mathcal{P}(z,u)| < 1/4e^{\alpha(p^{1+\sqrt{v}}+q^{1+\sqrt{v}})u^{1+\sqrt{v}}|z|^{1+\sqrt{v}}} + 1/4e^{\alpha|z|^{1+\sqrt{v}}} \\ (p^{1+\sqrt{v}} + q^{1+\sqrt{v}})u^{1+\sqrt{v}} = (1 - H\sqrt{v})(1 + O(v)) < 1 \end{cases}$$

VIII. ASYMPTOTIC ESTIMATES FOR THE EXTERNAL PATH LENGTH UNDER THE POISSON MODEL:

This section is devoted to an asymptotic development of the logarithm of the generating function $P(z,u)$. We also derive asymptotics for the mean and the variance of the external path length under the Poisson model.

Proposition 11: *For any cone $C_\theta, 0 < \theta < \frac{\pi}{2}$, there exists a neighbourhood of 1, $V(1)$, such that $L(z,u) = logP(z,u)$ be defined and analytic when $(z,u) \in C_\theta \times V(1)$.*

Proof: This follows immediately from Proposition 10, as $P(z,u) = \mathcal{P}(z,u)z$ is analytic and non zero.

Lemma 12: *Let $X(z)$ and $V(z)$ be the mean and the variance of the external path length under the Poisson model. They satisfy asymptotically:*

$$X(z) = \begin{cases} z\dfrac{\log z}{H} - \dfrac{1}{H}(\gamma + 1 + 1/2H2/H) + o(z), p \neq q \\ \dfrac{z\log z}{\log 2} - \dfrac{1}{\log 2}(\gamma + 1 + 1/2\log 2)z + zP(\{\log_2(z)\}) + O(1), p = q = 1/2 \end{cases}$$

and

$$V(z) = \begin{cases} \dfrac{z\log^2 z}{H^2} - (\dfrac{2H2}{H^3} + \dfrac{2(\gamma+1)}{H}) - \dfrac{1}{H}z\log z + o(z), p \neq q \\ \dfrac{z\log^2 z}{\log^2 2} - \dfrac{2}{\log 2}(\gamma + 1 + 1/2\log 2)z\log z + 2zP(\{\log_2 z\})\log z + O(z), p = q = 1/2 \end{cases}$$

with:$H = p\log\frac{1}{p} + q\log\frac{1}{q}, H2 = p\log^2 p + q\log^2 q$, and $P(u)$ a periodic function with mean 0.

Proof: The proof is given in Appendix. By differentiation of (1'), one gets functional equations for $X(z)$ and $V(z)$. Using Mellin transform yields the asymptotic expansions.

Proposition 12: *For z varying in a cone C_θ with $\theta \in]0, \frac{\pi}{2}[$ and $0 < -t < \frac{A}{|z|^{1/2}\log|z|}$, we have:*

$$L(z, e^t) = \frac{z \log z}{H} t + \frac{z \log^2 z}{H^2} t^2/2 + O(\frac{1}{logz}) .$$

Proof: The proof makes a large use of the Mellin transform method. We note:

$$\begin{cases} g(z,u) &= -\log(1 - \dfrac{e_b(z) - e_b(uz)}{\mathcal{P}(z,u)}) \\ \mathcal{L}(z,u) &= \log \mathcal{P}(z,u) \ = z + L(z,u). \end{cases}$$

From (1'), the function $L(z,u)$ satisfies the functional equation:

$$L(z,u) \ = \ L(puz,u) + L(quz,u) + g(z,u) - (1-u)z.$$

If we note:

$$\begin{cases} L_u^*(s) &= \displaystyle\int_0^\infty L(z,u) z^{s-1} dz \\ g_u^*(s) &= \displaystyle\int_0^\infty (g(z,u) - (1-u)z) z^{s-1} dz \end{cases}$$

the Mellin transforms of $L(z,u)$ and $g(z,u)$ when they exist, we get formally the functional equation:

$$L_u^*(s) \ = \ \frac{g_u^*(s)}{1 - (pu)^{-s} - (qu)^{-s}} .$$

As $g(z,u)$ is $O(z)$ when $z \to 0$ and $O(z^{-m})$ when $z \to \infty$, $g_u^*(s)$ is defined for $\Re(s) \in]-2,-1[$. The pole at $s = -1$ is simple, with residu $(1-u)$, thus $g_u^*(s)$ can be continued for $\Re(s) \in]-2, +\infty[$. The pole at $s = -1$ contributes by: $-z$. The smallest root of the equation:

$$(pe^t)^{-s} + (qe^t)^{-s} \ = \ 1$$

is, for t in a neighbourhood of 0:

$$s(t) \ = \ -1 - t/H + \beta t^2 + O(t^3).$$

with: $\beta = -1/2 \frac{2H + pq(\log p - \log q)^2}{H^3}$. Thus, $L_u^*(s)$ is defined in the strip $]-1, s(t)]$ and:

$$\mathcal{L}(z, e^t) \ \sim \ g_u^*(s(t)) z^{-s(t)} - z.$$

When $|t| < \frac{A}{|z|^{1/2}\log|z|}$, one can develop $z^{-s(t)}$ around 0:

$$z^{-s(t)} \ = \ z(1 + \frac{\log z}{H} t + 1/2(\beta \log z + \frac{\log^2 z}{H^2}) t^2 + O(z^{-1/2}).$$

Combining with: $g_u^*(s(t)) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \ldots$ yields $L(z, e^{-t})$. From the definition of the cumulants and the results in Lemma 1, "α_0, α_1 and α_2 are obtained by identification.

IX. BACK TO THE BERNOULLI CASE:

We can now finish the proof of our Theorem 1. To get the limiting distribution under the Bernoulli model, we proceed as in VI, making use of the Cauchy formula:

$$P_n(u) = \frac{n!}{2i\pi} \int \mathcal{P}(z,u) e^z \frac{dz}{z^{n+1}}.$$

The Bernoulli mean and variance will follow from this computation, as stated in Lemma 13 (see also [KP87]). Surprisingly, the Bernoulli asymptotic order of the variance is strictly smaller than the Poisson asymptotic order, i.e. $O(n)$ in the uniform case and $O(\log n)$ in the biased case, instead of $O(n \log^2 n)$.

We split the integral into two parts. We note $z = ne^{i\theta}$, and using the Stirling formula, the integrand becomes: $\sqrt{\frac{n}{2\pi}} e^{\phi(n,\theta)}$ with: $\phi(n,\theta) = L(ne^{i\theta}, e^t) + n(e^{i\theta} - 1) - in\theta$. As in Section VI, one can develop $X(ne^{i\theta})$ and $V(ne^{i\theta})$ for θ around 0, which yields:

$$\phi(n,\theta) = X(n)t + \frac{t^2}{2}[V(n) - nX'^2(n)] - 1/2n(\theta - X'(n)t)^2 + \psi(n,\theta).$$

with:$\psi(n,\theta) = -nX'(n)\theta^2 t/2 + O(n\theta^2 t + n\log n\theta^3 t + n\log n t^2)$. This gives, in passing, the formulae for the Bernoulli mean and variance. One can state an equivalent to Lemma 8.

Lemma 13: *The mean $X_B(n)$ and the variance $V_B(n)$ under the Bernoulli model satisfy:*

$$\begin{cases} X_B(n) \sim X(n) \\ V_B(n) \sim V(n) - nX'^2(n) \\ \quad \alpha n \leq V_B(n) \leq \beta n \ , p = q = 1/2 \\ \alpha n \log n \leq V_B(n) \leq \beta n \log n \ , p \neq q \end{cases}$$

for some positive reals α and β.

Normalizing t as $t/\sqrt{n}$ (or $t/\sqrt{n \log n}$) yields by a dominated convergence theorem:

$$\sqrt{\frac{n}{2\pi}} \int_{-\theta_{max}}^{\theta_{max}} e^{\phi(n,\theta)} d\theta = e^{tX(n)/\sqrt{V_B(n)} - t^2/2} . (1 + O(\frac{1}{\log n})).$$

Applying the Levy Continuity Theorem yields Theorem 1B.

IX CONCLUSION

We have considered the distribution of the size of the tries , and of the external path length, in all cases, -the uniform and biased cases, under the Poisson and Bernoulli model-. We proved the convergence to the normal distribution. The method makes a systematic use of generating functions and complex analysis, which prove to be a valuable tool. Our approach enables us to asymptotically solve non linear bivariate difference equations. Another nice feature is the derivation "in passing" of the growth of the moments of any order. Thus, it generalizes and completes previous results for mean and variance. In particular, there comes out a rather surprising difference between asymptotics of Poisson and Bernoulli variances.

One expects this scheme to have other applications. Our results can be extended to *Markov* models. Markov models are adequate when the keys come from textual data: the transition matrix is then formed with the transition probabilities from one letter to another [24]. Our analysis is a particular case of a stationary process over a binary alphabet.

This work can be put in parallel with our results on other parameters on tries. In [11], we studied the height of tries and the depth of insertion of a random record (also independently analyzed in [21]), with a different method. We extensively used the Mellin transform ([9]) and showed the limiting distributions to be either gaussian or periodic doubly exponential. .

References:

[1] J. CAPETANAKIS "The Multiple Access Broadcast Channel:Protocol and Capacity Considerations", *PhD-Thesis,MIT* , (1977).

[2] PH. FLAJOLET AND C. PUECH "Tree Structure for Partial Match Retrieval" in *Proc. 24-th I.E.E.E. Symp. on FOCS* (1983) and in *JACM* **33,9** (1986) 371–407.

[3] R. FAGIN, J. NIEVERGELT, N. PIPPENGER AND H.R. STRONG "Extendible Hashing:A Fast Access Method for Dynamic Files" in *ACM TODS* **4,3** (1979) 315–344 .

[4] G. FAYOLLE, PH. FLAJOLET, M. HOFRI AND PH. JACQUET "Analysis of a Stack Algorithm for Random Multiple-Access Communication", *IEEE Trans. on Information Theory* **IT-31,2**, (1985), 244-254.

[5] W. FELLER *An Introduction to Probability Theory and its Applications, Vol. II*, Wiley,Third edition-1971,(1957).

[6] PH. FLAJOLET. "On the Performance Evaluation of Extendible Hashing and Trie Searching," *Acta Informatica* **20**, (1983), 345–369.

[7] PH. FLAJOLET, M. RÉGNIER AND D. SOTTEAU "Algebraic Methods for Trie Statistics" in *Annals of Discrete Mathematics* **25** (1985) 145–188.

[8] PH. FLAJOLET, M. RÉGNIER AND R. SEDGEWICK *"Some Uses of the Mellin Transform Techniques in the Analysis of Algorithms"* in Combinatorial Algorithms on Words, Springer NATO ASI SEr. F12, (1985) 241–254.

[10] PH. FLAJOLET AND J.M. STEYAERT "A Branching Process Arising in Dynamic Hashing, Trie Searching and Polynomial Factorization", in *Proc. ICALP 82, Lecture Notes in Computer Science*, **140** (1982) 239-251.

[11] PH. JACQUET AND M. RÉGNIER "Trie Partitioning Process:Limiting Distributions" in *Proc. CAAP'86*, Lecture Notes in Computer Science **214** (1986) 194–210.

[12] D. KNUTH, *The Art of Computer Programming. Vol 3:Sorting and Searching*, Addison-Wesley,Reading,Mass.,(1973).

[13] P.A. LARSON "Dynamic Hashing" in *BIT* **18** (1978) 184–201.

[14] D. LAZARD "On Polynomial Factorization," in *Proc. EUROCAM 82, Lecture Notes in Computer Science*, Springer-Verlag, Marseille (1982).

[15] W. LITWIN "Trie Hashing" in *Proc. ACM-SIGMOD Conf. on MOD*, Ann Arbor, Mich. (1981).

[16] J.L. MASSEY "Collision-Resolution Algorithms and Random-Access Communications", in *Multi-User Communication Systems*, .Longo(ed), CISM Courses and Lectures (1981).

[17] P. MATHYS AND PH. FLAJOLET "Q-ary Collision-Resolution Algorithms with Free Access", in *IEEE Trans. on Information Theory* **IT-31,2** (1985), 217-243.

[18] V. MIKHAILOV AND B. TSYBACHOV "Free Synchronous Packet Access in a Broadcast Channel with Feedback", in *Problemy Peredachi Informatsii* **14** (1978) 32-59.

[19] J.NIEVERGELT, H HINTERBERGER AND K.C. SEVCIK "The Grid-File:an Adaptable Symmetric Multi-Key File Structure" in *ACM TODS* **9,1**, (1984).

[20] N. E. NÖRLUND. *Vorlesungen über Differenzenrechnung*, Chelsea Publishing Company, New York (1954).

[21] B. PITTEL "Paths in a Random Digital Tree: Limiting Distributions" in *Adv. Appl. Prob.* 18 (1986), 139–155.

[22] M. RÉGNIER "Evaluation des performances du hachage dynamique," Thèse de 3ème cycle, Univ de Paris-Sud, (1983).

[23] M. RÉGNIER AND PH. JACQUET "New results on the Size of Tries", *to appear in IEEE Trans. on Information Theory* (1987).

[24] M. RÉGNIER "Trie Hashing Analysis", in *Proc. 4-th Int. Conf. on Data Engin.*, Los Angeles, USA (1988).

[25] R. SEDGEWICK, *Algorithms, Ch. 22*, Addison-Wesley, (1983).

APPENDIX

This appendix is devoted to the derivation of asymptotic expansions of the mean and the variance of the external path length, under the Poisson model. More precisely, we want to prove the Lemma 12. We make use of the Mellin transform [8]. The mean and variance can be defined by the derivatives of the bivariate generating function. Derivating the functional equation satisfied by $\mathcal{P}(z,u)$ yields the functional equations satisfied by $X(z)$ and $V(z)$:

$$\begin{cases} X(z) = X(pz) + X(qz) + z(1 - f_{b-1}(z)) \\ V(z) = V(pz) + V(qz) + 2[pzX'(pz) + qzX'(qz)] + z(1 - f_{b-1}(z)) \\ \qquad\qquad + 2zf_{b-1}(z)X(z) + z^2[f_{b-1}(z)^2 + \dfrac{z^{b-1}}{(b-1)!}e^{-z}] \end{cases},$$

with: $f_{b-1}(z) = (1 + z/1! + \cdots + z^b/b!)e^{-z}$.

From the first equation follows easily a functional equation satisfied by $X'(z)$ or $Q(z) = zX'(z)$. Then, the Mellin transform of $X(z)$, $Q(z)$ and $V(z)$ are:

$$\begin{cases} X^*(s) = -1/(s+1).\dfrac{\Gamma(s+b+1)}{(b-1)!}.\dfrac{1}{1-(p^{-s}+q^{-s})} \\ Q^*(s) = \dfrac{s\Gamma(s+1)}{1-(p^{-s}+q^{-s})} \\ V^*(s)\dfrac{= 2Q^*(s)(p^{-s}+q^{-s}) + g_1^*(s) + g_2^*(s)}{1-(p^{-s}+q^{-s})} \end{cases}$$

with: $g_1(z) = z(1 - f_{b-1}(z), g_1^*(s) = -1/s.\frac{\Gamma(s+b+2)}{b!}$ and $g_2(z) = 2zf_{b-1}(z)X(z) + z^2[f_{b-1}(z)^2 + \frac{z^{b-1}}{(b-1)!e^{-z}}]$. The function $X^*(s)$ has a double pole at $s = -1$. Computing the residues yields the terms $\alpha z \log z + \beta z$ in the asymptotic expansion. In the uniform case, the poles $s_l = -1 + \frac{2il\pi}{\log 2}$ contribute by a periodic term $zP(\{\log_2(z)\})$ to the mean.

V^* is the sum of three terms. Considering the order of $g_2(z)$ at 0 and ∞ ($O(z^2)$ and $O(z^{-m})$), one sees that g_2 is analytic in $]-2, +\infty[$ and notably around $s = -1$. Thus, $\frac{g_2^*(s)}{1-(p^{-s}+q^{-s})}$ contributes by a linear term αz while $\frac{Q^*(s)}{1-(p^{-s}+q^{-s})}$ (resp. $\frac{g_1^*(s)}{1-(p^{-s}+q^{-s})}$) have a triple (resp. a double) pole at $s = -1$ and thus contributes by a term $\alpha z \log^2 z + \beta z \log z + \gamma z$ (resp. $\alpha z \log z + \beta z$). Computing the residues gives the coefficients as stated in Lemma 12.

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