Average profiles, from tries to suffix-trees

Pierre Nicodème

LIX École polytechnique, 91128 Palaiseau, France, `nicodeme@lix.polytechnique.fr`
and INRIA, Algorithms Project, 78153, Rocquencourt Le Chesnay, France

We build upon previous work of Fayolle (2004) and Park and Szpankowski (2005) to study asymptotically the average internal profile of tries and of suffix-trees. The binary keys and the strings are built from a Bernoulli source (p, q). We consider the average number $p_{k,\mathcal{P}}(\nu)$ of internal nodes at depth k of a trie whose number of input keys follows a Poisson law of parameter ν. The Mellin transform of the corresponding bivariate generating function has a major singularity at the origin, which implies a phase reversal for the saturation rate $p_{k,\mathcal{P}}(\nu)/2^k$ as k reaches the value $2\log(\nu)/(\log(1/p) + \log(1/q))$. We prove that the asymptotic average profiles of random tries and suffix-trees are mostly similar, up to second order terms, a fact that has been experimentally observed in Nicodème (2003); the proof follows from comparisons to the profile of tries in the Poisson model.

Keywords: tries, suffix-trees, profile, asymptotics, Mellin transform, saddle-point method.

1 Introduction

We consider tries and suffix-trees built upon binary keys and strings generated by a Bernoulli source (p, q) with $p \geq 1/2 \geq q$. Park and Szpankowski (2005) recently studied the external profile (or sequence with index k of number of external nodes at depth k) of random tries. We use the same approach, Mellin transform and inverse Mellin transform by saddle-point method, in the Poisson model, to study the average internal profile (that counts internal nodes) of tries. The position of the saddle-point is function of the value of $k/\log(\nu)$ where ν is the (Poisson) number of keys; this implies that, depending upon this position, the inverse integral counts, up to the sign, the number of present or missing nodes at depth k. Following from this analysis, and using an approach similar to Fayolle (2004), we bound the distance between the average number of nodes at depth k in tries in the Poisson model and in suffix-trees in the fixed (number of keys) model; we relate this to the case of tries in the fixed model. Since we only consider in this article internal nodes, we generally do not further specify that the nodes that we consider are internal nodes.

2 Average internal profile of tries

We consider the average number of nodes $p_k^{(T)}(n)$ at depth k in a random trie built on exactly n binary keys (called further n-fixed model). This is equivalent to counting the average number of urns containing at least two balls in an urn model with 2^k urns, where urn ω is indexed by a word ω of size k, and such that the probability that a ball falls in urn ω is $\pi_\omega = \mathbf{P}(\omega)$.

Poissonization. We use the classical poissonization method, where the number of balls thrown in the system is not a fixed number n but follows a Poisson law of parameter ν. We note $p_{k,\mathcal{P}}^{(T)}(\nu)$ † the average number of nodes at depth k in this model. We have.

Lemma 1. *When the number of keys of a random trie follows a Poisson model of parameter ν the expectation $p_{k,\mathcal{P}}(\nu)$ of number of nodes at depth k of the trie verifies*

$$p_{k,\mathcal{P}}(\nu) = \sum_{|\omega|=k} 1 - (1 + \pi_\omega \nu)e^{-\pi_\omega \nu} \qquad (\omega \in \{0,1\}^\star)\,. \tag{1}$$

Proof. As shown by elementary algebra, the number of balls falling in urn ω follows a Poisson law of parameter $\pi_\omega \nu$, which implies that the urns behave independently of each other. The random variable Y_ω

† We omit in the rest of this section the notation (T).

counting 1 if there are more than two balls in the urn ω, and 0 elsewhere, has generating function

$$Y_\omega(u) = e^{-\pi_\omega \nu}\left(1+\pi_\omega\nu + u\left(\frac{(\pi_\omega\nu)^2}{2!}+\frac{(\pi_\omega\nu)^3}{3!}+\dots\right)\right) = u+(1-u)(1+\pi_\omega\nu)e^{-\pi_\omega\nu}.$$

Let Z be the random variable counting the number of urns with at least two balls and $F_Z(u)$ be its generating function. Since the urns are independent of each other, we have

$$F_Z(u) = \prod_{|\omega|=k} u+(1-u)(1+\pi_\omega\nu)e^{-\pi_\omega\nu} \Rightarrow \mathbf{E}(Z) = p_{k,\mathcal{P}}(\nu) = \left.\frac{\partial F_Z(u)}{\partial u}\right|_{u=1} = \sum_{|\omega|=k} 1-(1+\pi_\omega\nu)e^{-\pi_\omega\nu}. \tag{2}$$

□

By algebraic depoissonization, we also obtain

Corollary 1. *The expected number of nodes $p_k(n)$ in the n-fixed model verifies*

$$p_k(n) = \sum_{|\omega|=k} 1-(1-\pi_\omega)^n - n\pi_\omega(1-\pi_\omega)^{n-1}. \tag{3}$$

2.1 Mellin transform of $p_{k,\mathcal{P}}(\nu)$

The quantity $p_{k,\mathcal{P}}(\nu)$ is given in Equation 1 as a sum of 2^k terms. By use of direct and inverse Mellin transform it is possible to obtain asymptotically an expression of $p_{k,\mathcal{P}}(\nu)$ that is equivalent to $\nu^\zeta/\sqrt{\log\nu}$ for a $\zeta < 1$ that depends upon k, for a wide range of values of k. This will further allow comparisons with the profile of tries in the fixed model and with the profile of suffix-trees.

The Mellin transform $\mathcal{M}[g(\nu);s]$ of a function $g(\nu)$ is defined by

$$\mathcal{M}[g(\nu);s] = \int_{\nu=0}^{\infty} g(\nu)\nu^{s-1}d\nu. \tag{4}$$

We refer to Flajolet et al. (1995) for an overview about Mellin transform and its applications.

We obtain the following fundamental result (see also Park and Szpankowski (2005)).

Theorem 1. *The Mellin transform $\mathcal{M}[p_{k,\mathcal{P}}(\nu);s]$ of the number of nodes at depth k of a trie in the Poisson model verifies*

$$\mathcal{M}[p_{k,\mathcal{P}}(\nu);s] = -(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k. \tag{5}$$

The inverse Mellin transform of this function is defined in the strip $\Re(s)\in]-2,0[$.

Proof. We consider the function $g(\nu) = 1-(1+\nu)e^{-\nu}$. We have $g(\nu) = O(\nu^0)$ as $\nu\to\infty$ and $g(\nu) = O(\nu^2)$ as $\nu \to 0$. Let $|\omega|_1$ and $|\omega|_0$ count respectively the number of 1 and 0 of the word ω. Using the basic properties of the Mellin transform, since $\pi_\omega = \mathbf{P}(\omega) = p^{|\omega|_1}q^{|\omega|_0}$, we find that

$$\mathcal{M}[p_{k,\mathcal{P}}(\nu);s] = -(1+s)\Gamma(s)\sum_{j=0}^{k}\binom{k}{j}p^{-js}q^{(k-j)s} = -(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k$$

□

2.2 Inverse Mellin transform and saddle point integration

From Equation 5 we obtain by inverse Mellin transform

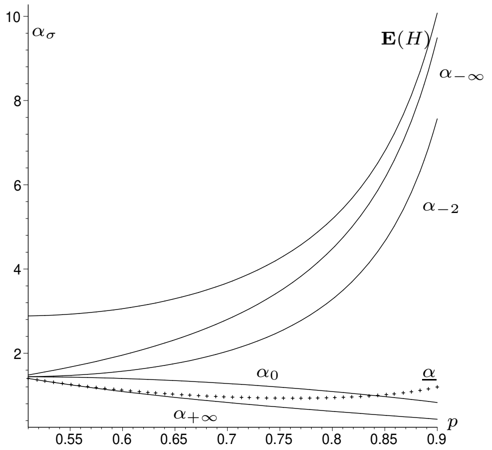
$$p_{k,\mathcal{P}}(\nu) = \frac{1}{2i\pi}\int_{x-i\infty}^{x+i\infty} -(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k\nu^{-s}ds = \frac{1}{2i\pi}\int_{x-i\infty}^{x+i\infty}F(s)ds \quad \text{with} \quad x\in\,]-2,0\,[. \tag{6}$$

We remark that $F(s)$ is analytic on $\mathbb{C}-\{0,-2,-3,-4,\dots\}$. We use a saddle-point method (see Flajolet and Sedgewick (2005) or De Bruijn (1981) for an introduction) to compute the inverse Mellin integral. We write $F(s) = e^{f(s)}$ in the following. The saddle-point s_0 verifies by definition $f'(s_0) = 0$. We have

$$f'(s) = \frac{1}{1+s}+\psi(s) - k\frac{p^{-s}\log p + q^{-s}\log q}{p^{-s}+q^{-s}} - \log\nu, \tag{7}$$

where $\psi(s)$ is the logarithmic derivatives of $\Gamma(s)$. In a first step, we consider the moduli of the terms $1/(1+s)$ and $\psi(s)$ as $O(1)$. The variables k and ν tend both to infinity. We therefore consider

$$k \times \frac{p^{-s}\log 1/p + q^{-s}\log 1/q}{p^{-s}+q^{-s}} = \log \nu \qquad \Longleftrightarrow \qquad \left(\frac{p}{q}\right)^s = \frac{\log \nu - k\log 1/p}{k\log 1/q - \log \nu}. \tag{8}$$



$k = \alpha \times \log \nu$.

From bottom to top, the plain curves are.
1) $\alpha_{+\infty}(p,q) = \dfrac{1}{\log 1/q}$
2) $\alpha_0(p,q) = \dfrac{2}{\log 1/p + \log 1/q}$
3) $\alpha_{-2}(p,q) = \dfrac{p^2+q^2}{p^2\log 1/p + q^2 \log 1/q}$
4) $\alpha_{-\infty}(p,q) = \dfrac{1}{\log 1/p}$
5) $\mathbf{E}(H) = \dfrac{2}{\log(1/(p^2+q^2))}$

Fig. 1: Range of values of k and corresponding saddle-point σ (index of α). The top curve represents the expectation of the height H of the trie. The dotted curve plots the values of the lower bound $\underline{\alpha}$ that is defined in Definition 4 and used in Theorem 5 and 6. It is well known that, when $p = q = 1/2$, almost all leaves are at depth close to the fill-up (saturation) level which is well approximated by $\alpha_{+\infty}$. This explains the shape of the region for which the saddle-point σ is real ($\alpha \in]\alpha_{+\infty}, \alpha_{-\infty}[$).

2.2.1 Parametrization of the problem and geometry of the saddle-point

Considering the right member of the last equation of Formula 8, it appears naturally that the saddle-point will be a function of the ratio $k/\log\nu$. This is not surprising, since parameters such as the average depth of insertion or the height of random tries of n keys are $O(\log n)$.

More precisely, we have.

Lemma 2. *As ν tends to infinity and $k = \alpha \log \nu$ with $\frac{1}{\log(1/q)} < \alpha < \frac{1}{\log(1/p)}$, the function $F(s) = -(1+s)\Gamma(s)\left(p^{-s}+q^{-s}\right)^k \nu^{-s}$ has a real saddle-point σ that verifies*

$$\sigma = \sigma(\alpha) = \log\left(\frac{1-\alpha\log 1/p}{\alpha \log 1/q - 1}\right) \Big/ \log(p/q). \tag{9}$$

Proof. We adopt here a parametrization of the problem inverse to this used in Park and Szpankowski (2005) and set $k = \alpha \times \log \nu$. We consider the solution $s = s'$ of the right equation of Formula 8 . By taking an expansion of $f'(s)$ in the neighborhood of s', we find that $\sigma = s' + o(1)$ where the development is easily made more precise. We neglect in the following the $o(1)$ term and consider that $\sigma = s'$. □

We remark that $\sigma(\alpha)$ is real and decreases monotonically from $+\infty$ to $-\infty$ as α increases from $1/\log(1/q)$ to $1/\log(1/p)$. We consider $\alpha(\sigma) = \alpha_\sigma = k/\log\nu$ with $\sigma \in \mathbb{R}$, where $\alpha(\sigma)$ follows from Equation 8 and is the inverse function of $\sigma(\alpha)$ of Equation 9. The set of poles of $F(s)$ is $L = \{0, -2, -3, -4, \dots\}$ and these poles correspond to values α_{-j} of $k/\log(\nu)$ given by

$$\alpha_{-j} = \frac{p^j + q^j}{p^j \log 1/p + q^j \log 1/q} \qquad (-j \in L), \tag{10}$$

Restricting σ to the positive axis gives

$$\alpha_{+\infty}(p,q) = \frac{1}{\log 1/q} < \frac{k}{\log \nu} < \frac{2}{\log 1/p + \log 1/q} = \alpha_0(p,q).$$

See the plots of $\alpha_{+\infty}, \alpha_0, \alpha_{-2}$ and $\alpha_{-\infty}$ on Figure 1 and the plots of $\sigma(\alpha)$ for $p = 0.6$ and $p = 0.9$ on Figure 2.

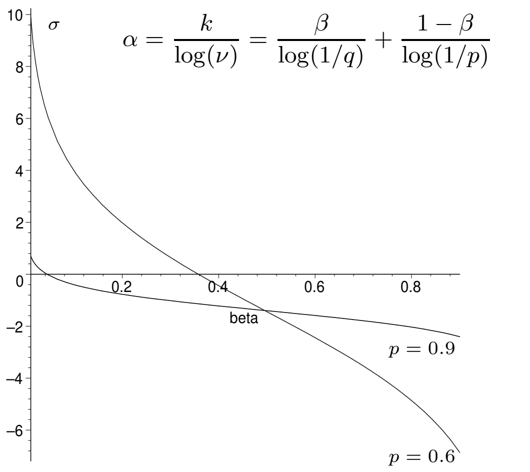


Fig. 2: The saddle-point σ as a function of β, where β is a barycentric weight varying from 0 to 1. The curves correspond to $p = 0.6$ and $p = 0.9$

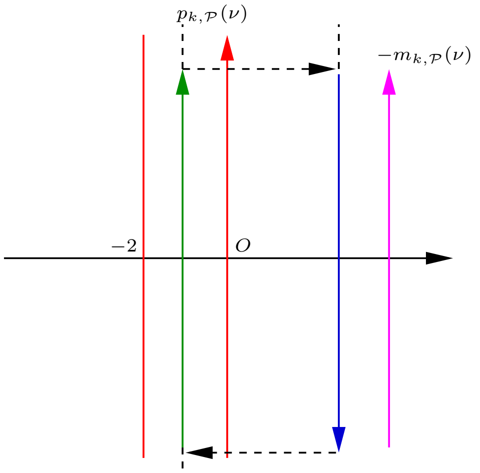


Fig. 3: The inverse Mellin integral gives $p_{k,\mathcal{P}}(\nu)$ when $\sigma \in]-2, 0[$ (number of nodes present at depth k), and $-m_{k,\mathcal{P}}(\nu) = -2^k + p_{k,\mathcal{P}}(\nu)$ when $\sigma \in]0, +\infty]$ (number of nodes missing at depth k).

2.2.2 Probabilistic consequences of the position of the saddle-point

We consider now the meaning of the inverse Mellin integral outside of the fundamental strip $]-2, 0[$. We note $\int_x$ the value of the inverse Mellin integral of Equation 6 for $x \in \mathbb{R} - L$. We consider the Cauchy contour shown in Figure 3, and let the ordinates of the horizontal segments of integration tend respectively to $-\infty$ and $+\infty$. Since the residue of $F(s)$ at $s = 0$ is -2^k and the winding number is -1, we have

$$\int_{x\in]-2,0[} - \int_{\sigma\in]0,\infty[} = 2^k \quad \Longrightarrow \quad -\int_{\sigma\in]0,\infty[} = 2^k - \int_{x\in]-2,0[} = 2^k - p_{k,\mathcal{P}}(\nu) = m_{k,\mathcal{P}}(\nu), \tag{11}$$

where $m_{k,\mathcal{P}}(\nu)$ is the number of missing nodes at depth k. The validity of Equation 11 follows from the exponential decrease of the function $\Gamma(s)$ for large imaginary values of s. By integrating in the strip $]0, \infty[$ we therefore obtain the number of missing nodes at level k (up to the sign). Using a similar contour with winding number $+1$ when $\sigma \in \bigcup\,]-j-1, -j[$ with an integer $j \geq 2$, we have

$$\int_{x\in]-2,0[} - \int_{\sigma\in]-j-1,-j[} = \sum_{j\in L\cap]\sigma,-2[} \mathrm{Res}(F(s); j) \quad \Longrightarrow \quad p_{k,\mathcal{P}}(\nu) = \int_{\sigma\in]-j-1,-j[} + \sum_{j\in L\cap]\sigma,-2[} \mathrm{Res}(F(s); j),$$

which provides a way to compute $p_{k,\mathcal{P}}(\nu)$ when $\sigma \in]-\infty, -2[$.

2.2.3 Detailed analysis of the saddle-point integral

We compute in this section the inverse Mellin integral of Equation 6

$$I(\nu) = \frac{1}{2i\pi} \int_{x-i\infty}^{x+i\infty} -(1+s)\Gamma(s) \left(p^{-s} + q^{-s}\right)^k \nu^{-s} ds = \frac{1}{2i\pi} \int_{x-i\infty}^{x+i\infty} F(s) ds. \tag{12}$$

We consider the behavior of $F(s)$ on the vertical line $\Re s = \sigma$. We write $t = \log \nu$ and $A(s) = (p^{-s} + q^{-s})^\alpha$ where $\alpha = k/t = k/\log \nu$; this gives

$$F(s) = -(1+s)\Gamma(s)(p^{-s} + q^{-s})^k \nu^{-s} = \phi(s)\Theta(s)^t \quad \text{with} \quad \Theta(s) = e^{-s} \times (p^{-s} + q^{-s})^\alpha = e^{-s} A(s) \tag{13}$$

The function $\Theta(s)$ is periodic on the vertical line $\Re s = \sigma$, and its first derivative is null at the saddle-point. On the other side the function $\phi(s)$ mostly behaves like the function Gamma, which implies exponential decrease for large imaginary values. The dominant part of the integral is concentrated upon a small neighborhood of the saddle-point, where it is approximated by a gaussian integral. This gives in first approximation $I(\nu) \approx \dfrac{F(\sigma)}{\sqrt{2\pi |f''(\sigma)|}}$, with $e^{f(s)} = F(s)$. This is the result of Park and Szpankowski

(2005). We consider also here the perturbation term corresponding to the other maxima of the function $\Theta(s)$ and we bound this perturbation.

We have

$$\Theta(\sigma + ir) = e^{-\sigma - ir} A(\sigma + ir) = e^{-\sigma - ir} \left(p^{-\sigma - ir} + q^{-\sigma - ir}\right)^{\alpha} \qquad (r \in \mathbb{R}).$$

The function $\Im(F(\sigma + ir))$ is an odd function of r; therefore all imaginary terms cancel in the integral, which corresponds to the combinatorial origin of the problem. We consider the local maxima of the function $|\Theta(\sigma + ir)| = e^{-\sigma}|A(\sigma + ir)|$.

On the vertical line $\Re s = \sigma$, with $\sigma \in \mathbb{R} - \{0\}$, the function $A(s)$ is periodic, $|A(\sigma + ir)|$ is an even function of r, we have

$$\min \left|(p^{-s} + q^{-s})^{\alpha}\right| = \left|\frac{1}{q^{\sigma}} - \frac{1}{p^{\sigma}}\right|^{\alpha}, \qquad \max \left|(p^{-s} + q^{-s})^{\alpha}\right| = \left(\frac{1}{q^{\sigma}} + \frac{1}{p^{\sigma}}\right)^{\alpha},$$

and $|A(s)|$ attains its maximum each time p^{-s} and q^{-s} are in phase. This corresponds for a given θ to

$$|r| \log 1/p = \theta + 2j_p\pi \quad \text{and} \quad |r| \log 1/q = \theta + 2j_q\pi \quad \text{with} \quad 0 < \theta < 2\pi, \quad j_p, j_q \in \mathbb{N}, \quad j_p < j_q,$$
$$\implies |r| > \rho = 2\pi \times v(p) \tag{14}$$

where $v(p)$ is defined as follows.

Definition 2. *Let $v(p) = 1/\log(p/q)$ when $p \in]1/2, p_2[$, and $v(p) = j/\log(1/q)$ when $p \in [p_j, p_{j+1}[$, where p_j is the real positive root of the equation $p^j + p - 1 = 0$ for any integer $j \geq 2$. We define by $\rho(p) = 2\pi \times v(p)$ the minimum ordinate of perturbation of the inverse Mellin integral.*

For $r = l \times \rho$ with $l \in \mathbb{Z} - \{0\}$ we have $|\Theta(\sigma + ri)| = |\Theta(\sigma)|$ but the corresponding contributions to the integral are small, since $|\Gamma(\sigma + ir)|$ decreases exponentially as $|r|$ increases. Figure 4 plots the value of ρ as function of p.

We remark that $|\Gamma(\sigma + ir)| = O(e^{-|r|})$ as $|r| \to \infty$ and $\sigma = O(1)$. (See Andrews et al. (1999), Corollary 1.4.4, for a more precise result). As results from the preceding analysis, on the vertical line $\Re(s) = \sigma$, the continuous function $|\Theta(s)/\Theta(\sigma)|$ attains its maximum value 1 at the saddle-point σ and approach it at secondary maxima $\sigma + \rho_j i$ where $\rho_j = j\rho$ and ρ is defined in Equation 14. We consider now a small δ and the intervals $V_j =]\rho_j - \delta, \rho_j + \delta[$.

By the preceding considerations, for $r \in \overline{\mathbb{R}}_\delta = \mathbb{R} - \bigcup_{j\in\mathbb{Z}} V_j$, there exists $\kappa < 1$ such that

$$\left|\frac{\Theta(\sigma + ir)}{\Theta(\sigma)}\right| < \kappa.$$

Since $|(1 + s)\Gamma(s)|$ decreases exponentially as $|\Im(s)| \to \infty$, the function $-(1 + s)\Gamma(s)$ is integrable on the line $\Re s = \sigma$. This gives

$$H_\delta = \left|\int_{r\in\overline{\mathbb{R}}_\delta} (1 + \sigma + ir)\Gamma(\sigma + ir)\Theta(\sigma + ir)^t dr\right| = \Theta(\sigma)^t O(\kappa^t) \qquad (\kappa < 1), \tag{15}$$

where κ will be later defined as a function of δ. We consider now

$$B_0 = \left|\int_{r=-\delta}^{\delta} (1 + \sigma + ir)\Gamma(\sigma + ir)\Theta(\sigma + ir)^t dr\right| \qquad \text{and} \qquad B_j = \left|\int_{r\in V_j} (1 + \sigma + ir)\Gamma(\sigma + ir)\Theta(\sigma + ir)^t dr\right|.$$

When α is bounded away from $1/\log(1/q)$ and $1/\log(1/p)$, we find as approximation for B_j and $\sum B_j$

$$B_j = B_0 \times O(e^{-|\rho_j|}), \quad \text{with} \quad |\rho_j| = |j| \times \rho \quad \implies \sum_{j\,\in\,\mathbb{Z}-\{0\}} B_j = B_0 \times O(e^{-\rho}).$$

We consider now the dominant term B_0 of the integral. By a Taylor expansion in the neighborhood of σ, we have

$$B_0 = \frac{F(\sigma)}{2\pi} \int_{r=-\delta}^{\delta} e^{-tr^2 f''(\sigma)/2 - tr^3 i f^{(3)}(\sigma + i\lambda(r)\delta)/3!} \qquad \text{with} \qquad |\lambda(r)| < 1. \tag{16}$$

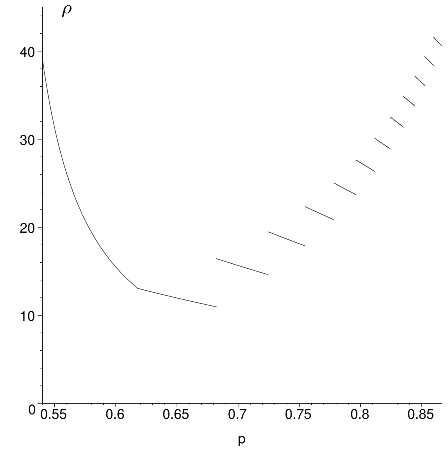


Fig. 4: Minimum ordinate of perturbation ρ. See Definition 2 and Theorem 2. The perturbation is at most $O(e^{-\rho})$ times the dominant part of the integral. Remark that $e^{-11} < 2 \times 10^{-5}$.

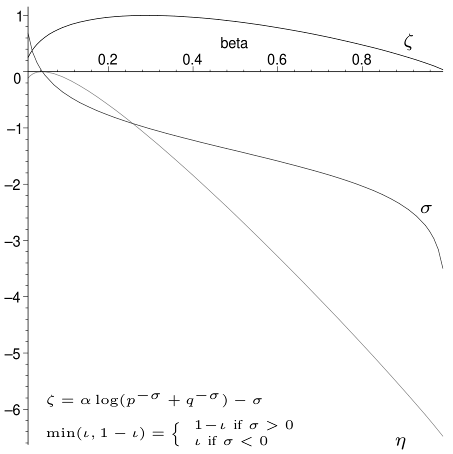


Fig. 5: Saddle-point σ, exponent ζ and η of ν respectively in $I(\nu)$ and $I(\nu)/2^k$ (see Equation 18) as a function of β (defined as in Figure 2). We have here $p = 0.9$.

We use the classical analysis described in Flajolet and Sedgewick (2005) and choose δ such that $t\delta^2$ is large and $t\delta^3$ is small; by completing the tails of the Gaussian integral and using the asymptotic approximation $\mathrm{Erf}(x) < x^{-1}\mathbf{n}(x)$, where $\mathbf{n}(x)$ is the density of the Gaussian distribution, we have

$$t^{-1/2} < \delta < t^{-1/3} \Longrightarrow B_0 = \frac{F(\sigma)}{2\pi}\left(\int_{r=-\delta}^{\delta} e^{-tr^2 f''(\sigma)/2}\right)\big(1+O(t\delta^3)\big) = \frac{F(\sigma)}{\sqrt{2\pi|f''(\sigma)|}}\Big(1+O(t\delta^3)\Big).$$

In Equation 15 we now have $\kappa = O(e^{-\delta^2})$; we obtain therefore $H_\delta = B_0 \times O(t^{1/2}e^{-\delta^2 t})$ since $|f''(\sigma)| = O(t)$. We obtain

$$M(\nu) = \frac{1}{2i\pi}\int_{s=\sigma-\infty}^{\sigma+\infty} F(s)ds = \frac{F(\sigma)}{\sqrt{2\pi|f''(\sigma)|}}\Big(1 + O(t\delta^3) + O(e^{-\rho}) + O(t^{1/2}e^{-\delta^2 t})\Big). \tag{17}$$

The following theorem summarizes the results obtained in the Poisson model. It is expressed in a different manner and without perturbation term in Park and Szpankowski (2005).

Theorem 2. *Considering the function $\rho(p)$ of Definition 2 and $k = \alpha\log\nu$, where $\alpha \in]1/\log(1/q), 1/\log(1/p)[$, let σ (saddle-point) verifies $\sigma = \log\left(\frac{1-\alpha\log(1/p)}{\alpha\log(1/q)-1}\right)\Big/\log(p/q)$.*

We have asymptotically, as ν tends to infinity, for a random trie in the Poisson model of parameter ν.

- *The dominant part of the inverse Mellin integral verifies*

$$J(\nu) = \frac{F(\sigma)}{\sqrt{2\pi|f''(\sigma)|}} = \frac{-(1+\sigma)\Gamma(\sigma)\nu^{\alpha\log(p^{-\sigma}+q^{-\sigma})-\sigma}}{\sqrt{2\pi\alpha\log(\nu)\times U(\sigma)}} \tag{18}$$

$$\text{where} \quad U(\sigma) = \frac{p^{-\sigma}\log^2 p + q^{-\sigma}\log^2 q}{p^{-\sigma}+q^{-\sigma}} - \left(\frac{p^{-\sigma}\log p + q^{-\sigma}\log q}{p^{-\sigma}+q^{-\sigma}}\right)^2.$$

- $m_{k,\mathcal{P}}(\nu) = -J(\nu)(1+O(e^{-\rho(p)}))$ *when* $\alpha \in \left]\frac{1}{log(1/q)}, \frac{2}{\log(1/p)+\log(1/q)}\right[$,

  where $m_{k,\mathcal{P}}(\nu)$ is the average number of missing nodes at depth k.

- $p_{k,\mathcal{P}}(\nu) = J(\nu)(1+O(e^{-\rho(p)}))$ *when* $\alpha \in \left]\frac{2}{\log(1/p)+\log(1/q)}, \frac{p^2\log(1/p)+q^2\log(1/q)}{p^2+q^2}\right[$,

  where $p_{k,\mathcal{P}}(\nu)$ is the average number of present nodes at depth k.

We obtain as corollary.

Corollary 3. *Let us consider the rate of saturation $\iota = \iota(\nu, \alpha) = p_{k,\mathcal{P}}(\nu)/2^k = p_{k,\mathcal{P}}(\nu)/\nu^{\alpha \log 2}$. We have*

$$\alpha \to \alpha_0^- \Rightarrow \iota \to 1 - \frac{1}{\sqrt{2\pi}} \qquad \text{and} \qquad \alpha \to \alpha_0^+ \Rightarrow \iota \to \frac{1}{\sqrt{2\pi}} \qquad \left(\alpha_0 = \frac{2}{\log(1/p) + \log(1/q)}\right). \tag{19}$$

Proof. This follows from the following equations,

$$\frac{\Gamma(\sigma)}{\sqrt{f''(\sigma)}} = 1 + o(1) \quad \text{and} \quad \alpha(\sigma)\log(p^{-\sigma} + q^{-\sigma}) - \sigma = \xi\sigma + O(\sigma^2) \qquad \text{as} \quad \nu \to \infty,$$

where we have $\sigma = o(1/\log \nu)$, the constant ξ depends only on p and q, and $\alpha(\sigma)$ is the inverse function of $\sigma = \sigma(\alpha)$. □

We observe therefore a phase reversal of ι when α goes from α_0^- to α_0^+.

Shape of the exponents. Considering Figure 5 we observe that the exponent η of ν in $\min(1-\iota, \iota)$ attains its maximum 0 at $\sigma = 0$, which is the point of phase reversal. We also observe, in the range $\sigma \in]-2, 0[$ that the exponent ζ of ν in $p_{k,\mathcal{P}}(\nu)$ attains its maximum 1 at $\sigma = -1$, which corresponds to $k = \log(\nu)/h(p,q)$ where $h(p,q)$ is the entropy of the alphabet. This has been previously observed by Park and Szpankowski (2005).

Similarly, it is possible to make a detailed study of the value of $p_{k,\mathcal{P}}(\nu)$ when $\alpha < \alpha_{-2}$ by taking in account the residues of $F(s)$ at the negative integers smaller than -2. These points correspond to minor discontinuities of the function $p_k(\nu, \alpha)$.

2.3 Average profile of tries, exact model versus Poisson model

We use here a simplify version of entry 10 of Ramanujan's notebook, Part I (see Berndt (1985) page 57).

Theorem 3. *Let $h(x)$ denote a function of at most polynomial growth as x (real) tends to ∞. Suppose that there exists a constant $A \geq 1$ and a function $a(x)$ of at most polynomial growth as x tends to ∞ such that for each nonnegative integer m and all sufficiently large x, the derivatives $h^{(m)}(x)$ exist and satisfy $\left|\frac{h^{(m)}(x)}{m!}\right| \leq a(x)\left(\frac{A}{x}\right)^m$. Put $h_\infty(x) = e^{-x}\sum_{k=2}^{\infty}\frac{x^k h(k)}{k!}$. Then,*

$$h_\infty(x) = h(x) + xh''(x) + O(a(x)x^{-2}),$$

as x tends to ∞.

Applying this theorem with $h(x) = x^\zeta/\sqrt{\log x}$ and $a(x) = A = 1$ for large integer $x = n$ and then reasoning by contradiction gives the following depoissonization result.

Theorem 4. *As $n \to \infty$ we have, for any small $\epsilon > 0$, $m_k^{(T)}(n) = m_{k,\mathcal{P}}^{(T)}(n)(1 + O(n^{-(1-\epsilon)}))$ when $\alpha \in]1/\log(1/q), \alpha_0[$ and $p_k^{(T)}(n) = p_{k,\mathcal{P}}^{(T)}(n)\left(1 + O\left(n^{-(1-\epsilon)}\right)\right)$ when $\alpha \in]\alpha_0, \alpha_{-2}[$.*

3 Average internal profile of suffix-trees

We consider now languages $\mathcal{L} \subseteq \{0,1\}^\star$ and the corresponding weighted generating functions $\mathcal{L}(z) = \sum_{\omega \in \mathcal{L}} \pi_\omega z^{|\omega|} = \sum_{n \geq 0} l_n z^n$, where π_ω is the probability of the word ω and l_n is the probability that a random word of size n belongs to the language.

We also consider the autocorrelation set $\mathcal{A}_\omega$ of a word ω, defined as

$$\mathcal{A}_\omega = \{h; \qquad \omega.h = u.\omega \quad \text{and} \quad |h| < |\omega|\}.$$

We define in the following by $\mathcal{O}_\omega^{(r)}$ (resp. $\mathcal{O}_\omega^{(r+)}$) the language of words with exactly (resp. at least) r (possibly overlapping) occurrences of ω and remark that $\mathcal{O}_\omega^{(2+)} = \Sigma^\star - \mathcal{O}_\omega^{(0)} - \mathcal{O}_\omega^{(1)}$.

We consider the suffix-tree built on the n first suffixes of an infinite string U, further referred to as suffix-tree with n keys. The number of nodes $s_k^{(S)}(n)$ at depth k in the suffix-tree is equal to the number of words ω of size k that occur at least twice in the prefix τ of length $n+k-1$ of U. We note $p_k^{(S)}(n)$ the average number of nodes at depth k. We have

$$s_k^{(S)}(n) = \sum_{|\omega|=k} \mathbf{1}_{\{\omega \text{ occurs at least twice in } \tau\}} \qquad \text{and} \qquad p_k^{(S)}(n) = \sum_{|\omega|=k} \mathbf{P}(\omega \text{ occurs at least twice in } \tau).$$

Multiplying by z^n and summing over n gives

$$P_k^{(S)}(z) = \sum_{n\geq 0} p_k^{(S)}(n) z^n = \sum_{|\omega|=k} \mathcal{O}_\omega^{(2+)}(z) = \frac{2^k}{1-z} - \sum_{|\omega|=k} \left(\mathcal{O}_\omega^{(0)}(z) + \mathcal{O}_\omega^{(1)}(z)\right). \tag{20}$$

It follows from an extension of Guibas and Odlyzko analysis of $\mathcal{O}_\omega^{(0)}$ (see Sedgewick and Flajolet (1996) p. 374) or from the Bernoulli case in Régnier and Szpankowski (1998) that

$$P_k^{(S)}(z) = \frac{2^k}{1-z} - \sum_{|\omega|=k} \left(\frac{\mathcal{A}_\omega(z)}{K_\omega(z)} + \frac{\pi_\omega z^{|\omega|}}{K_\omega(z)^2}\right) \qquad \text{with} \qquad \frac{1}{K_\omega(z)} = \frac{1}{\pi_\omega z^{|\omega|} + (1-z)\mathcal{A}_\omega(z)}. \tag{21}$$

We follow Fayolle (2004) and consider the dominant pole ρ_ω of $1/K_\omega(z)$; for $|\omega|$ large we have ρ_ω close to 1. Considering a suitable $R \in]1, 1/p[$ we perform a Cauchy integration along a circle of radius R of $P_k^{(S)}(z)/z^{n+1}$ and use a bootstrapping method (see Fayolle (2004)); this provides asymptotically

$$p_k^{(S)}(n) = [z^n]P_k^{(S)}(z) = \sum_{|\omega|=k} 1 - \left(1 + \frac{n\pi_\omega}{\mathcal{A}_\omega(1)}\right) e^{-n\pi_\omega/\mathcal{A}_\omega(1)} + \sum_{|\omega|=k} O\left(n\pi_\omega^2 e^{-n\pi_\omega/\mathcal{A}_\omega(1)}\right) + \sum_{|\omega|=k} O(kn\pi_\omega^2);$$

Let S_2 and S_3 represent the second and third sums of the right member of this equation. We have $\pi_\omega \leq p^{|\omega|} = p^k = n^{-\alpha \log(1/p)}$ for all ω and therefore $S_2 = p_{k,\mathcal{P}}^{(T)}(n) O(n^{-\alpha \log(1/p)})$.

Let $\zeta = \zeta(\alpha) = \alpha \log(p^{-\sigma(\alpha)} + q^{-\sigma(\alpha)}) - \sigma(\alpha)$ where n^ζ is the dominant asymptotic term of $p_{k,\mathcal{P}}^{(T)}(n)$ in Equation 18. We have

$$S_3/p_{k,\mathcal{P}}^{(T)}(n) = O\left(kn(p^2+q^2)^k/n^\zeta\right) = O\left(n^{1-\alpha\log(1/(p^2+q^2))-\zeta}\right)\log n.$$

This leads to the following definition.

Definition 4. *Let $\upsilon(\alpha) = -(1 - \alpha\log(1/(p^2+q^2)) - \zeta(\alpha))$ and $\underline{\alpha}$ be the solution of the equation $\upsilon(\alpha) = 0$.*

We have $\alpha > \underline{\alpha} \Rightarrow \upsilon(\alpha) > 0$, which implies

Lemma 3. *Considering $\underline{\alpha}$ defined in Definition 4 and $\alpha \in]\max(\underline{\alpha}, \alpha_0), \alpha_{-2}[$, as $n \to \infty$, the number of nodes at depth $k = \alpha \log n$ of a suffix-tree of n keys verifies for a positive υ*

$$p_k^{(S)}(n) = a_k^{(S)}(n) + p_{k,\mathcal{P}}^{(T)}(n)\left(O(n^{-\upsilon}) + O\left(n^{-\alpha\log(1/p)}\right)\right) \text{ where } a_k^{(S)}(n) = \sum_{|\omega|=k} 1 - \left(1 + \frac{n\pi_\omega}{\mathcal{A}_\omega(1)}\right) e^{-n\pi_\omega/\mathcal{A}_\omega(1)}.$$

See Figure 1 for a plot of $\underline{\alpha}(p)$. A result similar to Lemma 3 holds for the missing nodes when $\alpha \in]\underline{\alpha}, \alpha_0[$ and $p \in [0.5, 0.83[$.

4 Comparison of the suffix-tree and the trie

We compare now the internal profiles of a random suffix-tree with n keys and of a trie in the Poisson model of parameter n. We consider again the case of the present nodes, with $\sigma \in]-2, 0[$, but a similar proof applies to the case of missing nodes.

We have.

Theorem 5. *Let $k = \alpha \log n$ where $\alpha \in]\max(\underline{\alpha}, \alpha_0), \alpha_{-2}[$ and $\underline{\alpha}$ is defined in Definition 4. As $n \to \infty$ the numbers of nodes at depth k*

- *$p_k^{(S)}(n)$ of a suffix-tree of n keys*
- *and $p_{k,\mathcal{P}}^{(T)}(n)$ of a trie whose number of keys is Poisson of parameter n*

verify

$$\left| p_k^{(S)}(n) - p_{k,\mathcal{P}}^{(T)}(n) \right| = p_{k,\mathcal{P}}^{(T)}(n) \times O(n^{-\lambda})$$

for a positive λ.

Proof. Building upon Lemma 3 we analyze the difference $\left| a_k^{(S)}(n) - p_{k,\mathcal{P}}^{(T)}(n) \right|$, where

$$a_k^{(S)}(n) = \sum_{|\omega|=k} 1 - \left(1 + \frac{n\pi_\omega}{\mathcal{A}_\omega(1)}\right) e^{-n\pi_\omega/\mathcal{A}_\omega(1)} \quad \text{and} \quad p_{k,\mathcal{P}}^{(T)}(n) = \sum_{|\omega|=k} 1 - (1+n\pi_\omega)e^{-n\pi_\omega}. \quad (22)$$

We follow the spirit of Fayolle (2004), improving upon the worst case by use of Mellin transforms.

The basic period d of a word ω is the size of the smallest word u such that $\omega = u^h v$ where v is a prefix of u and h is a positive integer.

We have as previously $g(x) = 1-(1+x)e^{-x}$. Let $\mathcal{D}_k$ (resp. $\mathcal{D}_k^c$) be the set of *periodic* (resp. *aperiodic*) words of size k such that $d \le k/2$ (resp. $d > k/2$); we split the sums of Equation 22 with respect to these sets, and consider bounds for $\mathcal{A}_\omega(1)$ in these sets;

$$a_k^{(S)}(n) = \left(\sum_{\omega\in\mathcal{D}_k} + \sum_{\omega\in\mathcal{D}_k^c} \right) g\left(\frac{n\pi_\omega}{\mathcal{A}_\omega(1)}\right) = S_{\mathcal{D}_k} + S_{\mathcal{D}_k^c} \quad \text{and} \quad p_{k,\mathcal{P}}^{(T)}(n) = \left(\sum_{\omega\in\mathcal{D}_k} + \sum_{\omega\in\mathcal{D}_k^c} \right) g(n\pi_\omega) = T_{\mathcal{D}_k} + T_{\mathcal{D}_k^c},$$

$$1 \le \mathcal{A}_\omega(1) \le \frac{1}{1-p} \quad (\omega \in \mathcal{D}_k) \qquad \text{and} \qquad 1 \le \mathcal{A}_\omega(1) \le 1 + \frac{p^{k/2}}{1-p} = c_k(p) \quad (\omega \in \mathcal{D}_k^c).$$

We also consider in the following $B_{\mathcal{D}_k} = \sum_{\omega\in\mathcal{D}_k} g(n\pi_\omega / c_k(p))$.

We use repetitively the fact that $\mathcal{M}[g(x\pi_\omega/\chi); s] = (1/\chi)^{-s}\mathcal{M}[g(x\pi_\omega); s]$ and remark that the function $g(x) = 1-(1+x)e^{-x}$ is increasing on $[0,\infty[$.

We consider first the periodic words and write $\omega = (ab)^r a$ where $d = |ab|$, $r = \lfloor k/d \rfloor > 1$, and $|a| = k - rd$. The Mellin transform $\mathcal{M}[p_{k,\mathcal{P}}^{(d)}(n); s]$ of the expectation of the number of nodes at depth k labeled by a word of period d, in a trie and within the Poisson model of parameter n, is

$$\mathcal{M}[p_{k,\mathcal{P}}^{(d)}(n); s] = -(1+s)\Gamma(s)\left(p^{-(r+1)s} + q^{-(r+1)s}\right)^{|a|}\left(p^{-rs} + q^{-rs}\right)^{|b|} \implies \frac{\mathcal{M}[p_{k,\mathcal{P}}^{(d)}(n); \sigma]}{\mathcal{M}[p_{k,\mathcal{P}}(n); \sigma]} = O(\psi_1^k), \quad (23)$$

where we have $\psi_1 < 1$ and σ verifies Equation 9. The last equation follows by separately handling the cases where $r = O(1)$, in which case $|a+b|$ is of the order of $\log(n)$, and where r tends to infinity.

We perform now the inverse Mellin integral of $\mathcal{M}[p_{k,\mathcal{P}}^{(d)}(n); s]$ on the vertical line $\Re s = \sigma$; the point $s = \sigma$ is no more a saddle-point, but the analysis follows the same lines as in Section 2.2 and uses Equation 23 to upper bound the dominant part of the integral. Summing over d and using the inequalities $1/\mathcal{A}_\omega(1) \le 1$ and $1/c_k(p) < 1$ provide for $\psi_1 < \psi_2 < 1$ and k large enough

$$T_{\mathcal{D}_k} = p_{k,\mathcal{P}}^{(T)}(n) \times O(\psi_2^k), \qquad S_{\mathcal{D}_k} = p_{k,\mathcal{P}}^{(T)}(n) \times O(\psi_2^k) \qquad \text{and} \qquad B_{\mathcal{D}_k} = p_{k,\mathcal{P}}^{(T)}(n) \times O(\psi_2^k). \quad (24)$$

We consider now the non-periodic words. By expanding $(1/c_k(p))^{-\sigma}$, where σ is the saddle-point of Equation 9, we have, up to second order terms, for k large enough,

$$\left(1 - \frac{2|\sigma|p^{k/2}}{1-p}\right) p_{k,\mathcal{P}}^{(T)}(n) \le S_{\mathcal{D}_k^c} + B_{\mathcal{D}_k} \le p_{k,\mathcal{P}}^{(T)}(n)\left(1 + O(\psi_2^k)\right).$$

This gives, with $k = \alpha \log n$ and $\alpha > \underline{\alpha}$

$$\left| p_k^{(S)}(n) - p_{k,\mathcal{P}}^{(T)}(n) \right| = p_{k,\mathcal{P}}^{(T)}(n) \times O(n^{-\lambda})$$

where $\lambda = \min(\upsilon, \alpha\log(1/\psi_2), \alpha\log(1/p)/2)$ and υ satisfies Definition 4. $\square$

From there and Section 2.3 where we compared the Poisson model and the "fixed" model follows the theorem.

Theorem 6. *As $n \to \infty$ and $k = \alpha \log n$ with $\alpha \in]\max(\underline{\alpha}, \alpha_0), \alpha_{-2}[$, where $\underline{\alpha}$ and α_{-2} are defined as previously, the number of present nodes at depth k in a suffix-tree $p_k^{(S)}(n)$ and a trie $p_k^{(T)}(n)$ of n keys verify for a positive λ*

$$\left| p_k^{(S)}(n) - p_k^{(T)}(n) \right| = p_k^{(T)}(n) \times O\left(n^{-\min(\lambda,1)}\right).$$

A similar result holds for the missing nodes when $p < 0.83$ and α belongs to the range $]\underline{\alpha}, \alpha_0[$ (see Figure 1). We conjecture that these results extend to a larger range of values of α.

Acknowledgements

We thank Julien Fayolle, Philippe Flajolet and Bruno Salvy for precious hints and helpful discussions.

References

G. E. Andrews, R. Askey, and R. Roy. *Special Functions*. Cambridge University Press, 1999.

B. C. Berndt. *Ramanujan's Notebook, Part I*. Springer Verlag, 1985.

N. G. De Bruijn. *Asymptotic Methods in Analysis*. Dover, 1981.

J. Fayolle. An average-case analysis of basic parameters of the suffix tree. In M. Drmota, P. Flajolet, D. Gardy, and B. Gittenberger, editors, *Mathematics and Computer Science*, pages 217–227. Birkhäuser, 2004. Proceedings of a colloquium organized by TU Wien, Vienna, Austria, September 2004.

P. Flajolet, X. Gourdon, and P. Dumas. Mellin Transforms and Asymptotics: Harmonic Sums. *Theoretical Computer Science*, 144(1-2):3–58, 1995.

P. Flajolet and R. Sedgewick. Analytic combinatorics. Book to appear, 2005.

P. Nicodème. q-gram analysis and urn models. In *Discrete Mathematics and Theoretical Computer Science AC*, pages 243–258, 2003. Proceedings of the colloquium Discrete Random Walks DRW2003, organized at IHP, Paris, France, September 2003.

G. Park and W. Szpankowski. Towards a complete characterization of tries. In *SIAM-ACM Symposium on Discrete Algorithms (SODA2005), Vancouver*, pages 33–42, 2005.

M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4):631–649, 1998.

R. Sedgewick and P. Flajolet. *An Introduction to the Analysis of Algorithms*. Addison-Wesley Publishing Company, 1996.