Distribution of inter-node distances in digital trees

Rafik Aguech[1] and Nabil Lasmar[2] and Hosam Mahmoud[3]

Faculté des Sciences de Monastir, Département de mathématiques, 5019 Monastir, Tunisia. E-mail: rafikaguech@ipeit.rnu.tn
Institut préparatoire aux études d'ingénieurs de Tunis, Département de mathématiques, IPEIT, Rue lelnahrou-Montfleury, Tunis, Tunisia. E-mail: nabillasmar@yahoo.fr
Department of Statistics, The George Washington University, Washington, D.C. 20052, U.S.A. E-mail: hosam@gwu.edu

We investigate distances between pairs of nodes in digital trees (digital search trees (DST), and tries). By analytic techniques, such as the Mellin Transform and poissonization, we describe a program to determine the moments of these distances. The program is illustrated on the mean and variance. One encounters delayed Mellin transform equations, which we solve by inspection. Interestingly, the unbiased case gives a bounded variance, whereas the biased case gives a variance growing with the number of keys. It is therefore possible in the biased case to show that an appropriately normalized version of the distance converges to a limit. The complexity of moment calculation increases substantially with each higher moment; A shortcut to the limit is needed via a method that avoids the computation of all moments. Toward this end, we utilize the contraction method to show that in biased digital search trees the distribution of a suitably normalized version of the distances approaches a limit that is the fixed-point solution (in the Wasserstein space) of a distributional equation. An explicit solution to the fixed-point equation is readily demonstrated to be Gaussian.

Keywords: Random trees, recurrence, Mellin transform, poissonization, fixed point, contraction method.

1 Introduction

Various types of distances in random trees have lately become a topic of interest, as can be seen in half a dozen or so of recent papers. The distance between pairs of nodes in a recursive tree was investigated in Neininger (2002). The distance between pairs of nodes in a random binary search tree was investigated in Mahmoud and Neininger (2003) and in Devroye and Neininger (2004); Panholzer and Prodinger (2004) give a generalization. This extended abstract reports on results in Aguech, Lasmar and Mahmoud (2005a, 2005b) where distances between random pairs in digital trees are studied.

The standard data model for digital search trees is the Bernoulli probability distribution (infinitely long independent keys of independent bits). The probability model should ideally be unbiased. In practice this unbias is not guaranteed. So, our study is not limited to the unbiased Bernoulli case, and puts in good perspective the contrast between biased and unbiased data models.

2 Digital trees

There are two flavors of naturally grown digital trees, the digital search tree and the trie. We consider both. Digital trees are suited for digital data, which abound in science, engineering and technology. For instance, they are the building blocks of computer files. For ease of exposition, we shall deal with the binary case. Generalization to larger alphabets should not be hard. For example, for DNA strands one uses a 4-letter alphabet of protein nucleotides.

Let δ_n be the depth of a randomly selected node in a random digital tree of size n, with *random* meaning that all nodes are equally likely choices. Let Δ_n be the distance (i.e. the number of tree edges) between two randomly selected keys in a random digital tree of size n, where all $\binom{n}{2}$ pairs of keys are equally likely. The recurrence equations for Δ_n will use δ_n.

2.1 Digital search trees

The digital search tree was invented in Coffman and Eve (1970). In addition to the uses already mentioned as a data structure, the digital search tree provides a model for the analysis of several important algorithms, such as the Lempel-Ziv parsing algorithm (see Louchard and Szpankowski (1995)), and Conflict Resolution (see Mathys and Flajolet (1985)).

The binary DST grows according to an algorithm. The keys $K_1, K_2 \ldots, K_n$ come in serially. Initially we have an empty tree. For the first key, a root is allocated. The key K_2 is guided to the left subtree, where it becomes a left child of the root, if its first bit is 0, otherwise it goes to the right subtree, where it is linked as a right child of the root. Subsequent keys are treated similarly, they are taken into the left or right subtree according as whether the first bit is 0 or 1, and in the subtree the algorithm is applied recursively, but at level ℓ of the recursion the $(\ell + 1)$st bit is used for guiding the search.

2.2 Tries

The trie was invented independently by De La Briandais (1959) and Fredkin (1960) for information re*trie*val. A binary trie is a digital tree consisting of internal nodes that each has one or two children, and leaves that hold data. The trie grows from n keys according to a construction algorithm. If $n = 0$, the insertion algorithm terminates. If $n = 1$, a leaf is allocated for the key given. If $n \geq 2$, an internal node is allocated as a root of the tree; keys starting with 0 go to the left subtree, and keys starting with 1 go to the right. The construction proceeds recursively in the subtrees, but at level ℓ the $(\ell + 1)$st bit of the key is used for branching. When the algorithm terminates, each key is in a leaf by itself, and the root-to-leaf paths correspond to minimal prefixes sufficient to distinguish the keys.

In addition to the uses already mentioned as a data structure, the trie provides a model for the analysis of several important algorithms, such as Radix Exchange Sort (see Knuth 1998), and Extendible Hashing (see Fagin, Nievergelt, Pippenger and Strong (1979)).

A main distinction between the algorithm for digital search trees and that for tries is that all the nodes of the digital tree hold keys, whereas in tries the keys reside only in leaves.

3 Notation and methodology

The Mellin transform of a function $f(x)$ is

$$\int_0^\infty f(x) x^{s-1}\, ds,$$

and will be denoted by $f^*(s)$. For a survey of the Mellin transform in the context of the analysis of algorithms we refer the reader to the comprehensive survey in Flajolet, Gourdon and Dumas (1995).

Another tool we rely on in the analysis is depoissonization. This method is now standard and we shall not produce the details in any great length. We refer the reader to an original source such as Jacquet and Szpankowski (1998), or a textbook such as Szpankowski (2001).

Instrumental to our presentation is the functions

$$Q_k(s) = \prod_{j=0}^{k} (1 - p^{j-s} - q^{j-s}),$$

with $q = 1 - p$, and the data entropy

$$h_p = -p \ln p - q \ln q.$$

We shall also need the two functions

$$\tilde{h}_p = p \ln^2 p + q \ln^2 q \qquad \text{and} \qquad \hat{h}_p = p^2 \ln p + q^2 \ln q.$$

In the sequel the symbol γ is Euler's constant.

We remark in passing that some of the intermediate steps in the forthcoming derivation may be reachable via the binomial transform (see Poblete, Papadakis and Munro (1995).).

4 Distances in DST

In a random DST let L_n and R_n be respectively the number of keys residing in the left and right subtrees, among the n keys stored in the tree (so, $L_n + R_n = n - 1$). Given L_n, the distance Δ_n satisfies the recurrence

$$\Delta_n \mid L_n = \begin{cases} \Delta_{L_n}, & \text{with probability } \dfrac{\binom{L_n}{2}}{\binom{n}{2}}; \\[2ex] \tilde{\Delta}_{R_n}, & \text{with probability } \dfrac{\binom{R_n}{2}}{\binom{n}{2}}; \\[2ex] (\delta_{L_n}+1) + (\tilde{\delta}_{R_n}+1), & \text{with probability } \dfrac{L_n R_n}{\binom{n}{2}} \\[1ex] \delta_{L_n}+1, & \text{with probability } \dfrac{L_n}{\binom{n}{2}} \\[1ex] \tilde{\delta}_{R_n}+1, & \text{with probability } \dfrac{R_n}{\binom{n}{2}}\ . \end{cases} \tag{1}$$

Here $\tilde{\Delta}_{R_n}$ is conditionally independent of Δ_{L_n} and $\tilde{\delta}_{R_n}$ is conditionally independent of δ_{L_n}.

4.1 Functional equations

From the conditional recursion (1), we obtain for t real

$$\begin{aligned} \binom{n}{2}\mathbf{E}\big[e^{\Delta_n t}\big] &= \mathbf{E}\left[\binom{L_n}{2}e^{\Delta_{L_n}t}\right] + \mathbf{E}\left[\binom{R_n}{2}e^{\tilde{\Delta}_{R_n}t}\right] \\ &+ \ e^{2t}\mathbf{E}\big[L_n R_n e^{\delta_{L_n}t}e^{\tilde{\delta}_{R_n}t}\big] + e^t\mathbf{E}\big[L_n e^{\delta_{L_n}t}\big] + e^t\mathbf{E}\big[R_n e^{\tilde{\delta}_{R_n}t}\big]. \end{aligned}$$

We handle this recurrence, via poissonization—Let $N(z)$ be distributed like a Poisson random variable with mean z, and put

$$\Psi(t,z) := \mathbf{E}\left[\binom{N(z)}{2}e^{\Delta_{N(z)}t}\right] = e^{-z}\sum_{n=0}^{\infty}\binom{n}{2}\mathbf{E}\big[e^{\Delta_n t}\big]\frac{z^n}{n!}.$$

The poissonized function $\Psi(t,z)$ satisfies the equation

$$\begin{aligned} \frac{\partial}{\partial z}\Psi(t,z) + \Psi(t,z) &= \Psi(t,pz) + \Psi(t,qz) + e^{2t}\psi(t,pz)\psi(t,qz) \\ &\quad + e^t\psi(t,pz) + e^t\psi(t,qz), \end{aligned}$$

with $\psi(t,z) = e^{-z}\sum_{k=0}^{\infty} k\mathbf{E}[e^{\delta_k t}]z^k/k!$.

4.2 Moments

One can routinely show that the first derivative

$$X(z) = \frac{\partial}{\partial t}\Psi(t,z)\Big|_{t=0} - \frac{z^2}{2} = \mathbf{E}\left[\binom{N(z)}{2}\Delta_{N(z)}\right] - \frac{z^2}{2}$$

satisfies

$$X'(z) + X(z) = X(pz) + X(qz) + pzx(qz) + qzx(pz) + x(pz) + x(qz) + pqz^2, \tag{2}$$

with $x(z) = \sum_{n=1}^{\infty} n\mathbf{E}[\delta_n] z^n e^{-z}/n!$, and $x^*(s)$ is its Mellin transform. Functional equations for $x(z)$, and an explicit expression for $x^*(s)$ can be gleaned from a number of sources, such as Louchard and Szpankowski (1995):

$$x^*(s) = \frac{Q_\infty(-2)}{Q_\infty(s)}\,\Gamma(s).$$

Note that we defined $X(z)$ as a poissonized average with a shift to ensure the existence of its Mellin transform.

An informative rearrangement of (2) is helpful to our purpose:

$$\begin{aligned}\Big(X'(z) - \frac{z^2}{2}\Big) + X(z) &= X(pz) + X(qz) + pzx(qz) + qzx(pz) \\ &\quad + \Big(x(pz) - \frac{p^2z^2}{2}\Big) + \Big(x(qz) - \frac{q^2z^2}{2}\Big).\end{aligned}$$

Taking the Mellin transform of the latter equation, we obtain

$$\begin{aligned}-(s-1)X^*(s-1) + X^*(s) &= (p^{-s}+q^{-s})X^*(s) + (pq^{-s}+qp^{-s})x^*(s+1) \\ &\quad + (p^{-s}+q^{-s})x^*(s), \end{aligned} \tag{3}$$

existing in the strip $-3 < \Re s < -2$.

We now find a closed form expression for $X^*(s)$ by inspection. Put

$$X^*(s) = \Gamma(s)\lambda(s).$$

By (3), $\lambda(s)$ must satisfy

$$-\lambda(s-1) = \big(p^{-s}+q^{-s}-1\big)\lambda(s) + \Big(\frac{\big(pq^{-1-s}+qp^{-1-s}\big)}{\big(1-p^{-1-s}-q^{-1-s}\big)}s + p^{-s}+q^{-s}\Big)\frac{Q_\infty(-2)}{Q_\infty(s)}.$$

After some tedious iterative algebra we get

$$\lambda(s) = \frac{Q_\infty(-2)}{Q_\infty(s)}\Big(2\kappa_\infty - \frac{1}{2pq} + \sum_{k=0}^{\infty}\frac{T(s-k)}{1-p^{k-s}-q^{k-s}}\Big),$$

where

$$\begin{aligned}T(s) &= \frac{(pq^{-1-s}+qp^{-1-s})}{1-p^{-1-s}-q^{-1-s}}s + p^{-s}+q^{-s}, \\ \kappa_\infty &= -\sum_{k=1}^{\infty}\frac{T(-2-k)}{2(1-p^{k+2}-q^{k+2})}.\end{aligned}$$

Putting it together, the complete Mellin transform of $X(z)$ is

$$X^*(s) = \frac{Q_\infty(-2)}{Q_\infty(s)}\Big(2\kappa_\infty - \frac{1}{2pq} + \sum_{k=0}^{\infty}\frac{T(s-k)}{1-p^{k-s}-q^{k-s}}\Big)\,\Gamma(s).$$

It suffices to compute the residue of $X^*(s)z^{-s}$ at poles located on the line $\Re s = -2$ to obtain an asymptotic expression for $X(z)$. The required result then follows by depoissonization. The variance can be computed by similar methods, starting from second derivatives of $\Psi(t,z)$. It involves a significantly more elaborate residue computation.

Proposition 1 *In a random digital search tree of n random keys, the average distance between two randomly selected keys is*

$$\begin{aligned}\mathbf{E}[\Delta_n] &= \frac{2}{h_p}\ln n + \frac{\hat{h}_p}{pqh_p} + \frac{\tilde{h}_p}{h_p^2} - \frac{2(1-\gamma)}{h_p} + \frac{\ln(pq)}{h_p} \\ &\quad - \frac{2\alpha_\infty}{h_p} + 2 - 2\xi_p(\ln n) + O\Big(\frac{1}{n}\Big),\end{aligned}$$

where $\xi_p(.)$ is a small oscillating function. The variance is

$$\mathbf{Var}[\Delta_n] = 2\sigma_p^2 \ln n + O(1),$$

where $\sigma_p^2 = (\tilde{h}_p - h_p^2)h_p^{-3}$.

Remark: Except for the symmetric case, the variance grows logarithmically with the number of keys inserted in the tree. In the symmetric case the variance is $O(1)$ (oscillating but uniformly bounded), showing the stiff resistance of inter-node distances in digital search trees to change with the number of keys. In either case we have a concentration law as an immediate corollary (by Chebyshev's inequality).

Corollary 1 *As $n \to \infty$,*

$$\frac{\Delta_n}{\ln n} \xrightarrow{\mathcal{P}} \frac{2}{h_p}.$$

4.3 Limit laws

In principle, one can continue pumping higher moments by the methods utilized for the mean and variance, and aspire to determine limit distributions by a method of recursive moments (see Chern, Hwang and Tsai (2002), for example). However, as already mentioned, the explosive complexity is forbidding.

The contraction method offers a shortcut. Let

$$\Delta_n^* := \frac{\Delta_n - \mathbf{E}[\Delta_n]}{\sqrt{\ln n}}.$$

Based on some heuristics in the structure of the problem, a solution is guessed for the limit distribution of Δ_n^*. The guess is then verified by showing convergence of the distribution function to that of the guessed limit in some metric space. The contraction method was introduced by Rösler (1991). Rachev and Rüschendorf (1995) added several useful extensions. Recently general contraction theorems and multivariate extensions were added by Rösler (2001), and Neininger (2001). Rösler and Rüschendorf (2001) provide a valuable survey.

We start from the recursive decomposition (1), adapted in the form

$$\Delta_n = \Delta_{L_n} I_n + \tilde{\Delta}_{R_n} J_n + (\delta_{L_n} + \tilde{\delta}_{R_n} + 2)K_n + (\delta_{L_n} + 1)M_n + (\tilde{\delta}_{R_n} + 1)S_n,$$

where I_n, J_n, K_n, M_n, S_n are indicators of the mutually exclusive events that pick the right action and truncate all other. For example, I_n is the indicator of the event that both keys come from the left subtree. For $n \geq 2$, we can rewrite the latter relation in terms of the standardized variables:

$$\Delta_n^* = \Delta_{L_n}^* I_n \sqrt{\frac{\ln L_n}{\ln n}} + \tilde{\Delta}_{R_n}^* J_n \sqrt{\frac{\ln R_n}{\ln n}} + Y_n^* K_n + G_n, \tag{4}$$

where

$$Y_n^* := \frac{\delta_{L_n} - \mathbf{E}[\delta_{L_n}]}{\sqrt{\ln L_n}} \times \sqrt{\frac{\ln L_n}{\ln n}} + \frac{\tilde{\delta}_{R_n} - \mathbf{E}[\tilde{\delta}_{R_n}]}{\sqrt{\ln R_n}} \times \sqrt{\frac{\ln R_n}{\ln n}},$$

and

$$\begin{aligned} G_n \quad := \quad & \frac{1}{\sqrt{\ln n}} \Big(\mathbf{E}[\Delta_{L_n}] I_n + \mathbf{E}[\tilde{\Delta}_{R_n}] J_n + \big(\mathbf{E}[\delta_{L_n}] + \mathbf{E}[\tilde{\delta}_{R_n}] + 2\big) K_n \\ & + (\delta_{L_n} + 1)M_n + (\tilde{\delta}_{R_n} + 1)S_n - \mathbf{E}[\Delta_n] \Big). \end{aligned}$$

We first argue heuristically the existence of a limit for Δ_n^*. We then confirm our guess by an inductive proof in the Wasserstein metric space. By the sharp concentration of the binomial distribution of L_n we have

$$\frac{L_n}{n} \xrightarrow{a.s.} q, \qquad \frac{R_n}{n} \xrightarrow{a.s.} p, \tag{5}$$

and consequently

$$\sqrt{\frac{\ln L_n}{\ln n}} \xrightarrow{a.s.} 1, \qquad \sqrt{\frac{\ln R_n}{\ln n}} \xrightarrow{a.s.} 1. \tag{6}$$

If Δ_n^* converges to a limit, so would the ancillary variables $\Delta_{L_n}^*$ and $\tilde{\Delta}_{R_n}^*$, because both L_n and R_n grow to infinity almost surely, and these limits would be eventually independent. The limit variable δ^* of $(\delta_n - \mathbf{E}[\delta_n] \ln n) \ln^{-\frac{1}{2}} n$ is known to be $\mathcal{N}(0, \sigma_p^2)$ for biased digital search trees (it does not exist in unbalanced digital search trees); see Louchard and Szpankowski (1995). Similarly, $(\delta_{L_n} - \mathbf{E}[\delta_{L_n}] \ln n) \ln^{-\frac{1}{2}} L_n$ and $(\tilde{\delta}_{R_n} - \mathbf{E}[\delta_{R_n}] \ln n) \ln^{-\frac{1}{2}} R_n$, albeit dependency, would eventually be independent copies of $\mathcal{N}(0, \sigma_p^2)$.

The indicators (I_n, J_n, K_n) also tend to a vector (I, J, K) of three jointly distributed Bernoulli random variables on the nonzero vertices of the unit simplex in three dimensions, with marginals

$$I_n \xrightarrow{a.s.} I = \mathrm{Ber}(q^2), \qquad J_n \xrightarrow{a.s.} J = \mathrm{Ber}(p^2), \qquad K_n \xrightarrow{a.s.} K = \mathrm{Ber}(2pq). \tag{7}$$

The indicators M_n and S_n are much less probable than the former three, and we have $M_n \to 0$, and so does S_n.

Lemma 1 *As $n \to \infty$,*

$$G_n \xrightarrow{\mathcal{P}} 0.$$

Proof. Omitted. □

In view of (4)–(7) and Lemma 1, if Δ_n^* converges to a limit, say Δ^*, that limit would satisfy the distributional equation

$$\Delta^* \overset{\mathcal{L}}{=} \Delta^* I + \tilde{\Delta}^* J + Y^* K, \tag{8}$$

with $Y^* \overset{\mathcal{L}}{=} \delta^* + \tilde{\delta}^*$, and $(\Delta^*, \tilde{\Delta}^*, Y^*)$ independent of (I, J, K). This can be rigorously justified by showing that the sequence of distribution functions corresponding to the random variables Δ_n^* converges in the Wasserstein metric space to the distribution of Δ^*.

Theorem 1 *In a digital search tree of n random keys following the biased Bernoulli model, the distance Δ_n between two randomly selected keys satisfies*

$$\frac{\Delta_n - \frac{2}{h_p} \ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\sigma_p^2).$$

Proof. Let $\phi_W(t)$ be the moment generating function of W. The limiting random variable Δ^* of Δ_n^* has a distribution that satisfies the distributional equation (8). Conditioning on $\mathbf{M} = (I, J, K)$, we find the representation

$$\begin{aligned}
\phi_{\Delta^*}(t) &= \mathbf{E}\big[e^{t(I\Delta^* + J\tilde{\Delta}^* + Y^* K)}\big] \\
&= \mathbf{E}\big[e^{t(I\Delta^* + J\tilde{\Delta}^* + Y^* K)} \,|\, \mathbf{M} = (1,0,0)\big]\, \mathbf{P}\big(\mathbf{M} = (1,0,0)\big) \\
&\quad + \mathbf{E}\big[e^{t(I\Delta^* + J\tilde{\Delta}^* + Y^* K)} \,|\, \mathbf{M} = (0,1,0)\big]\, \mathbf{P}\big(\mathbf{M} = (0,1,0)\big) \\
&\quad + \mathbf{E}\big[e^{t(I\Delta^* + J\tilde{\Delta}^* + Y^* K)} \,|\, \mathbf{M} = (0,0,1)\big]\, \mathbf{P}\big(\mathbf{M} = (0,0,1)\big) \\
&= q^2 \phi_{\Delta^*}(t) + p^2 \phi_{\Delta^*}(t) + 2pq\phi_{Y^*}(t).
\end{aligned}$$

Thus,

$$\phi_{\Delta^*}(t) = \phi_{Y^*}(t),$$

and $\Delta^* = \delta^* + \tilde{\delta}^*$; both δ^* and $\tilde{\delta}^*$ are independent copies of the limit of the (normalized) random depth, which is known to be $\mathcal{N}(0, \sigma_p^2)$, see Louchard and Szpankowski (1995). That is, $\Delta^* \overset{\mathcal{L}}{=} \mathcal{N}(0, 2\sigma_p^2)$. □

5 Tries

In a random trie, let L_n and R_n be respectively the number of keys residing in the left and right subtrees, among the n keys stored in the tree (so, $L_n + R_n = n$). Given L_n, Δ_n can be Δ_{L_n} we have the conditional recurrence

$$\Delta_n \mid L_n = \begin{cases} \Delta_{L_n}, & \text{with probability } \dfrac{\dbinom{L_n}{2}}{\dbinom{n}{2}}; \\[2ex] \tilde{\Delta}_{R_n}, & \text{with probability } \dfrac{\dbinom{R_n}{2}}{\dbinom{n}{2}}; \\[2ex] (\delta_{L_n}+1)+(\tilde{\delta}_{R_n}+1), & \text{with probability } \dfrac{L_n R_n}{\dbinom{n}{2}}, \end{cases} \qquad (9)$$

with boundary condition $\Delta_0 = \Delta_1 = \delta_0 = \delta_1 = 0$. Here $\tilde{\Delta}_{R_n}$ is conditionally independent of Δ_{L_n} and $\tilde{\delta}_{R_n}$ is conditionally independent of δ_{L_n}.

5.1 Functional equations

We begin by deriving a functional equation for the moment generating function Δ_n from the basic conditional recurrence (9):

$$\begin{aligned} \binom{n}{2}\phi_{\Delta_n}(t) := \binom{n}{2}\mathbf{E}\big[e^{\Delta_n t}\big] &= \mathbf{E}\left[\binom{L_n}{2}e^{\Delta_{L_n}t}\right] + \mathbf{E}\left[\binom{R_n}{2}e^{\Delta_{R_n}t}\right] \\ &\quad + e^{2t}\mathbf{E}\left[L_n R_n e^{(\delta_{L_n}+\tilde{\delta}_{R_n})t}\right]. \end{aligned}$$

Direct work with $\Phi(t,z) = e^{-z}\sum_{k=0}^{\infty}\binom{n}{2}\mathbf{E}[e^{\Delta_n t}]z^n/n!$ gives rise to technical difficulty in the Mellin transform. To ensure the existence of the transform, we shift $\Phi(t,z)$ down by $e^{2t}\frac{z^2}{2}$, and define

$$P(t,z) = \Phi(t,z) - e^{2t}\frac{z^2}{2}.$$

We can now express the recurrence in the form

$$P(t,z) = P(t,pz) + P(t,qz) + e^{2t}\big[\big(Q(t,pz)+pz\big)\big(Q(t,qz)+qz\big)\big] - e^{2t}pqz^2, \qquad (10)$$

where $Q(t,z)$ is the shifted poissonized function $\mathbf{E}[N_z e^{\delta_{N_z}t}] - z$ for the random depth.

5.2 The Mean

The kth derivative of (10) yields a functional equation for the (shifted poissonized) kth moment of Δ_n. The first derivative gives

$$A(z) := \frac{\partial}{\partial t}P(t,z)\Big|_{t=0} = \mathbf{E}\left[\binom{N_z}{2}\Delta_{N_z}\right] - z^2.$$

And so,

$$A(z) = A(pz) + A(qz) + pza(qz) + qza(pz),$$

where $a(z) := \frac{\partial}{\partial t}Q(t,z)|_{t=0} = \mathbf{E}[N_z\delta_{N_z}]$. The Mellin transform of $A(z)$ is

$$A^*(s) = \frac{(qp^{-1-s}+pq^{-1-s})a^*(s+1)}{1-p^{-s}-q^{-s}},$$

where $a^*(s)$ is the Mellin transform of $a(z)$, which can be found in a number of sources. It is developed in Szpankowski (2001) via a shortcut argument that avoids recurrence and uses poissonization as a paradigm. These references give

$$a^*(s) = -\frac{\Gamma(s+1)}{1 - p^{-s} - q^{-s}}.$$

Upon plugging in $a^*(s)$ we get the Mellin transform

$$A^*(s) = -\frac{(pq^{-(s+1)} + qp^{-(s+1)})\Gamma(s+2)}{(1 - p^{-s} - q^s)(1 - p^{-(s+1)} - q^{-(s+1)})},$$

existing in $-3 < \Re\, s < -2$. Inverting the Mellin transform and going through depoissonization we obtain the following result.

Proposition 2 *In a trie of n random keys following the Bernoulli model, the average distance between two randomly selected keys is*

$$\begin{aligned}\mathbf{E}[\Delta_n] &= \frac{2}{h_p}\ln n + \frac{2\gamma}{h_p} + 2 - \frac{1}{pqh_p^2}\big(p^3 \ln^2 p + 2pq \ln p \ln q + q^3 \ln^2 q\big) \\ &\quad +4\beta(n) + o(1),\end{aligned}$$

where $\beta(n)$ is a small oscillating function. The variance is

$$\mathbf{Var}[\Delta_n] = 2\frac{pq}{h_p^3}(\ln p - \ln q)^2 \ln n + O(1) := 2\tilde{\sigma}_p^2 \ln n + O(1).$$

The case $p = q$ presents a degeneracy, which was handled in Christophi and Mahmoud (2005), where the details of the $O(1)$ term are specified, and where it is proved that no limit exists.

Corollary 2

$$\frac{\Delta_n}{\ln n} \xrightarrow{\mathcal{P}} \frac{2}{h_p}.$$

Proof. By Chebyshev's inequality. □

By an argument similar in its general gist to the one we used for the DST, but differing in many of its details we arrive at the main result for inter-distance in random tries.

Theorem 2 *In a trie of n random keys following the biased Bernoulli model, the distance Δ_n between two randomly selected keys satisfies*

$$\frac{\Delta_n - \frac{2}{h_p}\ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\tilde{\sigma}_p^2).$$

References

Aguech, R., Lasmar, N., and Mahmoud, H. (2005a). Distances in random digital search trees (submitted).

Aguech, R., Lasmar, N., and Mahmoud, H. (2005b). Limit distribution of distances in biased random tries (submitted).

Chern, H., Hwang, H. and Tsai, T. (2002). An asymptotic theory for Cauchy-Euler differential equations with applications to the analysis of algorithms. *Journal of Algorithms*, **44**, 177–225.

Christophi, C. and Mahmoud, H. (2005). The oscillatory distribution of distances in random tries. *The Annals of Applied Probability* (to appear).

Coffman, E. and Eve, J. (1970). File structures using hashing functions. *Communications of the ACM*, **13**, 427–432, and 436.

De La Briandais, R. (1959). File searching using variable length keys. *Proceedings of the Western Joint Computer Conference*, 295–298, AFIPS, San Francisco, California.

Devroye, L. and Neininger, R. (2004). Distances and finger search in random binary search trees. *SIAM Journal on Computing*, **33**, 647–658.

Fagin, R., Nievergelt, J., Pippenger, N., and Strong, H. (1979). Extendible hashing—a fast access method for dynamic files. *ACM Transactions on Database Systems*, **4**, 315–344.

Flajolet, P., Gourdon, X., Dumas, P. (1995). Mellin transform and asymptotic harmonic sums. *Theoretical Computer Science*, **144**, 3–58.

Fredkin, E. (1960). Trie memory. *Communications of the ACM*, **3**, 490–499.

Jacquet, P. and Szpankowski, W. (1998). Analytical depoissonization and its applications. *Theoretical of Computer Science*, **201**, 1–62.

Knuth, D. (1998). *The Art of Computer Programming*, Vol. 3: *Sorting and Searching*, 2nd ed. Addison-Wesley, Reading, Massachusetts.

Louchard, G. and Szpankowski, W. (1995). Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm. *IEEE Transactions on Information Theory*, **41**, 478–488.

Mahmoud, M. and Neininger, R. (2003). Distribution of distances in random binary search trees. *The Annals of Applied Probability*, **13**, 253–276.

Mathys, P. and Flajolet, P. (1985). Q-ary collision resolution algorithms in random-access systems with free and blocked channel access. *IEEE Transactions on Information Theory*, **31**, 217–243.

Neininger, R. (2001). On a multivariate contraction method for random recursive structures with applications to Quicksort. *Random Structures Algorithms*, **19**, 498–524.

Neininger, R. (2002). The Wiener index of random trees. *Combinatorics, Probability and Computing*, **11**, 587–597.

Panholzer, A. and Prodinger, H. (2004). Spanning tree size in random binary search trees. *The Annals of Applied Probability*, **14**, 718–733.

Poblete, P., Papadakis, T. and Munro, I. (1995). The binomial transform and its applications to the analysis of skip lists. In *Proc. ESA 95, Lecture Notes in Computer Science*, **979**, 554-569. Springer-Verlag, New York.

Rachev, S. and Rüschendorf, L. (1995). Probability metrics and recursive algorithms. *Advances in Applied Probability*, **27**, 770–799.

Rösler, U. (1991). A limit theorem for "Quicksort". *RAIRO Inform. Théor. Appl.*, **25**, 85–100.

Rösler, U. (2001). On the analysis of stochastic divide and conquer algorithms. *Algorithmica*, **29**, 238–261.

Rösler, U. and Rüschendorf, L. (2001). The contraction method for recursive algorithms. *Algorithmica*, **29**, 3–33.

Szpankowski, W. (2001). *Average Case Analysis of Algorithms on Sequences*. Wiley, New York.