

Distances in random digital search trees

Rafik Aguech · Nabil Lasmar · Hosam Mahmoud

Received: 29 July 2005 / Revised: 12 April 2006 /
Published online: 22 September 2006
© Springer-Verlag 2006

Abstract Distances between nodes in random trees is a popular topic, and several classes of trees have recently been investigated. We look into this matter in digital search trees. 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. In addition to various asymptotics, we give an exact expression for the mean and for the variance in the unbiased case. 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; it is prudent to seek a shortcut to the limit 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 of a distributional equation (distances being measured in the Wasserstein metric space). 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

R. Aguech

Département de mathématiques, Faculté des Sciences de Monastir, 5019 Monastir, Tunisia
e-mail: rafikaguech@ipeit.rnu.tn

N. Lasmar

Département de mathématiques, Institut préparatoire aux études d'ingénieurs de Tunis, IPEIT,
Rue Ielnahrou-Montfleury, Tunis, Tunisia
e-mail: nabillasmar@yahoo.fr

H. Mahmoud (✉)

Department of Statistics, The George Washington University, Washington, DC 20052, USA
e-mail: hosam@gwu.edu

AMS Subject Classifications Primary: 05C05 · 60C05; secondary: 60F05 · 68P05 · 68P10 · 68P20

1 Introduction

The behavior of distances between nodes in random trees has lately become a topic of interest, as can be seen in half a dozen or so of recent papers. Neininger [21] looked at these distances in recursive trees, Mahmoud and Neininger [18] studied these distances in binary search trees. The method used in these two papers was contraction. Devroye and Neininger [5] revisited the subject of binary search trees and refined the results using more elementary arguments. Their paper also takes up several other types of distances not considered in [18]. Employing generating functions and ensuing functional equations, Panholzer and Prodinger [23] generalized the result to the size of spanning trees for a randomly chosen set of nodes. Christophi and Mahmoud [3] considered the unbiased “tries” and demonstrated that the distribution of distances is oscillatory, with no possible nontrivial limit. Aguech, Lasmar and Mahmoud [1] considered the contrast that one encounters in biased tries, showing the existence of a Gaussian limit for a (normalized) version of the distance. A disparate variety of methods has been used in these studies.

Distances in yet another natural class of digital trees remain uninvestigated. It is the class of digital search trees, a class similar to tries, but the keys are kept in the nodes, instead of using them only as indexes for branching as in tries. The construction algorithm of the digital search tree seems more natural than that of the trie, and has an advantage of putting an upper bound on the size of the tree, whereas tries can degenerate into near linear structures with very long paths (with low probability of course).

We wish to study distances in a digital search tree growing on n keys. Some tough recurrences appear in the study and seem to be very challenging. The idea of poissonization is well-suited for such a recurrence. The method has become a popular tool. It involves a poissonization-Mellin-inverse Mellin-depoissonization program. The method is beginning to appear as a chapter on standard techniques in information science; see [31] for example, and numerous references within. Broadly speaking, the program works as follows. If a Poisson number of keys is assumed instead, the functional equations involved can asymptotically be solved by the Mellin transform and its inverse. The solution is a good approximation (with ignorable errors) for the fixed-population problem, when the Poisson parameter is taken to be n , as $n \rightarrow \infty$.

The complexity of such a Mellin approach increases considerably for higher moments, which strongly invites a consideration of some shortcut. The contraction method provides such a direct bridge to the limit.

The standard data model for digital search trees is the Bernoulli probability distribution, which should ideally be unbiased. In practice this unbiasedness is not guaranteed in view of the sensitivity of data generators. So, our study is not limited to the unbiased Bernoulli case, and puts in good perspective the contrast between biased and unbiased data models. We assume that keys of infinite precision are obtained from a memoryless source that emits independent bits, with $\mathbf{P}(\text{Bit} = 1) = p$, and $\mathbf{P}(\text{Bit} = 0) = q = 1 - p$. We say the Bernoulli model (or the resulting digital search tree) is unbiased if $p = q = \frac{1}{2}$, otherwise it is biased. It is desirable, but not guaranteed,

to have an unbiased Bernoulli model to have a data sample distributed like a sample from a standard uniform distribution, a realistic assumption under hashing schemes, whose primary goal is to achieve uniformity.

Several aspects of the depth of a randomly selected node (its distance from the root) in a random digital search tree can be found in various sources. The subject was revisited again and again in [6,11,12,14–16,24,30] etc. In this paper we study the distance between randomly selected pairs of nodes. The main result of this paper shows that a Gaussian limit exists for appropriately centered and scaled distances in biased digital search trees.

Here is a plan for the remaining parts. In Sect. 2, we give an overview of digital search trees and their algorithmic aspects. We give there an illustrating example. At the end of Sect. 2, notation used throughout is explained. In Sect. 3 the key results are presented, so that they can be found without much effort. However, several important corollaries (particularly relating to the special important unbiased case) are relegated to later sections. In Sect. 4, we show how the moments can be derived by a poissonization-Mellin-inverse Mellin-depoissonization program. In Sect. 4.1 a functional equation for the moment generating function is set up. From this functional equation, the mean is derived in Sect. 4.2, and the variance is set up in Sect. 4.3. The required residue calculation is taken up in Sect. 4.4, where exact and asymptotic results on variance are reported. The computational complexity increases considerably with every higher moment. Though doable in principle, it is not feasible to continue pumping the higher moments. A shortcut toward the limit distribution is needed. A Gaussian limit distribution is derived by the contraction method in Sect. 5, where we also add a few words on the origin of the method, its use, and references to it. Section 6 concludes the paper with a brief discussion on the contrast between biased and unbiased digital search trees. The appendix gives the preparatory work on the random depth necessary for distances between pairs. The required functional equations for the random depth are developed there.

2 Digital search trees

The digital search tree was invented in [4]. It is a natural data structure for digital data when their digital composition is available. This type of data abounds in science, engineering and technology. For instance, they are the building blocks of computer files. DNA strands are basically strings on a 4-letter alphabet of protein nucleotides, and so forth. It is therefore natural for an easy-to-use and easy-to-maintain supporting data structure, such as the digital search tree, to be quite popular. The digital search tree also provides a model for the analysis of several important algorithms, such as the Lempel-Ziv parsing algorithm (see [15]), and conflict resolution (see [19]).

For ease of exposition, we shall deal with the binary case. Generalization to larger alphabets should not be hard. The binary digital search tree grows according to an algorithm. The keys K_1, K_2, \dots, 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 for a position. A main

distinction between this 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.

As an illustration, suppose that $n = 5$, and that the data are

- $K_1 = 10101 \dots$
- $K_2 = 10011 \dots$
- $K_3 = 00010 \dots$
- $K_4 = 00110 \dots$
- $K_5 = 10001 \dots$

The algorithm places K_1 in the root; and K_2 and K_3 respectively go into the right and left subtrees as right and left children of the root. When K_4 joins, it goes to the left of the root, then to the left of K_3 , etc.

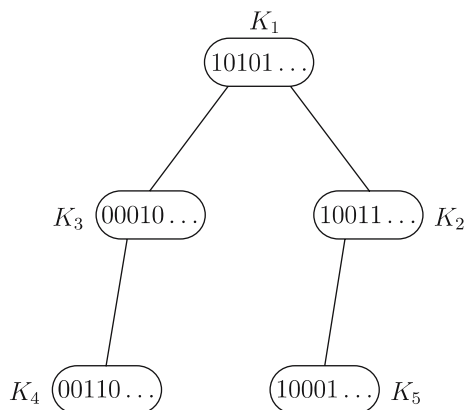
Binary data can always be scaled to be in the interval $[0, 1]$, by just viewing the string of bits in a key as the binary expansion of some real number in that interval (placing a binary point immediately before the leftmost digit).

Let δ_n be the depth of a randomly selected node in a random digital search tree of size n , with *random* meaning that all nodes are equally likely choices. For instance, in the trie T_5 of Fig. 1 the depth of K_3 is 1, $\mathbf{E}[\delta_5 | T_5] = (0+1+1+2+2)/5 = \frac{6}{5}$. Likewise, let Δ_n be the distance between two randomly selected nodes in a digital search tree of size n , where all $\binom{n}{2}$ pairs of keys are equally likely. For instance, the distance between K_2 and K_3 is 2, and $\mathbf{E}[\Delta_5 | T_5] = (1+1+2+2+2+3+1+1+3+4)/\binom{5}{2} = 2$. In both definitions note the double randomness (random tree, and a random distance in it).

A few additional symbols will facilitate our exposition. The symbol $\stackrel{\mathcal{L}}{=}$ will mean equality in law, while $\xrightarrow{\mathcal{L}}$ will denote convergence in law. Likewise, $\xrightarrow{\mathcal{P}}$, and $\xrightarrow{a.s.}$ will respectively denote convergence in probability and almost surely. A tilded variable will refer to an independent copy of an untilded random variable having the same distribution. For example, \tilde{Y} will mean a random variable independent of Y , with $\tilde{Y} \stackrel{\mathcal{L}}{=} Y$.

The Bernoulli random variable with success probability p will be denoted by $\text{Ber}(p)$. Similarly, the binomial random variable arising on n independent trials, with rate of success p per trial, will be denoted by $\text{Bin}(n, p)$, and the normal distribution with mean

Fig. 1 A digital search tree on 5 keys



μ and variance σ^2 will be denoted by $\mathcal{N}(\mu, \sigma^2)$. We shall let $\phi_X(t) = \mathbf{E}[e^{Xt}]$ be the moment generating function of a generic random variable X .

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

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

and will be denoted by $f^*(s)$. The Mellin transform usually exists in vertical strips in the s complex plane of the form

$$a < \Re s < b$$

for real numbers $a < b$. We shall denote this strip by $\langle a, b \rangle$. The function $f(x)$ can be recovered from its transform by a line integral

$$f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} f^*(s)x^{-s} ds$$

for any $c \in \langle a, b \rangle$. For a survey of the Mellin transform in the context of the analysis of algorithms we refer the reader to the comprehensive survey in [7].

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 standard sources such as Szpankowski [31].

The two functions

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

and the data entropy

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

are instrumental to our presentation. We shall also need the two functions

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

3 Main results

Our results have two flavors, some are exact and some are asymptotic as n , the number of keys in the tree, grows very large.

Theorem 1 *In a random digital search tree grown on $n \geq 3$ keys, the average distance between a randomly selected pair of nodes is given by*

$$\mathbf{E}[\Delta_n] = 1 + \frac{2}{n(n-1)} \sum_{k=3}^n \binom{n}{k} (-1)^k \lambda(-k),$$

where $\lambda(\cdot)$ is given by (6).

Theorem 2 *In a random digital search tree of n random keys, asymptotically, as $n \rightarrow \infty$:*

(a) The average distance between two randomly selected keys is

$$\mathbf{E}[\Delta_n] = \frac{2}{h_p} \ln n + \frac{\hat{h}_p}{pqh_p} + \frac{\tilde{h}_p}{h_p^2} - \frac{2(1-\gamma)}{h_p} + \frac{\ln(pq)}{h_p} - \frac{2\alpha_\infty}{h_p} + 2 - 2\xi_q(\ln n) + O\left(\frac{1}{n^{0.49999}}\right),$$

where $\xi_q(\cdot)$ is the oscillating function given in (7).

(b) The variance is

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

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

The main result, proved by the contraction method, is the following.

Theorem 3 *In a digital search tree of n random keys following the biased Bernoulli model, the distance Δ_n between two randomly selected keys satisfies*

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

4 Moments of the random distance

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, $L_n + R_n = n - 1$). Recall that the independent keys are generated from a Bernoulli model, where the individual bits have the probabilities $\mathbf{P}(\text{Bit} = 1) = p$, and $\mathbf{P}(\text{Bit} = 0) = q = 1 - p$, or in other words $L_n \stackrel{\mathcal{L}}{=} \text{Bin}(n - 1, q)$. The independence of the keys and the bits within, and the recursive manner of insertion of keys in the tree induce a probability structure in the subtrees similar to that of the whole tree, but defined on the respective sizes of the subtrees.

Given L_n , the distance Δ_n can be Δ_{L_n} with probability $\binom{L_n}{2} / \binom{n}{2}$ when both keys come from the left subtree, can be $\tilde{\Delta}_{R_n}$ with probability $\binom{R_n}{2} / \binom{n}{2}$ when both keys come from the right subtree, or can be $(\delta_{L_n} + 1) + (\tilde{\delta}_{R_n} + 1)$ with probability $L_n R_n / \binom{n}{2}$ when the two keys come from different subtrees. There is also a possibility that the root and one node from the left subtree are the two nodes chosen, with probability $L_n / \binom{n}{2}$. Likewise, the two chosen nodes can be the root and a node from the right subtree, with probability $R_n / \binom{n}{2}$. We have a conditional distribution (given L_n):

$$\Delta_n | L_n = \begin{cases} \Delta_{L_n} & \text{with probability } \frac{\binom{L_n}{2}}{\binom{n}{2}}, \\ \tilde{\Delta}_{R_n} & \text{with probability } \frac{\binom{R_n}{2}}{\binom{n}{2}}, \\ (\delta_{L_n} + 1) + (\tilde{\delta}_{R_n} + 1) & \text{with probability } \frac{L_n R_n}{\binom{n}{2}}, \\ \delta_{L_n} + 1 & \text{with probability } \frac{L_n}{\binom{n}{2}}, \\ \tilde{\delta}_{R_n} + 1 & \text{with probability } \frac{R_n}{\binom{n}{2}}. \end{cases} \tag{1}$$

We used the tilde notation (explained in Section 2) for Δ_{L_n} and $\tilde{\Delta}_{R_n}$ to emphasize their *conditional independence*—it is true that these random variables are dependent through the dependency of L_n and R_n , but given the value of L_n (and hence R_n) the two are conditionally independent. That is to say Δ_j and $\tilde{\Delta}_k$ are independent for any specified pair of indexes j and k . This follows from our remark on the stochastic independence in the subtrees. The same applies to δ_{L_n} and $\tilde{\delta}_{R_n}$.

4.1 Functional equations

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

$$\binom{n}{2} \mathbf{E}[e^{\Delta_n t} | L_n] = e^{\Delta_{L_n} t} \binom{L_n}{2} + e^{\tilde{\Delta}_{R_n} t} \binom{R_n}{2} + e^{(\delta_{L_n}+1)t+(\tilde{\delta}_{R_n}+1)t} L_n R_n + e^{(\delta_{L_n}+1)t} L_n + e^{(\tilde{\delta}_{R_n}+1)t} R_n.$$

From the above equation we get

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

By the fact that L_n and R_n are binomially distributed and the aforementioned conditional independence, we obtain

$$\begin{aligned} \binom{n}{2} \phi_{\Delta_n}(t) &= \sum_{\ell=0}^{n-1} \binom{\ell}{2} \phi_{\Delta_\ell}(t) \binom{n-1}{\ell} p^{n-1-\ell} q^\ell \\ &+ \sum_{r=0}^{n-1} \binom{r}{2} \phi_{\tilde{\Delta}_r}(t) \binom{n-1}{r} p^r q^{n-1-r} \\ &+ e^{2t} \sum_{\ell=0}^{n-1} \binom{n-1}{\ell} \ell (n-1-\ell) \phi_{\delta_\ell}(t) \phi_{\tilde{\delta}_{n-1-\ell}}(t) p^{n-1-\ell} q^\ell \\ &+ e^t \sum_{\ell=0}^{n-1} \ell \binom{n-1}{\ell} \phi_{\delta_\ell}(t) p^{n-1-\ell} q^\ell \\ &+ e^t \sum_{r=0}^{n-1} r \binom{n-1}{r} \phi_{\tilde{\delta}_r}(t) p^r q^{n-1-r}. \end{aligned} \tag{2}$$

To handle this recurrence, we introduce a super moment generating function for any real t , and $z \in \mathbb{C}$:

$$\Phi(t, z) = \sum_{n=0}^{\infty} \binom{n}{2} \phi_{\Delta_n}(t) \frac{z^n}{n!}.$$

Upon multiplying both sides of the recurrence (2) by $z^{n-1}/(n-1)!$, then summing over n , one formulates a functional equation:

$$\begin{aligned} \frac{\partial}{\partial z} \Phi(t, z) &= e^{qz} \Phi(t, pz) + e^{pz} \Phi(t, qz) + e^{2t} \phi(t, pz) \phi(t, qz) \\ &+ e^t e^{pz} \phi(t, qz) + e^t e^{qz} \phi(t, pz), \end{aligned} \tag{3}$$

where $\phi(t, z)$ is a super moment generating function for the random depth δ_n . The appendix gives functional equations for this poissonized function.

To obtain asymptotically the mean and the variance, we use poissonization techniques and the Mellin transform. Let $N(z)$ be distributed like a Poisson random variable with mean z , and put

$$\Psi(t, z) = e^{-z} \Phi(t, z) = \mathbf{E} \left[\binom{N(z)}{2} e^{\Delta_{N(z)} t} \right].$$

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 $\psi(t, z) = e^{-z} \phi(t, z)$.

4.2 The mean

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

$$X'(z) + X(z) = X(pz) + X(qz) + pz x(qz) + qz x(pz) + x(pz) + x(qz) + pqz^2, \quad (4)$$

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 $x^*(s)$ are given in the appendix. Note that we defined $X(z)$ as a poissonized average with a shift to ensure the existence of its Mellin transform.

Lemma 1 *The Mellin transforms of $X(z)$ and $X'(z) - \frac{z^2}{2}$ exist in the strip $\langle -3, -2 \rangle$, and $(X'(z) - \frac{z^2}{2})^*(s) = -(s - 1)X^*(s - 1)$.*

Proof Being the parameter of a Poisson distribution, the variable z is positive real. It suffices to show that $X(z)$ is $O(z^2)$, as $z \rightarrow 0$, and is $O(z^3)$, as $z \rightarrow \infty$. Recalling that

$$X(z) = \sum_{n=0}^{\infty} \frac{z^n e^{-z} \mathbf{E}[\Delta_n]}{n!} \binom{n}{2} - \frac{z^2}{2},$$

an upper bound can be established from the fact that $\Delta_n < n$ in a digital search tree of size n :

$$|X(z)| < \sum_{n=0}^{\infty} \frac{z^n e^{-z} n}{n!} \binom{n}{2} = \frac{1}{2} z^3 + \frac{3}{2} z^2.$$

The argument for the Mellin transform of $X'(z) = e^{-z} \sum_{n=0}^{\infty} \frac{(n-z)z^{n-1} \mathbf{E}[\Delta_n]}{n!} \binom{n}{2} - z$, is quite similar to show that the Mellin transform of $X'(z)$ exists in $\langle -2, -1 \rangle$. As argued in [7], the adjustment $-z^2/2$ only shifts the fundamental strip of the Mellin transform without changing its value. That is, $(X'(z) - \frac{z^2}{2})^* = (X'(z))^*$, in the strip $\langle -3, -2 \rangle$, with $(X'(z))^*$ meaning the meromorphic continuation of the Mellin transform of $X'(z)$. □

An informative rearrangement of (4) is helpful to our purpose:

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

Taking the Mellin transform of the latter equation, we obtain

$$\begin{aligned} -(s-1)X^*(s-1) + X^*(s) &= (p^{-s} + q^{-s})X^*(s) + (pq^{-s} + qp^{-s})x^*(s+1) \\ &\quad + (p^{-s} + q^{-s})x^*(s), \end{aligned} \tag{5}$$

where, as derived in the appendix,

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

An argument for the Mellin transform of the shifted function $x(z) - z^2/2$, as the one we used in Lemma 1 for $X'(z) - \frac{1}{2}z^2$, shows that $x(z) - z^2/2$ also has the Mellin transform $x^*(s)$ in $\langle -3, -2 \rangle$, where $x^*(s)$ is the meromorphic continuation of the definition of $x^*(s)$ in $\langle -2, -1 \rangle$ (see [7]). And so, all the Mellin transforms involved in (5) exist in the strip $\langle -3, -2 \rangle$.

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

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

By (5), $\lambda(s)$ must satisfy

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

After some tedious iterative algebra we get

$$\lambda(s) = \frac{Q_\infty(-2)}{Q_\infty(s)} \left(2\kappa_\infty - \frac{1}{2pq} + \sum_{k=0}^\infty \frac{T(s-k)}{1 - p^{k-s} - q^{k-s}}\right), \tag{6}$$

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

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

We can obtain an exact expression for the mean, using the following lemma.

Lemma 2 *Let $\{f_n\}_{n=0}^\infty$ be a sequence of real numbers, and suppose that its Poisson transform $F(z) = \sum_{n \geq 0} f_n \frac{z^n}{n!} e^{-z}$ is an entire function. Furthermore, suppose the Mellin*

transform $F^*(s)$ has the factorization $F^*(s) = \Lambda(s)\Gamma(s)$, and assume $F^*(s)$ exists in the strip $\langle -3, -2 \rangle$ while $\Lambda(s)$ is analytic in the strip $\langle -\infty, -2 \rangle$. Then

$$\Lambda(-n) = \sum_{k=0}^n \binom{n}{k} (-1)^k f_k \quad \text{for } n \geq 2.$$

Proof The proof is similar to that of Louchard, Szpankowski and Tang [16]. □

We can use the binomial transform to invert the relation in Lemma 2 and represent f_n in terms of Λ_k . An exact expression for the mean ensues as stated in Theorem 1.

Of particular interest is an application of the result of Theorem 1 in the unbiased case.

Corollary 1 *In a symmetric digital search tree,*

$$\mathbf{E}[\Delta_n] = \frac{n+1}{3} + \frac{2}{n(n-1)} \sum_{k=4}^n (-1)^k \binom{n}{k} \left(1 - \frac{4+4k}{2^k}\right) Q_{k-4}(-2).$$

Proof (sketch) The result follows directly from Theorem 1 and the identity

$$1 + \sum_{j=0}^{\infty} \frac{1}{2^{j+k-1}-1} \left(1 - \frac{j+k}{1-2^{2-k-j}}\right) = 1 - \frac{k}{2^{k-2}-1}. \quad \square$$

Our interest in the sequel is to obtain an asymptotic equivalent for the mean. Let $s_{k,\ell}$, for $k, \ell \in \mathbb{Z}$ be the roots of $1 - p^{k-s} - q^{k-s} = 0$; observe that one solution is $s_{k,0} = k - 1$. The roots $s_{k,\ell}$ have been studied quite extensively. The following special case is well known (e.g., see [31], who attributes the result to [9,29]).

Lemma 3 *There are countably many solutions of $1 - p^{-s} - q^{-s} = 0$. Let b be a solution of $1 - p^{-s} - q^{-s} = 0$. Then*

- (i) $-1 \leq \Re b \leq \chi_0$, where χ_0 is a positive real solution of $1 + p^{-s} - q^{-s} = 0$.
- (ii) If $\Re b = -1$ and $\Im b \neq 0$, then $\frac{\ln p}{\ln q}$ must be rational. In fact, if $\frac{\ln p}{\ln q} = \frac{r}{m}$ (where $\gcd(r, m) = 1$ for integers r and m), there are infinitely many solutions b_k , that are in the form

$$b_k = -1 + \frac{2\pi i k m}{\ln q} \quad \text{for } k = 0, \pm 1, \pm 2, \dots,$$

there are no other solutions.

To proceed we need an inverse Mellin transform for $X^*(s)$. The line integral in the inverse transform can be computed on a shifted line to the right, after compensating for the shift with the residue of the poles between the two lines. The poles give the main contribution; the integration on the shifted line only introduces a negligible error. If two contributing poles are of the same multiplicity, the one with the closer real part to the original line of integration has the asymptotically dominant contribution among the two. By now all this is a well understood fact, that is well documented in books such as [31].

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 poles of $X^*(s)z^{-s}$ with -2 as real part are: a double pole $s_0 = -2$, and a simple pole at each of the points $s_k = -2 + \frac{2ikm\pi}{\ln q}$,

$k = \pm 1, \pm 2, \dots$. With the help of a computer algebra system, such as MAPLE, one obtains

$$\begin{aligned} \operatorname{Res}_{s=-2} X^*(s)z^{-s} &= -\frac{z^2 \ln z}{h_p} + \left(-\frac{p^2 \ln p + q^2 \ln q}{2pqh_p} - \frac{p \ln^2 p + q \ln^2 q}{2h_p^2} \right. \\ &\quad \left. + \frac{(1-\gamma)}{h_p} - \frac{\ln(pq)}{2h_p} + \frac{\alpha_\infty}{h_p} - \frac{1}{2} \right) z^2, \end{aligned}$$

where $\gamma = 0.77215\dots$ is Euler’s constant, and

$$\begin{aligned} \alpha_\infty &= \frac{d}{ds} \left(\frac{Q_\infty(-2)}{Q_\infty(s)} \right)_{s=-2} = -\sum_{k=0}^\infty \frac{p^{k+2} \ln p + q^{k+2} \ln q}{1 - p^{k+2} - q^{k+2}}, \\ \kappa_\infty &= -\sum_{k=1}^\infty \frac{T(-2-k)}{2(1 - p^{k+2} - q^{k+2})}. \end{aligned}$$

and

$$\operatorname{Res}_{s=s_k} X^*(s)z^{-s} = -\frac{\Gamma(s_k + 1)}{h_p} z^2.$$

Consequently, for $\varepsilon > 0$,

$$\begin{aligned} X(z) &= \frac{z^2 \ln z}{h_p} + \left(\frac{\hat{h}_p}{2pqh_p} + \frac{\tilde{h}_p}{2h_p^2} - \frac{1-\gamma}{h_p} + \frac{\ln(pq)}{2h_p} \right. \\ &\quad \left. - \frac{\alpha_\infty}{h_p} + \frac{1}{2} - \xi_q(\ln z) \right) z^2 + O(z^{1-\varepsilon}), \end{aligned}$$

where $\xi_q(\cdot)$ is the function

$$\xi_q(u) = \begin{cases} -\frac{1}{h_p} \sum_{k \neq 0}^\infty \Gamma\left(-1 + \frac{2\pi i k m}{\ln q}\right) e^{-\frac{2\pi i k m}{\ln q} u} & \text{if } \frac{\ln p}{\ln q} \in \mathbb{Q}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

In all cases, $|\xi_q(u)|$ is a small function for the typical values of q staying away from 0 and 1. For instance, when $\frac{r}{m} = \frac{1}{2}$, we have $p^2 = q$, or $p = \frac{1}{2}(\sqrt{5} - 1) \approx 0.618$, and $|\xi_q(u)| \leq 0.2 \times 10^{-9}$, uniformly in u .

The conditions of the depoissonization lemma ([10]) can be readily checked, transforming the poissonized result for $X(z)$ back into a result for Δ_n , introducing only a small additional error. This completes a proof for part (a) of Theorem 2.

For $p = q = 1/2$,

$$\alpha_\infty = \ln 2 \sum_{k=1}^\infty \frac{1}{2^k - 1} = 1.1136762149\dots$$

In the symmetric case we have the following result.

Corollary 2 *In the symmetric digital search tree, the mean of Δ_n is given by*

$$\mathbf{E}[\Delta_n] = 2 \log_2 n - 1 + \frac{2(\gamma - 1 - \alpha_\infty)}{\ln 2} - 2\xi_{\frac{1}{2}}(\ln n) + O\left(\frac{1}{n^{0.49999}}\right).$$

4.3 The variance

An asymptotic computation of the second and higher moments can follow the same route as that for the mean. However, the technical difficulty increases considerably, and the computation is bogged down with plentiful algebraic details. Even the variance is too complicated. For this reason, we pursue only the second moment, and leave the full characterization of all moments (i.e. the asymptotic distribution) to be dealt with by another method. The contraction method is adept for the task, as discussed in Sect. 5. However, the contraction method could not be started without at least a good guess on the mean and variance. The discussion here is merely a sketch.

The starting point for the computation of the mean involved the first derivative of (3) with respect to t , at $t = 0$, which gave the differential functional equation (4). Likewise, the second derivative of (3) gives us a differential functional equation for the shifted poissonized second moment

$$W(z) = \mathbf{E} \left[\binom{N(z)}{2} \Delta_{N(z)}^2 \right] - \frac{z^2}{2}.$$

One obtains

$$\begin{aligned} W'(z) + W(z) &= W(pz) + W(qz) \\ &+ 4pzx(qz) + 4qzx(pz) + pzw(qz) + qzw(pz) \\ &+ 2\left(x(pz) - \frac{p^2z^2}{2}\right) + 2\left(x(qz) - \frac{q^2z^2}{2}\right) \\ &+ \left(w(pz) - \frac{p^2z^2}{2}\right) + \left(w(qz) - \frac{q^2z^2}{2}\right) \\ &+ 2x(pz)x(qz) + \frac{3}{2}z^2, \end{aligned} \tag{8}$$

where $w(z) = \mathbf{E}[N(z)\delta_{N(z)}^2]$ (see the appendix for a derivation via functional equations and Mellin transforms with delay); here $x(z)$ and all the other constants and functions are being reused from Sect. 4.2.

Let $W_0(z) = W(z) - x^2(z)$. Using the functional equation of $x(z)$ (from the appendix) to replace the cross-product $x(pz)x(qz)$, we find that $W_0(z)$ satisfies the following equation:

$$\begin{aligned} W'_0(z) + W_0(z) &= W_0(pz) + W_0(qz) \\ &+ 4pzx(qz) + 4qzx(pz) + pzw(qz) + qzw(pz) \\ &+ 2\left(x(pz) - \frac{p^2z^2}{2}\right) + 2\left(x(qz) - \frac{q^2z^2}{2}\right) \\ &+ \left(w(pz) - \frac{p^2z^2}{2}\right) + \left(w(qz) - \frac{q^2z^2}{2}\right) \\ &- z(x'(z) - z) - 2zx(z) + \frac{3}{2}z^2 \\ &+ (x'(z))^2 - zx'(z). \end{aligned}$$

To handle this equation, we decompose $W_0(z)$ into a dominant part $W_1(z)$, and an asymptotically negligible part $W_2(z)$, with $W_1(z) + W_2(z) = W_0(z)$, where $W_1(z)$ and

$W_2(z)$ respectively satisfy the following equations:

$$\begin{aligned} W'_1(z) + W_1(z) &= W_1(pz) + W_1(qz) \\ &\quad + 4pzx(qz) + 4qzx(pz) + pzw(qz) + qzw(pz) \\ &\quad + 2\left(x(pz) - \frac{p^2z^2}{2}\right) + 2\left(x(qz) - \frac{q^2z^2}{2}\right) \\ &\quad + \left(w(pz) - \frac{p^2z^2}{2}\right) + \left(w(qz) - \frac{q^2z^2}{2}\right) \\ &\quad - z(x'(z) - z) - 2zx(z) + \frac{3}{2}z^2, \end{aligned}$$

and

$$W'_2(z) + W_2(z) = W_2(pz) + W_2(qz) + (x'(z))^2 - zx'(z).$$

The rationale behind this partition is that an exact Mellin transform is not easy to obtain in the presence of $(x'(z))^2$. So, we subsume $(x'(z))^2$ into a part that is demonstrably negligible asymptotically, with respect to the other part which is clean for a complete Mellin transform. Toward this end we need to show the following asymptotic result for $W_2(z)$, in a form that is suitable for later dePoissonization.

Lemma 4 For all $\theta \in (0, \frac{\pi}{2})$,

$$W_2(z) = O(z^{3/2}) \quad \text{as } |z| \rightarrow \infty,$$

for $z \in S_\theta$, where $S_\theta = \{z \in \mathbb{C} : |\arg(z)| \leq \theta\}$.

Proof For positive integer m , let $D_m = \{z; \operatorname{Re}(z) \in [1, v^{-m}]\}$, and $v \leq \max(p, q)$. Write $W_2(z)$ as

$$W'_2(z) + W_2(z) = W_2(pz) + W_2(qz) + h(z),$$

where $h(z) = (x'(z))^2 - zx'(z) = O(z \ln z)$. We conclude that the above equation has the following solution:

$$W_2(z) = W_2(vz)e^{-z(1-v)} + e^{-z} \int_{vz}^z e^x (W_2(px) + W_2(qx) + h(x)) \, dx.$$

let V_m be the upper bound on $|\frac{W_2(z)}{z^{3/2}}|$ over the compact domain D_m in the convex cone S_θ , and V'_m be the upper bound on $|\frac{W_2(z)}{z^{3/2}}|$ over the domain $(D_{m+1} - D_m) \cap S_\theta$. Then

$$V_{m+1} \leq \max\{V_m, V'_m\},$$

showing that, for $z \in (D_{m+1} - D_m) \cap S_\theta$,

$$\begin{aligned} \left| \frac{W_2(vz)e^{-z(1-v)}}{z^{3/2}} \right| &\leq V_m v^{3/2} e^{-v^{-m}(1-v)}, \\ \left| e^{-z} \int_{vz}^z e^x \left(\frac{W_2(px) + W_2(qx)}{z^{3/2}} \right) \, dx \right| &\leq (p^{3/2} + q^{3/2}) V_m, \\ \left| e^{-z} \int_{vz}^z e^x \left(\frac{h(x)}{z^{3/2}} \right) \, dx \right| &\leq v^{\frac{m-1}{2}} (m-1) \ln\left(\frac{1}{v}\right). \end{aligned}$$

Then, V'_m satisfies

$$V'_m \leq V_m(1 + v^{3/2}e^{-v^{-m(1-v)}}) + v^{\frac{m-1}{2}}(m-1)\ln\left(\frac{1}{v}\right).$$

Hence, by induction V_m are uniformly bounded. □

We next deal with $W_1(z)$. This function has a Mellin transform in the strip $\langle -3, -2 \rangle$. It suffices to show that the entire quantity $W(z)$ is $O(z^2 \ln^2 z)$. This ensues from the following. The notations Δ_n and δ_n are of course meant as functions of sample space points in a probability space, which involves a product measure of trees and the set $\{1, \dots, n\} \times \{1, \dots, n\}$ (for the generation of a pair of keys). We need here a notation for the depth of a specific node and the distance between a given pair. Let $\delta_n(a) = \delta_n(a)(T)$ be the depth of node a in the tree T , and let $\delta_n(a, b) = \delta_n(a, b)(T)$ be the distance between the nodes a and b in T . We have $\Delta_n^2(a, b) \leq (\delta_n(a) + \delta_n(b))^2$, for any two nodes a and b in the tree. We can write $\mathbf{E}[\Delta_n^2] = O(\ln^2 n)$ (by the Cauchy–Schwartz inequality and the fact that $\mathbf{E}[\delta_n^2] = O(\ln^2 n)$). Then by the same argument used in Lemma 4, $W_1^*(s)$, the Mellin transform of $W_1(z)$, exists in the strip $\langle -3, -2 \rangle$ (we can show that $\mathbf{E}[\Delta_3^2] = 2$, then $W'_1(z) - \frac{3}{2}z^2 = O(z^3)$, as $z \rightarrow 0$). The Mellin transform $W_1^*(s)$ satisfies

$$\begin{aligned} W_1^*(s)(1 - p^{-s} - q^{-s}) &= (s - 1)W_1^*(s - 1) \\ &\quad + 2(2pq^{-1-s} + 2qp^{-1-s} - 1)x^*(s + 1) \\ &\quad + 2(p^{-s} + q^{-s})x^*(s) + (pq^{-1-s} + qp^{-1-s})w^*(s + 1) \\ &\quad + (p^{-s} + q^{-s})w^*(s) + sx^*(s), \end{aligned} \tag{9}$$

where

$$w^*(s) = \frac{Q_\infty(-2)}{Q_\infty(s)} \Gamma(s) \left(2 \sum_{k=0}^\infty \frac{q^{k-s} + p^{k-s}}{1 - p^{k-s} - q^{k-s}} + 1 - 2\beta_\infty \right),$$

and

$$\beta_\infty = \sum_{k=1}^\infty \frac{p^{k+1} + q^{k+1}}{1 - p^{k+1} - q^{k+1}}.$$

Writing $W_1^*(s) = \theta(s)\Gamma(s)$. By (9) the function θ must satisfy:

$$\begin{aligned} \theta(s)(1 - p^{-s} - q^{-s}) &= \theta(s - 1) + \frac{2(2pq^{-1-s} + 2qp^{-1-s} - 1)s\rho(s)}{1 - p^{-1-s} - q^{-1-s}} \\ &\quad + (s + 2(p^{-s} + q^{-s}))\rho(s) + (p^{-s} + q^{-s})\beta(s) \\ &\quad + (pq^{-1-s} + qp^{-1-s})s\beta(s + 1). \end{aligned}$$

4.4 Residue computation

What remains to be done for the variance is inverting the Mellin transform of the second moment. We start with the inversion formula, in its integral form, shift the line of integration to the right of the existence strip, compensate for the poles between the

two lines, etc. This results in the following residue calculation:

$$\begin{aligned} \operatorname{Res}_{s=-2} W_1^*(s)z^{-s} &= \operatorname{Res}_{s=-2} \frac{2(2pq^{-1-s} + 2qp^{-1-s} - 1)s\rho(s)}{(1 - p^{-1-s} - q^{-1-s})(1 - p^{-s} - q^{-s})} \Gamma(s)z^{-s} \\ &\quad + \operatorname{Res}_{s=-2} \frac{(s + 2(p^{-s} + q^{-s}))\rho(s)}{1 - p^{-s} - q^{-s}} \Gamma(s)z^{-s} \\ &\quad + \operatorname{Res}_{s=-2} \frac{(p^{-s} + q^{-s})\beta(s)}{1 - p^{-s} - q^{-s}} \Gamma(s)z^{-s} \\ &\quad + \operatorname{Res}_{s=-2} \frac{(pq^{-1-s} + qp^{-1-s})s\beta(s + 1)}{1 - p^{-s} - q^{-s}} \Gamma(s)z^{-s} + O(z^2). \end{aligned}$$

With the help of a computer algebra system, we obtain

$$\begin{aligned} \operatorname{Res}_{s=-2} W_1^*(s)z^{-s} &= -\frac{z^2 \ln^2(z)}{h_p^2} \\ &\quad + z^2 \ln(z) \left(-\frac{2\gamma}{h_p^2} + \frac{2\alpha_\infty}{h_p^2} - \frac{2\tilde{h}_p}{h_p^3} + \frac{2}{h_p^2} + \frac{2}{pqh_p} - \frac{3}{h_p} \right) + O(z^2), \end{aligned}$$

and, for $k \neq 0$,

$$\operatorname{Res}_{s=s_k} W_1^*(s)z^{-s} = -2 \frac{\Gamma(s_k + 1)}{h_p^2} z^2 \ln(z) + O(z^2).$$

Lemma 5

$$\begin{aligned} W(z) &= 2 \frac{z^2 \ln^2(z)}{h_p^2} \\ &\quad + z^2 \frac{\ln(z)}{h_p} \left(4 \frac{\gamma}{h_p} - 4 \frac{\alpha_\infty}{h_p} - \frac{4}{h_p} + \frac{3\tilde{h}_p}{h_p^2} + 3 - \frac{2}{pq} - 4\xi_q(\ln(z)) \right) + O(z^2). \end{aligned}$$

Part (b) of Theorem 2 is proved. In the unbiased case $p = q = \frac{1}{2}$, and the factor $\sigma_p^2 = 0$. In order not to leave the reader uniformed about what the variance is in this case, we completed a more detailed residue computation.

Theorem 4 *In an unbiased random digital tree of n random keys, the variance of the distance between two randomly selected keys is*

$$\begin{aligned} \operatorname{Var}[\Delta_n] &= \frac{6 + \pi^2}{3 \ln^2 2} + \frac{22}{3} - 2 \frac{\alpha_\infty + \psi_\infty}{\ln 2} + \frac{4(\gamma - 1)}{\ln 2} \xi_{\frac{1}{2}}(\ln n) \\ &\quad - 2\xi_{\frac{1}{2}}^2(\ln n) + \frac{4}{\ln 2} \tilde{\xi}(\ln n) + O\left(\frac{1}{n^{0.49999}}\right), \end{aligned}$$

where $\psi_\infty = \sum_{k=1}^\infty \frac{\ln 2}{(2^k - 1)^2}$, and

$$\tilde{\xi}(\ln z) = -\frac{1}{\ln 2} \sum_{\substack{k=-\infty \\ k \neq 0}}^\infty \Gamma' \left(-1 + \frac{2ik\pi}{\ln 2} \right) e^{2\pi ik \log_2 z}.$$

Remark Except for the symmetric case, the variance grows logarithmically with the number of keys inserted in the tree. In the symmetric case the coefficient of the logarithmic term is 0, leaving only $O(1)$ (oscillating but uniformly bounded) variance, showing the stiff resistance of 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 3 *As $n \rightarrow \infty$,*

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

5 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 [2]), 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. Recently, the Wasserstein and Zolotarev metrics have been popularized in the context of the contraction method.

The contraction method was introduced by Rösler [26]. Rachev and Rüschemdorf [25] added several useful extensions. Recently general contraction theorems and multivariate extensions were added by Rösler [27], Neininger [20], and Neininger and Rüschemdorf [22]. Rösler and Rüschemdorf [28] 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 is the indicator of the event that both keys are chosen from the left subtree, J_n is the indicator of the event that both keys are chosen from the right subtree, K_n is the indicator of the event that the keys are chosen from different subtrees, M_n is the indicator of the event that the root and a node from the left subtree are chosen, and S_n is the indicator of the event that the root and a node from the right subtree are chosen. The indicators are inserted to exclusively pick the right pair with appropriate probability. Thus, they are of course mutually exclusive ($I_n + J_n + K_n + M_n + S_n \equiv 1$). For $n \geq 2$, we can reorganize the latter relation as

$$\begin{aligned} \frac{\Delta_n - \mathbf{E}[\Delta_n]}{\sqrt{\ln n}} &\stackrel{\mathcal{L}}{=} \frac{\Delta_{L_n} - \mathbf{E}[\Delta_{L_n}]}{\sqrt{\ln L_n}} I_n \sqrt{\frac{\ln L_n}{\ln n}} + \frac{\tilde{\Delta}_{R_n} - \mathbf{E}[\tilde{\Delta}_{R_n}]}{\sqrt{\ln R_n}} J_n \sqrt{\frac{\ln R_n}{\ln n}} \\ &\quad + Y_n^* K_n + \frac{1}{\sqrt{\ln n}} \left(\mathbf{E}[\Delta_{L_n}] I_n + \mathbf{E}[\tilde{\Delta}_{R_n}] J_n \right. \\ &\quad \left. + (\mathbf{E}[\delta_{L_n}] + \mathbf{E}[\tilde{\delta}_{R_n}] + 2) K_n \right. \\ &\quad \left. + (\delta_{L_n} + 1) M_n + (\tilde{\delta}_{R_n} + 1) S_n - \mathbf{E}[\Delta_n] \right), \end{aligned}$$

where

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

This equation can be written in terms of the normed variables as

$$\Delta_n^* = \Delta_{L_n}^* I_n \sqrt{\frac{\ln L_n}{\ln n}} + \tilde{\Delta}_{R_n}^* J_n \sqrt{\frac{\ln R_n}{\ln n}} + Y_n^* K_n + G_n, \tag{10}$$

where

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

We first argue heuristically the existence of a limit for Δ_n^* . The argument elicits the nature of the limit. We then confirm our guess by an inductive proof in the Wasserstein metric space. The additive ingredients of (10) are dependent. For instance, Δ_{L_n} and $\tilde{\Delta}_{R_n}$ are dependent through the dependency of L_n and R_n (though these copies are conditionally independent when L_n and R_n are given). Moreover, $I_n, J_n, K_n, M_n,$ and S_n are dependent, etc. However, the sharp concentration of the binomial distribution of L_n , namely

$$\frac{L_n}{n} \xrightarrow{a.s.} q, \quad \frac{R_n}{n} \xrightarrow{a.s.} p, \tag{11}$$

loosens the dependence. Because the logarithm is a slowly growing function, it is an immediate consequence of (11) that

$$\sqrt{\frac{\ln L_n}{\ln n}} \xrightarrow{a.s.} 1, \quad \sqrt{\frac{\ln R_n}{\ln n}} \xrightarrow{a.s.} 1. \tag{12}$$

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 [15]. Similarly, $(\delta_{L_n} - \mathbf{E}[\delta_{L_n}] \ln n) \ln^{-\frac{1}{2}} L_n$ and $(\tilde{\delta}_{R_n} - \mathbf{E}[\tilde{\delta}_{R_n}] \ln n) \ln^{-\frac{1}{2}} R_n$, albeit dependency, would eventually be independent copies of $\mathcal{N}(0, \sigma_p^2)$.

Each of the indicators I_n, J_n, K_n is a *conditional* Bernoulli random variable. For instance, for any $n \geq 2$,

$$I_n = \text{Ber}\left(\frac{L_n(L_n - 1)}{n(n - 1)}\right),$$

which is to be interpreted as $\text{Ber}(\ell(\ell - 1)/(n(n - 1)))$, whenever $L_n = \ell$. The indicators (I_n, J_n, K_n) would 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

$$I_n \xrightarrow{a.s.} I = \text{Ber}(q^2), \tag{13}$$

$$J_n \xrightarrow{a.s.} J = \text{Ber}(p^2), \tag{14}$$

$$K_n \xrightarrow{a.s.} K = \text{Ber}(2pq). \tag{15}$$

The indicators M_n and S_n are much less probable than the former three. For example

$$M_n = \text{Ber}\left(\frac{L_n}{n(n-1)}\right),$$

and since $L_n < n$, we have $M_n \rightarrow 0$, and so does S_n .

Lemma 6 *As $n \rightarrow \infty$,*

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

Proof Utilizing Part(a) of Theorem 2 we can bound the term $\mathbf{E}[\Delta_{L_n}]$ by conditioning as follows:

$$\begin{aligned} \mathbf{E}[\Delta_{L_n}] &= \sum_{\ell=0}^n \mathbf{E}[\Delta_\ell] \binom{n}{\ell} p^{n-\ell} q^\ell \\ &= \sum_{\ell=2}^n \binom{n}{\ell} p^{n-\ell} q^\ell \left(\frac{2}{h_p} \ln \ell + O(1)\right) \\ &\leq \left(\frac{2}{h_p} \ln n + O(1)\right) \sum_{\ell=2}^n \binom{n}{\ell} p^{n-\ell} q^\ell \\ &< \frac{2}{h_p} \ln n + O(1). \end{aligned}$$

The asymptotic mean random depth from Louchard and Szpankowski [15] gives us (again by a conditional argument) $\mathbf{E}[\delta_{L_n}] < \frac{1}{h_p} \ln n + O(1)$. By symmetry, similar bounds hold in the right subtree for the terms $\mathbf{E}[\Delta_{R_n}]$ and $\mathbf{E}[\delta_{R_n}]$. The success rate of M_n and S_n is so weak that all the terms involving them converge to 0 in probability. For instance, from the growth rates of the mean and variance in biased cases, we have $\delta_{L_n} / \ln n \xrightarrow{\mathcal{P}} h_p$, and $M_n = O_P(1/n)$. Thus,

$$\frac{\delta_{L_n} M_n}{\sqrt{\ln n}} = O_P\left(\frac{\ln^{\frac{1}{2}} n}{n}\right).$$

So, now we can represent G_n as

$$\begin{aligned} G_n &= \frac{1}{\sqrt{\ln n}} \left(\left[\frac{2}{h_p} \ln n + O(1) \right] I_n + \left[\frac{2}{h_p} \ln n + O(1) \right] J_n \right. \\ &\quad + \left(\left[\frac{1}{h_p} \ln n + O(1) \right] + \left[\frac{1}{h_p} \ln n + O(1) \right] + 2 \right) K_n \\ &\quad \left. - (\delta_{L_n} + 1) M_n - (\delta_{L_n} + 1) S_n - \left(\frac{2}{h_p} \ln n + O(1) \right) \right) \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{\sqrt{\ln n}} \left(\frac{2}{h_p} (I_n + J_n + K_n) \ln n - \frac{2}{h_p} \ln n + O(1) \right) + O_P\left(\frac{\ln^{\frac{1}{2}} n}{n}\right) \\
 &= O_P\left(\frac{1}{\sqrt{\ln n}}\right).
 \end{aligned}$$

□

In view of (10) and the convergence relations (11)–(15) and Lemma 6, if Δ_n^* converges to a limit, say Δ^* , that limit would satisfy the distributional equation

$$\Delta^* \stackrel{\mathcal{L}}{=} \Delta^* I + \tilde{\Delta}^* J + Y^* K, \tag{16}$$

with $Y^* \stackrel{\mathcal{L}}{=} \delta^* + \tilde{\delta}^*$, and $(\Delta^*, \tilde{\Delta}^*, Y^*)$ independent of (I, J, K) .

Let F_n^* be the distribution function of Δ_n^* , and F^* be the that of Δ^* . To actually prove that a limit satisfying the fixed-point limit equation (16) exists in distribution for Δ_n^* , it suffices to show that the second-order Wasserstein distance

$$d_2(F_n^*, F^*) = \inf \|V_n - W\|_2$$

converges to 0, as $n \rightarrow \infty$; the infimum being taken over all pairs V_n, W of random variables with distributions F_n^* and F^* .

The L_2 -norm $\|V_n - W\|_2 = \sqrt{\mathbf{E}[(V_n - W)^2]}$ for any particular pair V_n and W gives an upper bound on $d_2(F_n^*, F^*)$. In particular we have

$$d_2^2(F_n^*, F^*) \leq v_n := \mathbf{E}[(\Delta_n^* - \Delta^*)^2].$$

Lemma 7 As $n \rightarrow \infty$,

$$v_n \rightarrow 0.$$

Proof A technical proof, almost identical to that in Aguech, Lasmar and Mahmoud [1] can be mimicked. We omit the details. □

Proof of Theorem 3 The limiting random variable Δ^* of Δ_n^* has a distribution that satisfies the distributional equation (16). Conditioning on $\mathbf{M} = (I, J, K)$, we find the representation

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

Thus,

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

and $\Delta^* \stackrel{\mathcal{L}}{=} \delta^* + \tilde{\delta}^*$; 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 [15]. That is, $\Delta^* \stackrel{\mathcal{L}}{=} \mathcal{N}(0, 2\sigma_p^2)$. □

6 Conclusion and discussion

Digital trees such as tries, digital search trees and Patricia trees have numerous applications in computer science, computational biology and several other areas. The books by Knuth [13], Mahmoud [17] and Szpankowski [31] survey numerous applications in sorting algorithms and data structures. Distances within a random structure such as random trees have practical significance. For example, the collective sum of all inter-node distances in the graph underlying a molecule is known in chemistry as the *Wiener index* (see [8,32]).

We showed that for random biased digital search trees the variance of the distance between a randomly selected pair grows logarithmically with the number of keys, whereas it remains bounded in unbiased digital trees. In the biased case the order of magnitude of the variance enables a limit (after appropriate normalization) to exist.

Acknowledgement The third author wishes to thank the Institute of Statistical Mathematics, Tokyo, for supporting this research, and Faculté des Sciences de Monastir, Tunisia, for sponsoring a research visit. The authors thank an anonymous referee for an effort that helped improve the exposition.

Appendix

The material of this appendix (in other related forms) can be found in some sources, such as [11]. However, the *raison-d'être* for this appendix is that we need a certain organization in functional equation form and a presentation of the Mellin transform as a delayed algebraic equation, suitable for a direct engagement in our setup for distances between pairs of nodes.

Let $\psi(t, z) = \sum_{n=1}^{\infty} \frac{n\mathbf{E}[e^{\delta_n t}]}{n!} z^n e^{-z}$. We have

$$\delta_n | L_n = \begin{cases} \delta_{L_n} + 1 & \text{with probability } \frac{L_n}{n}, \\ 0 & \text{with probability } \frac{1}{n}, \\ \tilde{\delta}_{R_n} + 1 & \text{with probability } \frac{R_n}{n}. \end{cases}$$

Then

$$\mathbf{E}[e^{\delta_n t} | L_n] = e^t e^{\delta_{L_n} t} \times \frac{L_n}{n} + e^t e^{\tilde{\delta}_{R_n} t} \times \frac{R_n}{n} + \frac{1}{n},$$

or

$$\begin{aligned} n\mathbf{E}[e^{\delta_n t}] &= e^t \sum_{\ell=0}^{n-1} \ell \mathbf{E}[e^{\delta_{\ell} t}] \binom{n-1}{\ell} q^{\ell} p^{n-1-\ell} \\ &\quad + e^t \sum_{r=0}^{n-1} r \mathbf{E}[e^{\delta_r t}] \binom{n-1}{r} q^{n-1-r} p^r + 1. \end{aligned}$$

So,

$$\sum_{n=1}^{\infty} \frac{n\mathbf{E}[e^{\delta n t}]}{(n-1)!} z^{n-1} = e^t \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} \frac{k\phi_{\delta k}(t)q^k z^k}{k!} \times \frac{p^{n-1-k} z^{n-1-k}}{(n-1-k)!} + e^t \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} \frac{k\phi_{\delta k}(t)p^k z^k}{k!} \times \frac{q^{n-1-k} z^{n-1-k}}{(n-1-k)!} + \sum_{n=1}^{\infty} \frac{z^n}{n!}.$$

Hence $\psi(t, z)$ satisfies the functional equation

$$\psi'(t, z) + \psi(t, z) = e^t \psi(t, pz) + e^t \psi(t, qz) + z;$$

the prime here indicates the derivative with respect to z . Further, let $x(z) = \frac{\partial}{\partial t} \psi(t, z)|_{t=0}$ and $w(z) = \frac{\partial^2}{\partial t^2} \psi(t, z)|_{t=0}$. This pair of functions satisfies the equations

$$\begin{aligned} x'(z) + x(z) &= x(pz) + x(qz) + z, \\ w'(z) + w(z) &= 2(x(pz) + x(qz)) + w(pz) + w(qz) + z. \end{aligned}$$

These differential equations give Mellin transform equations with delay:

$$\begin{aligned} -(s-1)x^*(s-1) + x^*(s) &= (p^{-s} + q^{-s})x^*(s), \\ -(s-1)w^*(s-1) + w^*(s)(1 - p^{-s} - q^{-s}) &= 2(p^{-s} + q^{-s})x^*(s). \end{aligned}$$

By an argument just like that in the proof of Lemma 1 we can show that $x(z)$ is $O(z)$, as $z \rightarrow 0$, and is $O(z^2)$, as $z \rightarrow \infty$, and that $x'(z)$ is $O(1)$, as $z \rightarrow 0$, and is $O(z)$, as $z \rightarrow \infty$. Thus the domain of existence of the Mellin transform of both $x(s)$ and $x'(s-1)$ is $(-2, -1)$.

Put $x^*(s) = \rho(s)\Gamma(s)$ and $w^*(s) = \beta(s)\Gamma(s)$. After some tedious algebra, we obtain

$$\begin{aligned} \rho(s) &= \frac{Q_{\infty}(-2)}{Q_{\infty}(s)}, \\ \beta(s) &= \frac{\beta(s-1)}{1 - p^{-s} - q^{-s}} + 2 \frac{(p^{-s} + q^{-s})Q_{\infty}(-2)}{Q_{\infty}(s)(1 - p^{-s} - q^{-s})}. \end{aligned}$$

then

$$\beta(s) = \rho(s) \left(2 \sum_{k=0}^{\infty} \frac{p^{k-s} + q^{k-s}}{1 - p^{k-s} - q^{k-s}} + 1 - 2\beta_{\infty} \right),$$

and

$$x(z) = \frac{z \ln(z)}{h_p} + \frac{1}{h_p} \left(\gamma - 1 - \alpha_{\infty} + \frac{\tilde{h}_p}{2h_p} \right) z - z\xi_q(\ln(z)).$$

References

1. Aguech, R., Lasmar, N., Mahmoud, H.: Limit distribution of distances in biased random tries. *J. Appl. Probab.* **43**, 1–14 (2006)
2. Chern, H., Hwang, H., Tsai, T.: An asymptotic theory for Cauchy-Euler differential equations with applications to the analysis of algorithms. *J. Algorithms* **44**, 177–225 (2002)
3. Christophi, C., Mahmoud, H.: The oscillatory distribution of distances in random tries. *Ann. Appl. Probab.* **15**, 1536–1564 (2005)

4. Coffman, E., Eve, J.: File structures using hashing functions. *Commun. ACM* **13**, 427–432, and 436 (1970)
5. Devroye, L., Neininger, R.: Distances and finger search in random binary search trees. *SIAM J. Comput.* **33**, 647–658 (2004)
6. Flajolet, P., Sedgewick, R.: Digital search trees revisited. *SIAM J. Comput.* **15**, 748–767 (1986)
7. Flajolet, P., Gourdon, X., Dumas, P.: Mellin transform and asymptotic harmonic sums. *Theor. Comput. Sci.* **144**, 3–58 (1995)
8. Gutman, I., Polansky, O.: *Mathematical Concepts in Organic Chemistry*. Springer, Berlin Heidelberg New York (1986)
9. Jacquet, P.: Contribution de l'Analyse d'Algorithmes a l'Evaluation de Protocoles de Communication. Thèse Université de Paris Sud-Orsay, Paris (1989)
10. Jacquet, P., Szpankowski, W.: Analytical depoissonization and its applications. *Theor. Comput. Sci.* **201**, 1–62 (1998)
11. Kirschenhofer, P., Prodinger, H.: Further results on digital search trees. *Theor. Comput. Sci.* **58**, 143–154 (1988)
12. Kirschenhofer, P., Prodinger, H., Szpankowski, W.: Digital search trees again revisited: the internal path length perspective. *SIAM J. Comput.* **23**, 598–616 (1994)
13. Knuth, D.: *The Art of Computer Programming, Vol. 3: Sorting and Searching*, 2nd ed. Addison-Wesley, Reading, MA (1989)
14. Louchard, G.: Exact and asymptotic distributions in digital and binary search trees. *RAIRO: Theor. Informat. Appl.* **21**, 479–495 (1987)
15. Louchard, G., Szpankowski, W.: Average profile and limiting distribution for a phrase size in the Lempel-Ziv parsing algorithm. *IEEE Trans. Informat. Theory* **41**, 478–488 (1995)
16. Louchard, G., Szpankowski, W., Tang, J.: Average profile of the generalized digital-search tree and the generalized Lempel-Ziv algorithms. *SIAM J. Comput.* **28**, 935–954 (1999)
17. Mahmoud, H.: *Evolution of Random Search Trees*. Wiley, New York (1992)
18. Mahmoud, M., Neininger, R.: Distribution of distances in random binary search trees. *Ann. Appl. Probab.* **13**, 253–276 (2003)
19. Mathys, P., Flajolet, P.: Q-ary collision resolution algorithms in random-access systems with free and blocked channel access. *IEEE Trans. Informat. Theory* **31**, 217–243 (1985)
20. Neininger, R.: On a multivariate contraction method for random recursive structures with applications to Quicksort. *Random Struct. Algorithms* **19**, 498–524 (2001)
21. Neininger, R.: The Wiener index of random trees. *Combinat. Probab. Comput.* **11**, 587–597 (2002)
22. Neininger, R., Rüschemdorf, L.: A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab.* **14**, 378–418 (2004)
23. Panholzer, A., Prodinger, H.: Spanning tree size in random binary search trees. *Ann. Appl. Probab.* **14**, 718–733 (2004)
24. Pittel, B.: Asymptotical growth of a class of random trees. *Ann. Probab.* **13**, 414–427 (1985)
25. Rachev, S., Rüschemdorf, L.: Probability metrics and recursive algorithms. *Adv. Appl. Probab.* **27**, 770–799 (1995)
26. Rösler, U.: A limit theorem for “Quicksort”. *RAIRO Inform. Théor. Appl.* **25**, 85–100 (1991)
27. Rösler, U.: On the analysis of stochastic divide and conquer algorithms. *Algorithmica*, **29**, 238–261 (2001)
28. Rösler, U., Rüschemdorf, L.: The contraction method for recursive algorithms. *Algorithmica* **29**, 3–33 (2001)
29. Schachinger, W.: Beiträge zur Analyse von Datenstrukturen zur Digitalen Suche. Dissertation Technische Universität Wien, Vienna (1993)
30. Szpankowski, W.: A characterization of digital search trees from the successful search viewpoint. *Theor. Comput. Sci.* **85**, 117–134 (1991)
31. Szpankowski, W.: *Average Case Analysis of Algorithms on Sequences*. Wiley, New York (2001)
32. Trinajstić, N.: *Chemical Graph Theory*. CRC Press, Boca Raton, FL (1992)