Probability in the Engineering and Informational Sciences, **22**, 2008, 133–149. Printed in the U.S.A.
DOI: 10.1017/S0269964808000089

ON CLIMBING TRIES

COSTAS CHRISTOPHI
Cyprus International Institute for the Environment and Public Health in Association with Harvard School of Public Health, 1105, Nicosia, Cyprus
and
Biostatistics Center, The George Washington University, Rockville, MD 20852
E-mail: cchristophi@cyprusinstitute.org

HOSAM MAHMOUD
Department of Statistics, The George Washington University
Washington, DC 20052
E-mail: hosam@gwu.edu

To sample a typical key in a "trie," an appropriate climbing might consider generating random edges in the same manner as the data are generated. In the absence of the probability generating the keys, an uninformed random choice among the children still provides an alternative. We are also interested in extremal sampling, achieved by following a leftmost (or a rightmost) path. Each of these climbing strategies always generates a key, but one that might not necessarily be in the database. We investigate the altitude of the position at which climbing is terminated. Analytical techniques, including poissonization and the Mellin transform, are used for the accurate calculation of moments. In all strategies, the mean is always logarithmic. For typical and uninformed climbing, the variance is bounded in unbiased tries but grows logarithmically in biased tries. Consequently, in the biased case, one can find appropriate centering and scaling to produce a limit distribution for these two climbing strategies; the limit is normal. For extremal climbing, the variance is always bounded for both biased and unbiased cases, and no nontrivial limit exists under any scaling.

1. INTRODUCTION

The random climbing of trees has been a subject that authors revisit from time to time. It was considered in Moon [10] and in Meir and Moon [11, 12]. The subject has been revisited recently by Panholzer [13], who considered several classes of random trees, including simply generated families and Pólya trees. In these investigations, a class of trees is considered, and a type of random walk on it is exercised. Starting at the root,

 0269-9648/08 $25.00

certain nodes are accessed, and at each node, a randomly selected edge emanating from it is chosen *at random* (all edges coming out of a node being equally likely). The process is perpetuated until it is no longer possible to proceed. When the process is stopped, the path inscribed in the tree by climbing reaches a leaf.

We would like to consider the climbing of a class of random digital trees called the "trie." However, we think the "model of randomness" should be changed from the usual uniform choice of edges to a climbing model that conforms with the manner in which these tries are randomly generated. The trie emerges from keys that are taken from a data generator that emits bits of data, with 1's having probability p and 0's having probability $q = 1 - p$. So, at each node of the trie, a simple Bernoulli random variable will govern the direction of the next turn. The process might not necessarily end on a leaf, as it might terminate at a null node, but it always generates a key (not necessarily in the trie).

The general interpretation of this climbing is that a typical key is being "sampled" from the database. Hence, we call this strategy *typical climbing*. In the absence of knowledge of the key generating probability, we consider an alternative strategy called *uninformed climbing*, in which we follow the right and left nexus with equal probability. Motivated by these sampling schemes, we also consider the case of sampling extremal data. In all of the cases, we develop asymptotic distributions for the length of the total distance climbed. One might be able to get exact expressions, too. For example, by purely combinatorial arguments, we compute the exact distribution of the climbing pathlength in extremal sampling to show that such exact expressions are possible in principle.

The plan of the article is as follows. For economy of notation, variables are reused in sections that are self-contained. For example, in Section 4, S_n, with moment generating function $\phi_n(t)$, will be the number of nodes on the path of typical climbing. These two symbols will be reused for the number of nodes on the path of uninformed and extremal climbing, and the corresponding moment generating functions, in Sections 5 and 6. Section 2 gives an overview of the trie structure and other terminology. In Section 3 the methodology is outlined. Analysis of typical climbing is pursued in Section 4, where in the biased case, we find a Gaussian limit for an appropriately normalized version of the climbing pathlength. Technicalities for asymptotically accurate mean and variance calculation are discussed in Subsection 4.1. Analysis of uninformed climbing is pursued in Section 5, where in the biased case, we also find a Gaussian limit. By contrast, we demonstrate in Section 6 that the extremal climbing pathlength does not possess a nontrivial limit under any scaling. We devote Subsection 6.1 to a combinatorial derivation of the exact distribution of the pathlength of extremal climbing.

2. TRIES

The trie is a data structure suitable for digital data (bits, hexadecimal strings, words, DNA strands, etc.), which are prevalent in science and technology. The trie was invented independently by De La Briandais [2] and Fredkin [5] for information re*trie*val. Tries have numerous applications as a data structure for computer files,

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

telecommunication signals, DNA, and so forth because of the digital nature of these data. Tries also provide a model for the analysis of several important algorithms, such as Radix Exchange Sort (see Knuth [8]) and Extendible Hashing (see Fagin, Nievergelt, Pippenger, and Strong [3]).

A *binary trie* is a digital tree consisting of *internal nodes*, each having one or two children, and *leaves* that hold data (*keys*). 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 zero 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. Figure 1 illustrates a trie with five keys:

$$X_1 = 00111\ldots,$$
$$X_2 = 11011\ldots,$$
$$X_3 = 00011\ldots,$$
$$X_4 = 01010\ldots,$$
$$X_5 = 11111\ldots.$$

For ease of exposition, we will assume our data to be in binary representation of numbers in [0, 1]. We can always insert a binary point to the left of a binary string to turn it into a number from this range. The binary case lays out the methodology for any size alphabet. The case of a larger alphabet can be handled similarly; it just involves more details.

Suppose we have $n \geq 0$ keys, each given as an (infinite) string representing its expansion into binary bits. We assume the Bernoulli(p) model of randomness, according to which the bits within a key are independent with probability p of a bit being 1 and probability q of it being 0 (with $p + q = 1$), and the keys themselves are independent. The data entropy in this model is

$$h_p = -(p \ln p + q \ln q).$$

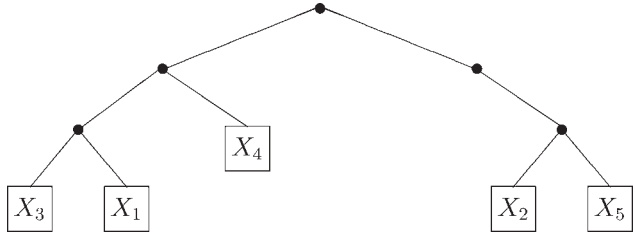

FIGURE 1. A trie with five keys.

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

The ideal unbiased Bernoulli model is equivalent to sampling from a uniform distribution, a realistic assumption under hashing schemes, where the primary goal is to achieve uniformity.

3. METHODOLOGY

Two main tools in the ensuing analysis are the Mellin transform and poissonization–depoissonization. These methods are now standard and we will not produce the details in any great length, but refer the reader to standard sources on such material.

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 will 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 $c \in (a, b)$. Usually, such an integral is computed asymptotically (as $x \to \infty$) by shifting the line of integration an arbitrary distance to the right of the existence strip and compensating for the shift by the residues of the poles between the two lines of integration. There often is a small residual error of the form $O(x^{-M})$ for an arbitrary large positive number M. For a survey of the uses of the Mellin transform in the analysis of algorithms, see Flajolet, Gourdon, and Dumas [4].

Certain complicated types of functional equation appear in the formulation of recurrence for pathlength under all climbing strategies. These types of recurrence equation are not easy to solve. However, poissonized versions are amenable to asymptotic analysis via the Mellin transform. In this context, poissonization means considering an analogous problem, but with a Poisson random number of keys instead of fixed n. The number of keys is taken to be a Poisson random variable with parameter z. The required asymptotic results for the fixed population are then extracted from the poissonized model by depoissonization, which usually means using the same results for the poissonized model after replacing z with n. This operation is justified by checking some regularity conditions, but it also introduces an asymptotically negligible error. We consider this as a standard program and will not give details, but refer the reader to the original work of Jacquet and Szpankowski [7] or its presentation in textbook style in Szpankowski [15].

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

4. TYPICAL CLIMBING

In typical sampling, we climb a trie by following an algorithm that emulates the natural frequency of bits. We start at the root and access nodes. At each node accessed, we generate an independent Bernoulli(p) random variable. If this variable yields zero, we follow the left edge if it exists (otherwise, the climbing is stopped), and if the value generated is 1, we follow the right edge if it exists (otherwise, the climbing is stopped).

Let S_n be the number of nodes on the path inscribed in the trie by the typical climbing. For example, given the trie of Figure 1, typical climbing might produce the key X_4 in two steps with conditional probability pq, in which case $S_5 = 3$ (counting the node containing X_4). It might also reach the only null node in the trie (left of the root's right child) with probability pq, in which case $S_5 = 2$. If the null node is reached, we take our typical sample to be 0.100000 ....

Note that S_n can be linked to the depth D_n of a randomly chosen key, after an additional key is added to the initial n keys. If we are inserting the $(n+1)$st key, this will follow the path of S_n. If the climbing terminates at an empty node, S_n and D_{n+1} are the same, but if the climbing terminates at a key, we need to insert a number of additional nodes. The number of additional nodes is geometrically distributed because we have to break the tie—there will be a number (zero or more) of bits in the new key agreeing with bits in the key colliding with it past the point of collision: Each agreement occurs with probability $p^2 + q^2$, but sooner or later, one disagreement (with probability $1 - p^2 - q^2$) will break the tie. This geometric random variable is independent of the structure of the trie; it is distributed like the depth of two keys in a trie of size 2, but the total amount of modification to S_n to become D_{n+1} is dependent on S_n. More precisely, if $X^{(r)}$ is the prefix of length r of a digital key X, then

$$D_{n+1} = S_n + \tilde{D}_2 \mathbf{1}_{\{\cup_{j=1}^{n}\{X_{n+1}^{(S_n)} = X_j^{(S_n)}\}\}} = S_n + \tilde{D}_2 \sum_{j=1}^{n} \mathbf{1}_{\{X_{n+1}^{(S_n)} = X_j^{(S_n)}\}},$$

where $\mathbf{1}_E$ is the indicator function of an event E and $\tilde{D}_2$ is an independent copy of D_2. The dependence in the sum introduces complications, but the analytic approach that follows is transparent and systematic enough to cover cases that cannot be linked easily to the depth, such as the case of climbing without the knowledge of p, where we generate right and left moves with equal probability.

Let $\phi_n(t)$ be the moment generating function of S_n. Let L_n and R_n be respectively the number of keys in the left and right subtrees (so $L_n + R_n = n$). In view of the Bernoulli model, $L_n \stackrel{\mathcal{L}}{=} \text{Binomial}\ (n, q)$. The subtrees themselves are random tries on their respective order, which follows from the independence structure assumed in the data. The variable S_n satisfies a basic recurrence:

$$S_n | L_n = \begin{cases} 1 + S_{L_n} & \text{with probability } q \\ 1 + \tilde{S}_{R_n} & \text{with probability } p. \end{cases}$$

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

Here and in the sequel, a tilded random variable stands for a random variable distributed like the untilded version and is conditionally independent of it. We have the conditional expectation

$$\mathbf{E}[e^{S_n t}|L_n] = e^{(1+S_{L_n})t}q + e^{(1+\tilde{S}_{R_n})t}p.$$

By a standard double expectation, we get

$$\mathbf{E}[e^{S_n t}] = \mathbf{E}[e^{(1+S_{L_n})t}q] + \mathbf{E}[e^{(1+\tilde{S}_{R_n})t}p].$$

By the binomial distribution of L_n, we get by conditioning

$$\begin{aligned}\phi_n(t) &= \mathbf{E}[e^{S_n t}] \\ &= \sum_{\ell=0}^{n} \mathbf{E}[e^{(1+S_\ell)t}]q\binom{n}{\ell}q^\ell p^{n-\ell} + \sum_{\ell=0}^{n} \mathbf{E}[e^{(1+S_{n-\ell})t}]p\binom{n}{\ell}q^\ell p^{n-\ell}.\end{aligned}$$

Toward poissonization we construct the supergenerating function $A(z,\ t) = \sum_{n=0}^{\infty}(\phi_n(t)/n!)z^n$. First, we multiply both sides of the latter equality by z^n and sum over all possible values of n to get

$$\begin{aligned}\sum_{n=2}^{\infty}\frac{\phi_n(t)}{n!}z^n &= qe^t\sum_{n=2}^{\infty}\sum_{\ell=0}^{n} z^n\phi_\ell(t)\frac{1}{\ell!\,(n-\ell)!}q^\ell p^{n-\ell} \\ &\quad + pe^t\sum_{n=2}^{\infty}\sum_{\ell=0}^{n} z^n\phi_{n-\ell}(t)\frac{1}{\ell!\,(n-\ell)!}q^\ell p^{n-\ell}.\end{aligned}$$

The sums in this last equation are then extended to start from $n = 0$, after introducing the necessary adjustments for $n = 0$ and $n = 1$. Subsequently,

$$\begin{aligned}A(z,\ t) &= qe^t e^{pz}A(qz,\ t) + pe^t e^{qz}A(pz,\ t) \\ &\quad + 1 - e^t + ze^t - (p^2+q^2)e^{2t}z - 2pqe^t z.\end{aligned}$$

Direct poissonization of $A(z,\ t)$ runs into a problem in the existence of the Mellin transform. The difficulty is that when we multiply both sides of the equation by e^{-z}, the right-hand side has the loose term $(1 - e^t)\,e^{-z}$. Formally, its Mellin transform is $(1 - e^t)\Gamma(s)$ and its domain of existence is $\Re s > 0$. The remaining terms, however, impose domains of existence in a strip with a negative real part. Hence, there will be no common domain of intersection.

To circumvent the difficulty, we deal with a shifted supermoment generating function. As we will see shortly, in this way the right-hand side has a difference of exponential terms, for which a domain of existence is consistent with the rest of the expression. Set

$$B(z,\ t) = e^{-z}(A(z,\ t) - 1).$$

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

The function $B(z, t)$ has the interpretation:

$$B(z, t) = \mathbf{E}[e^{S_{N(z)}t}] - e^{-z},$$

where $N(z)$ is a Poisson random variable with parameter z; that is, $B(z, t)$ is the poissonized moment generating function of the climbing pathlength, modified by an exponentially negligible error, from which we obtain the functional equation

$$\begin{aligned} B(z, t) &= e^t(pe^{-pz}A(pz, t) + qe^{-qz}A(qz, t) + (p^2 + q^2)ze^{-z}(1 - e^t) - e^{-z}) \\ &= e^t(pB(pz, t) + qB(qz, t) + p(e^{-pz} - e^{-z}) + q(e^{-qz} - e^{-z}) \\ &\quad + (p^2 + q^2)ze^{-z}(1 - e^t)). \end{aligned}$$

The Mellin transform of this function is

$$B^*(s, t) = \frac{e^t\Gamma(s)(p^{-s+1} + q^{-s+1} - 1 + (p^2 + q^2)(1 - e^t)s)}{1 - e^t(p^{-s+1} + q^{-s+1})}, \tag{1}$$

where we treated a term like $e^{-pz} - e^{-z}$ as $e^{-pz} - 1 - (e^{-z} - 1)$, with each shifted exponential function having a Mellin transform representation in terms of the Cauchy–Saalschütz gamma function in the domain $\langle -1, 0\rangle$; that is, such a term has the Mellin transform $(p^{-s} - 1)\Gamma(s)$ in that domain. Now, the Mellin transform $B^*(s, t)$ exists in the domain

$$-1 < \Re s < s_0(t),$$

where $s_0(t)$ is the only real solution to the equation

$$p^{-s+1} + q^{-s+1} = e^{-t}. \tag{2}$$

Observe that $s_0(t)$ is a continuous function of t, with value 0 at $t = 0$. We thus can find a neighborhood around $t = 0$ for which $s_0(t)$ is arbitrarily close to zero. We will keep $|t|$ small enough for the entire strip $\langle s_0(t), -s_0(t)\rangle$ to be contained in $\langle -1/4, 1/4\rangle$.

4.1. The Mean Climbing Pathlength in Typical Climbing

We will need a few technicalities for the proof, and we discuss them first. The inverse Mellin transform involved in the mean requires a computation of residues at the roots of the *characteristic equation*

$$1 - p^{-s+1} - q^{-s+1} = 0. \tag{3}$$

The roots of this equation have been studied. The following special case is known (e.g., see Szpankowski [15], who attributed the result to Jacquet [6] and Schachinger [14]). We present the result as written in Drmota, Reznik, Savari, and Szpankowski [private communication].

LEMMA 1: *Let $p < q$. There are countably infinitely many simple solutions (characteristic roots) of $1 - p^{-s+1} - q^{-s+1} = 0$. The roots satisfy the following:*

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

(i) $s_0 = 0$ *is always a root.*

(ii) *If* b *is a root, then* $0 \le \Re b \le \rho$, *where* ρ *is the unique real positive solution of* $1 - p^{-s+1} + q^{-s+1} = 0$. *Moreover, for every integer* k, *there exists a unique root* s_k *with imaginary part* $(2k-1)\pi/|\ln p| < \Im s_k \le (2k+1)\pi/|\ln p|$. *Consequently* s_k, *for* $k = 0, \pm 1, \pm 2, \ldots$, *are all the roots.*

(iii) *If* $\ln p/\ln q = m/r$ (*where* $\gcd(m, r) = 1$ *for positive integers* m *and* r), *there are* $m - 1$ *roots,* $s_1, s_2, \ldots, s_{m-1}$, *with real part greater than zero. The rest of the roots are in the form*

$$s_k = s_{k \bmod m} + \frac{2\pi i(k - k \bmod m)}{\ln p} \quad \text{for } k \ge m \text{ and } k < 0.$$

(iv) *If* $\ln p/\ln q$ *is irrational, then* $\Re s_k > 0$ *for all* $k \ne 0$.

A symmetrical statement applies when $p > q$, *but in this case,* ρ *is defined as the positive root of* $1 + p^{-s+1} - q^{-s+1} = 0$.

THEOREM 1: *Let* S_n *be the number of nodes on the path of typical climbing of a trie on* n *keys from the* Bernoulli(p) *model. Then*

$$\begin{aligned}\mathbf{E}[S_n] &= \frac{\ln n}{h_p} + \frac{1}{h_p}(\gamma - 1 - \ln p + 2pq - \ln q) \\ &\quad - \frac{1}{2h_p^2}(p \ln^2 p + 2 \ln p \ln q + q \ln^2 q) + \eta_1(\ln n) \\ &\quad + o(1), \\ \mathbf{Var}[S_n] &= \frac{pq(\ln p - \ln q)^2}{h_p^3} \ln n + o(\ln n),\end{aligned}$$

where $\eta_1(\cdot)$ *is the function given by the Fourier expansion*

$$\eta_1(u) = \begin{cases} -\dfrac{1}{h_p} \displaystyle\sum_{\substack{k=-\infty \\ k \ne 0}}^{\infty} (1 + (p^2 + q^2)s_{mk})\Gamma(s_{mk})e^{-s_{mk}u} \\ 0 \end{cases}$$

if $\dfrac{\ln p}{\ln q} = \dfrac{m}{r}$ *is rational with* $\gcd(m, r) = 1$

otherwise

(with s_{mk} *a nonzero solution of* $p^{-s+1} + q^{-s+1} = 1$ *with real part zero). In either case,* η_1 *is uniformly bounded by a small number.*[1] *The* $o(\ln n)$ *term in the variance might also have small bounded oscillations.*

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

PROOF: In order to calculate the mean, we take the first derivative of (1) with respect to t and evaluate for $t = 0$, yielding

$$\frac{\partial}{\partial t} B^*(s, t)|_{t=0} = -\frac{\Gamma(s)(1 + (p^2 + q^2)s)}{1 - q^{-s+1} - p^{-s+1}}.$$

This is the Mellin transform of the expected poissonized pathlength $S_{N(z)}$ and exists in $\langle 0, 1\rangle$.

The roots of the characteristic equation (3) determine the asymptotics of the mean. According to Lemma 1, in the case when $\ln p/\ln q$ is rational, there are roots aligned and equispaced on the vertical axis of the s complex plane, and the rest of the characteristic roots fall in the right half of the s complex plane, whereas in the case where $\ln p/\ln q$ is irrational, all of the roots fall in the right half of the s complex plane except for $s_0 = 0$.

The inverse Mellin transform is

$$\mathbf{E}[S_{N(z)}] = O(z^{-M}) + \sum_{k=-\infty}^{\infty} \operatorname*{Res}_{s=s_k} \left[\frac{z^{-s}\Gamma(s)(1 + (p^2 + q^2)s)}{1 - q^{-s+1} - p^{-s+1}}\right],$$

where $M > \rho \geq a$ is any arbitrary large positive number.

The main contribution comes from $s_0 = 0$, as it is the only double pole; the rest are simple. We obtain

$$\mathbf{E}[S_{N(z)}] = \frac{\ln z}{h_p} + \frac{1}{h_p}(\gamma - 1 - \ln p + 2pq - \ln q) - \frac{1}{2h_p^2}(p \ln^2 p + 2 \ln p \ln q + q \ln^2 q) + \eta_1(\ln z) + o(1).$$

By standard depoissonization we arrive at the same expression for $\mathbf{E}[S_n]$, and only the error term is modified by the depoissonization error of $O(n^{-1} \ln n)$, which comes on top of the Mellin inversion error of $o(1)$.

In order to calculate the second moment, we take the second derivative of (1) with respect to t and evaluate at $t = 0$. We have

$$\frac{\partial^2}{\partial t^2} B^*(s, t)|_{t=0} = -\frac{\Gamma(s)}{(1 - q^{-s+1} - p^{-s+1})^2}[1 + 3(p^2 + q^2)s - (1 - (p^2 + q^2)s)(-q^{-s+1} - p^{-s+1})].$$

This is the Mellin transform of the expected value of $S^2_{N(z)}$. In the inverse Mellin transform, the main contribution comes from $s_k = 0$. After depoissonization, we get

$$\mathbf{E}[S_n^2] \sim \frac{1}{h_p^2} \ln^2 n + \frac{1}{h_p^3}((1 - p^2) \ln^2 q - (p^2 - 2p) \ln^2 p - 2pq \ln p \ln q - (4p^3 - 8p^2 + 6p + 2\gamma q - 2) \ln q - (4p^2 q + 2\gamma p - 2p) \ln p) \ln n.$$

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

The variance follows from the first two moments, after straightforward algebraic simplification. ■

Curiously, the variance in the unbiased case is $O(1)$ (in this case, all of the poles lie on the vertical axis of the s complex plane). In the biased case ($p \neq q$), we have growth in the variance with the number of keys, which admits the existence of an asymptotic distribution for the typical climbing pathlength (after an appropriate normalization).

THEOREM 2: *Let S_n be the number of nodes on the path of typical climbing of a trie on n keys from a biased* Bernoulli(p) *model. Then*

$$\frac{S_n - \left(\frac{1}{h_p}\right)\ln n}{\sqrt{\ln n}} \xrightarrow{D} \mathcal{N}\left(0, \frac{pq}{h_p^3}(\ln p - \ln q)^2\right).$$

PROOF: We take $|t|$ small enough so that $\langle s_0(t), -s_0(t)\rangle \subseteq \langle -(1/4), (1/4)\rangle$. The inverse Mellin transform of (1) yields

$$B(z, t) = -\sum_{k=-\infty}^{\infty} \operatorname*{Res}_{s=s_k(t)} [B^*(s, t) z^{-s}] + O(z^{-M})$$

for any fixed $M > \rho$. Hence,

$$\begin{aligned}\mathbf{E}[e^{S_{N(z)}t}] = &-\frac{\Gamma(s_0(t))(e^{-t} - 1 + (p^2 + q^2)(1 - e^t)s_0(t))z^{-s_0(t)}}{p^{-s_0(t)+1}\ln p + q^{-s_0(t)+1}\ln q} \cdot \\ &-\frac{1}{p^{-s_0(t)+1}\ln p + q^{-s_0(t)+1}\ln q} \\ &\times \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} (\Gamma(s_k(t))(e^{-t} - 1 + (p^2 + q^2)(1 - e^t)s_k(t))z^{-s_k(t)}) + O(z^{-M}).\end{aligned}$$

We isolated the role of $s_0(t)$ because, as we will see shortly, it provides the dominant asymptotics when t is in a neighborhood of zero (where the gamma function also becomes very large), contrasting the finite limit of $\Gamma(s_k(t))$, as $t \to 0$, for each $k \neq 0$. Depoissonization gives

$$\mathbf{E}[e^{S_n t}] \sim -\frac{\Gamma(s_0(t))(e^{-t} - 1 + (p^2 + q^2)(1 - e^t)s_0(t))n^{-s_0(t)}}{p^{-s_0(t)+1}\ln p + q^{-s_0(t)+1}\ln q}.$$

The essential root $s_0(t)$ is a continuous infinitely differentiable function of t. For $t \to 0$, $s_0(t)$ has the expansion

$$s_0(t) = s_0(0) + s_0'(0)t + s_0''(0)\frac{t^2}{2} + O(t^3).$$

It is clear from (2) that $s_0(0) = 0$. Also, $s_0'(0) = -\frac{1}{h_p}$ and $s_0''(0) = -\frac{pq}{h_p^3}(\ln p - \ln q)^2$, as can be seen from the derivatives of (2). Further, we use the local expansions $1 -$

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

$e^x = -x + O(x^2)$ and $\Gamma(x) = \frac{1}{x} + \gamma + O(x)$, near $x = 0$. After substituting t by $v/\ln n$, for fixed v, we obtain

$$\mathbf{E}[e^{S_n(v/\sqrt{\ln n})}] \sim \frac{n^{-\left(s_0(0)\left(v/\sqrt{\ln n}\right)+s''_0(0)(v^2/2\ln n)+O(1/\ln^{3/2} n)\right)}}{s'(0)\left(p^{-s_0\left(v/\sqrt{\ln n}\right)+1}\ln p + q^{-s_0\left(v/\sqrt{\ln n}\right)+1}\ln q\right)}$$

$$\sim \frac{e^{\ln n\left((v/h_p\sqrt{\ln n})-(s''(0)v^2/2\ln n)\right)}}{s'_0(0)(p\ln p + q\ln q)}.$$

Therefore,

$$\mathbf{E}\left[e^{(S_n-((\ln n)/h_p))\left(v/\sqrt{\ln n}\right)}\right] \to e^{-(s''(0)/2)v^2},$$

with the right-hand side being the moment generating function of a normal random variate with mean 0 and variance $-s''_0(0)$. ■

Note that Theorem 2 and its proof can stand alone without the need for the development of the mean and variance of Theorem 1. However, the mean and variance given by the shortcut in Theorem 2 are only the leading terms in the full expansion provided by the more elaborate residue calculation of Theorem 1. One would not even detect the oscillations in the mean and variance with the method used in Theorem 2.

5. CLIMBING WITH THE LACK OF KNOWLEDGE OF *p*

If one is uninformed about p, one might be inclined to plead ignorance and simply generate moves in the random walk to the right and left subtrees with equal probability, hoping that this will average good and bad cases, achieving a sampling strategy that is not too much worse than typical climbing.

The result presented next indicates that the average speed of climbing is improved in uninformed climbing on average. Of course, the two strategies coincide when $p = q = 1/2$, but uninformed climbing requires less time than typical climbing as p gets away from 1/2, and the uninformed strategy speeds up considerably near the extremal values $p = 0$ and $p = 1$. However, the improved performance in the uninformed search comes at the expense of the quality of sampling, as less probable keys are given more weight than their actual probability.

The techniques are much the same as in typical climbing. We will only set up the problem, show the salient intermediate steps, and state the analogous results without proof.

Let S_n be the number of nodes on the path inscribed in the trie by the uninformed climbing and let $\phi_n(t)$ be its moment generating function. For example, given the trie of Figure 1, uninformed climbing might produce in two steps the key X_4, with $S_5 = 3$, or the null node (corresponding to a key value 0.100000. . .), in which case, $S_n = 2$. In either case, the key is generated with probability 1/4.

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

The length S_n satisfies a basic recurrence:

$$S_n|L_n = \begin{cases} 1 + S_{L_n} & \text{with probability } \frac{1}{2} \\ 1 + \tilde{S}_{R_n} & \text{with probability } \frac{1}{2}. \end{cases}$$

By a standard double expectation, we get

$$\mathbf{E}[e^{S_n t}] = \frac{1}{2}\mathbf{E}[e^{(1+S_{L_n})t}] + \frac{1}{2}\mathbf{E}[e^{(1+\tilde{S}_{R_n})t}].$$

Toward poissonization, we reintroduce the supergenerating function $A(z, t) = \sum_{n=0}^{\infty}(\phi_n(t)/n!)z^n$, and after manipulation similar to that in the case of typical climbing (mutatis mutandis, of course), we reach

$$A(z, t) - 1 = \frac{1}{2}e^t(e^{pz}A(qz, t) + e^{qz}A(pz, t) - 2 + z - ze^t).$$

As earlier, we reintroduce

$$B(z, t) = e^{-z}(A(z, t) - 1)$$

to work with a shifted function possessing a Mellin transform. We first obtain the functional equation

$$B(z, t) = \frac{1}{2}e^t(B(pz, t) + B(qz, t) + ze^{-z}(1 - e^t) + e^{-pz} + e^{-qz} - 2e^{-z}),$$

the Mellin transform of which is

$$B^*(s, t) = \frac{e^t\Gamma(s)(p^{-s} + q^{-s} - 2 + (1 - e^t)s)}{2 - e^t(p^{-s} + q^{-s})},$$

existing in the strip $\langle -1, s_0(t)\rangle$, and $s_0(t)$ being the only real root of the equation

$$p^{-s(t)} + q^{-s(t)} = 2e^{-t}.$$

We take $|t|$ small enough so that $\langle s_0(t), -s_0(t)\rangle \subseteq \langle -1/4, 1/4\rangle$. After all of the manipulation and the residue calculation, we reach the result for this random walk.

THEOREM 3: *Let S_n be the number of nodes on the path of uninformed climbing of a trie on n keys from the* Bernoulli(p) *model. Then*

$$\mathbf{E}[S_n] = 2\log_{1/pq} n + \frac{\ln^2 p + (1 - 2\gamma)\ln(pq) + \ln^2 q}{\ln^2(pq)} + \eta_2(\ln n) + o(1),$$

$$\mathbf{Var}[S_n] \sim \frac{2(\ln p - \ln q)^2}{\ln^3 \frac{1}{pq}} \ln n,$$

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

where $\eta_2(\cdot)$ is a function given by the Fourier expansion

$$\eta_2(u) = \begin{cases} -\frac{1}{h_p} \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \Gamma(s_{mk})(2+s_{km})e^{-s_{mk}u} & \text{if } \frac{\ln p}{\ln q} = \frac{m}{r} \text{ is rational with } \gcd(m, r) = 1 \\ 0 & \text{otherwise} \end{cases}$$

(with s_{mk} a nonzero solution of $p^{-s} + q^{-s} = 2$ with real part equal to zero), which is bounded by a very small number. The lower-order term in the variance can also have small bounded oscillations. Moreover, in the biased case,

$$\frac{S_n - 2\log_{1/pq} n}{\sqrt{\ln n}} \xrightarrow{D} \mathcal{N}\left(0, \frac{2(\ln p - \ln q)^2}{\ln^3 \frac{1}{pq}}\right).$$

6. EXTREMAL SAMPLING

To develop a sense for the extremes of the data present in the trie, a sampler might take after the extremal strategy of following a leftmost (for smallest) or a rightmost (for largest) path. Of course, the two strategies are symmetric with respect to the roles of p and q, and we only analyze one of them.

Let us reintroduce S_n as the number of nodes on the leftmost path and let $\phi_n(t)$ be its moment generating function. For instance, given the trie of Figure 1, the extremal leftmost climbing samples the key X_3, and $S_5 = 4$. If the leftmost path reaches a null node, we augment the corresponding prefix of zeros with a 1 to construct a representative sample of the smallest data.

The problem can be thought of in terms of the longest run of consecutive zeros. This case is connected to the maximum and second largest of independent and identically distributed geometric variables—let Z_i be the number of initial zeros in the key X_i and let $Z_{(i)}$ be the ith order statistics of $Z_1, \ldots, Z_n$. Then Z_i are identically distributed geometric variables and

$$S_n = \begin{cases} Z_{(2)} + 1 & \text{if } Z_{(1)} = Z_{(2)} \\ Z_{(2)} + 2 & \text{otherwise.} \end{cases}$$

One might be able to obtain a result for S_n from this representation. However, order statistics of independent and identically distributed discrete random variables are somewhat intricate because of the possible ties. We will proceed with our systematic analytic method.

The basic conditional recurrence is

$$S_n | L_n = 1 + S_{L_n},$$

for $n \geq 2$, giving

$$\mathbf{E}[e^{S_n t} | L_n] = e^{(1+S_{L_n})t}.$$

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

Hence,

$$\mathbf{E}[e^{S_n t}] = \sum_{\ell=0}^{n} \mathbf{E}\left[e^{(1+S_\ell)t}\right] \binom{n}{\ell} q^\ell p^{n-\ell}.$$

Toward poissonization, we reintroduce the generating function $A(z,\ t) = \sum_{n=0}^{\infty} z^n (\phi_n(t)/n!)$, where $\phi_n(t) = \mathbf{E}[e^{S_n t}]$. By steps similar to previous derivations in the other two strategies, we can easily establish the relation

$$A(z,\ t) - 1 - ze^t = e^t e^{pz} A(qz,\ t) - e^t(1 + pz + qze^t).$$

As we did in Section 4 for typical sampling, we do not poissonize $A(z,\ t)$ directly, but we poissonize the shifted version $A(z,\ t) - 1$, for the same technical reason to overcome existential problems of the Mellin transform. The routine is pretty much the same and we omit its details. One obtains the Mellin transform

$$B^*(s,\ t) = \frac{e^t \Gamma(s)(q^{-s} - 1 + sq(1 - e^t))}{1 - q^{-s}e^t}. \tag{4}$$

Theorem 4: *Let S_n be the number of nodes on the path of extremal climbing of a trie on n keys from the* Bernoulli(p) *model. Then*

$$\mathbf{E}[S_n] = \log_{1/q} n + \frac{2q + \ln q - 2\gamma}{2 \ln q} + \eta_3(\ln n) + o(1),$$

$$\mathbf{Var}[S_n] = \frac{1}{12} + \frac{\pi^2}{6 \ln^2 q} + \frac{2q}{\ln q} - \frac{q^2}{\ln^2 q} + o(1),$$

where $\eta_3(\cdot)$ is the function given by the Fourier expansion

$$\eta_3(u) = \begin{cases} \frac{1}{\ln q} \sum\limits_{\substack{k=-\infty \\ \neq 0}}^{\infty} \Gamma(s_k)(1 + s_k q) e^{-s_k u} & \text{if } \frac{\ln p}{\ln q} \text{ is rational} \\ 0 & \text{otherwise} \end{cases}$$

(with $s_k = 2\pi i k/\ln q$), which is bounded by a very small number. The $o(1)$ term in the variance might also have small bounded oscillations. Furthermore, $S_n - \lfloor \log_{1/q} n \rfloor$ does not have a nontrivial limit in distribution under any scaling.

Proof: The mean and variance are computed by the same poissonization–depoissonization routine, aided by the Mellin transform and residue calculation as was done for typical and uninformed climbing.

We restrict $|t| < 1/\ln(1/q)$. The distribution is found from the inverse of the Mellin transform (4). The poles of the transform are the roots of the equation

$$q^{-s}e^t = 1;$$

that is, they are at the points $s_k(t) = \left(\frac{1}{\ln q}\right)(t + 2\pi i k)$, for $k = 0,\ \pm 1,\ \pm 2, \ldots$.

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

So,

$$B(z,\, t) = -\sum_{k=-\infty}^{\infty} \operatorname*{Res}_{s=s_k(t)} [B^*(s,\, t) z^{-s}] + O(z^{-M}),$$

for any fixed $M > -1/\ln q$. Hence,

$$\mathbf{E}[e^{S_{N(z)}t}] = -\frac{1}{\ln q} \sum_{k=\infty}^{\infty} (\Gamma(s_k(t)) \times (1 - q^{s_k(t)} + q^{s_k(t)+1}(1 - e^t) s_k(t)) z^{-s_k(t)}) + O(z^{-M}).$$

It helps put the result in a concise form to define the function

$$\{x\} = x - \lfloor x \rfloor.$$

Depoissonization gives the same expression with n replacing z and an adjustment in the error term. We then have

$$\mathbf{E}[e^{(S_n - \lfloor \log_{1/q} n \rfloor)t}] \sim \frac{1}{\ln \frac{1}{q}} \sum_{k=-\infty}^{\infty} \Gamma\left(\frac{t + 2\pi i k}{\ln q}\right) \times \left(1 - q^{(t+2\pi i k/\ln q)} + q^{(t+2\pi i k/\ln q)+1}(1 - e^t) \frac{t + 2\pi i k}{\ln q}\right) \times n^{-(2\pi i k/\ln q)} e^{\{\log_{1/q} n\} t}.$$

We can write this as

$$\mathbf{E}[e^{(S_n - \lfloor \log_{1/q} n \rfloor)t}] \sim (g(t) + h_n(t)) e^{\{\log_{1/q} n\} t},$$

with $g(t)$ being equal to the zeroth term in the sum and $h_n(t)$ collecting all of the remaining terms. It is clear that no increasing scale of t will give a nontrivial limit.

It is well known that the function $\{\log_{1/q} n\}$ is dense in the interval [0, 1); see, for example, Kuipers and Niederreiter [9]. For any fixed t in the range $|t| < 1/\ln(1/q)$, the function $h_n(t)$ provides additional oscillations around $g(t)$, and, of course, the small error term can be made smaller than any arbitrary fixed number. Hence, no limit distribution exists. For a more detailed account of how such arguments work, see Christophi and Mahmoud [1]. ■

6.1. The Exact Distribution

Some of the exact distributions within the scope of this research might be amenable to direct combinatorial methods. We illustrate this for extremal climbing, to show that it can be done in principal.

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

THEOREM 5: *Let S_n be the number of nodes on the path of leftmost climbing of a trie on $n \geq 2$ keys from the* Bernoulli(p) *model. Then, for $k \geq 2$,*

$$\mathbf{Prob}(S_n = k) = nq^{k-1}(q(1-q^{k-1})^{n-1} - (1-q^{k-2})^{n-1}) + (1-q^k)^n - (1-q^{k-1})^n$$

and $\mathbf{Prob}(S_n = 0) = 0$ *and* $\mathbf{Prob}(S_n = 1) = p^n$.

PROOF: The boundary cases $\mathbf{Prob}(S_n = k)$, for $k = 1, 2$, are trivial. We will develop the result in terms of the number of edges $S'_n = S_n - 1$. Let $k \geq 2$. We dissect the event $\{S'_n = k\}$ into two disjoint subsets. One of the two subsets, A_1, corresponds to the case where the tree goes down the left path k edges and then turns right, with all of the keys having a string of k zeros as a prefix continuing with 1 at position $k+1$ (there must be at least two such keys). This construction leaves a null node dangling at the leftmost position in the tree. This event can occur by having r keys, $r = 2, \ldots, n$, in the subtree, the root of which is a sibling of the leftmost null node; the probability for any specific r to have this particular key structure is $(q^k p)^r$. The rest of the $n - r$ keys are not allowed to have a prefix of k zeros, otherwise they would disturb the pattern. The probability for these other keys not to have the forbidden prefix is $(1-q^k)^{n-r}$. The r keys can be chosen in $\binom{n}{r}$ ways. Hence,

$$\mathbf{Prob}(A_1) = \sum_{r=2}^{n} \binom{n}{r} (pq^k)^r (1-q^k)^{n-r}.$$

The second event, A_2, corresponds to the case where there is exactly one key at the end of a leftmost path with k internal vertices on it. By combinatorial arguments similar to that for A_1, we see that

$$\mathbf{Prob}(A_2) = \sum_{r=1}^{n-1} (r+1) \binom{n}{r+1} (pq^{k-1})^r q^k (1-q^{k-1})^{n-r-1}.$$

Now,

$$\begin{aligned}\mathbf{Prob}(S'_n = k) &= \mathbf{Prob}(A_1 \cup A_2) \\ &= \sum_{r=2}^{n} \binom{n}{r} (pq^k)^r (1-q^k)^{n-r} \\ &\quad + \sum_{r=1}^{n-1} (r+1) \binom{n}{r+1} (pq^{k-1})^r q^k (1-q^{k-1})^{n-r-1}.\end{aligned}$$

The sums can be reduced via the binomial theorem. ■

https://doi.org/10.1017/S0269964808000089 Published online by Cambridge University Press

Note

1. As an instance, when $(\ln p)/(\ln q) = (2/3)$, $\eta_1(\ln n)$ is bounded uniformly in n by 0.752×10^{-14}.

Acknowledgment

The second author wishes to thank the Institute of Statistical Mathematics, Tokyo, for supporting this research.

References

1. Christophi, C. & Mahmoud, H. (2005). The oscillatory distribution of distances in random tries. *Annals of Applied Probability* 15: 1536–1564.
2. De La Briandais, R. (1959). File searching using variable length keys. In *Proceedings of the Western Joint Computer Conference*, AFIPS, San Francisco, pp. 295–298.
3. Fagin, R., Nievergelt, J., Pippenger, N., & Strong, H. (1979). Extendible hashing: A fast access method for dynamic files. *ACM Transactions on Database Systems* 4: 315–344.
4. Flajolet, P., Gourdon, X., & Dumas, P. (1995). Mellin transform and asymptotic harmonic sums. *Theoretical Computer Science* 144: 3–58.
5. Fredkin, E. (1960). Trie memory. *Communications of the ACM* 3: 490–499.
6. Jacquet, P. (1989). *Contribution de l'Analyse d'Algorithmes a l'Evaluation de Protocoles de Communication.* Thèse Universitè de Paris Sud-Orsay, Paris, France.
7. Jacquet, P. & Szpankowski, W. (1998). Analytical depoissonization and its applications. *Theoretical Computer Science* 201: 1–62.
8. Knuth, D. (1998). *The art of computer programming*, Vol. 3: *Sorting and searching*, 2nd ed. Reading, MA: Addison-Wesley.
9. Kuipers, L. & Niederreiter, H. (1974). *Uniform distribution of sequences*. New York: Wiley.
10. Moon, J. (1970). Climbing random trees. *Aequationes Mathematicae* 5: 68–74.
11. Meir, A. & Moon, J. (1975). Climbing certain types of rooted trees I. In *Proceedings of the Fifth British Combinatorial Conference*, pp. 461–469.
12. Meir, A. & Moon, J. (1978). Climbing certain types of rooted trees II. *Acta Mathematica Academia Scientiarum Hungaricae* 31: 43–54.
13. Panholzer, A. (2005). The climbing depth of random trees. *Random Structures and Algorithms* 26: 84–109.
14. Schachinger, W. (1993). Beiträge zur Analyse von Datenstrukturen zur Digitalen Suche. Dissertation, Technische Universität Wien, Vienna.
15. Szpankowski, W. (2001). *Average case analysis of algorithms on sequences*. New York: Wiley.