J. Appl. Prob. **45,** 888–900 (2008)
Printed in England

AVERAGE-CASE ANALYSIS OF COUSINS IN *m*-ARY TRIES

HOSAM M. MAHMOUD,* *The George Washington University*
MARK DANIEL WARD,** *Purdue University*

Abstract

We investigate the average similarity of random strings as captured by the average number of 'cousins' in the underlying tree structures. Analytical techniques including poissonization and the Mellin transform are used for accurate calculation of the mean. The string alphabets we consider are m-ary, and the corresponding trees are m-ary trees. Certain analytic issues arise in the m-ary case that do not have an analog in the binary case.

Keywords: Analysis of algorithms; random tree; recurrence; Mellin transform; poissonization; combinatorics on words; similarity of strings

2000 Mathematics Subject Classification: Primary 05C05; 60C05
Secondary 68P05; 68P10; 68P20

1. Introduction

The similarity of strings is an important area, with numerous applications in data processing (comparison of computer files and various types of text) and computational biology (similarity of species on the hereditary scale by comparing DNA strands).

The trie is a data structure suitable for the storage, representation, and retrieval of—as well as supporting algorithms on—strings or digital data keys (bits, hexadecimal strings, words, DNA strands, etc.), which abound in science and technology. The trie was introduced in [3] and [11] for information retrieval. Tries also provide a model for the analysis of several important algorithms, such as radix exchange sort [14, Section 5.2.5] and extendible hashing [8].

Tries are usually defined recursively over a collection of strings composed of symbols from an alphabet

$$\mathcal{A} = \{a_1, \ldots, a_m\}.$$

If a node of a trie contains zero strings then the node is empty and does not appear in the trie structure. If a node contains exactly one string then the node is called a leaf or external node. If a node contains more than one string then the node is internal to the trie and the node has one or more descendants, each of which is also a trie. When splitting the strings from a node on the jth level of a trie into their proper locations in various nodes on the $(j+1)$th level, the splitting depends on the $(j+1)$th characters of the strings; the ℓth subtree contains all of the node's strings having the form $x_1x_2x_3\cdots$ such that $x_{j+1} = a_\ell \in \mathcal{A}$.

Received 7 February 2008; revision received 13 May 2008.
* Postal address: Department of Statistics, The George Washington University, 2140 Pennsylvania Avenue NW, Washington, DC 20052, USA. Email address: hosam@gwu.edu
** Postal address: Department of Statistics, Purdue University, 150 North University Street, West Lafayette, IN 47907-1451, USA. Email address: mdw@purdue.edu
Supported by NSF grant number 0603821.

Now we present the usual recursive definition of a trie. For a collection $\mathcal{C}$ of words with characters from $\mathcal{A}$, we write $\mathcal{C}_a$ to denote the set of all words in $\mathcal{C}$ that begin with the letter a. The set $\mathcal{C}_a \setminus a$ denotes the collection of words from $\mathcal{C}_a$ with the initial a removed from each word. The recursive definition of a trie is

$$\text{trie}(\mathcal{C}) = \begin{cases} \phi & \text{if } |\mathcal{C}| = 0, \\ \Box & \text{if } |\mathcal{C}| = 1, \\ \langle \text{trie}(\mathcal{C}_{a_1} \setminus a_1), \text{trie}(\mathcal{C}_{a_2} \setminus a_2), \ldots, \text{trie}(\mathcal{C}_{a_m} \setminus a_m) \rangle & \text{if } |\mathcal{C}| > 1. \end{cases}$$

The notation '$\Box$' represents an external node (one that stores a key), and $\langle ., ., \ldots, . \rangle$ stands for a tree rooted at an internal node and having the subtrees given in the list of arguments.

An example is helpful to illustrate the definition. A trie built from the twenty strings

$$\begin{aligned} S_1 &= 101010100111011\ldots, & S_{11} &= 100010100100000\ldots, \\ S_2 &= 001111001101010\ldots, & S_{12} &= 010010100111001\ldots, \\ S_3 &= 100100000100101\ldots, & S_{13} &= 000001101011011\ldots, \\ S_4 &= 000010111011100\ldots, & S_{14} &= 011010001000101\ldots, \\ S_5 &= 100100110011010\ldots, & S_{15} &= 100001101001111\ldots, \\ S_6 &= 000100001001101\ldots, & S_{16} &= 110101111000100\ldots, \\ S_7 &= 010001000011011\ldots, & S_{17} &= 111110111110001\ldots, \\ S_8 &= 111111001011101\ldots, & S_{18} &= 110101111100100\ldots, \\ S_9 &= 111011000010111\ldots, & S_{19} &= 001000100101000\ldots, \\ S_{10} &= 111110010011000\ldots, & S_{20} &= 110111011011110\ldots, \end{aligned}$$

has the form given in Figure 1.

In this paper we consider tries consisting of n independent keys each drawn from $\mathcal{A}^*$ (the set of all words on $\mathcal{A}$). We write

$$\mathrm{P}(a_j) = p_j$$

as the probability of selecting the jth letter of the alphabet (of course, $\sum_{j=1}^{m} p_j = 1$). For convenience, we assume, without loss of generality, that the probabilities are arranged in increasing order, i.e. $p_1 \le p_2 \le \cdots \le p_m$. In other words, we build a trie from n strings; the rth such string has the form

$$Y_{r,1} Y_{r,2} Y_{r,3} \cdots,$$

where $Y_{r,j} \in \mathcal{A}$ for each pair r, j and all of the $Y_{r,j}$s are selected independently, so

$$\mathrm{P}(Y_{r,1} Y_{r,2} Y_{r,3} \cdots Y_{r,j} = a_{\ell_1} a_{\ell_2} a_{\ell_3} \cdots a_{\ell_j}) = p_{\ell_1} p_{\ell_2} p_{\ell_3} \cdots p_{\ell_j}.$$

We will assume that p_j is positive for $j = 1, \ldots, m$. This avoids degeneracy and superfluous situations: if p_j is 0 for some j, it will mean that the jth branch of the tree never receives any keys, and there is no jth subtree, i.e. no internal node will have a jth subtree, so the tree is actually an $(m-1)$-ary branching structure.

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

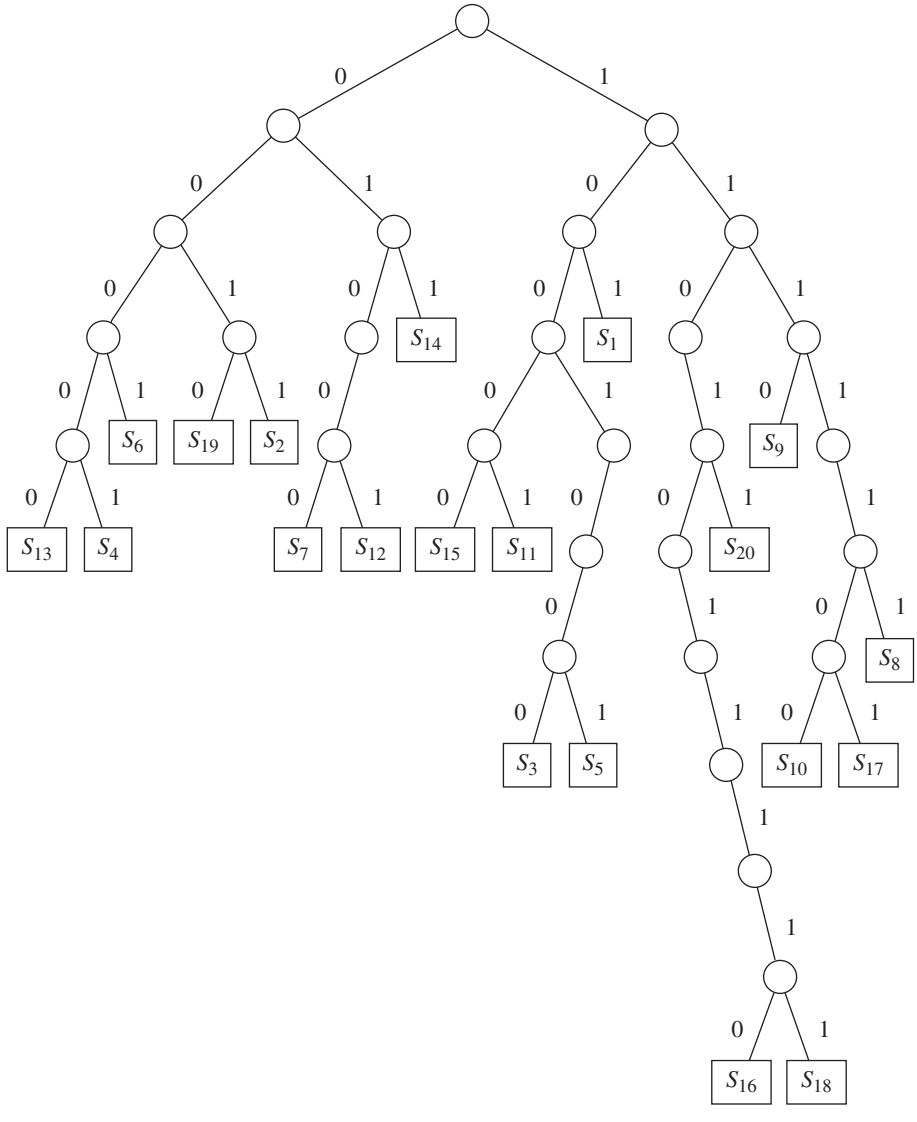

FIGURE 1: Example of a trie.

2. Scope

In tree structures built on n keys we enumerate the number $X_{n,k}$ of subtrees on the fringe that each contain $k > 1$ keys. A subtree with k keys is 'on the fringe' if it has no proper subtree that also has k keys. For instance, if $k = 2$ then $X_{n,2}$ denotes the number of pairs of siblings (i.e. 2-cousins). If $k = 3$ then $X_{n,3}$ denotes the number of families containing a pair of siblings and a nearest cousin; we refer to the two siblings and their nearest cousin collectively as a set of 3-cousins. In general, $X_{n,k}$ denotes the number of k-cousins. In the example given in Figure 1 there are seven sets of 2-cousins ($\{S_{13}, S_4\}$, $\{S_{19}, S_2\}$, $\{S_7, S_{12}\}$, $\{S_{15}, S_{11}\}$, $\{S_3, S_5\}$, $\{S_{16}, S_{18}\}$, $\{S_{10}, S_{17}\}$), four sets of 3-cousins ($\{S_{13}, S_4, S_6\}$, $\{S_7, S_{12}, S_{14}\}$,

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

$\{S_{16}, S_{18}, S_{20}\}, \{S_{10}, S_{17}, S_8\}$), two sets of 4-cousins ($\{S_{15}, S_{11}, S_3, S_5\}, \{S_9, S_{10}, S_{17}, S_8\}$), two sets of 5-cousins ($\{S_{13}, S_4, S_6, S_{19}, S_2\}, \{S_{15}, S_{11}, S_3, S_5, S_1\}$), etc. So $X_{20,2} = 7$, $X_{20,3} = 4$, $X_{20,4} = 2$, $X_{20,5} = 2$, etc. in this example.

We consider the expected value of $X_{n,k}$ in tries constructed from independent strings. Our analysis uses generating functions, poissonization and depoissonization, the Mellin transform, and singularity analysis. It is customary in this type of problem to set up a functional equation for the exponential (poissonized) generating function and solve it asymptotically via the Mellin transform. We use here an alternative approach derived from combinatorics on words, in which we find the poissonized generating function directly and in explicit form.

The results contain the data entropy function

$$h = h(p_1, \ldots, p_m) = -\sum_{j=1}^{m} p_j \ln p_j.$$

The main result of this paper is the following.

Theorem 1. *Let $X_{n,k}$ denote the number of k-cousins in an m-ary trie built over n independent keys from an m-ary alphabet $\{a_1, \ldots, a_m\}$ with probabilities $\mathrm{P}(a_j) = p_j$. Then*

$$\mathrm{E}[X_{n,k}] = \frac{1 - \sum_{j=1}^{m} p_j^k}{k(k-1)h} n + nQ_k(n) + o(n),$$

where $Q_k(\cdot)$ is a small oscillating function (possibly 0).

The rest of this paper is organized as follows. In Section 3 the general methodology is overviewed, where we briefly discuss the Mellin transform and its inverse, and the poissonization–depoissonization operation. In Section 4 the analysis of cousins is carried out at a high level, relegating the details to Appendix A. Subsection 4.1 is dedicated to the very transparent uniform alphabets, where one can specify lower-order terms more explicitly, and Subsection 4.2 is for the nonuniform case. In Appendix A we study the location of the characteristic roots that govern the average number of cousins.

3. Methodology

Two main tools in the forthcoming analysis are the Mellin transform and poissonization–depoissonization. These methods are by now standard, so we will not present lengthy details, but rather we refer the reader to standard sources on such material.

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

$$\mathcal{M}[f(x), s] := \int_0^\infty f(x) x^{s-1} \, \mathrm{d}s,$$

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

$$a < \mathrm{Re}(s) < b$$

for real numbers $a < b$. We will denote this strip by $\langle a, b \rangle$. If $f(x) = O(x^\alpha)$ as $x \to 0$ and $f(x) = O(x^\beta)$ as $x \to \infty$, the Mellin transform of $f(x)$ is defined for all s in the strip $\langle -\alpha, -\beta \rangle$; this is referred to as the *fundamental strip*.

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

The function $f(x)$ can be recovered from its transform by a line integral

$$f(x) = \frac{1}{2\pi \mathrm{i}} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty} f^*(s) x^{-s}\, \mathrm{d}s \quad \text{for any } c \in (a, b).$$

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 is often a small residual error of the form $O(x^{-\theta})$ for an arbitrarily large positive number θ. For a survey of the uses of the Mellin transform in the analysis of algorithms, see [10].

The expression for the average number of cousins among n strings is complicated for direct analysis; however, a poissonized version is amenable to asymptotic analysis via the Mellin transform. In this context, poissonization means considering an analogous problem, but with a Poisson random number of strings, 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 rather refer the reader to the original work [13] or its presentation in textbook style [18, Chapter 10].

4. Cousins in tries

Each set of k-cousins corresponds to a unique subtree consisting of k strings that have a unique longest string $w \in \mathcal{A}^*$ appearing at the start of all k strings in the subtree. Conversely, each $w \in \mathcal{A}^*$ uniquely denotes such a subtree if the following $m+1$ conditions are satisfied.

Condition 0: Exactly k of the n keys inserted in the trie have w as a prefix.

Condition j: For $1 \le j \le m$, fewer than k of the n keys inserted in the trie have wa_j as a prefix.

In other words, w is the unique longest common prefix for exactly k strings, and the splitting of the k strings occurs exactly at w (as opposed to further down the tree). So, $w \in \mathcal{A}^*$ is the unique longest common prefix of exactly k strings with probability

$$\binom{n}{k} \left(1 - \sum_{j=1}^{m} p_j^k\right) \mathrm{P}(w)^k (1 - \mathrm{P}(w))^{n-k}. \tag{1}$$

The binomial coefficient accounts for the number of ways to choose which of the n keys will be in the k-cousin, and the factor $(1 - \sum_{j=1}^{m} p_j^k)$ is inserted to exclude the case of a string that is a common prefix of k keys, but not the longest possible for them. Let $\mathbf{1}_{\mathcal{E}}$ be the indicator of the event $\mathcal{E}$, that is, the Bernoulli random variable that assumes the value 1 when $\mathcal{E}$ occurs and the value 0 otherwise. Let $\mathcal{E}_{n,k}(w)$ be the event that the word w is the unique longest common prefix of exactly k of the n strings; this event occurs with the probability in (1).

The count $X_{n,k}$ has a representation as a sum of indicators,

$$X_{n,k} = \sum_{w \in \mathcal{A}^*} \mathbf{1}_{\mathcal{E}_{k,n}(w)},$$

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

with average

$$\mathrm{E}[X_{n,k}] = \sum_{w\in\mathcal{A}^*} \mathrm{E}[\mathbf{1}_{\mathcal{E}_k(w)}] = \sum_{w\in\mathcal{A}^*} \binom{n}{k}\left(1-\sum_{j=1}^{m} p_j^k\right)\mathrm{P}(w)^k(1-\mathrm{P}(w))^{n-k}.$$

This yields the exponential generating function

$$F_k(z) := \sum_{n\geq 0}\mathrm{E}[X_{n,k}]\frac{z^n}{n!} = \sum_{w\in\mathcal{A}^*}\frac{1}{k!}\left(1-\sum_{j=1}^{m} p_j^k\right)\mathrm{P}(w)^k z^k \mathrm{e}^{z(1-\mathrm{P}(w))}.$$

For simplicity, in the remainder of this paper we use the notation

$$\rho := \frac{1}{k!}\left(1-\sum_{j=1}^{m} p_j^k\right).$$

Note that $\tilde{F}_k := \mathrm{e}^{-z}F_k(z)$ has the following poissonization interpretation:

$$\begin{aligned}\tilde{F}_k(z) &= \sum_{n\geq 0}\mathrm{E}[X_{n,k}]\frac{z^n}{n!}\mathrm{e}^{-z}\\ &= \sum_{n\geq 0}\mathrm{E}[X_{n,k}]\,\mathrm{P}(N_z=n)\\ &= \sum_{n\geq 0}\mathrm{E}[X_{N_z,k}\mid N_z=n]\,\mathrm{P}(N_z=n)\\ &= \mathrm{E}[X_{N_z,k}],\end{aligned}$$

where N_z is a random variable with a Poisson distribution with mean z.

The poissonized version is just like the original fixed population problem, only the fixed value n is replaced by a random variable N_z; the associated generating function is

$$\tilde{F}_k(z) = \sum_{n\geq 0}\mathrm{E}[X_{n,k}]\frac{z^n}{n!}\mathrm{e}^{-z} = \sum_{w\in\mathcal{A}^*}\rho\,\mathrm{P}(w)^k z^k \mathrm{e}^{-z\,\mathrm{P}(w)}.$$

As $z\to 0$, we have $\tilde{F}_k(z) = O(z^k)$; this is straightforward, since $\mathrm{e}^{-z\,\mathrm{P}(w)} = 1 + O(z\,\mathrm{P}(w))$. For each (fixed) ε with $0<\varepsilon<1$, as $z\to\infty$, we have $\tilde{F}_k(z) = O(z^{1+\varepsilon})$; to see this, we first observe that

$$z^{k-1-\varepsilon}\mathrm{e}^{-z\,\mathrm{P}(w)} \leq \left(\frac{k-1-\varepsilon}{\mathrm{P}(w)}\right)^{k-1-\varepsilon}\mathrm{e}^{-(k-1-\varepsilon)} \quad \text{for all } z\geq 0$$

(note that, for each constant $c>0$, the function $z^{k-1-\varepsilon}\mathrm{e}^{-cz}$ has a maximum value over all $z\geq 0$; the maximum occurs exactly at $z=(k-1-\varepsilon)/c$). Thus,

$$\begin{aligned}\tilde{F}_k(z) &= O(z^{1+\varepsilon})\sum_{w\in\mathcal{A}^*}\mathrm{P}(w)^k\left(\frac{k-1-\varepsilon}{\mathrm{P}(w)}\right)^{k-1-\varepsilon}\mathrm{e}^{-(k-1-\varepsilon)}\\ &= O(z^{1+\varepsilon})\sum_{w\in\mathcal{A}^*}\mathrm{P}(w)^{1+\varepsilon}\\ &= O(z^{1+\varepsilon}).\end{aligned}$$

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

Of course, we emphasize that the constants hidden in the $O(\cdot)$ terms in the lines above depend on (the fixed values of) k and ε.

So, the fundamental strip of the Mellin transform of $\tilde{F}_k(z)$ is $\langle -k, -1\rangle$. (The real parts in the fundamental strip fall in the open interval $(-k, -1)$.)

Using the well-known properties of the Mellin transform (see, e.g. [9], [10], and [18, Chapter 10]), we compute

$$\begin{aligned}
\tilde{F}_k^*(s) &= \sum_{w\in\mathcal{A}^*} \rho\,\mathrm{P}(w)^k \mathcal{M}[z^k \mathrm{e}^{-z\,\mathrm{P}(w)}, s] \\
&= \sum_{w\in\mathcal{A}^*} \rho\,\mathrm{P}(w)^k \mathcal{M}[\mathrm{e}^{-z\,\mathrm{P}(w)}, s+k] \\
&= \sum_{w\in\mathcal{A}^*} \rho\,\mathrm{P}(w)^k\,\mathrm{P}(w)^{-(s+k)} \mathcal{M}[\mathrm{e}^{-z}, s+k] \\
&= \rho\Gamma(s+k) \sum_{w\in\mathcal{A}^*} \mathrm{P}(w)^{-s}.
\end{aligned}$$

Finally, if we use $b_1, \ldots, b_m$ to denote the number of occurrences of $a_1, \ldots, a_m$, respectively, found in w, it follows that

$$\begin{aligned}
\sum_{w\in\mathcal{A}^*} \mathrm{P}(w)^{-s} &= \sum_{b_1,\ldots,b_m} \binom{b_1+\cdots+b_m}{b_1,\ldots,b_m} (p_1^{b_1}\cdots p_m^{b_m})^{-s} \\
&= \sum_{b_1,\ldots,b_m} \binom{b_1+\cdots+b_m}{b_1,\ldots,b_m} (p_1^{-s})^{b_1}\cdots(p_m^{-s})^{b_m} \\
&= \frac{1}{1-\sum_{\ell=1}^m p_\ell^{-s}}.
\end{aligned}$$

So the Mellin transform of $\tilde{F}_k(z)$ is

$$\tilde{F}_k^*(s) = \rho\Gamma(s+k)\frac{1}{1-\sum_{\ell=1}^m p_\ell^{-s}}.$$

It is well known that the poles of the Mellin transform play an important role in determining the character of the transformed function, and the equation

$$1-\sum_{j=1}^m p_j^{-s} = 0$$

will be a deciding element for the inverse Mellin transform. We will call this relation the *characteristic equation*, and we refer to its roots as the *characteristic roots*. It is clear that -1 is a root. This root provides dominant asymptotic terms, and we will call it the *dominant pole*. The characteristic equation has many other poles. We study the location of the poles in Appendix A.

4.1. Uniform alphabets

The case of a uniform alphabet, where all the symbols are equally likely, has a transparent structure for the inverse Mellin transform. In this case $p_j = 1/m$ for all $j = 1, \ldots, m$, and the

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

Mellin transform simplifies to

$$\tilde{F}_k^*(s) = \frac{1}{k!}(1 - m^{-k+1})\Gamma(s+k)\frac{1}{1-m^{s+1}}.$$

This Mellin transform has simple poles at the roots of the characteristic equation

$$m^{s+1} = 1 = \mathrm{e}^{2\pi \mathrm{i} j}$$

for any integer j, that is, the roots are

$$s_j = -1 + \frac{2\pi \mathrm{i} j}{\ln m}, \qquad j = \ldots, -2, -1, 0, 1, 2, \ldots;$$

the pole $s_0 = -1$ is the dominant pole, as we will see. We can invert the Mellin transform using Cauchy's residue theorem. We integrate counterclockwise over a large rectangle with corners $-\frac{3}{2} \pm \mathrm{i}\lambda$ and $\theta \pm \mathrm{i}\lambda$ for large λ and θ (the numbers λ and θ are chosen so that the sides do not cross any poles). As λ grows arbitrarily large, the integrals on the top and bottom sides of the integration box vanish (owing to the rapid decrease in the magnitude of the gamma function), and what is left is the integration on the vertical lines at $\mathrm{Re}(s) = -\frac{3}{2}$ (which is the negative of the desired inverse transform) and an error of magnitude $O(z^{-\theta})$ as $z \to \infty$, corresponding to the integration on the vertical line $\mathrm{Re}(s) = \theta$. But, when $\lambda \to \infty$ and $\theta = 0$, the integration box grows to encompass all the poles to the right of the vertical line $\mathrm{Re}(s) = -\frac{3}{2}$. And so,

$$\tilde{F}_k(z) = \frac{1}{2\pi \mathrm{i}} \int_{-3/2-\mathrm{i}\infty}^{-3/2+\mathrm{i}\infty} \tilde{F}_k^*(s) z^{-s}\, \mathrm{d}s = \sum_{j \in \mathbb{Z}} -\mathrm{Res}[\tilde{F}_k^*(s) z^{-s}; s = z_j] + O(1).$$

If we define

$$Q_k(z) = \frac{1 - m^{-k+1}}{k! \ln m} \sum_{j \in \mathbb{Z} \setminus \{0\}} \Gamma\left(k - 1 + \frac{2\pi \mathrm{i} j}{\ln m}\right) \exp(-2\pi \mathrm{i} j \log_m z),$$

we have the poissonized representation

$$\mathrm{E}[X_{N_z,k}] = \tilde{F}_k(z) = \frac{1 - m^{-k+1}}{k(k-1)\ln m} z + z Q_k(z) + O(1).$$

Standard depoissonization gives a similar result, with n replacing z, but a small $O(1)$ depoissonization error appears. For the uniform case, we obtain the following version of Theorem 1, with slightly refined error terms:

$$\mathrm{E}[X_{n,k}] = \frac{1 - m^{-k+1}}{k(k-1)\ln m} n + n Q_k(n) + O(1).$$

Remarks. (i) The error is not reduced even if we take θ much larger, and thus obtain a smaller inversion error of order $O(z^{-\theta})$. There will ultimately still be a residual $O(1)$ error from depoissonization. For example, if we take $\theta = 5$, the inversion error is $O(z^{-5})$, and when depoissonized it is reflected into an $O(n^{-5})$ error, which the $O(1)$ depoissonization error subsumes.

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

TABLE 1: Uniform bounds on the oscillations for $m = 4$ and some small values of k.

k	A uniform bound on $\lvert Q_k(z)\rvert$
2	0.002 339 1
3	0.004 528 5
4	0.005 900 5
5	0.006 511 3
6	0.006 610 5
7	0.006 418 9
8	0.006 084 3
9	0.005 694 0

(ii) The function Q_k is an oscillating function that is absolutely bounded uniformly in z. In Table 1 we present an absolute uniform bound on the oscillations for $m = 4$ and a few small values of k. It is quite remarkable that the oscillations here are relatively large. For instance, with $m = 4$, $k = 8$, $Q_8(200) \approx -0.005\,300\,145\,28$, and $|Q_8(z)| < 0.006\,084\,3$ for all z, while $(1 - m^{-k+1})/k(k-1)\ln m \approx 0.012\,880\,419\,52$.

(iii) In numerous problems on tries, the oscillations are of a much smaller order of magnitude; see, e.g. [1], [2], [5], and [6], where oscillations of the typical order 10^{-5}, and sometimes as small as 10^{-14}, are reported.

(iv) For depoissonization, versions of [4] and [18, Chapter 10] can be helpful to our purposes.

4.2. Nonuniform alphabets

The presentation for nonuniform alphabets involves some technicalities. The presentation of the lower-order terms will not always be as refined (as compared to the uniform alphabet scenario above).

We still go through the Mellin transform inversion. The main contribution comes from $s = z_0 = -1$, which is

$$-\text{Res}[\tilde{F}_k^*(s)z^{-s}; s = z_0 = -1] = \rho\Gamma(k-1)\frac{z}{h} = \frac{1 - \sum_{j=1}^m p_j^k}{k(k-1)h}z,$$

where $h = h(p_1, \ldots, p_m) = -\sum_{j=1}^m p_j \ln p_j$ is the data entropy.

As for the rest of the poles, all combined they contribute only $o(z)$ if none lie on the vertical line $\text{Re}(s) = -1$, which is a corollary of the Wiener–Ikehara theorem [15] (all that is required in this case is that the Mellin transform, with the singularity at -1 removed, can be analytically continued to a domain to the right of the line $\text{Re}(s) = -1$, which is the case). If some poles fall on the line $\text{Re}(s) = -1$, they introduce small oscillations at the linear level, and the rest contribute only $o(z)$. As shown in Appendix A, there is a number $\Delta_m \geq -1$, at which—or to the left of which—all the poles lie. In carrying out the inversion by shifting the line of integration, we take that line to the right of Δ_m, and that will account for all the poles. Upon completing the residue calculation and performing depoissonization, we prove Theorem 1.

Remark. Some of the poles may lie on the vertical line $\text{Re}(s) = -1$, such as, for example, the uniform case for any m or the case in which $m = 3$, $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{4}$, and $p_3 = \frac{1}{4}$. In the latter case we have two sets of poles: a group lined up at $\text{Re}(s) = -1$ and another at $\text{Re}(s) = 0$. For every integer j, the complex number $-1 + 4j\pi\mathrm{i}/\ln 4$ is a pole lined up vertically with the

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

dominant pole, and the complex number $(4j+2)\pi\mathrm{i}/\ln 4$ is a pole lined up vertically with the imaginary axis. In this example we can find an explicit representation:

$$\mathrm{E}[X_{n,k}] = \left(\frac{1-4^{-k+1}}{k(k-1)\ln 4} + \frac{1-4^{-k+1}}{k!\ln 4}\sum_{j\in\mathbb{Z}\setminus\{0\}}\Gamma\left(k-1+\frac{4\pi\mathrm{i}j}{\ln 4}\right)\exp(-4\pi\mathrm{i}j\log_4 z)\right)z + O(1).$$

The $O(1)$ term itself contains the oscillations

$$\frac{1-4^{-k+1}}{k!\ln 4}\sum_{j\in\mathbb{Z}\setminus\{0\}}\Gamma\left(k-1+\frac{(4j+2)\pi\mathrm{i}}{\ln 4}\right)\exp(-(4j+2)\pi\mathrm{i}j\log_4 z).$$

Appendix A. Location of the poles

We will study the location of the roots of the characteristic equation in this appendix via a number of small technical lemmas. We first give an overview of the plan of the proof. In the following lemmas we will use the definition of the minimum symbol probability: let $p := \min_{1\le j\le m} p_j$; there may be several symbols of the same minimal probability $p = p_1 = p_2 = \cdots = p_\nu$, $\nu \ge 1$ (recall that the probabilities are arranged in increasing order). For positive constants K and a, we will call a function like $K\mathrm{e}^{ax}$ an *exponential function with index a*.

If we cut up the complex s-plane into horizontal slices each of height $2\pi\mathrm{i}/|\ln p|$, every slice will contain exactly one characteristic root. Each root's real part falls between -1 and a fixed positive real number Δ_m. Thus, all the characteristic roots fall in the vertical strip

$$-1 \le \mathrm{Re}(s) \le \Delta_m.$$

We prove this general picture in the next few lemmas. The proof follows and generalizes some of the lines in [7]. Origins of this argument can be found in [17] (see also [16]). There are no direct references to the versions we need for our proof, and we thought that drawing the full picture would be helpful for the exposition.

Lemma 1. ([17].) *Let s be a characteristic root. Then*

$$-1 \le \mathrm{Re}(s).$$

Proof. Suppose, toward a contradiction, that there is a root s that lies to the left of the dominant pole $s_0 = -1$ (i.e. $\mathrm{Re}(s) < -1$). For such a root,

$$\left|\sum_{j=1}^{m} p_j^{-s}\right| \le \sum_{j=1}^{m} |p_j^{-s}| = p_1^{-\mathrm{Re}(s)} + \cdots + p_m^{-\mathrm{Re}(s)} < p_1 + \cdots + p_m = 1;$$

thus, s cannot be a root of the characteristic equation, which is a contradiction.

Lemma 2. *Let s be a characteristic root. There exists a real number Δ_m (which depends only on m and the symbol probabilities) such that*

$$\mathrm{Re}(s) \le \Delta_m.$$

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

Proof. Let us exclude the case in which $\nu = m$ (the uniform alphabet), since Δ_m is clearly -1 in this case. So, we take $\nu < m$. Compare the two functions $f(x) = 1+p_{\nu+1}^{-x}+p_{\nu+2}^{-x}+\cdots+p_m^{-x}$ and $g(x) = \nu p^{-x}$ for real x. The function $g(x)$ is an exponential function with index $|\ln p|$ and $f(x)$ is 1 plus a linear combination of exponential functions with indices that are all less than that in $g(x)$, as p is the minimal symbol probability. The function $f(x)$ rises from 1 at $-\infty$ to ∞ at $+\infty$, and the function $g(x)$ rises from 0 at $-\infty$ to ∞ at $+\infty$. According to the indices (which govern rates of increase), the two functions intersect at a point $x = \Delta_m$. The curve of $f(x)$ stays above that of $g(x)$ until Δ_m, where the two functions become equal, then $f(x)$ passes below $g(x)$ for $x > \Delta_m$.

Now, if s is a root of the characteristic equation, it satisfies

$$p_1^{-s} + p_2^{-s} + \cdots + p_m^{-s} = 1$$

and

$$\begin{aligned} g(\mathrm{Re}(s)) &= \nu p^{-\mathrm{Re}(s)} \\ &= \nu |p^{-s}| \\ &= |1 - (p_{\nu+1}^{-s} + \cdots + p_m^{-s})| \\ &\le 1 + p_{\nu+1}^{-\mathrm{Re}(s)} + \cdots + p_m^{-\mathrm{Re}(s)} \\ &= f(\mathrm{Re}(s)). \end{aligned}$$

So, the real part of s must be at most Δ_m.

We know that there are poles with real part -1, such as s_0 (and possibly many others); we see that

$$-1 \le \mathrm{Re}(s) \le \Delta_m,$$

and the number Δ_m must be at least -1. By further partitioning the vertical strip $-1 \le \mathrm{Re}(s) \le \Delta_m$ into 'cells' of height $2\pi\mathrm{i}/|\ln p|$ each, next we will demonstrate that each cell contains exactly one characteristic root. It is sufficient for our purpose to consider the strip

$$-2 \le \mathrm{Re}(s) \le \Delta_m + 1.$$

We define the cells B_j to be

$$B_j = \left\{ s\colon\ -2 \le \mathrm{Re}(s) \le \Delta_m + 1,\ \frac{(2j-1)\pi}{|\ln p|} \le \mathrm{Im}(s) \le \frac{(2j+1)\pi}{|\ln p|} \right\} \quad \text{for } j \in \mathbb{Z}.$$

To prove that each cell contains exactly one root, we resort to Rouché's theorem [12, p. 280], a good aid in locating the 0s of an entire function. We state that theorem for the reader's convenience.

Theorem 2. (Rouché's theorem [12, p. 280].) *Let the complex-valued functions $f(z)$ and $g(z)$ be holomorphic inside and on some closed contour C, with $|g(z)| < |f(z)|$ on C. Then $f(z)$ and $f(z) + g(z)$ have the same number of 0s inside C (each 0 is counted according to its multiplicity).*

The general idea in the application of Rouché's theorem is to replace a complicated function with a dominating simpler one, with easy to calculate roots, and state something about the number of 0s in a certain closed domain.

https://doi.org/10.1239/jap/1222441835 Published online by Cambridge University Press

Lemma 3. *For each integer j, the cell B_j contains exactly one characteristic root.*

Proof. We apply Rouché's theorem with the two functions $f(s) = \nu p^{-s} - 1$ and $g(s) = p_{\nu+1}^{-s} + \cdots + p_m^{-s}$. Both are entire (hence holomorphic, as required). The function $f(s)$ has the roots $(\ln \nu + 2\pi \mathrm{i} j)/\ln p$ for integer j. Within the cell B_j, there is only one of these (at the middle of the vertical line segment defined by the intersection of the line $\mathrm{Re}(s) = \ln \nu / \ln p$ and the cell). It is then sufficient to show that $|g(s)| < |f(s)|$ on the boundary of B_j, as it will then follow that $f(s) + g(s) = p_1^{-s} + p_2^{-s} + \cdots + p_m^{-s} - 1$ has exactly one root in B_j. We take up the following four sides.

- *The right side.* On this side $\mathrm{Re}(s) = \Delta_m + 1$ and

$$\begin{aligned} |g(s)| &= |p_{\nu+1}^{-s} + \cdots + p_m^{-s}| \\ &\le |p_{\nu+1}^{-s}| + \cdots + |p_m^{-s}| \\ &\le p_{\nu+1}^{-\Delta_m - 1} + \cdots + p_m^{-\Delta_m - 1} \\ &< \nu p^{-\Delta_m - 1} - 1 \\ &\le |f(s)|. \end{aligned}$$

- *The left side.* On this side $\mathrm{Re}(s) = -2$. We have $(p_1 + \cdots + p_m)^2 = 1$, i.e. $p_1^2 + \cdots + p_m^2 + 2\sum_{1 \le j,\, \ell \le m} p_j p_\ell = 1$, and $p_1^2 + \cdots + p_m^2 < 1$. It follows that

$$|g(s)| \le |p_{\nu+1}^{-s}| + \cdots + |p_m^{-s}| \le p_{\nu+1}^2 + \cdots + p_m^2 < 1 - \nu p^2.$$

  We also have

$$\nu p^2 \le \frac{\nu}{m^2} \le \frac{m}{m^2} < 1,$$

  and, therefore,

$$|g(s)| < |1 - \nu p^{-s}| \le |f(s)|.$$

- *The top side.* On this side $\mathrm{Im}(s) = (2j + 1)\pi/|\ln p|$. So, $f(s) = \nu p^{-s} - 1 = -\nu p^{-\mathrm{Re}(s)} - 1$ and $|f(s)| = 1 + \nu p^{-\mathrm{Re}(s)}$, and

$$|g(s)| \le |p_{\nu+1}^{-s}| + \cdots + |p_m^{-s}| \le p_{\nu+1}^{-\mathrm{Re}(s)} + \cdots + p_m^{-\mathrm{Re}(s)} < 1 + \nu p^{-\mathrm{Re}(s)} = |f(s)|.$$

  The strict equality is according to the exponentiality index, as p is minimal.

- *The bottom side.* On this side $\mathrm{Im}(s) = (2j - 1)\pi/|\ln p|$, and the argument is similar to that on the top side.

The proof is complete.

Acknowledgement

The authors are indebted to Michael Drmota for several conversations and advice.

References

[1] Aguech, R., Lasmar, N. and Mahmoud, H. (2006). Distances in random digital search trees. *Acta Informatica* **43,** 243–264.

[2] Aguech, R., Lasmar, N. and Mahmoud, H. (2006). Limit distribution of distances in biased random tries. *J. App. Prob.* **43,** 1–14.

[3] Briandais, R. D. L. (1959). File searching using variable length keys. In *Proc. Western Joint Comput. Conf.*, AFIPS, San Francisco, CA, pp. 295–298.
[4] Bruss, F. T., Louchard, G. and Ward, M. D. (2008). Injecting unique minima into random sets. To appear in *ACM Trans. Algorithms.*
[5] Christophi, C. and Mahmoud, H. (2005). The oscillatory distribution of distances in random tries. *Ann. Appl. Prob.* **15,** 1536–1564.
[6] Christophi, C. and Mahmoud, H. (2008). On climbing tries. *Prob. Eng. Inf. Sci.* **22,** 133–149.
[7] Drmota, M., Reznik, Y., Savari, S. and Szpankowski, W. (2008). Analysis of variable-to-fixed length codes. Submitted.
[8] Fagin, R., Nievergelt, J., Pippenger, N. and Strong, H. (1979). Extendible hashing—a fast access method for dynamic files. *ACM Trans. Database Systems* **4,** 315–344.
[9] Flajolet, P. and Sedgewick, R. (1995). Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoret. Comput. Sci.* **144,** 101–124.
[10] Flajolet, P., Gourdon, X. and Dumas, P. (1995). Mellin transforms and asymptotics: harmonic sums. *Theoret. Comput. Sci.* **144,** 3–58.
[11] Fredkin, E. (1960). Trie memory. *Commun. ACM* **3,** 490–499.
[12] Henrici, P. (1986). *Applied and Computational Complex Analysis*. John Wiley, New York.
[13] Jacquet, P. and Szpankowski, W. (1998). Analytical depoissonization and its applications. *Theoret. Comput. Sci.* **201,** 1–62.
[14] Knuth, D. E. (1998). *The Art of Computer Programming*, Vol. 3, 2nd edn. Addison-Wesley, Reading, MA.
[15] Korevaar, J. (2002). A century of complex Tauberian theory. *Bull. Amer. Math. Soc. (N. S.)* **39,** 475–531.
[16] Schachinger, W. (1992). Beitrage zur analyse von datenstrkturen zur digitalen suche. Doctoral Thesis, Technical University of Vienna.
[17] Schachinger, W. (2000). Limiting distributions for the costs of partial match retrievals in multidimensional tries. *Random Structures Algorithms* **17,** 428–459.
[18] Szpankowski, W. (2001). *Average Case Analysis of Algorithms on Sequences*. John Wiley, New York.