ANALYSIS OF RANDOM LC-TRIES

Luc Devroye
School of Computer Science
McGill University
Montreal, Canada H3A 2K6
luc@cs.mcgill.ca

June 13, 2001

ABSTRACT. LC tries were introduced by Andersson and Nilsson in 1993. They are compacted versions of tries or PATRICIA tries in which, from the top down, maximal height complete subtrees are level compressed. Andersson and Nilsson (1993) showed that for i.i.d. uniformly distributed input strings, the expected depth of the LC PATRICIA trie is $\Theta(\log^* n)$. In this paper, we refine and extend this result. We analyze both kinds of LC tries for the uniform model, and study the depth of a typical node and the height H_n. For example, we show that H_n is in probability asymptotic to $\log_2 n$ and $\sqrt{2 \log_2 n}$ for the LC trie and the LC PATRICIA trie, respectively, and that for both tries, the depth of a typical node is asymptotic to $\log^*(n)$ in probability and in expectation.

KEYWORDS AND PHRASES. Trie, PATRICIA tree, probabilistic analysis, law of large numbers, LC trie, height of a tree.

CR CATEGORIES: 3.74, 5.25, 5.5.

1991 MATHEMATICS SUBJECT CLASSIFICATIONS: 60D05, 68U05.

Authors' address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6. The authors' research was sponsored by NSERC Grant A3456 and FCAR Grant 90-ER-0291.

Introduction

Tries are efficient data structures that were initially developed and analyzed by Fredkin (1960) and Knuth (1973). The tries considered here are constructed from n independent numbers $X_1, \ldots, X_n$, each drawn uniformly from $[0, 1]$. The binary expansion of X_i gives rise to an infinite binary string $(X_{i1}, X_{i2}, \ldots)$ which in turn defines an infinite path in a binary tree. From the root, we take the X_{i1}-st child, then its X_{i2}-nd child, and so forth. The collection of nodes and edges visited by the union of the n paths is the infinite trie. If the X_i's are different, then each infinite path ends with a suffix path that is traversed by that string only. If this suffix path for X_i starts at node u, then we may trim it by cutting away everything below node u. This node becomes the leaf representring X_i. If this process is repeated for each X_i, we obtain a finite tree with n leaves, called the trie. PATRICIA is a space efficient improvement of the classical trie discovered by Morrison (1968) and first studied by Knuth (1973). It is simply obtained by removing from the trie all internal nodes with one child. Thus, it necessarily has n leaves. Each non-leaf (or internal) node has two or more children.

The LC trie is a further compactification of the trie or the PATRICIA tree. The following operation is repeated recursively: at the root of the trie or PATRICIA tree T, find the highest complete subtree C, and let h be its height. (This means that all 2^h nodes at distance h from the root exist in T, but not all 2^{h+1} nodes at distance $h + 1$.) Let $T_i, 1 \leq i \leq 2^h$ be the subtrees rooted at the 2^h nodes at distance h from the root. Replace T by the root of T and 2^h child subtrees, $T_i, 1 \leq i \leq 2^h$. Repeat the above path compression procedure recursively to each T_i. The resulting trie is called the LC trie. It is called an LC trie if T is a trie, and an LC PATRICIA trie if T is a PATRICIA tree. Note that in LC tries, the number of children of each node is a power of 2. For compact and simple array implementations, we refer to the work of Andersson and Nilsson. The idea of level compression was proposed by Andersson and Nilsson (1993). LC tries are first defined there, and an early average case analysis may be found in that paper and in Andersson and Nilsson (1994). LC tries were suggested by Andersson and Nilsson (1995) for string searching, as improvements of suffix trees. Nilsson and Karlsson (1998, 1999) noted their usefulness for fast address look-up for internet routers and IP address look-up. Experimental comparisons can be found in Iivonen, Nilsson and Tikkanen (1999) and Nilsson and Tikkanen (1998).

The purpose of this note is to analyze random LC tries under the uniform model of randomness: $X_1, \ldots, X_n$ are i.i.d. and uniformly distributed on $[0, 1]$, and consequently, the bits in the binary expansions of each X_i are i.i.d. Bernoulli $(1/2)$ random variables. Other models of randomness, such as the Bernoulli model (the bits in the binary expansion of X_1 are i.i.d., taking the value 1 with probability p) and the density model (X_1 has a density f on $[0, 1]$) will be treated elsewhere.

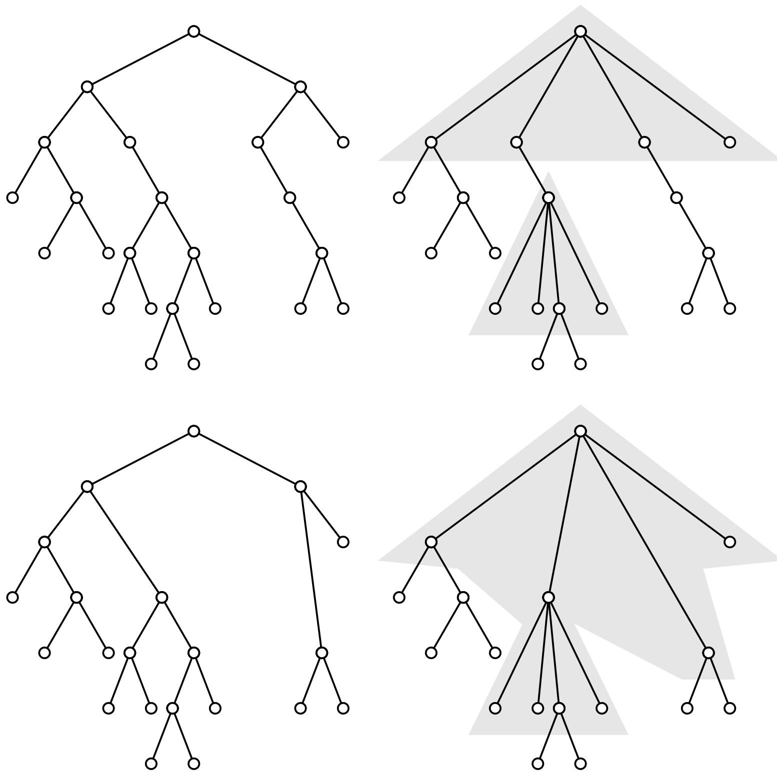

FIGURE 1. At the top, from left to right: a binary trie, and the corresponding LC trie. At the bottom, left to right: the PATRICIA trie, and the LC PATRICIA trie obtained by level compaction from the PATRICIA trie. The compacted parts in the LC trie are highlighted.

The quantities of interest to us in a trie or LC trie are D_n, the depth of the n-th string (which is thus also the depth of a typical string, as all tries considered here are permutation-invariant); A_n, the average leaf depth, $(1/n)\sum_{i=1}^{n} D_i$; and H_n, the height of the trie.

Andersson and Nilsson (1993) considered the LC PATRICIA tries and showed that for the uniform model and for the density model with bounded density f, $\mathbf{E}\{A_n\} = \Theta(\log^* n)$, where $\log^*(n)$ is the log star function, defined as the minimum positive integer i such that $\log_2 \log_2 \cdots \log_2 n \le 1$, where the logarithms are iterated i times. Also, in the Bernoulli model with $p \ne 1/2, p \in (0,1)$, $\mathbf{E}\{A_n\} = \Theta(\log\log n)$. These results confirm and strengthen the good experimental results obtained with the data structure, as the expected height and expected average leaf depth for random tries and PATRICIA tries are $\Theta(\log n)$ for these random models.

We will first improve the Andersson-Nilsson result to $\mathbf{E}\{A_n\} \sim \log^*(n)$, and extend it to include both LC tries and LC PATRICIA tries. We will also show that $D_n/\log^*(n) \to 1$ in probability, a law of large numbers. We conclude with a study of H_n, which has escaped scrutiny in the literature of LC tries. In particular, we note that the height and average depth in random LC tries are very different. For example, for LC PATRICIA tries, $H_n \sim \sqrt{2\log_2 n}$ in probability, which is much larger than the $\log^* n$ average leaf depth. All of the above is shown in this note for the uniform model. The density and Bernoulli models will be treated elsewhere.

THE MAIN PARAMETERS FOR RANDOM TRIES. The asymptotic behavior of tries under the uniform model is well-known. The height is studied by Régnier (1981), Mendelson (1982), Flajolet and Steyaert (1982), Flajolet (1983), Devroye (1984), Pittel (1985, 1986), and Szpankowski (1988,1989). For the depth of a node, see e.g. Pittel (1986), Jacquet and Régnier (1986), Flajolet and Sedgewick (1986), Kirschenhofer and Prodinger (1986), and Szpankowski (1988). For example, it is known that

$$H_n / \log_2 n \to 2 \text{ in probability .}$$

The limit law of H_n was obtained in Devroye (1984), and laws of the iterated logarithm for the difference $H_n - 2\log_2 n$ can be found in Devroye (1990). For other models, we refer to Devroye (1982, 1984), Régnier (1988), Szpankowski (1988) and Pittel (1985).

THE MAIN PARAMETERS FOR PATRICIA TRIES. As the PATRICIA trie is simply obtained by removing from the trie all internal nodes with one child, it necessarily has n leaves and $n-1$ internal nodes. The trie from which it is deduced is called the associated trie. All parameters of the PATRICIA trie such as H_n improve over those for the associated trie: Pittel (1985) has shown that for the uniform model, $H_n / \log_2 n \to 1$ in probability, which constitutes a 50% improvement over the trie. For other properties, see Knuth (1973), Flajolet and Sedgewick (1986), Kirschenhofer and Prodinger (1986), Szpankowski (1988), and Kirschenhofer, Prodinger and Szpankowski (1989). Pittel and Rubin (1990) and Pittel (1991) showed that for the uniform model,

$$\frac{H_n - \log_2 n}{\sqrt{2\log_2 n}} \to 1 \text{ in probability.}$$

Aldous and Shields (1988) showed that the same property holds true for the digital search tree, another modification of the trie with properties typically similar to those of PATRICIA trees. The following results will be required further on.

PROPOSITION 1 (DEVROYE, 1992A). *For a* PATRICIA *trie under the uniform model,*

$$\frac{H_n - \log_2 n}{\sqrt{2\log_2 n}} \to 1$$

in probability, and

$$\frac{F_n - \log_2 n}{\log_2 \log n} \to -1$$

in probability, where F_n (the fill-up level) is the number of consecutive levels in the trie that are full.

VARIABLE FANOUT TRIES. Variable fanout tries are tries in which on level i, a fanout of size f_i is used. For example, if at level 0, a fanout of size $f_0 = n$ is used, and similarly for all subtrees (fanout = subtree size), then we obtain the so-called N-trees (Ehrlich, 1982; Tamminen, 1983). For $f_i \equiv 2$, we have classical tries. It is easy to see that with f_i depending upon i and n in intricate ways, any structure interpolating between hash structures and tries is obtainable. There are many disadvantages of such structures. First of all, n must be known beforehand, so this requires some (modest) work to be made adaptive. Also, the space requirements could become important, whereas LC PATRICIA tries have carefully controlled space usage. With $f_1 \approx n/\log_2 n$, $f_1 \approx \log_2 n / \log_2 \log_2 n$, $f_2 \approx \log_2 \log_2 n / \log_2 \log_2 \log_2 n$ and so forth, with n referring to the total input size; or if $f_i \approx N/\log_2 N$, where N is the size of the subtree (varying from point to point), we should expect to see properties similar to those of the LC tries studied here. We will report on variable fanout tries elsewhere.

LC tries: depth for the uniform model

We introduce the notation

$$\mathcal{L}_i(n) = \log_2 \cdots \log_2(n) \ ,$$

where the logarithms are repeated i times. For positive n, the smallest i such that $\mathcal{L}_i(n) \le 1$ is denoted by $\log^*(n)$, the log star function. Thus, for any $k \ge 1$,

$$\mathcal{L}_{\log^*(n)}(n) \le 1 \le 2^{k-1} \le \mathcal{L}_{\log^*(n)-k}(n) \ .$$

In this section, we prove the following result.

THEOREM 1. *For a random LC trie under the uniform model,*

$$\lim_{n\to\infty} \frac{\mathrm{E}\{A_n\}}{\log^*(n)} = \lim_{n\to\infty} \frac{\mathrm{E}\{D_n\}}{\log^*(n)} = 1 \ .$$

Before proceeding with the proof, we will give an intuitive explanation of the result. It is well known (Devroye, 1992a) that the fill-up level (number of consecutive full levels starting at the root) in a random PATRICIA trie and a random trie is about $\log_2 n - \log_2 \log_2 n + O(1)$ in probability (Proposition 1 above, Lemma 1 below). This means that, barring a multiplicative factor, the root of the LC trie has about $n/\log_2 n$ children, each of which is the root of a new LC trie of size about $\log_2 n$ (because in the uniform model, subtrees are again distributed as for the uniform model). The nodes in the LC trie at distance one have fanout about $\log_2 n/\log_2 \log_2 n$, and the subtrees at distance two in the LC trie have sizes about $\log_2 \log_2 n$. If we repeat this argument until we run out of nodes, the number of levels in the LC trie is about $\log^*(n)$. The proof of Theorem 1 makes this argument more rigorous and attempts to identify, in fact, the positions of the various levels of the LC trie in terms of depths of nodes in the original trie.

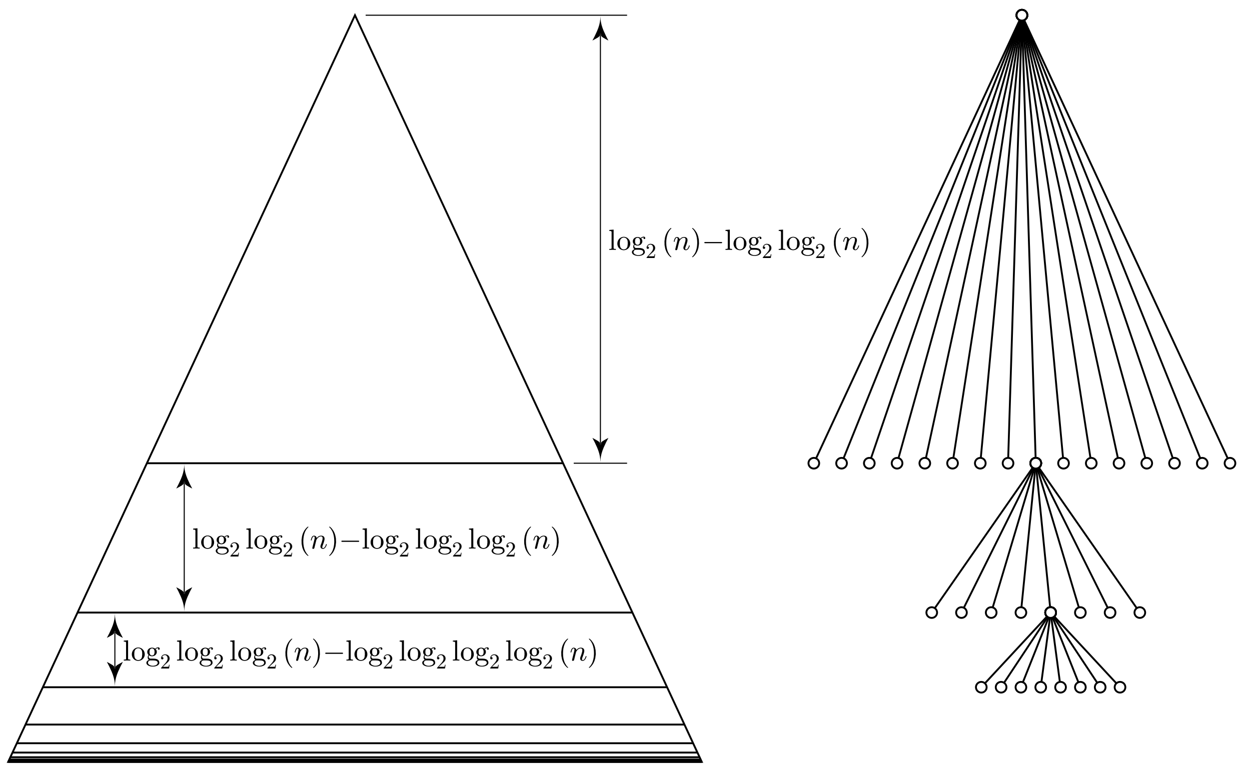

FIGURE 2. The left figure shows the top $\log_2 n$ levels of a random trie. Roughly speaking, such tries get compacted to LC tries (on the right) in the manner shown, with only minor random variations. The trie on the left has about $\log^*(n)$ horizontal sections of nodes, each section corresponding to one level in the random LC trie. Therefore, most nodes in the LC trie are at distance about $\log^*(n)$ from the root. However, the random fluctuations are such that the height of the LC trie is actually much larger.

PROOF OF THEOREM 1. We begin by simplifying matters greatly. Let x be a fixed infinite binary string, and let $D_n(x)$ denote the depth of the leaf for x in the LC trie for $x, X_1, \ldots, X_n$. Then, under the uniform model, we have $\mathbb{E}\{D_n(x)\}$ is independent of x. Thus, in particular, $\mathbb{E}\{D_n\} = \mathbb{E}\{D_{n-1}(x)\}$ with x arbitrary and fixed (like the string of all zeroes). We also have $\mathbb{E}\{A_n\} = \mathbb{E}\{D_n\}$. For these reasons, we study $\mathbb{E}\{D_n(x)\}$ for the zero-string x. We will write D_n instead of $D_n(x)$.

In what follows, $\epsilon > 0$ is always an arbitrary small positive number, and $M = M(\epsilon)$ is a finite positive integer depending upon ϵ. Consider a path for x in a random trie for $x, X_1, \ldots, X_n$. On this path, we define complete trees $T_1, T_2, \ldots$ with roots $u_1, u_2, \ldots$, where u_i is a leaf of T_{i-1} (see figure below). Also, the depth of u_{i+1} is ℓ_i:

$$\ell_i = \lfloor \mathcal{L}_1(n) \rfloor - \lfloor \mathcal{L}_{i+1}(n) \rfloor - M$$

for all $i \geq 0$. Since these numbers must be positive, we assume that n is so large that $\ell_1 > 0$. Further on, we will use the inequalities

$$\mathcal{L}_1(n) - \mathcal{L}_{i+1}(n) - M - 1 \leq \ell_i \leq \mathcal{L}_1(n) - \mathcal{L}_{i+1}(n) - M + 1 \ .$$

The height of tree T_i is given by

$$h_i = \ell_i - \ell_{i-1} = \lfloor \mathcal{L}_i(n) \rfloor - \lfloor \mathcal{L}_{i+1}(n) \rfloor \ .$$

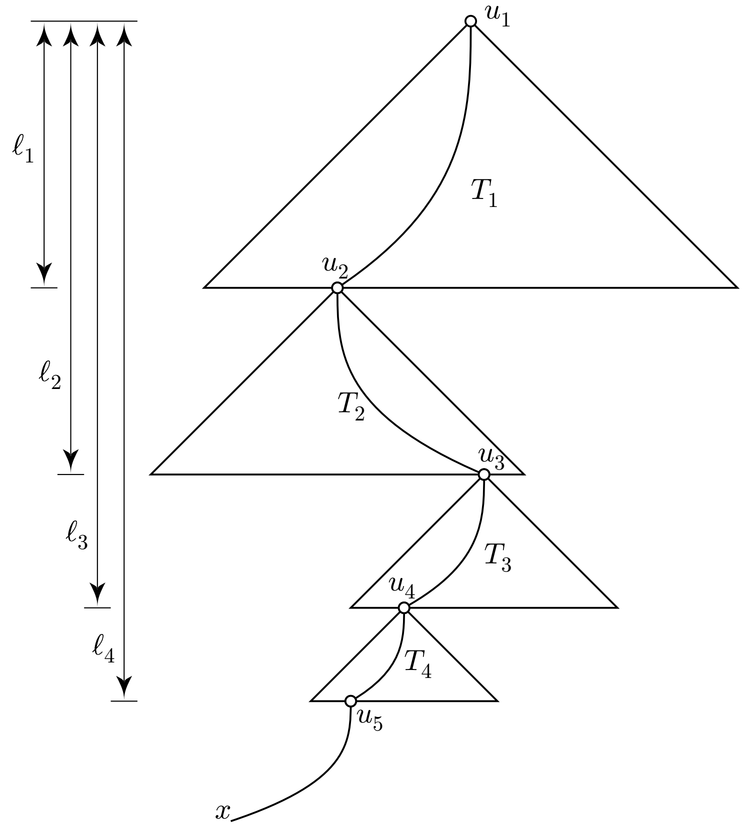

FIGURE 3. This figure illustrates the various quantities defined above.

We first bound the height of the LC trie from above. We say that T_i is busy if each of its 2^{h_i} leaf nodes is visited by at least two strings in the data (recalling that strings in the data correspond to infinite paths). We define B_n, the depth of the leaf of x in the original trie beyond $\lfloor \log_2 n \rfloor - 2M$ ($B_n = 0$ if this depth does not extend past that point), and C_n, the depth in the LC trie of y, the unique node on the path of x that is at distance $\lfloor \log_2 n \rfloor - 2M$ from the root in the original trie. The depth of x in the LC trie is bounded from above by $B_n + C_n$. Recall that all nodes on the path x get mapped to nodes in the LC trie by the following mapping: mark on x all nodes that correspond to leaves of complete subtrees as they are identified for LC trie compaction, and call them $v_1, v_2, \ldots$. Let v_0 be the root. A node v between v_i (inclusive) and the next deeper node v_{i+1} (exclusive) is mapped to v_i. Its depth in the LC trie is i. If T_i is busy then it contributes at most one to the depth of y in the LC trie. If T_i is not busy, then it contributes at most h_i. In calculating a bound for the depth of y in the LC trie, we need to consider all T_i's with $\ell_i \leq \lfloor \log_2 n \rfloor - 2M$. In other words, we consider all i with

$$\ell_i = \lfloor \mathcal{L}_1(n) \rfloor - \lfloor \mathcal{L}_{i+1}(n) \rfloor - M \leq \lfloor \mathcal{L}_1(n) \rfloor - 2M \ ,$$

or

$$\lfloor \mathcal{L}_{i+1}(n) \rfloor \geq M \ .$$

Call this collection of positive integers i, $\mathcal{I}$. Note that $|\mathcal{I}| \le \log^* n$, as $M \ge 1$. Thus,

$$C_n \le \sum_{i \in \mathcal{I}} 1_{[T_i \text{ busy}]} + \sum_{i \in \mathcal{I}} h_i 1_{[T_i \text{ not busy}]} \ .$$

Therefore,

$$\mathrm{E}\{C_n\} \le |\mathcal{I}| + \sum_{i \in \mathcal{I}} h_i \mathrm{P}\{T_i \text{ not busy}\} \ .$$

Now,

$$\begin{aligned}
\mathrm{P}\{T_i \text{ not busy}\} &\le 2^{h_i}\, \mathrm{P}\{\text{there is at most one string visiting the leftmost leaf of } T_i\} \\
&\le 2^{h_i} \left(1 - 1/2^{\ell_i}\right)^n + 2^{h_i} \left(1 - 1/2^{\ell_i}\right)^{n-1} n/2^{\ell_i} \\
&\le 2^{h_i} \left(1 - 1/2^{\ell_i}\right)^{n-1} (1 + n/2^{\ell_i}) \\
&\le 2^{h_i}\, e^{1-n/2^{\ell_i}} (1 + n/2^{\ell_i}) \ .
\end{aligned}$$

Note that $2^{h_i} \le 2^{\mathcal{L}_i(n)+1}/2^{\mathcal{L}_{i+1}(n)} = 2\mathcal{L}_{i-1}(n)/\mathcal{L}_i(n)$. By using the inequalities on ℓ_i mentioned earlier, we have

$$2^{M-1}\mathcal{L}_i(n) \le \frac{n}{2^{\ell_i}} \le 2^{1+M}\mathcal{L}_i(n) \ .$$

Thus, if $M \ge 3$,

$$\begin{aligned}
h_i \mathrm{P}\{T_i \text{ not busy}\} &\le h_i \frac{2e\mathcal{L}_{i-1}(n)}{\mathcal{L}_i(n)} \times e^{-2^{M-1}\mathcal{L}_i(n)} (1 + 2^{1+M}\mathcal{L}_i(n)) \\
&\le 2e(\mathcal{L}_{i-1}(n))^{1-2^M/\log 4} (1 + 2^{1+M}\mathcal{L}_i(n)) \\
&\le \frac{2e(1 + 2^{1+M}\mathcal{L}_i(n))}{(\mathcal{L}_{i-1}(n))^4} \\
&\le \frac{2^{4+M}}{(\mathcal{L}_{i-1}(n))^3} \\
&\le \frac{2^{4+M}}{2^{3\,2^M}} \ .
\end{aligned}$$

We conclude that

$$\mathrm{E}\{C_n\} \le |\mathcal{I}| \times \left(1 + \frac{2^{4+M}}{2^{3\,2^M}}\right) \le \log^*(n) \times \left(1 + \frac{2^{4+M}}{2^{3\,2^M}}\right) \ .$$

The last factor is less than $1 + \epsilon$ for any given $\epsilon > 0$ by choice of M. We now turn to B_n. The leaf for x is at distance $> k$ from the root of the trie if and only if there exists at least one other string in the input whose first k bits coincide with those of x. By the union bound, the probability of this event is not more than $n/2^k$. Thus, for $k \ge 0$,

$$\mathrm{P}\{B_n > k\} \le \frac{n}{2^{\lfloor \log_2 n \rfloor - M + k}} \le \frac{n}{2^{\log_2 n - M - 1 + k}} = \frac{2^{M+1}}{2^k}$$

Thus, as $\mathrm{E}B_n = \sum_{k=0}^{\infty} \mathrm{P}\{B_n > k\}$, we see that

$$\mathrm{E}B_n \le M + 1 + \sum_{k=0}^{\infty} \frac{1}{2^k} = M + 2 \ .$$

Putting everything together, we conclude that the expected depth of the leaf of x in the LC trie is not larger than

$$(1+\epsilon)\log^*(n) + M + 2 \ .$$

This concludes the proof of the upper bound.

We now turn to the lower bound. This time, we define

$$\ell_i = \lfloor \mathcal{L}_1(n) \rfloor - \lfloor \mathcal{L}_{i+1}(n) \rfloor + M \ .$$

Note the $+M$ in this definition, as opposed to $-M$. The h_i's are not affected by this change. We mark T_i if $h_i > 0$, if not all its $2^{h_i}-1$ leaves, the leaf on the path of x excepted, are visited by an input string, and if at least one of those leaves is visited. Observe that the depth of the leaf of x in the LC trie is bounded from below by the number of marked T_i's. Let the set of i with $h_i > 0$ be called $\mathcal{I}$. Note that $|\mathcal{I}| = \log^*(n) - 1$. Thus,

$$\mathrm{E}\{D_n\} \ge \sum_{i\in\mathcal{I}} \mathrm{P}\{T_i \text{ is marked}\} = |\mathcal{I}| - \sum_{i\in\mathcal{I}} \mathrm{P}\{T_i \text{ is not marked}\}$$

Let N_j, $1 \le j \le 2^{h_i}-1$, be the number of data strings that visit the j-th leaf of T_i (thus, for a given leaf, all these N_j input strings have identical prefixes of length ℓ_i). We have $[T_i \text{ is marked}] = [\max_j N_j \ge 1, \min_j N_j = 0]$. Thus,

$$\mathrm{P}\{T_i \text{ is not marked}\} = \mathrm{P}\{\max_j N_j = 0\} + \mathrm{P}\{\min_j N_j \ge 1\} \ .$$

By the union bound, since $h_i \ge 1$,

$$\begin{aligned}
\mathrm{P}\{\max_j N_j = 0\} &= \mathrm{P}\{\mathrm{binomial}(n, (2^{h_i}-1)/2^{h_i}2^{\ell_i}) = 0\} \\
&\le \mathrm{P}\{\mathrm{binomial}(n, 2^{-\ell_i-1}) = 0\} \\
&= (1-2^{-\ell_i-1})^n \\
&\le \exp\left(-n/2^{\ell_i+1}\right) \\
&\le \exp\left(-2^{-M-2}\mathcal{L}_i(n)\right) \\
&\le 2^{-2^{-M-2}\mathcal{L}_i(n)} \\
&= (\mathcal{L}_{i-1}(n))^{-2^{-M-2}} \ .
\end{aligned}$$

Next, by Mallows' inequality for the multinomial distribution (Mallows, 1968; see also Esary, Proschan and Walkup, 1967 and Joag-Dev and Proschan, 1983),

$$\begin{aligned}
\mathrm{P}\{\min_j N_j \ge 1\} &\le \prod_j \mathrm{P}\{N_j \ge 1\} \\
&= (1-\mathrm{P}\{N_1 = 0\})^{2^{h_i}-1} \\
&\le \exp\left(-(2^{h_i}-1)\mathrm{P}\{N_1 = 0\}\right) \ . \\
&\le \exp\left(1-2^{h_i}\mathrm{P}\{N_1 = 0\}\right) \ .
\end{aligned}$$

Observe that

$$2^{h_i} \ge 2^{\mathcal{L}_i(n)-1}/2^{\mathcal{L}_{i+1}(n)} = \mathcal{L}_{i-1}(n)/2\mathcal{L}_i(n) \ ,$$

and that

$$
\begin{aligned}
\mathrm{P}\{N_1 = 0\} &\geq \mathrm{P}\{\mathrm{binomial}(n, 1/2^{\ell_i}) = 0\} \\
&= \left(1 - 1/2^{\ell_i}\right)^n \\
&\geq \exp\left(-n/(2^{\ell_i} - 1)\right) \\
&\qquad (\text{use } 1 - u \geq \exp(-u/(1-u)), u \in [0,1)) \\
&= \exp\left(-n/2^{\ell_i}\right) \exp\left(-n/2^{\ell_i}(2^{\ell_i} - 1)\right) \\
&\geq \exp\left(-n/2^{\ell_i}\right) \exp\left(-2n/2^{2\ell_i}\right) \\
&\geq \exp\left(-2^{1-M}\mathcal{L}_i(n)\right) \exp\left(-2(2^{1-M}\mathcal{L}_i(n))^2/n\right) \\
&= \exp\left(-\mathcal{L}_i(n)/4\right) \exp\left(-(\mathcal{L}_i(n))^2/8n\right) \\
&\qquad (\text{if we take } M = 3) \\
&\geq 2^{-\mathcal{L}_i(n)/2} 2^{-(\mathcal{L}_i(n))^2/4n} \\
&\geq \frac{1}{\sqrt{\mathcal{L}_{i-1}(n)}} 2^{-(\log_2(n))^2/4n} .
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathrm{P}\{\min_j N_j \geq 1\} &\leq \exp\left(1 - \frac{2^{-(\log_2(n))^2/4n}\sqrt{\mathcal{L}_{i-1}(n)}}{2\mathcal{L}_i(n)}\right) \\
&\leq \exp\left(1 - \frac{\sqrt{\mathcal{L}_{i-1}(n)}}{3\mathcal{L}_i(n)}\right)
\end{aligned}
$$

for all n large enough. Putting things together, we have

$$
\mathrm{E}\{D_n\} \geq |\mathcal{I}| - \sum_{i \in \mathcal{I}} (\mathcal{L}_{i-1}(n))^{-1/32} - \sum_{i \in \mathcal{I}} e^{1 - \frac{\sqrt{\mathcal{L}_{i-1}(n)}}{3\mathcal{L}_i(n)}} .
$$

Split $\mathcal{I}$ into $\mathcal{I}_1$ (the collection of i's for which $\mathcal{L}_i \geq 1/\epsilon$, with $\epsilon > 0$ arbitrary and given), and the remainder, $\mathcal{I}_2$. Clearly, $|\mathcal{I}_2| \leq 1 + \log^*(1/\epsilon)$. Note that $\log(u) \leq (4/e)u^{1/4}$ for $u \geq 1$. Thus, on $\mathcal{I}$, we have $\mathcal{L}_i(n) \leq (4/e\log 2)(\mathcal{L}_{i-1}(n))^{1/4}$. Putting all this back into the bound, we have

$$
\begin{aligned}
\mathrm{E}\{D_n\} &\geq |\mathcal{I}| - (1+e)|\mathcal{I}_2| - \sum_{i \in \mathcal{I}_1} e^{-1/32\epsilon} - \sum_{i \in \mathcal{I}_1} e^{1 - (e\log 2/12)(\mathcal{L}_{i-1}(n))^{1/4}} \\
&\geq \log^*(n) - 1 - (1+e)(1 + \log^*(1/\epsilon)) - e^{-1/32\epsilon}\log^*(n) - e^{1-(e\log 2/12)2^{1/4\epsilon}}\log^*(n)
\end{aligned}
$$

and this is greater than $(1-\delta)\log^*(n)$ for all n large enough by choice of ϵ. This concludes the proof of the lower bound. □

REMARK: RANDOM LC PATRICIA TRIES. Clearly, the upper bounds for depths in tries are upper bounds for PATRICIA tries, so that $\limsup_{n\to\infty} \mathrm{E}\{A_n\}/\log^*(n) \leq 1$, and similarly for D_n. For matching lower bounds, some additional work is needed that is not presented here. The LC PATRICIA trie typically has smaller average depths for its leaves. However, the lower bound in the proof above in terms of the number of marked subtrees is also valid as a lower bound in case of LC PATRICIA tries. The definition of a marked T_i refers to the original trie, and not the PATRICIA trie, so new bounds are needed in that section of the proof. It is easy to see that $\mathrm{P}\{\max_j N_j = 0\}$ is smaller for the PATRICIA trie than for the original trie, so we need only bound $\mathrm{P}\{\min_j N_j \geq 1\}$.

A law of large numbers

THEOREM 2. *For a random LC trie under the uniform model,*

$$\frac{D_n}{\log^*(n)} \to 1$$

in probability. In fact, for $\delta > 0$, we show that

$$\lim_{n\to\infty} \mathrm{P}\{D_n \geq \log^*(n) + (1+\delta)\log_2\log_2\log^*(n)\} = 0 \ .$$

and

$$\lim_{n\to\infty} \mathrm{P}\{D_n \leq \log^*(n) - (1+\delta)\log^*(\log^*(n))\} = 0 \ .$$

PROOF. Note that D_n is distributed as D_n^*, the depth of the zero string x in the LC trie or LC PATRICIA trie formed by itself and the first $n-1$ data points. The technique used here is the following: we find random variables L_n and U_n (L for lower; U for upper) such that

$$L_n \leq D_{n+1}^* \leq U_n \ .$$

We show that $\liminf_{n\to\infty} \mathrm{P}\{L_n < \log^*(n) - (1+\delta)\log^*(\log^*(n))\} = 0$, and that $\limsup_{n\to\infty} \mathrm{P}\{U_n > \log^*(n) + (1+\delta)\log_2\log_2(\log^*(n))\} = 0$. These results together imply Theorem 2.

We inherit the notation of the proof of Theorem 1. Consider first L_n:

$$L_n = \sum_{i\in\mathcal{I}_1} 1_{[T_i \text{ is marked}]} \ .$$

The definition of $\mathcal{I}_1$ involves a parameter ϵ which we define here as $\epsilon = 1/32\log^*(n)$. One may verify in the proof of Theorem 1 that $|\mathcal{I}_2| \leq 1 + \log^*(32\log^*(n)) = o(\log^*(n))$. Clearly,

$$L_n \geq |\mathcal{I}_1| \geq \log^*(n) - 2 - \log^*(32\log^*(n))$$

if all T_i's with $i \in \mathcal{I}_1$ are marked. But as we showed in the proof of Theorem 1,

$$\mathrm{P}\{T_i \text{ is not marked}\} \leq e^{-1/32\epsilon} + e^{1-(e\log 2/12)2^{1/4\epsilon}} \ .$$

The upper bound reduces to

$$e^{-\log^*(n)} + e^{1-(e\log 2/12)2^{8\log^*(n)}} = o(1/\log^*(n))$$

uniformly over all $i \in \mathcal{I}_1$. Thus, by the union bound,

$$\mathrm{P}\{\cup_{i\in\mathcal{I}_1}[T_i \text{ is not marked}]\} \leq \sum_{i\in\mathcal{I}_1} \mathrm{P}\{T_i \text{ is not marked}\} = o(1) \ .$$

Therefore, for any $\delta > 0$,

$$\lim_{n\to\infty} \mathrm{P}\{L_n \leq \log^*(n) - (1+\delta)\log^*(\log^*(n))\} = 0.$$

This is more than we needed to show.

We now turn to the upper bound. Recall the definitions from the first part of the proof of Theorem 1, where we will make one change: M will be chosen as an appropriate positive integer-valued function of n. We set

$$U_n = B_n + |\mathcal{I}| + \sum_{i\in\mathcal{I}} h_i 1_{[T_i \text{ not busy}]} .$$

That this bounds D^*_{n+1} was established in the proof of Theorem 1. Recall $|\mathcal{I}| \le \log^* n$. Also, for positive integer k,

$$\mathbf{P}\{B_n \ge k\} \le 2^{M+1-k} .$$

Finally, we recall that for $M \ge 3$,

$$\mathbf{P}\{T_i \text{ not busy}\} \le \frac{2^{4+M}}{M 2^{M-2}} .$$

This suggests the following choice:

$$M = \lfloor \log_2(\log_2(\log^*(n))) \rfloor .$$

Thus, $U_n \ge \log^*(n) + k$ implies that either $B_n \ge k$ or that one of the T_i's, $i \in \mathcal{I}$, is not busy. By the union bound, we have

$$\begin{aligned}\mathbf{P}\{U_n \ge \log^*(n) + k\} &\le \mathbf{P}\{\cup_{i\in\mathcal{I}}[T_i \text{ not busy}]\} + \mathbf{P}\{B_n \ge k\} \\ &\le \frac{2^{4+M}\log^*(n)}{M 2^{M-2}} + 2^{M+1-k} \\ &= o(1) + 2^{M+1-k} .\end{aligned}$$

Thus, $\mathbf{P}\{U_n \ge \log^*(n) + M + \omega(1)\} \to 0$, where $\omega(1)$ represents any sequency diverging however slowly to ∞. This concludes the proof of Theorem 2. □

LC tries: the height for the uniform model

We show here that, remarkably, level compaction does not alter the height a lot—it improves over the ordinary random trie by about 50 percent. First, we establish a bound on the fill-up level of an ordinary trie:

LEMMA 1. *Let 2^K be the fanout of the root in an LC trie for the uniform model. Then*

$$\lim_{n\to\infty} \mathbf{P}\{K \le \log_2(n/\log n)\} = 0 .$$

PROOF. If all 2^k possible binary prefixes occur among $X_1, \ldots, X_n$, then clearly, $K \ge k$. Thus, by the union bound,

$$\mathbf{P}\{K < k\} \le 2^k(1 - 1/2^k)^n \le 2^k e^{-n/2^k} .$$

Now, take $k = \lfloor \log_2(n/\log n) \rfloor$ and conclude. □

THEOREM 3. *For a random LC trie under the uniform model,*

$$\frac{H_n}{\log_2 n} \to 1$$

in probability.

PROOF. Group the n data strings by their first k bits. The number of groups of strings is thus 2^k, one for each possible combination of k bits. Observe that $H_n \geq \ell$ if there is a group of cardinality two in which both strings in that group have their first $k + \ell$ bits coincide. We call such a group "good". The reason is that the LC trie will have an appendage of length at least ℓ rooted at a node at some distance to the root, this distance being an integer between 1 and k. Conditional on two strings in a group, the probability of a match of the last ℓ bits of the first $k + \ell$ bits is $1/2^\ell$. Let N be the number of groups of cardinality two, when $k = \lfloor \log_2 n \rfloor$. Define $\ell = \lfloor (1 - \epsilon) \log_2 n \rfloor$ for some $\epsilon \in (0, 1)$. Thus, by conditioning on N,

$$\mathbf{P}\{H_n < \ell\} \leq \mathbf{E}\left\{\left(1 - 1/2^\ell\right)^N\right\} \leq \mathbf{E}\left\{\exp\left(-N/2^\ell\right)\right\} \to 0$$

if $N/2^\ell \to \infty$ in probability, by the dominated convergence theorem. But $N/2^\ell \geq N/n^{1-\epsilon} = n^\epsilon (N/n)$. Clearly,

$$\mathbf{E}\{N\} = 2^k \binom{n}{2} (1/2^k)^2 (1 - 1/2^k)^{n-2} = C(n)n$$

where $0 < a \leq C(n) \leq b < \infty$ for some positive constants a, b. Also, if we replace one input string by another one, N changes by at most one. By McDiarmid's version of the bounded difference inequality (McDiarmid, 1989),

$$\mathbf{P}\{|N - \mathbf{E}\{N\}| \geq t\} \leq e^{-t^2/2n} \ .$$

Thus, $N/\mathbf{E}\{N\} \to 1$ in probability, and $N/2^\ell \to \infty$ in probability.

For the upper bound, we note that the height of the LC trie is at most $h_n - K$, where K is the fill-up level of the ordinary trie (see Lemma 1) and h_n is the height of the ordinary trie. Thus, for $\epsilon > 0$,

$$\mathbf{P}\{H_n \geq (1 + \epsilon) \log_2 n\} \leq \mathbf{P}\{h_n \geq (2 + \epsilon/2) \log_2 n\} + \mathbf{P}\{K \leq (1 - \epsilon/2) \log_2 n\} \ .$$

The first term on the right-hand side is $o(1)$ as pointed out earlier. The last term is $o(1)$ by Lemma 1. This concludes the proof of Theorem 4. □

LC PATRICIA tries: the height under the uniform model

THEOREM 4. *For a random LC PATRICIA trie, under the uniform model,*

$$\frac{H_n}{\sqrt{2 \log_2 n}} \to 1$$

in probability, i.e., for all $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbf{P}\left\{H_n \notin \left[\sqrt{(2 - \epsilon) \log_2 n}, \sqrt{(2 + \epsilon) \log_2 n}\right]\right\} = 0 \ .$$

PROOF. A lower bound for the height is obtained by the following simple argument. We pick $\ell = \lfloor\sqrt{(2-\epsilon)\log_2 n}\rfloor$, for arbitrary $\epsilon > 0$. Consider the original trie (before compacting it into a PATRICIA tree), and take the nodes at distance $k = \lfloor\log_2 n\rfloor$ from the root. We color such a node black if it has ℓ strings visiting it (thus, ℓ of the X_i's have their first k bits agreeing with those for the node), and if these ℓ strings have the following property: all agree in bit $k+1$ except one; of those agreeing, all agree in bit $k+2$ except one; and so forth. Given that there are ℓ strings, the probability of such a configuration is

$$\frac{1}{2^{\ell-1}} \times \frac{1}{2^{\ell-2}} \times \cdots \frac{1}{2^1} = \frac{1}{2^{\ell(\ell-1)/2}} \stackrel{\text{def}}{=} p\ .$$

For a black node, the trie and PATRICIA subtrees are identical as each internal node has degree two. If there is a black node, then the level compaction of the LC trie can at best make the root disappear and its two children become children of the root of the LC trie. Thus, the height of the LC trie is at least $\ell - 1$. We have

$$\mathrm{P}\{H_n < \ell - 1\} \leq \mathrm{P}\{\text{no node at distance } k \text{ is black}\}\ .$$

Let Y_j be the indicator that the j-th of the 2^k possible nodes at distance k from the root in the original trie exists and is black. Let $N = \sum_j Y_j$ be the number of black nodes. We will apply a version of the second moment method given, e.g., in Alon, Spencer and Erdős, 1992: $\lim_{n\to\infty} \mathrm{P}\{N = 0\} = 0$ if

A. $\lim_{n\to\infty} \mathrm{E}\{N\} = \infty$;

B. (Y_j, Y_m) are identically distributed for all $j \neq m$, and $\lim_{n\to\infty} \rho_n = 1$, where $\rho_n = \mathrm{E}\{Y_1Y_2\}/\mathrm{E}\{Y_1\}\mathrm{E}\{Y_2\}$.

But $\mathrm{P}\{H_n < \ell - 1\} \leq \mathrm{P}\{N = 0\} \to 0$ as $n \to \infty$, and we are done. To verify condition A, note that $\mathrm{E}\{N\} = 2^k p\mathrm{P}\{N_1 = \ell\}$, where N_1 is the number of strings starting with k zeroes, a binomial $(n, 1/2^k)$ random variable. Note that if $\ell = o(\sqrt{n})$, then

$$\mathrm{P}\{N_1 = \ell\} = \binom{n}{\ell}(1/2^k)^\ell(1-1/2^k)^{n-\ell} \sim \frac{n^\ell}{\ell!}(1/n)^\ell(1-1/n)^{n-\ell} \sim \frac{1}{e\ell!}$$

as $n \to \infty$. Then $p\mathrm{P}\{N_1 = \ell\} = \Omega(n^{\epsilon/2-1})$ and $\mathrm{E}\{N\} = np\mathrm{P}\{N_1 = \ell\} = \Omega(n^{\epsilon/2})$, which tends to infinity.

Let N_2 denote the number of data strings whose first k bits are $k-1$ zeroes followed by one 1 (we could have taken any k-string that is not all zero). To verify condition B, we see that as $n \to \infty$,

$$\begin{aligned}
\rho_n &= \mathrm{E}\{Y_1Y_2\} \\
&= p^2\mathrm{P}\{N_1 = N_2 = \ell\} \\
&= p^2\mathrm{P}\{N_1 = \ell\}\mathrm{P}\{N_2 = \ell | N_1 = \ell\} \\
&\sim (p\mathrm{P}\{N_1 = \ell\})^2 \\
&= \mathrm{E}\{Y_1\}\mathrm{E}\{Y_2\}
\end{aligned}$$

provided that $\mathrm{P}\{N_2 = \ell | N_1 = \ell\}/\mathrm{P}\{N_1 = \ell\} \to 1$ as $n \to \infty$. As $n \to \infty$, we have,

$$\begin{aligned}
&\frac{\mathrm{P}\{N_2 = \ell | N_1 = \ell\}}{\mathrm{P}\{N_1 = \ell\}} \\
&= \frac{\binom{n-\ell}{\ell}(1/(2^k-1))^\ell(1-(1/(2^k-1)))^{n-2\ell}}{\binom{n}{\ell}(1/2^k)^\ell(1-1/2^k)^{n-\ell}} \\
&= \frac{(n-\ell)!^2}{n!(n-2\ell)!} \times \frac{1}{(1-1/2^k)^\ell} \times (1-(1/(2^k-1)))^{-\ell} \times \left(\frac{(2^k-2)2^k}{(2^k-1)(2^k-1)}\right)^{n-\ell} \\
&\sim \frac{(n-\ell)!^2}{n!(n-2\ell)!} \sim \frac{(n-\ell)^{2n-2\ell}}{n^n(n-2\ell)^{n-2\ell}}
\end{aligned}$$

$$\sim \frac{(1-\ell/n)^{2n-2\ell}}{(1-2\ell/n)^{n-2\ell}} \sim \frac{(1-\ell/n)^{2n}}{(1-2\ell/n)^{n}} \sim \frac{e^{-2\ell}}{e^{-2\ell}} = 1 \ .$$

This concludes the proof of the lower bound for H_n.

For the upper bound, note that the height H_n of the LC PATRICIA trie is bounded from above by $P_n - F_n$, where F_n is the number of full levels of the PATRICIA trie, and P_n is the height of the PATRICIA tree. From proposition 4 of Devroye (1992a), we have for all $\epsilon > 0$,

$$\mathbf{P}\{P_n \ge \log_2 n + \sqrt{(2+\epsilon)\log_2 n}\} \le e^{9/2} n(n-1)^{-(2+\epsilon)/2} \to 0 \ .$$

From Proposition 1, for all $\epsilon > 0$,

$$\mathbf{P}\{F_n \le \log_2 n - (1+\epsilon)\log_2 \log n\} = o(1) \ .$$

Combining these results shows that

$$\mathbf{P}\{H_n \ge \sqrt{(2+\epsilon)\log_2 n} + (1+\epsilon)\log_2 \log n\} \to 0 \ .$$

This concludes the proof of Theorem 4. □

Conclusion

In future work, we intend to look at the density model, when the input strings are the bits in the binary expansions of random variables drawn independently from a given density f on $[0,1]$. The first term asymptotic behavior should be as for the model studied in this paper.

The asymmetric bit model (when the bits are independent but zeroes occur with probability $p > 1/2$) yields dramatically different behavior. The depth of a typical node is of the order of $\log\log n$. That behavior will be explored elsewhere.

Acknowledgment

We thank both referees for their feedback.

References

D. Aldous and P. Shields, "A diffusion limit for a class of randomly-growing binary trees," *Probability Theory and Related Fields*, vol. 79, pp. 509–542, 1988.

N. Alon, J. Spencer, and P. Erdős, *The Probabilistic Method*, John Wiley, New York, 1992.

A. Andersson and N. Nilsson, "Efficient implementation of suffix trees," *Software Practice and Experience*, vol. 25(2), pp. 129–141, 1995 .

A. Andersson and S. Nilsson, "Improved behaviour of tries by adaptive branching," *Information Processing Letters*, vol. 46, pp. 295–300, 1993.

A. Andersson and S. Nilsson, "Faster searching in tries and quadtrees – an analysis of level compression," in: *Proceedings of the Second Annual European Symposium on Algorithms*, pp. 82–93, 1994.

K. Azuma, “Weighted sums of certain dependent random variables,” *Tohoku Mathematical Journal*, vol. 37, pp. 357–367, 1967.

L. Devroye, “A note on the average depth of tries,” *Computing*, vol. 28, pp. 367–371, 1982.

L. Devroye, “A probabilistic analysis of the height of tries and of the complexity of triesort,” *Acta Informatica*, vol. 21, pp. 229–237, 1984.

L. Devroye, “A note on the probabilistic analysis of PATRICIA trees,” *Random Structures and Algorithms*, vol. 3, pp. 203–214, 1992a.

L. Devroye, “A study of trie-like structures under the density model,” *Annals of Applied Probability*, vol. 2, pp. 402–434, 1992b.

G. Ehrlich, “Searching and sorting real numbers,” *Journal of Algorithms*, vol. 2, pp. 1–12, 1982.

J. D. Esary, F. Proschan, and D. W. Walkup, “Association of random variables, with applications,” *Annals of Mathematical Statistics*, vol. 38, pp. 1466–1474, 1967.

P. Flajolet, “On the performance evaluation of extendible hashing and trie search,” *Acta Informatica*, vol. 20, pp. 345–369, 1983.

P. Flajolet and R. Sedgewick, “Digital search trees revisited,” *SIAM Journal on Computing*, vol. 15, pp. 748–767, 1986.

P. Flajolet and J. M. Steyaert, “A branching process arising in dynamic hashing, trie searching and polynomial factorization,” in: *Lecture Notes in Computer Science*, vol. 140, pp. 239–251, Springer-Verlag, New York, 1982.

E. H. Fredkin, “Trie memory,” *Communications of the ACM*, vol. 3, pp. 490–500, 1960.

W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

J.-P. Iivonen, S. Nilsson, and M. Tikkanen, “An experimental study of compression methods for functional tries,” in: *Workshop on Algorithmic Aspects of Advanced Programming Languages (WAAAPL’99)*, 1999.

P. Jacquet and M. Régnier, “Trie partitioning process: limiting distributions,” in: *Lecture Notes in Computer Science*, vol. 214, pp. 196–210, 1986.

P. Jacquet and W. Szpankowski, “Analysis of digital tries with Markovian dependency,” *IEEE Transactions on Information Theory*, vol. IT-37, pp. 1470–1475, 1991.

K. Joag-Dev and F. Proschan, “Negative association of random variables, with applications,” *Annals of Statistics*, vol. 11, pp. 286–295, 1983.

P. Kirschenhofer and H. Prodinger, “Some further results on digital trees,” in: *Lecture Notes in Computer Science*, vol. 226, pp. 177–185, Springer-Verlag, Berlin, 1986.

P. Kirschenhofer, H. Prodinger, and W. Szpankowski, "On the balance property of PATRICIA tries: external path length viewpoint," *Theoretical Computer Science*, vol. 68, pp. 1–17, 1989.

P. Kirschenhofer, H. Prodinger, and W. Szpankowski, "On the variance of the external path length in a symmetric digital trie," *Discrete Applied Mathematics*, vol. 25, pp. 129–143, 1989.

D. E. Knuth, *The Art of Computer Programming, Vol. 3 : Sorting and Searching*, Addison-Wesley, Reading, Mass., 1973.

V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov, *Random Allocations*, V.H. Winston and Sons, Washington, D.C., 1978.

C. L. Mallows, "An inequality involving multinomial probabilities," *Biometrika*, vol. 55, pp. 422–424, 1968.

C. McDiarmid, "On the method of bounded differences," in: *Surveys in Combinatorics*, (edited by J. Siemons), vol. 141, pp. 148–188, London Mathematical Society Lecture Note Series, Cambridge University Press, 1989.

H. Mendelson, "Analysis of extendible hashing," *IEEE Transactions on Software Engineering*, vol. 8, pp. 611–619, 1982.

D. R. Morrison, "PATRICIA — Practical Algorithm To Retrieve Information Coded in Alphanumeric," *Journal of the ACM*, vol. 15, pp. 514–534, 1968.

A. Nilsson and G. Karlsson, "Fast address look-up for Internet routers," in: *Proceedings IFIP 4th International Conference on Broadband Communications*, pp. 11–22, 1998.

S. Nilsson and G. Karlsson, "IP-address lookup using LC-tries," *IEEE Journal on Selected Areas in Communications*, vol. 17(6), pp. 1083–1092, 1999.

S. Nilsson and M. Tikkanen, "An experimental study of compression methods for dynamic tries," *Submitted to Algorithmica*, 1998.

B. Pittel, "Asymptotical growth of a class of random trees," *Annals of Probability*, vol. 13, pp. 414–427, 1985.

B. Pittel, "Path in a random digital tree: limiting distributions," *Advances in Applied Probability*, vol. 18, pp. 139–155, 1986.

B. Pittel, "On the height of PATRICIA search tree," ORSA/TIMS Special Interest Conference on Applied Probability in the Engineering, Information and Natural Sciences, Monterey, CA, 1991.

B. Pittel and H. Rubin, "How many random questions are necessary to identify n distinct objects?," *Journal of Combinatorial Theory*, vol. A55, pp. 292–312, 1990.

B. Rais, P. Jacquet, and W. Szpankowski, "A limiting distribution for the depth in Patricia tries," *SIAM Journal on Discrete Mathematics*, vol. 6, pp. 197–213, 1993.

M. Régnier, "On the average height of trees in digital searching and dynamic hashing," *Information Processing Letters*, vol. 13, pp. 64–66, 1981.

M. Régnier, “Trie hashing analysis,” in: *Proceedings of the Fourth International Conference on Data Engineering*, pp. 377–381, IEEE, Los Angeles, 1988.

W. Szpankowski, “Some results on V-ary asymmetric tries,” *Journal of Algorithms*, vol. 9, pp. 224–244, 1988.

W. Szpankowski, “Digital data structures and order statistics,” in: *Algorithms and Data Structures: Workshop WADS ’89 Ottawa*, vol. 382, pp. 206–217, Lecture Notes in Computer Science , Springer-Verlag, Berlin, 1989.

W. Szpankowski, “Patricia trees again revisited,” *Journal of the ACM*, vol. 37, pp. 691–711, 1990.

W. Szpankowski, “On the height of digital trees and related problems,” *Algorithmica*, vol. 6, pp. 256–277, 1991.

W. Szpankowski and C. Knessl, “Heights in generalized tries and PATRICIA tries,” in: *LATIN’2000*, pp. 298–307, Lecture Notes in Computer Science 1776, 2000.

M. Tamminen, “Analysis of N-trees,” *Information Processing Letters*, vol. 16, pp. 131–137, 1983.