Patricia Tries Again Revisited

WOJCIECH SZPANKOWSKI

Purdue University, West Lafayette, Indiana

Abstract. The Patricia trie is a simple modification of a regular trie. By eliminating unary branching nodes, the Patricia achieves better performance than regular tries. However, the question is: how much on the average is the Patricia better? This paper offers a thorough answer to this question by considering some statistics of the number of nodes examined in a *successful search* and an *unsuccessful search* in the Patricia tries. It is shown that for the Patricia containing n records the average of the successful search length S_n asymptotically becomes $1/h_1 \cdot \ln n + O(1)$, and the variance of S_n is either var $S_n = c \cdot \ln n + O(1)$ for an asymmetric Patricia or var $S_n = O(1)$ for a symmetric Patricia, where h_1 is the entropy of the alphabet over which the Patricia is built and c is an explicit constant. Higher moments of S_n are also assessed. The number of nodes examined in an unsuccessful search U_n is studied only for binary symmetric Patricia tries. We prove that the mth moment of the unsuccessful search length EU_n^m satisfies $\lim_{n\to\infty} EU_n^m/\log_2^m n = 1$, and the variance of U_n is var $U_n = 0.87907$. These results suggest that Patricia tries are very well balanced trees in the sense that a random shape of Patricia tries resembles the shape of complete trees that are ultimately balanced trees.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems—*computations on discrete structures, sorting and searching*; G.2.1 [**Discrete Mathematics**]: Combinatorics—*generating functions*; *recurrences and difference equations*; G.2.2 [**Discrete Mathematics**]: Graph Theory—*trees*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Balanced trees, data structures, digital search trees, Patricia tries, probabilistic analysis of algorithms, random shape of trees, successful search, unsuccessful search

1. *Introduction*

Algorithms designed from the worst-case perspective often have to cope efficiently with quite unrealistic, if not pathological, inputs. Sometimes there exist simpler algorithms that perform just as well, or even better, in practice. For example, in the extendable hashing algorithm [5] digital search trees are used to access keys (records). This algorithm is usually accompanied with another procedure to balance the tree in order to achieve good worst-case performance. The balancing procedure often restructures the entire tree, so it is rather an expensive operation. Therefore, the question arises whether or not we really need the balancing algorithm. In general, we ask to what extent simpler and more direct algorithms can be expected in practice to match the performance of more complicated, worst-case asymptotically better ones. These thoughts motivated our studies on the average complexity

This research was supported in part by the National Science Foundation under grants NCR 87-02115 and CCR 89-00305, in part by grant R01 LM05118 from the National Library of Medicine, and in part by AFOSR grant 90-0107.

Author's address: Department of Computer Science, Purdue University, West Lafayette, IN 47907.

of digital trees [22–24]. In this paper, we concentrate on Patricia tries, and ask how well on the average the Patricia is balanced. In other words, we inquire whether we really need to balance Patricia tries. In addition, we consider a question of how much on the average the Patricia is better than regular tries.

Digital searching is a well-known technique for storing and retrieving information using lexicographical (digital) structure of words. A *V-ary trie* or radix search trie is a digital search tree in which edges are labeled by elements from an alphabet of cardinality V, and leaves (external nodes) contain the records (keys) [2, 7, 8, 10, 12, 14, 19]. A key consists of a (possibly infinite) sequence of elements from the alphabet that is used to access a record. The access path from the root to a leaf is a minimal prefix of the information contained in the leaf. The radix trie has an annoying flaw: there is "one-way branching" that leads to the creation of extra nodes in the tree. D. R. Morrison (cf. [19]) discovered a way to avoid this problem in a structure which he named the *Patricia trie.* In such a tree, all nodes have branching degree greater than or equal to two. This is achieved by collapsing one-way branches on internal nodes, that is, by eliminating all unary nodes [8, 10, 14, 19]. The Patricia trie finds many applications [2, 10, 14, 19]. Among others we mention here lexicographical sorting [2, 16], dynamic hashing algorithms [5, 6], polynomial factorization, simulation, Huffman's algorithm, string matching [2, 14], and most recently conflict resolution algorithms for broadcast communications [9, 21].

In all searching algorithms, in particular those built over the Patricia tries, the main problem is to locate a record that contains given key. After the search is completed, two possibilities can occur: Either the search was *successful* and the record was located or it was *unsuccessful* and the record was not found. These two possibilities lead to two tree parameters that are of particular interest to us; namely, the *successful search length* and the *unsuccessful search length.* At this point, it is worth mentioning that the successful search length is equal to the depth of a leaf (external node) in the Patricia trie; that is, the number of internal nodes in the trie on the path from the root to a randomly selected external node that contains the chosen key. Some statistics of the depth are also used to assess the balancing property of the underlying tree [24]. An unsuccessful search also terminates at an external node, but the node does not contain the desired key. At last, we note that the unsuccessful search length is not simply related to the successful search length, since unsuccessful searches are more likely to occur at external nodes near the root.

The performance evaluation of the Patricia is very scarce (see [8], [13], [14] and [17]), and in fact restricted to the binary symmetric Patricia; that is, all letters from the alphabet occur with the same probability. In most analyses, only average values were studied, with the exception of [13], where Kirschenhofer and Prodinger computed the variance of the successful search length for the binary symmetric Patricia. These simplifications are dropped in this paper, and we shall analyze a general V-ary Patricia trie, that is, each element from the alphabet occurs independently and with different probability. Under these assumptions, we study all moments of the successful search length S_n, where n is the number of records stored in the tree. We prove that the average length ES_n of a successful search is equal to $1/h_1 \cdot \ln n + O(1)$, and the variance of S_n is either $c \cdot \ln n + O(1)$ for the asymmetric Patricia or $O(1)$ for the symmetric one, where h_1 is the entropy of the alphabet and c is an explicit constant. In addition, we show that the mth moment ES_n^m of S_n satisfies the following $\lim_{n\to\infty} ES_n^m/\ln^m n = 1/h_1^m$. These results extend the works of Knuth [14], Flajolet and Sedgewick [8], Jacquet and Regnier [12],

and Kirschenhofer and Prodinger [13]. They also suggest that the symmetric Patricia is very well balanced, since it achieves the asymptotically optimal value for the depth of a randomly selected leaf, and the variation of the depth is very small (see Remark 1(iii) for more details).

The results for the number of nodes inspected in an unsuccessful search U_n are even more scarce, and to the author's knowledge only the mean value of U_n was obtained by Knuth [14]. The problem is also much more intricate, and therefore only symmetric binary Patricia tries are analyzed. However, asymptotic analysis of all moments of the unsuccessful search length is discussed. It is proved that $EU_n = \log_2 n + O(1)$, the variance of U_n becomes $\operatorname{var} U_n = 0.8790$, while higher moments satisfy $\lim_{n\to\infty} EU_n^m/\log_2 n = 1$.

The paper is organized as follows: In the next section, we present our main findings concerning the lengths of a successful search and an unsuccessful search in a form of two propositions. In Section 3, we derive these results for the successful search length using a unifying approach through some general recurrences. In a similar way, we organize Section 4, where unsuccessful search for binary symmetric Patricia tries is investigated.

2. *Main Results*

Let us consider a family $\mathscr{T}_n$ of Patricia tries with n keys (records) built over an alphabet $\mathscr{A} = \{\omega_1, \ldots, \omega_V\}$. A key is a (possible infinite) string of elements from $\mathscr{A}$, such that the ith element $\omega_i \in \mathscr{A}$ occurs independently of other elements, and with probability p_i. The keys are stored in external nodes, while internal nodes determine branching strategy. The degree of each internal node is greater than or equal to two; that is, one-way branches are collapsed on internal nodes by including in the nodes the number of bits to skip over before making the next decision (for details, see [10], [13], [14], and [19]).

We study properties of successful and unsuccessful searches, as defined above, in a family of random Patricia tries $\mathscr{T}_n$. Let S_n and U_n (random variables) denote the successful search length and the unsuccessful search length in $\mathscr{T}_n$. The mth factorial moments of S_n and U_n are defined as

$$s_n^{\underline{m}} \stackrel{\text{def}}{=} E\{S_n(S_n - 1)(S_n - 2) \cdots (S_n - m + 1)\}, \tag{2.1}$$

$$u_n^{\underline{m}} \stackrel{\text{def}}{=} E\{U_n(U_n - 1)(U_n - 2) \cdots (U_n - m + 1)\}, \tag{2.2}$$

where the expectations in (2.1) and (2.2) are taken over all tries in $\mathscr{T}_n$ and over all external nodes in a given trie $t \in \mathscr{T}_n$. It is shown that these moments (as well as regular moments) are related to the mth derivatives of the generating function $H_n(z)$ with the coefficient at z^k being the expected number of external nodes at level k in our family of trees (cf. [14, 24]). There is no explicit formula for $H_n(z)$, but a rather sophisticated recurrence, as shown in Lemma 1 below. Let $\mathbf{j} = (j_1, j_2, \ldots, j_V)$ be a vector such that $j_1 + j_2 + \cdots + j_V = n$, and

$$\binom{n}{\mathbf{j}} \stackrel{\text{def}}{=} \frac{n!}{j_1! j_2! \cdots j_V!}$$

be a multinomial coefficient. By $\sum_{\{j_\Sigma = n\}} f(\mathbf{j})$ we denote a sum over all $\mathbf{j}$ such that $j_1 + j_2 + \cdots + j_V = n$ for a given function $f(\cdot)$. Then, the following recurrence on $H_n(z)$ may be established.

LEMMA 1. *For any natural n, the generating function $H_n(z)$ of the random family of Patricia tries $\mathcal{T}_n$ satisfies the recurrence*

$$H_0(z) = 0, \qquad H_1(z) = 1$$

$$H_n(z) = z \sum_{\{j_\Sigma = n\}} \binom{n}{\mathbf{j}} p_1^{j_1} \cdots p_V^{j_V} [H_{j_1}(z) + \cdots + H_{j_V}(z)]$$

$$- (z-1)[p_1^n + p_2^n + \cdots + p_V^n] H_n(z) \tag{2.3}$$

PROOF. Consider V subtrees of the root, each with $j_1, j_2, \ldots, j_V$ keys and $j_1 + j_2 + \cdots + j_V = n$. Then, for a given trie $\ell \in \mathcal{T}_n$ the generating function $H_n^\ell(z)$ for that particular trie satisfies

$$H_n^\ell(z) = [H_{j_1}^\ell(z) + \cdots + H_{j_V}^\ell(z)] \cdot [z + \delta_{j_1,n}(1-z) + \delta_{j_2,n}(1-z) + \cdots + \delta_{j_V,n}(1-z)],$$

where $\delta_{j,k}$ is the Kronecker delta. In the second expression enclosed in square brackets, the first z represents the subtrees that are one level below the root, and the other terms are responsible for avoiding one-way branches (cf. [14]). Taking the expectation of the above recurrence over all tries in $\mathcal{T}_n$, we finally obtain (2.3). □

Now we are in position to present our main results for the number of inspections made in a *successful search*, that is, the depth of a randomly selected leaf in the Patricia trie. We start with a lemma that shows a relationship between the mth derivatives of $H_n^{(m)}(z)$ of the generating function $H_n(z)$ and the mth factorial moment of $s_n^{\underline{m}}$ of the successful search length.

PROPERTY 1. *For integers n and m, the following relationship holds*

$$s_n^{\underline{m}} = \frac{H_n^{(m)}(1)}{n}, \tag{2.4}$$

where $H_n^{(m)}(1)$ is the mth derivative of $H_n(z)$ at $z = 1$.

PROOF. The same relationship holds for regular tries, and was proved in Szpankowski [24]. □

We note that a simple relationship holds also between the factorial moments and the regular moments. In particular, the variance var S_n of the successful search length can be computed from the first two factorial moments, as follows: var $S_n = s_n^{\underline{2}} + s_n^{\underline{1}} - (s_n^{\underline{1}})^2$. In Section 3, we prove the following proposition.

PROPOSITION 1

(*i*) *The average ES_n of the successful search length asymptotically becomes*

$$ES_n = \frac{1}{h_1}[ln\, n + \rho + F_1(n)] + O(n^{-1}), \tag{2.5}$$

where $ln(\cdot)$ denotes the natural logarithm,

$$h_k = (-1)^k \sum_{i=1}^{V} p_i ln^k p_i \quad \text{and} \quad \bar{h}_k = (-1)^k \sum_{i=1}^{V} p_i ln^k (1-p_i),$$

while $\rho \stackrel{def}{=} \gamma - \bar{h}_1 + h_2/(2h_1)$ and $F_1(n)$ is a fluctuating function with a small amplitude (see Section 3 for explicit definition of $F_1(n)$), $\gamma = 0.571\ldots$ is the Euler constant.

(*ii*) *The variance var* S_n *of* S_n *for large n satisfies*

$$var\ S_n = \frac{h_2 - h_1^2}{h_1^3} \ln n + \alpha - 2\beta + F_2(n) + O(n^{-1}), \tag{2.6}$$

where $F_2(n)$ *is a fluctuating function with a small amplitude* (*see Section* 3 *for details*), *and*

$$\alpha = \frac{1}{h_1^2}\left[\frac{\pi^2}{6} + \gamma^2 + \frac{3}{2}\frac{h_2^2}{h_1^3} + \frac{2\gamma h_2}{h_1} - \frac{2}{3}\frac{h_3}{h_1} + h_2 + \bar{h}_2 + 2h_1\bar{h}_1\right] - 2(h_1 + \bar{h}_1)\frac{\gamma h_1 + h_2}{h_1^3} + \frac{\rho}{h_1}\left(1 - \frac{\rho}{h_1}\right), \tag{2.7}$$

with

$$\beta = \frac{1}{h_1}\sum_{\lambda=1}^{V}\sum_{\nu=1}^{V} p_\nu p_\lambda \sum_{l=0}^{\infty}\sum_{\{i_\Sigma = l\}}\binom{l}{\mathbf{i}}\prod_{\mu=1}^{V} p_\mu^{i_\mu} \cdot \ln\left\{1 + p_\lambda(1 - p_\nu)\cdot(1 - p_\lambda)^{-1}\prod_{\mu=1}^{V} p_\mu^{i_\mu}\right\}. \tag{2.8}$$

In particular, for V-ary symmetric *Patricia tries* $h_2 = h_1^2$ *and* (2.6) *reduces to*

$$var\ S_n = \frac{\pi^2}{6\ln^2 V} + \frac{1}{12} - \frac{2}{\ln V}\ln\prod_{l=1}^{\infty}\left(1 + \frac{1}{V^l}\right) + F_2(n) + O(n^{-1}). \tag{2.9}$$

(*iii*) *The mth moment* ES_n^m *of the successful search length satisfies*

$$\lim_{n\to\infty}\frac{ES_n^m}{\ln^m n} = \frac{1}{h_1^m} \tag{2.10}$$

for all $m \geq 1$.

Remark 1

(i) *Comparison with Regular Tries.* We first compare the successful search length for regular tries and Patricia tries. Let $S_n^{[T]}$ and $S_n^{[P]}$ denote the lengths of successful search for the regular tries and the Patricia tries, respectively. Then, by Proposition 1 and the results from [23], we easily see that $ES_n^{[T]} - ES_n^{[P]} = \bar{h}_1/h_1 > 0$. For example, for binary symmetric case $ES_n^{[T]} - ES_n^{[P]} = 1$. On the other hand, in the symmetric case, the average height of the Patricia is $\log_V n + O(1)$ [17] whereas, for the regular tries, the height becomes $2\cdot\log_V n + O(1)$ [6], so it is twice as much. The variance of the depth, that is, the length of a successful search, for regular tries was derived by Szpankowski in [24] (for binary symmetric case, see also [13], and for binary asymmetric case, see also [12]) who proved that it satisfies our formula (2.6) with $\beta = 0$. In particular, for the symmetric case

$$\text{var}\ S_n^{[T]} - \text{var}\ S_n^{[P]} = \frac{1}{\ln V}\ln\prod_{l=1}^{\infty}\left(1 + \frac{1}{V^l}\right).$$

Table I compares these two variances. For small values of V the variance for the Patricia is substantially smaller than for the regular tries (for consequences of this, see Remark 1(iii)). Finally, these two data structures can be compared in terms of the number of internal nodes (i.e., the storage requirements). Naturally, the number of internal nodes in the Patricia is exactly equal to $n - V + 1$. In regular tries, the number of internal nodes may vary, and in general, it is a random variable. We can easily prove that the average number of internal nodes is asymptotically equal

TABLE I

V	var $S_n^{[T]}$	var $S_n^{[P]}$
2	3.507	1.000
3	1.446	0.630
4	0.939	0.500
5	0.718	0.430
6	0.596	0.387

to $n/h_1 + O(1)$. For example, a binary symmetric regular trie has on the average $1.41 \cdot n + O(1)$ internal nodes. But, every node in the Patricia contains a counter that indicates the number of digits to skip over before making the next test.

(ii) *The Successful Search Length S_n Converges in Probability to ES_n!* Applying our Proposition 1, we can show that $S_n/ES_n \to 1$ *in probability* as $n \to \infty$. Indeed, by Chebyshev's inequality,

$$\Pr\left\{\left|\frac{S_n}{ES_n} - 1\right| \geq \epsilon\right\} \leq \frac{\text{var } S_n}{\epsilon^2 (ES_n)^2} = O\left(\frac{1}{\ln n}\right) \to 0.$$

Note, however, that in the symmetric case the rate of convergence is better and equal to $O(1/\ln^2 n)$. Moreover, using the external path length approach and more sophisticated probabilistic tools (i.e., Borel–Cantelli lemma), we can prove that for the *symmetric* case the convergence *in probability* can be replaced by stronger convergence *with probability one* (see also [17]).

(iii) *How Well Is the Patricia Balanced?* A tree that is ultimately balanced is called a complete tree [2], and its depth is equal to $\log_V n$. Therefore, any tree with good balance property should have average depth equal to $\log_V n + O(1)$. For example, in a binary search tree, the depth is $1.41 \cdot \log_2 n + O(1)$ [2, 14], while for binary *digital* search trees (i.e., regular tries and Patricia tries) the successful search length, that is, the depth is $\log_2 n + O(1)$. Hence, *digital* search trees are better balanced than binary search trees. For the Patricia, even the height (maximum over all depths) is $\log_V n + O(1)$, so the shape of the Patricia resembles, on the average, a complete tree. In the asymmetric case, however, the situation is slightly different. The constant at $\ln n$ in formula (2.5) on the average successful search length is the reciprocal of the entropy h_1 of the alphabet, and the more asymmetric the alphabet is, the more skew the Patricia is. This can be even better characterized by considering the limiting distribution of the depth. Using the ideas of Jacquet and Regnier [12], one can prove that the limiting distribution of S_n is *normal* for the asymmetric case, but *not* for the symmetric one.[1] This proof and our discussion in Remark 1(ii) suggest that fluctuation of S_n around ES_n is bounded in probability in the symmetric case, and unbounded, with magnitude $\ln^{1/2} n$, in the asymmetric case. In conclusion, the Patricia is a well-balanced tree, and this is especially true for the symmetric alphabet. Therefore, in the asymmetric case, one may consider (efficiently) preprocessing the alphabet, before constructing the tree, in order to obtain a symmetric one. In this case, the Patricia *will not need* any additional construction to keep it balanced [2, 5]. □

Now, we turn our attention to the number of nodes examined in an *unsuccessful search*. The unsuccessful search length U_n is much harder to analyze, and, except for the average value of U_n, nothing is really known about the behavior of

[1] See Note Added in Proof.

unsuccessful search. Below, we present our results on U_n in the case of binary symmetric alphabet. To derive them, we shall use our unifying approach through the generating function $H_n(z)$ defined above in Lemma 1. We need the following property that couples the mth factorial moment of the unsuccessful search length and the mth derivatives of $H_n^{(m)}(\frac{1}{2})$ of the generating function $H_n(z)$ at $z = \frac{1}{2}$.

PROPERTY 2. *For any integer m, the following holds*

$$u_n^{\underline{m}} = 2^{-m} H_n^{(m)}\left(\frac{1}{2}\right), \tag{2.11}$$

where $H_n(\frac{1}{2}) = 1$.

PROOF. Let us set $V = 2$ and $p_1 = p_2 = 0.5$ in Lemma 1 formula (2.3). Then, putting $z = \frac{1}{2}$ in (2.3), one proves $H_n(\frac{1}{2}) = 1$. On the other hand, the average value of the unsuccessful search length, $u_n^{\underline{1}}$ is $\sum_{l=0}^{\infty} lH_l 2^{-l} = \frac{1}{2} H_n^{(1)}(\frac{1}{2})$, since we end up at a given external node on level l with probability 2^{-l} (by H_l we denote the number of external nodes at level l). For $m = 2$, we have

$$u_n^{\underline{2}} = \sum_{l=0}^{\infty} l(l-1)H_l 2^{-l} = \left(\frac{1}{2}\right)^2 H_n^{(2)}\left(\frac{1}{2}\right),$$

and so on. This proves Property 2. □

Now we can present our main results on the unsuccessful search length U_n. We prove them in Section 4.

PROPOSITION 2

(*i*) *The mean of the unsuccessful search length is*

$$EU_n = \lg n - \theta + G_1(n) + O(n^{-1}), \tag{2.12}$$

where $\lg x = \log_2 x$ *and*

$$\theta = \frac{\ln \pi - \gamma}{\ln 2} - \frac{1}{2} = 0.31875, \tag{2.13}$$

and $G_1(n)$ *is a fluctuating function with a small amplitude (see Section 4 for details).*

(*ii*) *The variance* $\operatorname{var} U_n$ *of* U_n *satisfies*

$$\begin{aligned} \operatorname{var} U_n &= 4(\alpha - \beta - \theta - 2) - \theta - \theta^2 + G_2(n) + O(n^{-1}) \\ &\approx 0.87904 + G_2(n), \end{aligned} \tag{2.14}$$

where

$$\alpha = \frac{1}{2}\theta + \frac{23}{24} + \frac{1}{\ln^2 2}\left[\frac{\pi^2}{24} + \frac{\gamma^2}{4} - \frac{\gamma \ln 2\pi}{2} - \zeta_2\right], \tag{2.15a}$$

with ζ_2 *being half of the second derivatives at zero of the Riemann zeta function* $\zeta(z)$ [1]. *Ramanujan proved that (see* [3, *p.* 204])

$$\zeta_2 = \frac{1}{2}\zeta''(0) = \frac{\gamma^2}{4} + \frac{c_1}{2} - \frac{\pi^2}{48} - \frac{\ln^2(2\pi)}{4}, \tag{2.15b}$$

with $c_1 = -0.0728158$. *Finally*

$$\beta = \theta + \frac{1}{2} + \frac{2}{\ln 2}\sum_{k=2}^{\infty}\frac{\zeta(k)2^{-k}}{k}\left\{\sum_{\lambda=2}^{\infty}2^{-k(\lambda-1)}\left[\sum_{i=1}^{2^{\lambda-1}-1} i^k - \frac{2^{\lambda-1}}{k+1} + \frac{1}{2}\right] - \frac{1}{2(k+1)}\right\}$$
$$\approx 0.48738, \tag{2.15c}$$

and $G_2(n)$ *is a fluctuating function with a small amplitude.*

(*iii*) *The mth moment* EU_n^m *of* U_n *satisfies*

$$\lim_{n\to\infty}\frac{EU_n^m}{\lg^m n} = 1 \tag{2.16}$$

for all $m \geq 1$.

Remark 2

(i) *Comparison with the Successful Search.* In Propositions 1 and 2, we have found evidence that the unsuccessful search is more likely to occur near the root since $ES_n - EU_n = \rho + \theta = 0.6515$ and var S_n − var $U_n = 0.121$.

(ii) *Convergence in Probability.* As in Remark 1(ii), we can prove that U_n/EU_n converges *in probability* to one as $n \to \infty$. Indeed, again by Chebyshev's inequality and Proposition 2(i), (ii) we find

$$\Pr\left\{\left|\frac{U_n}{EU_n} - 1\right| \geq \epsilon\right\} \leq \frac{0.87904}{\epsilon^2 \lg^2 n} \to 0;$$

therefore, $U_n = EU_n(1 + o(1))$ *in probability.*

3. *Successful Search*

In this section, we prove Proposition 1. To simplify the analysis, we shall consider only a binary model, that is, for $V = 2$ we set $p_1 = p$ and $p_2 = q = 1 - p_1$. By Property 1, the successful search length is related to the mth derivatives $H_n^{(m)}(1)$ of the generating function $H_n(z)$, and one can find a simple physical interpretation of $H_n^{(m)}(1)$. Indeed, let L_n denote an external path (i.e., sum of all paths from the root to all external nodes) in a family of tries $\mathscr{T}_n$, and $S_n(i)$ be a path from the root to the ith external node. For a given integer m, we define

$$L_n^{\underline{m}} = \sum_{i=1}^{n} S_n(i)[S_n(i) - 1][S_n(i) - 2]\cdots[S_n(i) - m + 1],$$

and let $l_n^{\underline{m}} = EL_n^{\underline{m}}$. We call $l_n^{\underline{m}}$ the mth semifactorial moment of the external path length. Then, it is easy to show [24] that also $l_n^{\underline{m}} = H_n^{(m)}(1)$. Therefore, by Property 1 $s_n^{\underline{m}} = L_n^{\underline{m}}/n$, and for simplicity we work on $l_n^{\underline{m}}$ instead of $s_n^{\underline{m}}$. Lemma 1 and the above lead to the following recurrences for the first two semifactorial moments $l_n^{\underline{1}}$ and $l_n^{\underline{2}}$

$$l_n^{\underline{1}} = n(1 - p^n - q^n) + \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}[l_k^{\underline{1}} + l_{n-k}^{\underline{1}}]. \tag{3.1}$$

$$l_n^{\underline{2}} = 2(1 - p^n - q^n)[l_n^{\underline{1}} - n] + \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}[l_k^{\underline{2}} + l_{n-k}^{\underline{2}}]. \tag{3.2}$$

Note that (3.1) and (3.2) is a system of recurrences, that is, to find l_n^2 we need l_n^1. Generalizing the above, we can prove the following lemma for general V-ary Patricia tries.

LEMMA 2. *For any integers m and n, the mth semifactorial moment l_n^m satisfies the following recurrence*

$$\begin{aligned} l_0^m &= l_1^m = 0 \\ l_n^m &= m!\left(1 - \sum_{i=1}^{V} p_i^n\right) \sum_{k=1}^{m} (-1)^{m-k} \frac{l_n^{k-1}}{(k-1)!} \\ &\quad + \sum_{\{j_\Sigma = n\}} \binom{n}{\mathbf{j}} p_1^{j_1} \cdots p_V^{j_V} [l_{j_1}^m + \cdots + l_{j_V}^m] \qquad n \geq 2 \end{aligned} \tag{3.3}$$

where by definition $l_n^0 = n$ for $n \geq 2$.

PROOF. The proof uses induction arguments applied to (2.3), and is left to the reader. □

As noted before, the recurrence established in Lemma 2 is a system of recurrences. To compute l_n^m we need $l_n^1, l_n^2, \ldots, l_n^{m-1}$ from the previous recurrences. But, the recurrences (3.3) just derived in Lemma 2 have a common pattern and they differ only by the first term, which we call the *additive term* and denote by a_n. This type of recurrence has been extensively analyzed by Szpankowski [21, 24] (see also [14]), and since we shall often use these recurrences we quote below some of the results of [21] and [24] but without proofs.

Let $x_0, x_1, \ldots, x_n$ be a sequence of numbers satisfying the following linear recurrence: $x_0 = x_1 = 0$, and for $n \geq 2$

$$x_n = a_n + \sum_{\{j_\Sigma = n\}} \binom{n}{\mathbf{j}} p_1^{j_1} \cdots p_V^{j_V} [x_{j_1} + \cdots + x_{j_V}], \tag{3.4}$$

where a_n is any sequence of numbers. To solve this recurrence we introduce the so-called *binomial inverse relations* [14, 18]. We define a new sequence $\hat{a}_n$ as

$$\hat{a}_n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} a_k \qquad a_n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \hat{a}_k. \tag{3.5}$$

(The second equation justifies the name binomial inverse relations.) For more details, see Riordan [18]. In particular, we note that [14]

$$a_n = \binom{n}{r} c^n \rightarrow \hat{a}_n = \binom{n}{r} (-c)^r (1-c)^{n-r}, \tag{3.6}$$

where r is an integer, and c is a constant. In fact, in our entire analysis we shall deal only with additive terms of the above form. In general, however, the solution to the recurrence (3.4) is derived in Szpankowski [21], and we repeat it below.

THEOREM 1. *The recurrence* (3.4) *possesses the following solution*

$$x_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{\hat{a}_k + k a_1 - a_0}{1 - \sum_{i=1}^{V} p_i^k} \tag{3.7a}$$

for all $n \geq 2$. *In addition, the inverse* $\hat{x}_n$ *becomes*

$$\hat{x}_n = \frac{\hat{a}_n + n \cdot a_1 - a_0}{1 - \sum_{i=1}^{V} p_i^n} \tag{3.7b}$$

for all $n \geq 2$. □

Now, a direct and simple application of Theorem 1, together with the fact that for $a_n = n(1 - \sum_{i=1}^{V} p_i^n)$ the inverse becomes $\hat{a}_n = n \sum_{i=1}^{V} p_i^{n-1}$ for $n \geq 2$ (cf. (3.6)), lead to the following solution for the first semifactorial moment l_n^1 of the external path length

$$l_n^1 = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \binom{k}{1} \frac{\sum_{i=1}^{V} p_i (1 - p_i)^{k-1}}{1 - \sum_{i=1}^{V} p_i^k}. \tag{3.8}$$

The second moment l_n^2 for binary asymmetric case satisfies recurrence (3.2), which falls into our general recurrence (3.4) with $a_n = (1 - p^n - q^n)(l_n^1 - n)$. By (3.6), the inverse sequence $\hat{a}_n$ is

$$\hat{a}_n = 2\hat{l}_n^1 + 2(\delta_{n,1} - npq^{n-1} - npq^{n-1}) - 2I(p^n l_n^1) - 2I(q^n l_n^1),$$

where $I(p^n l_n^1)$ is the inverse relation to $p^n l_n^1$. To compute $\hat{l}_n^1$, we use Theorem 1 formula (3.7b), and then simple algebra reveals

$$\hat{l}_k^1 = k \frac{\sum_{i=1}^{V} p_i (1 - p_i)^{k-1}}{1 - \sum_{i=1}^{V} p_i^k} \qquad k \geq 2. \tag{3.9}$$

To estimate $I(p^n l_n^1)$, we need the following lemma

LEMMA 3. *Let* a_n *and* $\hat{a}_n$ *are given, and let* $b_n = p^n a_n$, *where* $0 \leq p < 1$. *Then*

$$\hat{b}_n = \sum_{j=0}^{n} \binom{n}{j} \hat{a}_j p^j (1 - p)^{n-j}.$$

PROOF. Using well-known relationships for binomial coefficients (see Riordan [24]) we find

$$\begin{aligned} b_n &= \sum_{k=0}^{n} (-1)^k \binom{n}{k} a_k p^k = \sum_{k=0}^{n} (-1)^k \binom{n}{k} p^k \sum_{j=0}^{k} (-1)^j \binom{k}{j} \hat{a}_j \\ &= \sum_{j=0}^{n} \binom{n}{j} \hat{a}_j p^n \sum_{k=0}^{n-j} (-1)^k \binom{n-j}{k} \left(\frac{1}{p}\right)^{n-k-j} \\ &= \sum_{j=0}^{n} \binom{n}{j} \hat{a}_j p^j (1 - p)^{n-j}, \end{aligned}$$

and this completes the proof. □

Then, $I(p^n l_n^1)$ follows from (3.9) and (3.6). This leads to our final results of this subsection.

THEOREM 2. *The second semifactorial moment* l_n^2 *becomes* $l_n^2 = 2 \cdot A_n - 2 \cdot B_n$, *where*

$$A_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \binom{k}{1} \frac{\{\sum_{i=1}^{V} p_i(1-p_i)^{k-1}\}\{\sum_{i=1}^{V} p_i^k\}}{(1-\sum_{i=1}^{V} p_i^k)^2} \tag{3.10a}$$

$$B_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{1}{1-\sum_{i=1}^{V} p_i^k} \cdot \sum_{j=2}^{k} \binom{k}{j} \binom{j}{1} \frac{\{\sum_{i=1}^{V} p_i^j(1-p_i)^{k-j}\}\{\sum_{i=1}^{V} p_i(1-p_i)^{j-1}\}}{1-\sum_{i=1}^{V} p_i^j} \tag{3.10b}$$

PROOF. The details of algebraic manipulations are left to the reader. □

The explicit formulas derived above for the first two moments of the successful search length suggest that we need only asymptotics of the following two alternating sums

$$T_{n,r}(c) \stackrel{\text{def}}{=} \sum_{k=2}^{n} (-1)^k \binom{n}{k} \binom{k}{r} \frac{c^k}{1-\sum_{i=1}^{V} p_i^k} \tag{3.11a}$$

and

$$T_n^{(2)}(c) \stackrel{\text{def}}{=} \sum_{k=2}^{n} (-1)^k \binom{n}{k} \binom{k}{1} \frac{c^k}{(1-\sum_{i=1}^{V} p_i^k)^2}, \tag{3.11b}$$

where r is an integer, and c is a real number. Let $h_k = (-1)^k \sum_{i=1}^{V} p_i \ln^k p_i$. Then, in [21] and [24] (see also [23]), we have proved the following asymptotics for $T_{n,r}(c)$ and $T_n^{(2)}(c)$.

THEOREM 3

(*i*) *For any* r, c *and large* n, *the following holds*

$$T_{n,r}(c) = \begin{cases} nc\left\{\dfrac{\ln(nc)+\gamma-\delta_{n,0}}{h_1} + \dfrac{h_2}{2h_1^2} + (-1)^r f_r(nc)\right\} + O(1) & r = 0, 1, \\ (-1)^r nc\left\{\dfrac{1}{r(r-1)h_1} + f_r(nc)\right\} + O(1) & r \geq 2, \end{cases} \tag{3.12}$$

where $\gamma = 0.571\ldots$ *is the Euler constant, and* $f_r(n)$ *is a fluctuating function with a small amplitude defined as*

$$f_r(n) = -\sum_{\{z_k^r \neq 0, r-1\}} \frac{\Gamma(z_k^r) n^{r-z_k^r}}{\sum_{i=1}^{V} p_i^{r+1-z_k^r} \ln p_i}, \tag{3.13}$$

and z_k^r, $(k = 0, \pm 1, \pm 2, \ldots,)$ *are roots of the following equation*

$$1 - \sum_{i=1}^{V} p_i^{r-z} = 0. \tag{3.14}$$

It is shown in [6], [8], [10], [14], *and* [21] *that the function* $f_r(n)$ *has a very small amplitude and may be safely ignored in practice.*

(*ii*) *For large n the alternating sum* $T_n^{(2)}(c)$ *becomes*

$$T_n^{(2)}(c) = \frac{nc}{2h_1^2} \cdot \{ln^2 nc + \epsilon ln\, nc + \delta + f^{(2)}(nc)\} + O(1),$$

where

$$\epsilon = \frac{\gamma h_1 + h_2}{h_1^3}, \qquad \delta = \frac{1}{h_1^2}\left[\frac{\pi^2}{12} + \frac{\gamma^2}{2} + \frac{3h_2^2}{4h_1^2} + \frac{\gamma h_2}{h_1} - \frac{h_3}{3h_1}\right],$$

$$f^{(2)}(n) = \sum_{\substack{k \neq -\infty \\ k \neq 0}}^{\infty} \frac{\Gamma(z_k) n^{-z_k}}{h_1^2(z_k)}$$

and $f^{(2)}(x)$ *is a fluctuating function with a small amplitude* [21].

Using the above theorem, it is easy to establish our Proposition 1(i). We note only that the fluctuating function $F_1(n)$ in the formula (2.5) on the average value of the successful search length becomes $F_1(n) = \sum_{i=1}^{V} p_i f_1[n(1 - p_i)]$, and this completes the proof of Proposition 1(i).

The asymptotics for the second moment and the variance of the successful search length are more intricate. According to Theorem 2, we must analyze A_n and B_n given in (3.10). Using Theorem 3, we easily prove the following lemma:

LEMMA 4. *The coefficient* A_n *given in* (3.10*a*) *for large n becomes*

$$A_n = n\left\{\frac{1}{2h_1^2} ln^2 n + \left[\epsilon - \frac{h_1 + \bar{h}_1}{h_1^2}\right] ln\, n + \eta + F_A(n)\right\} + O(1), \tag{3.15}$$

where

$$\eta = \frac{1}{2h_1^2}(h_2 + \bar{h}_2 + 2h_1\bar{h}_1) - \epsilon(h_1 + \bar{h}_1) + \delta,$$

$$F_A(n) = \sum_{i=1}^{V} \sum_{j=1}^{V} p_i p_j f^{(2)}[np_j(1 - p_i)],$$

where $\bar{h}_n$ *is defined in Proposition* 1(*i*).

PROOF. To prove (3.15), it is enough to note that the first term A_n of l_n^2 can be equivalently represented in terms of $T_n^{(2)}(c)$ defined in (3.11b) as

$$A_n = \sum_{i=1}^{n} \frac{p_i}{1 - p_i} \sum_{j=1}^{V} T_n^{(2)}[p_j(1 - p_i)].$$

Applying Theorem 3 one immediately proves the lemma. □

To evaluate the second-term B_n in the formula on l_n^2 we express B_n in terms of $T_{n,r}(c)$ defined in (3.11a) and studied in Theorem 3. Let

$$B_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{B_k'}{1 - \sum_{i=1}^{V} p_i^k},$$

where, after some simple algebra, one obtains

$$B_k' = k \sum_{\lambda=1}^{V} \sum_{\nu=1}^{V} p_\nu p_\lambda \sum_{l=0}^{\infty} \sum_{\{i_\Sigma = l\}} \binom{l}{\mathbf{i}} c_{\mathbf{i}}\{[c_{\mathbf{i}} p_\lambda(1 - p_\nu) + 1 - p_\lambda]^{k-1} - (1 - p_\lambda)^{k-1}\},$$

where $\mathbf{i} = (i_1, i_2, \ldots, i_V)$ such that $i_1 + i_2 + \cdots + i_V = l$ (l is an integer) and $c_{\mathbf{i}} = \prod_{\mu=1}^{V} p_\mu^{i_\mu}$. Putting everything together, we show that B_n can be represented as

$$B_k = \sum_{\lambda=1}^{V} \sum_{\nu=1}^{V} p_\nu p_\lambda \sum_{l=0}^{\infty} \sum_{\{i_\Sigma = l\}} \binom{l}{\mathbf{i}} c_{\mathbf{i}} \left\{ \frac{T_{k,1}[c_{\mathbf{i}} p_\lambda (1-p_\nu) + 1 - p_\lambda]}{c_{\mathbf{i}} p_\lambda (1-p_\nu) + 1 - p_\lambda} - \frac{T_{k,1}(1-p_\lambda)}{1-p_\lambda} \right\}.$$

Hence, by Theorem 3, we finally obtain

$$B_n = n\beta = n\left\{ \frac{1}{h_1} \sum_{\lambda=1}^{V} \sum_{\nu=1}^{V} p_\lambda p_\nu \sum_{l=0}^{\infty} \sum_{\{i_\Sigma = l\}} \binom{l}{\mathbf{i}} c_{\mathbf{i}} \ln\left[1 + \frac{p_\lambda (1-p_\nu) c_{\mathbf{i}}}{1-p_\lambda}\right] + F_B(n) \right\},$$

where

$$F_B(n) = \sum_{\lambda=1}^{V} \sum_{\nu=1}^{V} p_\nu p_\lambda \sum_{l=0}^{\infty} \sum_{\{i_\Sigma = l\}} \binom{l}{\mathbf{i}} c_{\mathbf{i}} \{ f_1[n(c_{\mathbf{i}} p_\lambda (1-p_\nu) + 1 - p_\lambda)] - f_1(n(1-p_\lambda)) \},$$

and the constant β is defined as B_n/n. Summarizing, by Property 1, Lemma 4 and the above we have just proved

THEOREM 4. *The second factorial moment s_n^2 of the successful search length S_n is given by*

$$\begin{aligned} s_n^2 &= \frac{2A_n}{n} - \frac{2B_n}{n} \\ &= \frac{1}{h_1^2} \ln^2 n + 2\left[\epsilon - \frac{h_1 + \bar{h}_1}{h_1^2} \right] \ln n + 2\eta - 2\beta + 2F_A(n) - 2F_B(n) + O(n^{-1}) \end{aligned}$$

where η, β, and $F_A(n)$, $F_B(n)$ are defined above.

To compute the variance of S_n, we note that $\operatorname{var} S_n = s_n^2 + ES_n - (ES_n)^2$; hence, the Proposition 1(ii) follows with $\alpha = 2\eta + \rho/h_1 \cdot (1 - \rho/h_1)$, and $F_2(n) = 2F_A(n) - 2F_B(n) + F_1(n) - [F_1(n)]^2$. Note also that for the symmetric Patricia trie $h_1 = \ln V$ and $h_k = h_1^k$, so the coefficient at $\ln n$ in the variance disappears. The higher moments of the successful search length are analyzed in a similar manner. Details are left for the reader, and they can also be found in [20].

4. *Unsuccessful Search*

In this section, we prove Proposition 2. The unsuccessful search is neither simply related to the external path length of Patricia tries nor to successful search, since unsuccessful search is more likely to occur at external nodes near the root. This makes the analysis much more difficult, and hence we consider only binary symmetric Patricia tries ($p_1 = p_2 = 0.5$). However, we derive asymptotic approximations for all moments of the unsuccessful search length. The organization of this section follows the pattern adopted in Section 3.

Property 2 and Lemma 1 proved in Section 2 imply, in particular, that the first two moments of the unsuccessful search length satisfy the following recurrences: $u_0^1 = u_1^1 = u_0^2 = u_1^2 = 0$, and for $n \geq 2$

$$u_n^1 (2^n - 2) = 2^n - 2 + \sum_{k=1}^{n-1} \binom{n}{k} u_k^1, \tag{4.1}$$

$$u_n^2 (2^n - 2) = 2(2^n - 2)(u_n^1 - 1) + \sum_{k=1}^{n-1} \binom{n}{k} u_k^2. \tag{4.2}$$

Generalizing the above, we find

LEMMA 5. *For any integers n and m, the mth factorial moment of U_n satisfies*

$$\begin{aligned} u_0^m 0 &= u_1^m = 0 \\ u_n^m(2^n - 2) &= m(2^m - 2)\left[u_n^{m-1} + \sum_{k=1}^{m} (-1)^k (m-k) u_n^{m-k}\right] + \sum_{k=1}^{n-1} \binom{n}{k} u_k^m \end{aligned} \tag{4.3}$$

and $u_1^0 \stackrel{def}{=} 1$.

PROOF. The proof uses induction arguments and is left to the reader. □

Note that (4.3) is a system of recurrences, as in the case of successful search; however, this new recurrence is much harder to deal with (cf. [14, 22]). We note, that to compute u_n^m, we need $u_n^1, u_n^2, \ldots, u_n^{m-1}$. But, recurrences of type (4.3) have been extensively studied by Szpankowski [22], and we summarized these results below.

Let $x_0, x_1, \ldots, x_n$ be a sequence of numbers such that

$$\begin{aligned} &\textbf{given} \quad x_0 = x_1 = 0 \quad \text{and} \quad x_2, \ldots, x_N \\ &\textbf{solve} \quad x_n(2^n - 2) = 2^n a_n + \sum_{k=1}^{n-1} \binom{n}{k} x_k \qquad x > N, \end{aligned} \tag{4.4}$$

where N is an integer, and a_n is a given, but otherwise arbitrary, sequence. It turns out that the solution of (4.4) depends on the so-called *Bernoulli inverse relations* (see Riordan [24]). Define for an a_n a new sequence $\tilde{a}_n$ as

$$\tilde{a}_n = \sum_{k=0}^{n} \binom{n}{k} B_k a_{n-k} \longrightarrow a_n = \sum_{k=0}^{n} \binom{n}{k} \frac{\tilde{a}_{n-k}}{k+1} = \frac{1}{n+1} \sum_{k=0}^{n} \binom{n+1}{k} \tilde{a}_k, \tag{4.5}$$

where B_k are the Bernoulli numbers defined as the coefficients of the Taylor expansion of $z(e^z - 1)^{-1}$ [1, 15, 18, 21] (see also (4.9)). The sequence $\tilde{a}_n$ and a_n are called inverse pair since $\tilde{\tilde{a}}_n = a_n$. Then,

THEOREM 5. *Our general recurrence* (4.4) *possesses the following solution*

$$x_n = b_n + \frac{1}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{\tilde{b}_k}{2^{k-1} - 1}, \tag{4.6}$$

where

$$g_k = x_k(1 - 2^{-k}) - a_k - 2^{-k} \sum_{i=1}^{k} \binom{k}{i} x_i \qquad k = 1, 2, \ldots, N,$$

$$b_n = a_n + g_n \chi_{(n \le N)},$$

$$\tilde{b}_k = \tilde{a}_k - a_0 B_k + \sum_{i=1}^{N} \binom{k}{i} g_i B_{k-i},$$

and $\chi_{(n \le N)}$ *is the indicator function. In addition, the inverse solution* $\tilde{x}_n$ *of* x_n *satisfies*

$$\tilde{x}_n = \tilde{b}_n + \frac{\tilde{b}_n}{2^{n-1} - 1} \tag{4.7}$$

for $n \ge 2$.

Application of Theorem 5 depends on the satisfactory computation of the Bernoulli inverse relations. In particular, we need the following, which is proved in [18] and [22]

$$a_n = \binom{n}{r} q^n \rightarrow \tilde{a}_n = \binom{n}{r} q^r B_{n-r}(q), \tag{4.8}$$

where r is an integer, and $0 < q < 1$, while $B_n(q)$ denotes the Bernoulli polynomial defined as [1, 15, 22]

$$\frac{ze^{tz}}{e^z - 1} = \sum_{k=0}^{\infty} B_k(t) \frac{z^k}{k!}. \tag{4.9}$$

Now we are ready to compute the first two moments of the unsuccessful search length U_n. In particular, recurrence (4.1) and Theorem 5 lead directly to

$$u_n^{\underline{1}} = 2 - \frac{4}{n+1} + 2\delta_{n0} + \frac{2}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k}{2^{k-1} - 1}, \tag{4.10}$$

$$\tilde{u}_n^{\underline{1}} = 4B_n + 2\delta_{n1} - 4\delta_{n0} + (1 - \delta_{n0} - \delta_{n1}) \frac{2B_n}{2^{n-1} - 1}, \tag{4.11}$$

where (4.11) is used in the evaluation of the second moment.

The second factorial moment $u_n^{\underline{2}}$ satisfies recurrence (4.4), and can be equivalently represented as

$$u_n^{\underline{2}} = U_n^{(1)} - U_n^{(2)},$$

where $U_n^{(1)}$ is defined by the following recurrence

$$U_0^{(1)} = U_1^{(1)} = 0$$

$$(2^n - 2)U_n^{(1)} = 2^n \cdot 2[u_n^{\underline{1}} - 1 + 2^{1-n}] + \sum_{k=1}^{n-1} \binom{n}{k} U_k^{(1)}$$

and $U_n^{(2)}$ satisfies similar recurrence with the additive term equal to $4u_n^{\underline{1}}$. By Theorem 5 (see in particular (4.11)), the solution to the above is

$$U_n^{(1)} = \frac{8}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k}{2^{k-1} - 1} + \frac{4}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k}{(2^{k-1} - 1)^2}. \tag{4.12}$$

The analysis of $U_n^{(2)}$ is much more intricate. Note that the recurrence for $U_n^{(2)}$ *does not* fall into our general recurrence (4.4) since the additive term is $4u_n^{\underline{1}}$ (not $2^n a_n$ as required). But $u_n^{\underline{1}} = 2^n(2^{-n}u_n^{\underline{1}})$, and for the solution, we need the inverse sequence to $2^{-n}u_n^{\underline{1}}$. But

LEMMA 6. *Let $A_n = q^n a_n$, and $\tilde{a}_n$ is given. Then*

$$\tilde{A}_n = \sum_{j=0}^{n} \binom{n}{j} \tilde{a}_j q^{j-1} \frac{B_{n+1-j}(q)\ B_{n+1-j}}{n+1-j}. \tag{4.13}$$

PROOF. In the proof, we use identities from [1], [24], and (4.5). We have

$$\begin{aligned}
\tilde{A}_n &= \sum_{k=0}^{n} \binom{n}{j} B_{n-k} q^k a_k \\
&= \sum_{k=0}^{n} \binom{n}{k} B_{n-k} q^k \sum_{j=0}^{k} \binom{k}{j} \frac{1}{k+1-j} \tilde{a}_j \\
&= \sum_{j=0}^{n} \binom{n}{j} \tilde{a}_j q^j \sum_{k=0}^{n-1} \binom{n-j}{k} \frac{1}{k+1} B_{n-j-k} q^k \\
&= \sum_{j=0}^{n} \binom{n}{j} \tilde{a}_j q^{j-1} \frac{1}{n-j+1} \sum_{k=1}^{n-j+1} \binom{n+1-j}{k} B_{n+1-j-k} q^k \\
&= \sum_{j=0}^{n} \binom{n}{j} \tilde{a}_j q^{j-1} \frac{B_{n+1-j}(q) - B_{n+1-j}}{n+1-j},
\end{aligned}$$

and this proves the lemma. □

Let now $A_n = 2^{-n} u_n^{\frac{1}{2}}$. Then, using the above one shows

$$\tilde{A}_n = 2\left[B_n\left(\frac{1}{2}\right) - B_n\right] - 8\frac{B_{n+1}(1/2) - B_{n+1}}{n+1} + 2V_n,$$

where

$$V_n = \sum_{j=2}^{n} \binom{n}{j} \frac{B_j}{2^{j-1}-1} 2^{1-j} \frac{B_{n+1-j}(1/2) - B_{n+1-j}}{n+1-j}. \tag{4.14}$$

Finally, applying Theorem 5 and the above, we obtain

THEOREM 6. *The solution to u_n^2 given by recurrence* (4.2) *is $u_n^2 = U_n^{(1)} - U_n^{(2)}$, where $U_n^{(1)}$ is evaluated in* (4.12) *and*

$$\begin{aligned}
U_n^{(2)} &= 4 \cdot 2^{-n} u_n^{\frac{1}{2}} + \frac{8}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k(1/2) + B_k}{2^{k-1}-1} \\
&\quad - \frac{32}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_{k+1}(1/2) - B_{k+1}}{(k+1)(2^{k-1}-1)} \\
&\quad + \frac{8}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{1}{2^{k-1}-1} \\
&\quad \cdot \sum_{j=2}^{k} \binom{k}{j} \frac{B_j}{2^{j-1}-2} 2^{1-j} \frac{B_{k+1-j}(1/2) - B_{k+1-j}}{k+1-j}
\end{aligned} \tag{4.15}$$

for $n \geq 2$.

In the next analysis, we need the asymptotics of the following intricate alternating sums

$$R_{n,r}(q) = \frac{1}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \binom{k}{r} \frac{B_{k-r}(q)}{2^{k-1}-1}, \tag{4.16}$$

$$R_n^{(2)} = \frac{1}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{B_k}{(2^{k-1}-1)^2}. \tag{4.17}$$

In [22] (see also [23]), we have proved the following asymptotics for $R_{n,r}(q)$ and $R_n^{(2)}$.

THEOREM 7

(*i*) *For large n one finds*

$$R_{n,0}(q) = \left(\frac{1}{2} + \delta_{q,1} - q\right)\left(\lg n - \frac{1}{2} + \frac{\gamma}{\ln 2}\right) + \frac{\zeta'(1 - q + \delta_{q,1})}{\ln 2} + g_0(n) + O(n^{-1}), \tag{4.16a}$$

$$R_{n,1}(q) = \lg n - \frac{1}{2} + \frac{\gamma}{\ln 2} - \frac{\psi(1 - q + \delta_{q,1})}{\ln 2} + g_1(n) + O(n^{-1}), \tag{4.16b}$$

$$R_{n,r}(q) = \frac{1}{r \ln 2}\,\zeta(r, 1 - q + \delta_{q,1}) + \frac{1}{r!}\,g_r(n) + O(n^{-1}), \qquad r \geq 2, \tag{4.16c}$$

where $\psi(x)$ *is the psi function,* $\zeta(z, q)$ *is the generalized Riemann zeta function* ($\zeta(z) = \zeta(z, 1)$) [1, 4, 11, 25], *and*

$$g_r(n) = \frac{1}{\ln 2} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \zeta\left(r + \frac{2\pi i k}{\ln 2}\right)\Gamma\left(r + \frac{2\pi i k}{\ln 2}\right) exp[-2\pi i k \lg n].$$

The function $g_r(n)$ *is a fluctuating function with a small amplitude and may be safely ignored in practice* [14, 22].

(*ii*) *For large n, we have*

$$R_n^{(2)} = \tfrac{1}{4}\lg^2 n - \tfrac{1}{2}[2.5 + \theta]\lg n + \delta + G_0(n) + g_0(n) + O(n^{-1}), \tag{4.17a}$$

where

$$\delta = \frac{1}{2}\theta + \frac{23}{24} + \frac{1}{\ln^2 2}\left[\frac{\pi^2}{24} + \frac{\gamma^2}{4} - \frac{\gamma \ln 2\pi}{2} - \zeta_2\right],$$

$$G_0(n) = \frac{1}{\ln^2 2} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} [\zeta(z_k)\Gamma'(z_k) - \Gamma(z_k)\zeta'(z_k) - \ln n \Gamma(z_k)] \cdot exp[-2\pi i k \lg n], \tag{4.17b}$$

and $\zeta_2 = \frac{1}{2}\zeta''(0)$, *where the second derivatives at zero of the zeta function,* $\zeta''(0)$, *was computed by Ramanujan* [3, *p.* 204], *and it is repeated in* (2.15*b*).

Using Theorem 7, we easily evaluate the asymptotics of the first moment, namely

$$U_n^1 = 2R_{n,0}(1) = \lg n - \theta + g_0(n) + O(n^{-1}),$$

as needed in Proposition 2(i). Now, we shall concentrate on the second moment of U_n. As before, we evaluate it into two steps. First, we deal with $U_n^{(1)}$, defined in (4.12). But, $U_n^{(1)} = 8R_{n,0}(1) + 4R_n^{(2)}$; hence,

$$R_{n,0}(1) = \tfrac{1}{2}\lg n - \tfrac{1}{2}\theta - 1 + g_0(n) + O(n^{-1}).$$

The asymptotics for $U_n^{(2)}$ is harder to compute, but using Theorem 7 we can easily find tight upper and lower bounds for it, namely $0.48701 \leq U_n^{(2)} \leq 0.48748$ (see Appendix). Nevertheless, it is interesting to see if exact asymptotics for $U_n^{(2)}$

are available. This interest is motivated not only by a "pure mathematical whim," but such a solution extends the analysis of our general recurrence (4.4) to the case when the additive term is any sequence of numbers, not particularly $2^n a_n$. Such an extension will generalize the analysis from [22], and in addition, it finds substantially many applications in practice.

An asymptotic approximation of $U_n^{(2)}$ depends on finding an appropriate representation on

$$V_n^{(1)} \stackrel{\text{def}}{=} \frac{1}{n+1} \sum_{k=2}^{n} \binom{n+1}{k} \frac{V_k}{2^{k-1}-1}, \tag{4.18}$$

where V_k is given by (4.14), that is,

$$V_k \stackrel{\text{def}}{=} \sum_{j=2}^{k} \binom{k}{j} \frac{B_j}{2^{j-1}-1} 2^{1-j} \frac{B_{k+1-j}(1/2) - B_{k+1-j}}{k+1-j}.$$

Note that $V_n^{(1)}$ is the last term of $U_n^{(2)}$ in Theorem 6. In order to obtain the asymptotics for V_n, we must express it in terms of Bernoulli polynomials $B_n(q)$, so we can apply the asymptotics for $R_{n,r}(q)$ (see (4.16)) found in Theorem 7. Developing the denominator $(2^{j-1}-1)$ in a geometric series, we obtain another form for V_n, namely,

$$V_n = \sum_{\lambda=2}^{\infty} \sum_{j=2}^{n} \binom{n}{j} B_j 2^{-\lambda(j-1)} \frac{B_{n+1-j}(1/2) - B_{n+1-j}}{n+1-j}. \tag{4.19}$$

Then,

LEMMA 7. *Let $q = \frac{1}{2}$ and define*

$$T_\lambda^n = \sum_{j=0}^{n} \binom{n}{j} B_j q^{\lambda(j-1)} \frac{B_{n+1-j}(q) - B_{n+1-j}}{n+1-j}. \tag{4.20a}$$

Then

$$T_\lambda^n = \sum_{l=1}^{2^{\lambda-1}-1} B_n(lq^l) + B_n. \tag{4.20b}$$

PROOF. Let $T_\lambda(z)$ be the exponential generating function for T_λ^n. Then, multiplying both sides of (4.20a) by $z^k/k!$, one finds

$$T_\lambda(z) = \frac{z}{e^z - 1} \frac{e^{z/2} - 1}{e^{z2^{-\lambda}} - 1}.$$

The easiest way to show the above is by using the so-called generalized Bernoulli polynomials, as defined in [15]. Then, by consecutive applications of the following identity $(e^{z/2} - 1) = (e^{z/4} - 1)(e^{z/4} + 1) = (e^{z/8} - 1)(e^{z/8} + 1)(e^{z/4} + 1)$, we can obtain

$$T_\lambda(z) = \frac{z}{e^z - 1} \prod_{k=2}^{\lambda} (1 + e^{zq^k}).$$

Inverting this formula one proves the lemma. □

Now, applying Lemma 7 to our expression (4.19) on V_n, we show that

$$V_n = \sum_{\lambda=2}^{\infty} \left[T_\lambda^n - 2^\lambda \frac{B_{n+1}(1/2) - B_{n+1}}{n+1} + \frac{B_n(1/2) - B_n}{2} \right].$$

This allows us to represent $U_n^{(2)} = 4\beta$ as

$$\beta = 2\left[R_{n,0}\left(\frac{1}{2}\right) + R_{n,0}(1)\right] - 8r_n\left(\frac{1}{2}\right) + 2\sum_{\lambda=2}^{\infty}\left[R_{n,0}(\lambda 2^{-\lambda}) + \frac{1}{2}R_{n,0}(1) - 2^{\lambda}r_n\left(\frac{1}{2}\right) + \frac{1}{2}R_{n,0}\left(\frac{1}{2}\right)\right]$$

where $r_n(q)$ is defined as

$$r_n(q) = \frac{1}{n+1}\sum_{k=0}^{n}\binom{n+1}{k}\frac{B_{k+1}(q) - B_{k+1}}{(k+1)(2^{k-1}-1)}.$$

The asymptotic analysis of $R_{n,0}(q)$ is given in Theorem 7; hence, to compute β we need only asymptotics for $r_n(\frac{1}{2})$ given in the lemma below.

LEMMA 8. *For large n, the following holds*

$$r_n\left(\frac{1}{2}\right) = \frac{1}{8}\left(\lg n + \frac{\gamma}{\ln 2} - \frac{1}{2}\right) + \frac{1}{\ln 2}\int_0^{1/2}\zeta'(1-t)\,dt + \frac{1}{2}g_0(n) + O(n^{-1}),$$

where [4]

$$\int_0^{1/2}\zeta'(1-t)\,dt = \frac{\gamma}{8} - \frac{\ln 2\pi}{4} + \sum_{k=2}^{\infty}\frac{\zeta(k)2^{-k-1}}{k(k+1)}.$$

PROOF. Using some well-known identities for Bernoulli polynomials [1, 4], we come up with

$$r_n(q) = \int_0^q \frac{1}{n+1}\sum_{k=2}^{n}\binom{n+1}{k}\frac{B_k(t)}{2^{k-1}-1}\,dt = \int_0^q R_{n,0}(t)\,dt,$$

and then Theorem 7 proves the lemma. □

Applying Lemma 8, one finally proves the formula on β established in Proposition 2(ii) eq. (4.15c). The Proposition 2(ii) follows from the above and (4.18), if one notes that var $U_n = u_n^2 + u_n^1 - (u_n^1)^2$, and the fluctuating function $G_2(n)$ in (2.14) becomes $G_2(n) = 12_{g_0}(n)$ and $4G_0(n)$, where $g_0(n)$ and $G_0(n)$ are computed in Theorem 7. Finally, to prove the asymptotics for the higher moments, we proceed in the same manner as above. Details are left to the reader. In fact, the reader can find this derivation, and some others omitted in this paper, in [20].

Appendix. Upper and Lower Bounds for $U_n^{(2)}$

We estimate $U_n^{(2)}$ defined as

$$(2^n - 2)U_n^{(2)} = 4u_n^1 + \sum_{k=1}^{n-1}\binom{n}{k}U_k^{(2)} \qquad n \geq 2 \tag{A1}$$

with $U_0^{(2)} = U_1^{(2)} = 0$. This recurrence is not of type (4.4), and in this appendix, we give a tight lower bound and a tight upper bound on $U_n^{(2)}$. In fact, we prove that $U_n^{(2)} = O(1)$.

Note that by Proposition 2(i) $u_n^1 = \lg n + O(1)$; hence, we can find such constants ξ_0, ξ_1 and ξ_2 that $\xi_0 \leq u_n^1 \leq \xi_1 n + \xi_2$. This implies that upper and lower bounds for u_n^1 might be established through Theorems 5 and 7, since for the lower bound we assume $a_n = \xi_0 2^{-n}$ while for the upper bound we set $a_n = \xi_1 n 2^{-n} + \xi_2 2^{-n}$, and these fall into our recurrence for (4.4). The accuracy of our evaluation depends on a

good approximation of u_n^1 for small values of n, say $n \le N$. In fact, we assume that we know $u_0^1 = u_1^1 = 0$ and $u_2^1, \ldots, u_N^1$. Then

LEMMA A1. *For $n > N$, the following holds*

$$\xi_0 \le u_n^1 \le \xi_1 n + \xi_2 \tag{A2}$$

with $\xi_0 = u_{N+1}^1$, $\xi_1 = [(N+1)\ln 2]^{-1}$, $\xi_2 = \xi_0 - 1/\ln 2$.

PROOF. The proof uses induction applied to recurrence (4.1), and is left to the reader. □

Let us now define two sequences $\underline{x}_n$ and $\bar{x}_n$ as

$$\begin{aligned} \underline{x}_0 = \underline{x}_1 = 0, \qquad \underline{x}_2 = \frac{U_2^{(2)}}{4}, \quad \ldots, \underline{x}_N = \frac{U_N^{(2)}}{4}, \\ (2^n - 2)\underline{x}_n = \xi_0 + \sum_{k=1}^{n-1} \binom{n}{k} \underline{x}_k \qquad n > N, \end{aligned} \tag{A3}$$

and

$$\begin{aligned} \bar{x}_0 = \bar{x}_1 = 0, \qquad \bar{x}_2 = \frac{U_2^{(2)}}{4}, \quad \ldots, \bar{x}_N = \frac{U_N^{(2)}}{4}, \\ (2^n - 2)\bar{x}_n = \xi_2 + \xi_1 n + \sum_{k=1}^{n-1} \binom{n}{k} \bar{x}_k \qquad n > N. \end{aligned} \tag{A4}$$

Note that by Lemma 7 $4\underline{x}_n \le U_n^{(2)} \le 4\bar{x}_n$. The asymptotic approximations for (A3) and (A4) are available by Theorems 5 and 7 with $a_n = \xi_0 2^{-n}$ and $a_n = \xi_1 n 2^{-n} + \xi_2 2^{-n}$, respectively.

THEOREM A1. *For large n, the following holds*

$$\underline{x}_n = 0.5\xi_0\{\theta + 0.5\} + \frac{1}{\ln 2} \sum_{r=2}^{N} \frac{\zeta(r) G_r}{r} + O(n^{-1}), \tag{A5}$$

with

$$G_r = \underline{x}_r - 2^{-r}\left\{\underline{x}_k + \xi_0 + \sum_{i=1}^{r} \binom{r}{k} \underline{x}_i\right\}, \qquad r = 1, 2, \ldots, N$$

and

$$\bar{x}_n = \xi_1 + 0.5\xi_2(\theta + 0.5) + \frac{1}{\ln 2} \sum_{r=2}^{N} \frac{\zeta(r) g_r}{r} + O(n^{-1}), \tag{A6}$$

with

$$g_r = \bar{x}_r - 2^{-r}\left\{\bar{x}_r + \xi_1 r + \xi_2 + \sum_{i=1}^{N} \binom{r}{i} \bar{x}_i\right\} \qquad r = 1, 2, \ldots, N.$$

Note that by Theorem A1 we have proved that $U_n^{(2)} = O(1)$. Let $U_n^{(2)} = 4\beta$, and $\underline{\beta}$, $\bar{\beta}$ be the lower and the upper bound for β, that is, $\underline{x}_n = \underline{\beta}$ and $\bar{x}_n = \bar{\beta}$. The accuracy of β evaluation depends on N. Table II contains $\underline{\beta}$ and $\bar{\beta}$ for $2 \le N \le 6$.

In Section 4, we have proved that $\beta = 0.487385$, which confirms the above approximations. In fact, the method established here can be used to solve the recurrence (4.4) in the case when Theorems 5 and 7 are not applicable; that is, when the additive term is not of the form $2^n a_n$. For example, if the additive term in (4.4) is $\log_2 n$, then using our approach, we can prove that $0.4997 \le x_n \le 0.5001$.

TABLE II

N	$\underline{\beta}$	$\overline{\beta}$
2	0.46574	0.49869
3	0.48020	0.49031
4	0.48479	0.48824
5	0.486411	0.48766
6	0.48701	0.48748

NOTE ADDED IN PROOF. Recently, this was formally proved by B. Rais, P. Jacquet, and W. Szpankowski, A limiting distribution for the depth in Patricia tries. Tech. Rep. CSD TR-954. Purdue Univ., West Lafayette, Ind., 1989.

REFERENCES

1. ABRAMOWITZ, M., AND STEGUN, I. *Handbook of Mathematical Functions.* Wiley, New York, 1972.
2. AHO, A., HOPCROFT, J., AND ULLMAN, J. *Data Structures and Algorithms.* Addison-Wesley, Reading, Mass., 1983.
3. BERNDT, B. *Ramanujan's Notebooks.* Springer-Verlag, New York, 1985.
4. ERDELYI, A. *Higher Transcendental Functions.* McGraw-Hill, New York, 1953.
5. FAGIN, R., NIEVERGELT, J., PIPPENGER, N., AND STRONG, H. Extendible hashing: A fast access method for dynamic files. *ACM Trans. Datab. Syst. 4* (1979), 315–344.
6. FLAJOLET, PH. On the performance evaluation of extendible hashing and trie searching. *Acta Inf. 20* (1983), 345–369.
7. FLAJOLET, PH., AND SAHEB, N. The complexity of generating an exponentially distributed variate. *J. Algorithms 7* (1986), 463–488.
8. FLAJOLET, PH., AND SEDGEWICK, R. Digital search trees revisited. *SIAM J. Comput. 15* (1986), 748–767.
9. GALLAGER, R. Conflict resolution in random access broadcast networks. In *Proceedings of the AFOSR Workshop in Communication Theory and Applications.* Provincetown, Mass., 1978, pp. 74–76.
10. GONNET, G. *Handbook of Algorithms and Data Structures.* Addison-Wesley, Reading, Mass., 1984.
11. HENRICI, P. *Applied and Computational Complex Analysis.* Wiley, New York, 1977.
12. JACQUET, PH., AND REGNIER, M. Limiting distributions for trie parameters. In *Lecture Notes in Computer Science*, vol. 214. Springer-Verlag, New York, 1986, pp. 196–210.
13. KIRSCHENHOFER, P., AND PRODINGER, H. Some further results on digital trees. In *Lecture Notes in Computer Science*, vol. 226. Springer-Verlag, New York, 1986, pp. 177–185.
14. KNUTH, D. *The Art of Computer Programming. Sorting and Searching.* Addison-Wesley, Reading, Mass., 1973.
15. NÖRLUND, N. E. Memoire sur les polynomes de Bernoulli. *Acta Math. 43* (1923), 124–196.
16. PAIGE, R., AND TARJAN, R. Three efficient algorithms based on partition refinements, preprint.
17. PITTEL, B. Asymptotical growth of a class of random trees. *Ann. Prob. 13* (1985), 414–427.
18. RIORDAN, J. *Combinatorial Identities.* Wiley, New York, 1968.
19. SEDGEWICK, R. *Algorithms.* Addison-Wesley, Reading, Mass., 1983.
20. SZPANKOWSKI, W. Patricia tries again revisited. Tech. Rep. CSD-TR 625. Dept. Comput. Sci., Purdue Univ., West Lafayette, Ind., 1986.
21. SZPANKOWSKI, W. On the recurrence equation arising in the analysis of conflict resolution algorithms. *Stochastic Models* 3 (1987), 89–114.
22. SZPANKOWSKI, W. Solution of a linear recurrence equation arising in the analysis of some algorithms. *SIAM J. Alg. and Discr. Meth. 8* (1987), 233–250.
23. SZPANKOWSKI, W. The evaluation of an alternative sum with applications to the analysis of some data structures. *Inf. Proc. Lett. 28* (1988), 13–19.
24. SZPANKOWSKI, W. Some results in V-ary asymmetric tries. *J. Alg. 9* (1988), 224–244.
25. WHITTAKER, E., AND WATSON, G. *A Course of Modern Analysis.* Cambridge Press, London, 1935.

RECEIVED SEPTEMBER 1986; REVISED JUNE 1988 AND AUGUST 1989; ACCEPTED DECEMBER 1989