Towards a Complete Characterization of Tries*

June 26, 2004

Gahyun Park
Department of Computer Science
Purdue University
West Lafayette, IN 47907-2066
U.S.A.
gpark@cs.purdue.edu

Wojciech Szpankowski
Department of Computer Science
Purdue University
West Lafayette, IN 47907-2066
U.S.A.
spa@cs.purdue.edu

Abstract

Tries and suffix trees are the most popular data structures on words. Tries were introduced in 1960 by Fredkin as an efficient method for searching and sorting digital data. Since then myriad of novel trie applications were found such as dynamic hashing, conflict resolution algorithms, leader election algorithms, IP addresses lookup, coding, polynomial factorization, Lempel-Ziv compression schemes, and so on. Furthermore, various analyses of tries reveal new fundamental properties of strings appearing in those applications. Parameters of interest are the (partial) fillup level (the largest full level of the trie), shortest path, height (longest path), typical depth, and path length (sum of depths). All of these parameters are analyzed here in a unifying manner by studying the *external* and *internal* profiles. A *profile* of a tree at level k is the number of nodes (internal or external) at level k. We derive recurrences for both profiles and solve them asymptotically for various ranges of k when the strings stored in the trie are generated by a memoryless source (extension to a Markov source is possible). In particular, we present asymptotic results for the average profile, the variance and the limiting distribution. As consequences we find the height, shortest path, fillup level, and the depth. These results are derived here by methods of analytic algorithmics such as generating functions, Mellin transform, poissonization and depoissonization, and the saddle point method.

1 Introduction

A *trie* is a digital tree built over, say n, strings (the reader is referred to [10, 12, 18] for an in depth discussion of digital trees.) A string is stored in an external node of a trie and the path length to such a node is the shortest prefix of the string that is not a prefix of any other strings. Let us assume a binary alphabet. Then each branching node in a trie is a binary node. A special case of a trie structure is a *suffix trie* (tree) which is a trie built over suffixes of a *single* string.

Tries and suffix trees are the most popular data structures on words [3]. Since 1960 tries were used in many computer science applications such as searching and sorting, dynamic hashing, conflict resolution algorithms, leader election algorithms, IP addresses lookup, coding, polynomial factorization, Lempel-Ziv compression schemes, and molecular biology. For example, in the internet IP addresses lookup problem [13, 16] one needs a fast algorithm that directs an incoming packet with a given IP address to its destination. As a matter of fact, this is the *longest matching prefix* problem, and standard tries are well suited for it. However, the search time is too large. If there are n IP addresses in the database, the search time is $O(\log n)$, and this is not acceptable. In order to improve the search time, one may use a novel data structure called a *level compressed trie* or in short LC trie [13]. In the LC trie

*This work was supported by the NSF Grant CCR-0208709, and NIH grant R01 GM068959-01.

we replace the root with a node of degree that is equal to the size of the largest full subtree emanating from the root (the depth of such a subtree is called the *fillup level*). This is further carried on recursively throughout the whole trie. Knowing the distribution of the fillup we can analyze the LC tries and the search time for the IP address problem (cf. [1, 13, 14]).

From this brief description, one may conclude that better understanding of tries and suffix tries behaviors is quintessential to various applications. Consider a trie built from n strings generated by a binary memoryless source (extensions to Markov sources and dependent strings are possible). Parameters of interest include the fillup F_n (largest full level in a trie), shortest path S_n (shortest path from the root to an external node), height H_n (the longest path), depth D_n (the path length from the root to a randomly selected node), and the total path length L_n (sums of node depths). In passing we should mention that these parameters can be re-interpreted in terms of strings reflecting fundamental structural properties of sequences. For example, consider a suffix tree built from n suffixes of a string of length n. Then the fillup $F_n = k$ if *all* 2^k strings of length k appear as substrings of the original string; the height H_n is the length of the longest substrings that appears at least twice within the first n symbols of the original string. Finally, we should mention that as proved in [5] suffix trees resembles probabilistically random tries. Thus the first step to analyze suffix trees is to gain deep understanding of tries which we investigate in this paper.

Tries were analyzed over the last thirty years for memoryless and Markov sources (cf. [7, 4, 8, 9, 10, 12, 15, 17, 18]). Surprisingly enough, very little is known about the *profile* of a trie. The *external* (*internal*) profile is define as the number of external (internal) nodes at a given level. More precisely, let B_n^k be the number of external nodes at level k in a trie built from n strings. In a similar manner we define U_n^k as the internal profile, that is, the number of internal nodes at level k. Observe that in a binary trie the number of "empty node" at level k is equal to $2^k - B_n^k - U_n^k$. To the best of our knowledge there are no results in literature concerning the profile with an exception of the so called *total average* (and *total variance*) profile defined as $\sum_k kB_n^k$ (cf. [10, 18]). We remedy this situation and present a complete characterization of the profile by analyzing the average $\mathbf{E}[B_n^k]$, variance $\mathbf{Var}[B_n^k]$ and the distribution of B_n^k for *various* ranges of k and n.

Before presenting our main results, let us ponder about possible consequences of our findings. Observe that we can express all interesting parameters of tries in terms of the external and internal profiles. Indeed, we have

$$
\begin{aligned}
F_n &= \max\{k: \ U_n^k = 2^k\}, \\
S_n &= \min\{k: \ B_n^k > 0\}, \\
H_n &= \max\{k: \ B_n^k > 0\}, \\
L_n &= \sum_{k\geq 0} kB_n^k, \\
\mathsf{Pr}(D_n &= k) = \frac{\mathbf{E}[B_n^k]}{n}.
\end{aligned}
$$

Let us now look more carefully at mathematics behind the evaluation of the profile. Define $B_n^k(u) = \mathbf{E}[u^{B_n^k}]$ and $U_n^k(u) = \mathbf{E}[u^{U_n^k}]$ as the probability generating functions of the external profile B_n^k and the internal profile U_n^k. It is easy to see that $B_n^k(u)$ satisfies the following recurrence for $n \geq 2$ and $k \geq 1$

$$
B_n^k(u) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} B_i^{k-1}(u) B_{n-i}^{k-1}(u), \tag{1}
$$

where p (and $q = 1 - p$) is the probability of generating a “1” (a “0”). In fact, the internal profile generating function satisfies the same recurrence, but with different initial condition. For example, $B_n^0(u) = 1$ for all except $n = 1$ when $B_1^0(u) = u$, while $U_n^0(u) = u$ for all $n > 1$ and equal to 1 for $n = 0, 1$. We formulate precisely initial conditions below. In this conference version of the paper we only discuss the external profile, that we shall call the profile since it should be clear that once we develop a method to analyze (1) we easily adapt it to the internal profile.

Recurrence (1) is a challenging one since it is non-linear with three parameters to control, namely n, k and u. We first express it in terms of the Poisson transform, that is, we analyze B_n^k in the Poisson model in which n is replaced by the Poisson process n with $\mathbf{E}[N] = n$. This reduces (1) to a recurrence of *functional equations* for the bivariate generating function $B(u, z) = \sum_{n\geq 0} B_n(u)z^n$ that we are able to solve by the Mellin transform [2, 18]. In order to find asymptotics for the Poisson transform we invert the Mellin transform through the saddle point method. The final step is to convert Poissonized results into the original (i.e., Bernoulli) setting. This last step is done by the analytic depoissonization developed in [6]. We should mention that one of the difficulty is that we aspire to obtain results for wide ranges of k and n.

Let us now summarize our main findings. In Theorem 1 we present the average asymptotics $\mathbf{E}[B_n^k]$ for $k = \Theta(\log n)$. Variance of B_n^k is derived in Theorem 2 for the same range of k. Finally, in Theorem 3 we present our limiting distribution of the external profile. We provide some consequences of our findings in three corollaries. Among others we analyze the height and the shortest path in a trie.

The plan for this paper is as follows. We concentrate in this conference version of the external profile. In Section 2 we describe our main results. In Section 3 we briefly discuss methodology used to derive these results and present sketches of proofs. This work falls under the so called “analytic algorithmics” initiated in 1962 by D. E. Knuth [10, 11] (cf. also [12, 15, 18]). It deals with analysis of algorithms through analytic methods, that is, those in which complex analysis plays a pivotal role. In this paper, as already discussed above, we apply generating functions, poissonization, Mellin transform, the saddle point method, and analytic depoissonization.

2 Main Results

We let B_n^k be the number of external nodes of a trie at level k when n binary strings are generated by a biased memoryless source. In each string a ”1” occurs with probability $p \in (0, 1)$ and a ”0” occurs with probability $q = 1 - p$. Throughout this paper we assume that $p > q$. We let $B_n^k(u)$ be the probability generating function(PGF) of B_n^k, that is,

$$B_n^k(u) = \mathbf{E}[u^{B_n^k}] = \sum_{l\geq 0} \mathsf{Pr}(B_n^k = l) \cdot u^l.$$

Due to the recursive nature of tries, it is easy to see that $B_n^k(u)$ satisfies the following recurrence relation

$$B_n^k(u) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} B_i^{k-1}(u) B_{n-i}^{k-1}(u), \quad n \geq 2, k \geq 1 \tag{2}$$

with the initial conditions $B_n^0(u) = 1$ if $n \neq 1$ and equal to u if $n = 1$. Furthermore, $B_0^k(u) = 1,\ k \geq 0$ and $B_1^k(u) = 1$ for $k > 0$ and equal to u when $k = 0$.

For the internal profile U_n^k, we define in a similar manner the probability generating function $U_n^k(u) = \mathbf{E}[u^{U_n^k}]$. It satisfies the same recurrence (2) as the external profile, but with different initial conditions. In particular, $U_n^0(u) = u$ for all $n > 1$ and 1 for $n = 0, 1$. For $n = 0$ we have $U_0^k(u) = 1$ for all k and $U_1^k(u) = 1$ for all $k \geq 0$. It should be clear that any methodology that handles well the external profile can also deal with the internal profile.

From now on we only analyze the external profile that we shall call the profile. Since $\mathbf{E}[B_n^k] = \frac{d}{du} B_n^k(u)|_{u=1}$ we obtain the following recurrence for the average profile

$$\mathbf{E}[B_n^k] = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} (\mathbf{E}[B_i^{k-1}] + \mathbf{E}[B_{n-i}^{k-1}]) \quad n \geq 2,\ k \geq 1. \tag{3}$$

with initial values $\mathbf{E}[B_n^0] = 0$ for all $n \neq 1$ and 1 for $n = 1$. Furthermore, $\mathbf{E}[B_0^k] = 0, k \geq 0$ and $\mathbf{E}[B_1^k] = 0$ for $k > 1$ and equal to 1 when $k = 0$.

In passing we should point out that recurrences of type (3) are not quite typical for trie analysis due to the additional parameter k. Without this additional parameter, the resulting binomial recurrences were studied quite extensively in literature [10, 12, 18].

Throughout this paper, we denote $y = \frac{\log n}{k}$ and $r = p/q$. Before presenting our main results on the average profile $\mathbf{E}[B_n^k]$, we observe that for $k = O(1)$ one easily derives $\mathbf{E}[B_n^k] \sim q^k n(1-q^k)^{n-1}$. Hereafter, we assume $k = \Theta(\log n)$.

Theorem 1 *Consider a binary trie built over n sequences generated by a memoryless source with $p > q = 1 - p$ and $r = p/q > 1$. Define $y = \log n/k$. Then the average profile $\mathbf{E}[B_n^k]$ has the following asymptotic expansions when $n \to \infty$:*

(a) *If $y = \log q^{-1} + \alpha, \quad \alpha > 0$, then*

$$\mathbf{E}[B_n^k] \sim n^\nu \exp(-n^\nu)$$

where $\nu = \frac{\alpha}{\log q^{-1} + \alpha} < 1$.

(b) *If $y = \log q^{-1} + \frac{\omega \log(\log_r k) - \log r}{k}$ with $\omega = O(1)$, then*

$$\mathbf{E}[B_n^k] \sim \frac{S_0}{\sqrt{1 + k(\log r)^2 S_0 r^{-S_0}}} \exp\left(-S_0 + k(1 + S_0 \log r) r^{-S_0}\right),$$

where S_0 is defined implicitly as a solution of

$$s = \exp(\omega \log\log_r k - \log r + k(\log r) r^{-s}).$$

(c) *If $\frac{p^2 \log p^{-1} + q^2 \log q^{-1}}{p^2 + q^2} < y < \log q^{-1}$, then*

$$\mathbf{E}[B_n^k] = b \frac{n^\nu}{\sqrt{\log n}} \left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right)$$

where for

$$S_* = S_*(y) = -\frac{1}{\log r} \log\left(\frac{-y - \log q}{y + \log p}\right)$$

we have

$$b = b(S_*) = \frac{S_*(p^{-S_*} + q^{-S_*} - 1)\Gamma(S_*)(r^{S_*/2} + r^{-S_*/2})}{(p^{-S_*} + q^{-S_*})\sqrt{2\pi/y} \log r},$$
$$\nu = \nu(S_*) = -S_* + \log(p^{-S_*} + q^{-S_*})/y.$$

(**d**) *If* $y = \frac{p^2 \log p^{-1} + q^2 \log q^{-1}}{p^2+q^2} + \frac{\omega}{\sqrt{k}}$ *with* $\omega = O(1)$, *then*

$$\mathbf{E}[B_n^k] = bn^{\nu}\Big(1 + O\Big(\frac{1}{\sqrt{\log n}}\Big)\Big),$$

where

$$b = \frac{2pq}{(p^2+q^2)\sqrt{2\pi}} \int_{\omega'}^{\infty} e^{-u^2/2} du, \quad \omega' = \frac{(p^2+q^2)\omega}{pq \log r}$$
$$\nu = 2 + \log(p^2+q^2)/y.$$

(**e**) *If* $y < \frac{p^2 \log p^{-1} + q^2 \log q^{-1}}{p^2+q^2}$, *then*

$$\mathbf{E}[B_n^k] = bn^{\nu}(1 + o(1)),$$

where

$$b = \frac{2pq}{p^2+q^2}, \qquad \nu = 2 + \frac{\log(p^2+q^2)}{y},$$

where $\nu < 1$.

Theorem 1 suggests that for $k = O(1)$ the number of external nodes is exponentially small, then increases subexponentially for k around $\log q^{-1}$, becoming polynomial (i.e., $B_n^k = O(n^{\nu})$ with $\nu \le 1$ for $\frac{\log(p^2+q^2)^{-1}}{2} < y < \log q^{-1}$ (cf. Theorem 1(c)–(e)). This is illustrated in Figure 1.

In order to apply Theorem 1 and infer asymptotic behaviors of various trie parameters, we need better understanding of the profile. In particular, we must know variance of B_n^k for various ranges of k and n. This is presented in the next theorem.

Theorem 2 *Under the same condition as in Theorem 1, we find that for* $\frac{\log(p^2+q^2)^{-1}}{2} < y < \log q^{-1}$ *variance of* B_n^k *is asymptotically equivalent to the average value, that is, as* $n \to \infty$

$$\mathbf{Var}[B_n^k] = A\frac{n^{\nu}}{\sqrt{\log n}}\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right) = A \cdot \mathbf{E}[B_n^k] \tag{4}$$

where ν *is defined in Theorem 1 and* A *is an explicitly computable constant.*

A more precise information about B_n^k one can obtain around the most important scale, namely when $k = \frac{1}{h}\log n$ where $h = -p\log p - q \log q$ is the entropy of the memoryless source. We recall that the average depth is $\mathbf{E}[D_n] = \frac{1}{h}\log n + O(1)$. In fact, $\mathsf{Pr}(D_n = k) = \frac{\mathbf{E}[B_n^k]}{n}$ and recent distributional results for the depth D_n (cf. [17, 18]) lead to the following refinement of part (d) of Theorem 1.

Corollary 1 *Let* $h = -p\log p - q\log q$ *be the entropy. For* $y = h$ *(i.e.,* $k = \frac{\log n}{h}$*),* $\mathbf{E}[B_n^k]$ *and* $\mathbf{Var}[B_n^k]$ *achieve maximum*

$$\mathbf{E}[B_n^k] = \frac{h^{\frac{3}{2}}}{\log r\sqrt{2\pi pq \log n}} n^1\Big(1 + O\Big(\frac{1}{\sqrt{\log n}}\Big)\Big), \tag{5}$$

$$\mathbf{Var}[B_n^k] = A\frac{n^1}{\sqrt{\log n}}\Big(1 + O\Big(\frac{1}{\sqrt{\log n}}\Big)\Big) \tag{6}$$

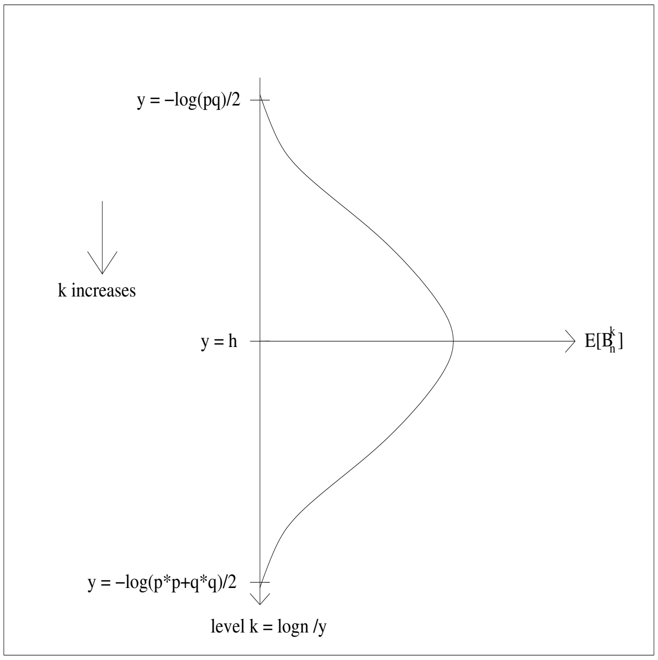

Figure 1: Average profiles as a function of k with in fixed.

where $A = \frac{h+pq(5+4pq)(p+1)^{-1}(q+1)^{-1}}{\log r\sqrt{2\pi pq/h}}$.
More precisely, for $k = \frac{\log n}{h} + x\sqrt{\mathbf{Var}[D_n]}$ *where* $x = O(1)$ *with* $\mathbf{Var}[D_n] = \frac{h_2 - h^2}{h^3}\log n + O(1)$ *and* $h_2 = p\log^2 p + q\log^2 q$, *we find*

$$\mathbf{E}[B_n^k] \sim \frac{n}{\sqrt{2\pi \mathbf{Var}[D_n]}} e^{-\frac{1}{2}x^2} = \frac{h^{\frac{3}{2}} n}{\log r\sqrt{2\pi pq\log n}} e^{-\frac{1}{2}x^2}$$

for large n.

Finally, we are able to provide some distributional results for B_n^k. In the next theorem we claim that appropriately normalized B_n^k is normal for the most interesting range of k, namely for $k = \Theta(\log n)$. This solves an open problem of [18].

Theorem 3 *For* $\frac{\log(p^2+q^2)^{-1}}{2} < y < \log q^{-1}$, *the limiting distribution of* B_n^k *is normal, that is,*

$$\frac{B_n^k - \mathbf{E}[B_n^k]}{\sqrt{\mathbf{Var}[B_n^k]}} \to N(0,1), \qquad \text{as } n \to \infty \tag{7}$$

where $N(0,1)$ *represents the standard normal distribution. More precisely,*

$$e^{-\tau\mu_n/\sigma_n} B_n^k(e^{\tau/\sigma_n}) = e^{\tau^2/2}(1 + O(1/\sqrt{\log n}))$$

for any complex τ, *where we abbreviated* $\mu_n = \mathbf{E}[B_n^k]$ *and* $\sigma_n^2 = \mathbf{Var}[B_n^k]$.

As mentioned in the introduction, knowing profile we can evaluate all trie parameters. For example, in the next two corollaries we estimate the height H_n and the shortest path S_n.

Corollary 2 ***(Hight of a Trie)*** *Let H_n be the hight of a trie. Then with high probability (whp) $H_n \sim \frac{2}{\log(p^2+q^2)^{-1}} \log n$ as $n \to \infty$.*

Sketch of Proof. We can easily derive it from Theorems 1 and 2. Let $\epsilon > 0$ and define $k_H = \frac{2}{\log(p^2+q^2)^{-1}} \log n$. First we estimate an upper bound for H_n as follows

$$\begin{aligned} \mathsf{Pr}(H_n \geq (1+\epsilon)k_H) &\leq \mathsf{Pr}(B_n^{(1+\epsilon)k_H} \geq 1) \\ &\leq \mathbf{E}[B_n^{(1+\epsilon)k_H}] \to 0 \end{aligned}$$

where the last implication follows from Theorem 1(e). For the lower bound, we use the second moment method [18] to find

$$\begin{aligned} \mathsf{Pr}(H_n \leq (1-\epsilon)k_H) &\leq \mathsf{Pr}(B_n^{(1-\epsilon)k_H} = 0) \\ &\leq \frac{\mathbf{Var}[B_n^{(1-\epsilon)k_H}]}{(\mathbf{E}[B_n^{(1-\epsilon)k_H}])^2} \\ &= O\left(\frac{1}{\mathbf{E}[B_n^{(1-\epsilon)k_H}]}\right) \to 0. \end{aligned}$$

as needed. ■

For the shortest path we find the following. Its proof is similar to one given above for the height, so we omit it here.

Corollary 3 ***(Shortest Path Length of a Trie)*** *Let S_n be the shortest path length of a trie. Then $S_n \sim \frac{\log n}{\log q^{-1}} - \frac{\log\log_r \frac{\log n}{\log q^{-1}}}{\log q^{-1}} + O\left(\frac{(\log\log\log n)^2}{\log n}\right)$ with high probability.*

3 Analysis and Proofs

Now we present our analytic approach for proving Theorems 1–3. In this conference version of the paper, we only sketch the main ideas and briefly explain our methodology.

We let $E_k(z)$ be the exponential generating function (EGF) of $\mathbf{E}[B_n^k]$, that is,

$$E_k(z) = \sum_{n=0}^{\infty} \mathbf{E}[B_n^k] \frac{z^n}{n!}$$

for z complex. We can directly compute $E_0(z) = z$ and $E_1(z) = z(pe^{qz} + qe^{pz} - 1)$ from the initial conditions of B_n^k. For $k \geq 2$ the recurrence (3) yields

$$E_k(z) = E_{k-1}(zp)e^{qz} + E_{k-1}(zq)e^{pz}.$$

By multiplying e^{-z} on both sides of the above equation, we have the following simple recurrence

$$\tilde{E}_k(z) = \tilde{E}_{k-1}(zp) + \tilde{E}_{k-1}(zq), \qquad k \geq 2, \tag{8}$$

where $\tilde{E}_k(z) = E_k(z)e^{-z}$ is the Poisson transform of the sequence $\mathbf{E}[B_n^k]$.

First, we assume that z is real. Then the Mellin transform (cf. [2, 18] for a detailed discussion on Mellin transforms) of $\tilde{E}_k(z)$ is defined as

$$\tilde{E}_k^*(s) := \int_0^{\infty} z^{s-1} \tilde{E}_k(z) dz.$$

Using recurrence (8), we have $\tilde{E}_k^*(s) = (p^{-s} + q^{-s})^{k-1}\tilde{E}_1^*(s)$. Since $\tilde{E}_1^*(s) = s(p^{-s} + q^{-s} - 1)\Gamma(s)$ (where $\Gamma(s)$ is the Euler's gamma function) for $\Re(s) \in (-2, \infty)$, we arrive at

$$\tilde{E}_k^*(s) = (p^{-s} + q^{-s})^{k-1} s(p^{-s} + q^{-s} - 1)\Gamma(s),$$

and its fundamental strip is $\Re(s) \in (-2, \infty)$ for any fixed k. We note, however, that for $k = O(\log n)$ and $n \to \infty$ the fundamental strip will be smaller.

To compute $\tilde{E}_k(z)$ from $\tilde{E}_k^*(s)$ we must apply the inverse Mellin transform, that is,

$$\begin{aligned} \tilde{E}_k(z) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} z^{-s}\tilde{E}_k^*(s)ds \\ &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} s(p^{-s} + q^{-s} - 1)\Gamma(s)z^{-s}(p^{-s} + q^{-s})^{k-1}ds \end{aligned} \tag{9}$$

where $-2 < c < \infty$ and z is real. We note that the integrand of (9) has simple poles at $s = -2, -3, -4, \ldots$.

3.1 Sketch of the Proof for Theorem 1

We establish Theorem 1 by asymptotically evaluating the integral in (9) using the saddle point method [18]. Since Laplace's method is an integral part of the saddle point method we recall it below.

Theorem 4 (Laplace's Method) *Let*

$$I(x) = \int_a^b f(t)e^{-xh(t)}dt$$

where $-\infty \le a < b \le \infty$ and x is a large positive parameter. Assume that
(i) $h(t)$ has only one minimum inside (a, b) at point $t = t_0$.
(ii) $h(t)$ and $f(t)$ are continuously differentiable in a neighborhood of t_0 with $f(t_0) \neq 0$.
(iii) The integral $I(x)$ exists for sufficiently large x.
Then the following holds

$$I(x) = f(t_0)\sqrt{\frac{2\pi}{xh''(t_0)}}e^{-xh(t_0)}\left(1 + O\left(\frac{1}{\sqrt{x}}\right)\right)$$

for $x \to \infty$.

In this conference paper we concentrate only on proving part (c) of Theorem 1. Recall that $k = \frac{\log n}{y}$ and in part (c) we assume that $\frac{p^2 \log p^{-1} + q^2 \log q^{-1}}{p^2+q^2} < y < \log q^{-1}$. We treat y as a constant. Writing $z^{-s}(p^{-s} + q^{-s})^k = \exp[-s \log z + k \log(p^{-s} + q^{-s})]$ we see that there are saddle points where

$$\frac{d}{ds}[-s \log z + k \log(p^{-s} + q^{-s})] = 0,$$

that is,

$$y \equiv \frac{\log z}{k} = \frac{p^{-s}\log(1/p) + q^{-s}\log(1/q)}{p^{-s} + q^{-s}}.$$

Solving the above explicitly we obtain

$$s = S_* \equiv \frac{-1}{\log r} \log\left(\frac{-y - \log q}{y + \log p}\right). \tag{10}$$

The saddle lies on the real axis for $p^{-k} < z < q^{-k}$ (i.e. $y \in (\log p^{-1}, \log q^{-1})$). For $\frac{p^2 \log p^{-1} + q^2 \log q^{-1}}{p^2+q^2} < y < \log q^{-1}$ we have $S_* \in (-2, \infty)$. By shifting the contour in (9) into another vertical contour that goes through the saddle, we obtain

$$\begin{aligned}\tilde{E}_k(z) \sim \tilde{E}^{saddle}(S_*) &\equiv \frac{1}{\sqrt{2\pi\Delta_0}} S_*(p^{-S_*} + q^{-S_*} - 1)\Gamma(S_*) z^{-S_*} (p^{-S_*} + q^{-S_*})^{k-1} \\ &= \frac{1}{\sqrt{2\pi\Delta_0}} S_*(p^{-S_*} + q^{-S_*} - 1)\Gamma(S_*) z^{\nu}, \quad \nu = -S_* + \frac{\log(p^{-S_*} + q^{-S_*})}{y} \end{aligned} \tag{11}$$

where

$$\Delta_0 = \frac{d^2}{ds^2}[-s \log z + k \log(p^{-s} + q^{-s})] = \frac{k(\log r)^2}{(r^{S_*}+1)(r^{-S_*}+1)}.$$

We observe that $\nu = \nu(s) = -s + \frac{\log(p^{-s}+q^{-s})}{y} = -s + \frac{(p^{-s}+q^{-s})(\log(p^{-s}+q^{-s}))}{-p^{-s}\log p - q^{-s}\log q}$ has maximum 1 at $s = -1$ and increases for $s < -1$ and decreases $s > -1$. In fact, $\nu'(s) = \frac{-p^{-s}q^{-s}(\log p - \log q)^2 \log(p^{-s}+q^{-s})}{(p^{-s}\log p + q^{-s}\log q)^2} = 0$ when $s = -1$ and $\nu'(s) > 0$ for $s < -1$ and $\nu'(s) < 0$ for $s > -1$. In other words, $\tilde{E}_k(z)$ achieves maximum growth rate when $y = y(-1) = h$, where $h = -p\log p - q \log q$ is the entropy (this leads to Corollary 1).

To complete the proof we must *depoissonize*, that is, to find $\mathbf{E}[B_n^k]$ from the asymptotic expansion (11) of the Poisson transform $\tilde{E}_k(z)$. But up to now we have assumed that $z \to \infty$ along real axis. In order to apply depoissonization we need extend the results to complex plane. This is accomplished through analytic continuation. Thus, in a cone $\mathcal{S}_\theta$ ($\theta < \pi/2$) around real axis we know the asymptotic expansion of $\tilde{E}_k(z)$. Is it enough to recover asymptotics of $\mathbf{E}[B_n^k]$? In general, the answer is in the negative, but in this particular case we can apply the so called *analytic depoissonization lemma* developed in [6]. We recall it below.

Theorem 5 ***(Jacquet and Szpankowski, 1998)*** *Let $\tilde{G}(z)$ be the Poisson transform of a sequence g_n that is assumed to be an entire function of z. We postulate that in a linear cone $S_\theta (\theta < \pi/2)$ the following two conditions simultaneously hold: (I) For $z \in S_\theta$ and some reals $B, R > 0, \nu$*

$$|z| > R \Rightarrow |\tilde{G}(z)| \le B|z|^\nu \Psi(|z|), \tag{12}$$

where $\Psi(x)$ is a slowly varying function, that is , such that for fixed t $\lim_{x\to\infty} \frac{\Psi(tx)}{\Psi(x)} = 1$; (O) For $z \notin S_\theta$ and $A, \alpha < 1$

$$|z| > R \Rightarrow |\tilde{G}(z) e^z| \le A \exp(\alpha |z|). \tag{13}$$

Then for every nonnegative integer m

$$\begin{aligned} g_n &= \sum_{i=0}^{m} \sum_{j=0}^{i+m} b_{ij} \tilde{G}^{(j)}(n) + O(n^{\nu-m-1}\Psi(n)) \\ &= \tilde{G}(n) + \sum_{k=1}^{m} \sum_{i=1}^{k} b_{i,k+i} n^i \tilde{G}^{k+i}(n) + O(n^{\nu-m-1}\Psi(n)), \end{aligned} \tag{14}$$

where b_{ij} *are the coefficients of* $\exp(x\log(1+y)-xy)$ *at* x^iy^j*, that is,*

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} b_{ij}x^iy^j = \exp(x\log(1+y)-xy).$$

with $b_{ij}=0$ *for* $j<2i$.

Condition (I) in Theorem 5 is satisfied in a cone $\mathcal{S}_\theta$ by (11) and analytic continuation. The outside condition (O) can be proved by induction on k: It is easy to verify that $|E_1(z)| < |z|e^{\beta|z|} < e^{\alpha|z|}$, for $\beta=\cos(\theta)>0, \beta<\alpha<1$ since $E_1(z)=z(pe^{qz}+qe^{pz}-1)$. For $k>1$, we have $|E_k| = |E_{k-1}(zp)e^{qz}+E_{k-1}(zq)e^{pz}| \le |z|pe^{\beta|z|p}e^{\beta|z|q}+|z|qe^{\beta|z|q}e^{\beta|z|p} = |z|e^{\beta|z|}\le e^{\alpha|z|}$. Thus $\mathbf{E}[B_n^k]=\tilde{E}(n)+O(n^{\nu-1})$. This proves Theorem 1(c).

3.2 Sketch of the Proof for Theorem 2

For simplicity, we denote $\mathbf{E}[B_n^k]$ and $\mathbf{E}[B_n^k(B_n^k-1)]$ as b_n^k and c_n^k, respectively. Thus $b_n^k = \frac{d}{du}B_n^k(u)|_{u=1}$ and $c_n^k=\frac{d^2}{du^2}B_n^k(u)|_{u=1}$.

Differentiating twice with respect to u in (2) and setting $u=1$, we obtain for $n\ge 2$

$$c_n^k = \sum_{i=0}^{n}\binom{n}{i}p^iq^{n-i}(c_i^{k-1}+c_{n-i}^{k-1}+2b_i^{k-1}b_{n-i}^{k-1}), \qquad n\ge 0,\ k\ge 1. \tag{15}$$

where $c_0^k=c_1^k=0$ for all $k\ge 0$. Let $c_k(z)$ be the EGF of c_n^k and $\tilde{c}_k(z)$ be the Poisson transform, that is

$$c_k(z)=\sum_{n=0}^{\infty}c_n^k\frac{z^n}{n!},$$

and $\tilde{c}_k(z)=c_k(z)e^{-z}$.

Recurrence (15) implies

$$c_k(z)=c_{k-1}(zp)e^{qz}+c_{k-1}(zq)e^{pz}+2E_{k-1}(pz)E_{k-1}(qz), \qquad k\ge 1$$

and hence

$$\tilde{c}_k(z)=\tilde{c}_{k-1}(zp)+\tilde{c}_{k-1}(zq)+2\tilde{E}_{k-1}(pz)\tilde{E}_{k-1}(qz), \qquad k\ge 1. \tag{16}$$

We introduce the Poisson variance $\tilde{V}_k(z)$ as

$$\tilde{V}_k(z)=\tilde{c}_k(z)+\tilde{E}_k(z)-\left(\tilde{E}_k(z)\right)^2. \tag{17}$$

Then using (16) and the fact that $\tilde{E}_k(z)=\tilde{E}_{k-1}(zp)+\tilde{E}_{k-1}(zq)$ for $k\ge 2$ one obtains

$$\begin{aligned}\tilde{V}_k(z) &= \tilde{c}_{k-1}(zp)+\tilde{c}_{k-1}(zq)+\tilde{E}_{k-1}(zp)+\tilde{E}_{k-1}(zq)-\left(\tilde{E}_{k-1}(zp)\right)^2-\left(\tilde{E}_{k-1}(zq)\right)^2\\ &= \tilde{V}_{k-1}(zp)+\tilde{V}_{k-1}(zq).\end{aligned} \tag{18}$$

In order to recover $\mathbf{Var}[B_n^k]$ we must depoissonize $\tilde{V}_k(z)$. However, this not that straightforward since in general $\mathbf{Var}[B_n^k]$ and $\tilde{V}(n)$ are not equivalent as the following lemma indicates [18].

Lemma 1 *If* $\frac{\log(p^2+q^2)^{-1}}{2}<y<\log q^{-1}$*, then for* $n\to\infty$

$$\mathbf{Var}[B_n^k]=\tilde{V}_k(n)-n(\tilde{E}_k'(n))^2+O\left(\frac{n^{\nu-1}}{\sqrt{\log n}}\right),$$

and the error estimate does not depend on k.

The remaining proof follows the footsteps of the proof of Theorem 1.

3.3 Sketch of Proof for Theorem 3

Now we prove Theorem 3, that is, we establish the limiting distribution of the profile B_n^k. We show that the appropriately normalized variable B_n^k tends to the normal distribution by proving that the sequence of the normalized moment generating functions converges to the moment generating function of standard normal distribution, that is, for τ complex, $e^{-\tau\mu/\sigma}B_n^k(e^{\tau/\sigma_n}) \to e^{\tau^2/2}$. By the Levy Continuity Theorem this implies the central limit theorem. First we establish asymptotic expansion of the probability generating function (PGF) $B_n^k(u) = \mathbf{E}[u^{B_n^k}]$.

We recall from (2)

$$B_n^k(u) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} B_i^{k-1}(u) B_{n-i}^{k-1}(u) + (1-u)\delta_{n=1,k=1}.$$

Let $B_k(u,z) := \sum_{n=0}^{\infty} B_n^k(u)\frac{z^n}{n!}$ be the EGF of $B_n^k(u)$. The above equation leads to the following recurrence for $B_k(u,z)$

$$B_k(u,z) = B_{k-1}(u,pz)B_{k-1}(u,qz), \quad k \geq 2, \tag{19}$$

The Poisson transform $\tilde{B}_k(u,z) = B_k(u,z)e^{-z}$ becomes

$$\tilde{B}_k(u,z) = \tilde{B}_{k-1}(u,pz)\tilde{B}_{k-1}(u,qz), \quad k \geq 2, \tag{20}$$

with initial condition $\tilde{B}_1(u,z) = 1 + (u-1)z(pe^{-pz} + qe^{-qz} - e^{-z}) + pq(u-1)^2 z^2 e^{-z}$. The recurrence (20) involves multiplication. To overcome this problem, we take log on both sides. Let $\tilde{l}_k(u,z) = \log \tilde{B}_k(u,z)$. Then

$$\tilde{l}_k(u,z) = \tilde{l}_{k-1}(u,pz) + \tilde{l}_{k-1}(u,qz), \quad k \geq 2 \tag{21}$$

holds. Let $\tilde{l}_k^*(u,s) := \int_0^\infty z^{s-1} l_k(u,z)dz$ for $\Re(s) \in (-2,-\nu)$ be the Mellin transform of $\tilde{l}_k(u,z)$. Then $\tilde{l}_k^*(u,s) = (p^{-s}+q^{-s})^{k-1}\tilde{l}_1^*(u,s)$. Using the Taylor expansion around $t=0$ of $\log(1+t) = t - t^2/2 + O(t^3)$ we find

$$\tilde{l}_1^*(u,s) = (u-1)\tilde{E}_1^*(s) + \frac{(u-1)^2}{2}(\tilde{c}_1^*(s) - (\tilde{E}_1^2)^*(s)) + (u-1)^3 g(u,s)$$

where $g(u,s)$ is analytic in the region $|u-1| < \delta$, $\Re(s) \in (-2,\infty)$.

By the inverse Mellin transform, we obtain $\tilde{l}_k(u,z) = (u-1)\tilde{E}_k(z) + \frac{(u-1)^2}{2}(\tilde{V}_k(z) - \tilde{E}_k(z)) + (u-1)^3 O(n^\nu/\sqrt{\log n})$. Hence

$$\tilde{l}_k(e^{\tau/\sigma_n}, z) = \frac{\tau}{\sigma_n}\tilde{E}_k(z) + \frac{\tau^2}{2\sigma_n^2}\tilde{V}_k(z) + \frac{\tau^3}{\sigma_n^3}O(n^\nu/\sqrt{\log n}). \tag{22}$$

The next step is depoissonization. But Theorem 5 is of no use this time. Much care must be taken for the depoissonization process of $\tilde{B}_k(u,z) = e^{\tilde{l}(z,u)}$ since condition (I) is violated. However we prove that we can estimate $B_n^k(u)$ asymptotically from $\tilde{B}_k(u,z)$ leading to the following result.

Lemma 2 *Let* $\frac{p^2\log p^{-1} + q^2 \log q^{-1}}{p^2+q^2} < y < \log q^{-1}$. *Then* $B_n^k(e^{\tau/\sigma_n}) = \mathbf{E}[e^{\tau B_n^k/\sigma_n}]$, *for* τ *complex, has the following asymptotic*

$$B_n^k(e^{\tau/\sigma_n}) = \exp\left(\frac{\tau}{\sigma_n}\tilde{E}_k(n) + \frac{\tau^2}{2\sigma_n^2}\tilde{V}_k(n)\right)(1 + O(1/\sqrt{\log n})) \tag{23}$$

for large n.

A proof of this lemma is based on the saddle point method applied to $B_n^k(u) = \frac{n!}{2\pi i}\oint \frac{\tilde{B}_k(u,z)e^z}{z^{n+1}}dz$. We omit details proof in this conference paper. The final step is to observe that $e^{\tau\mu_n/\sigma_n}B_n^k(e^{\tau/\sigma_n}) \to e^{\tau^2/2}$ completing the proof of Theorem 3.

References

[1] L. Devroye, Analysis of Random LC Tries, *Random Structures & Algorithms*, 19, 359–375, 2001.

[2] P. Flajolet, X. Gourdon, and P. Dumas, Mellin transforms and asymptotics: Harmonic sums, *Theoretical Computer Science* **144** (1995), 3–58.

[3] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, Cambridge, 1997.

[4] P. Jacquet, and W. Szpankowski, Analysis of Digital Tries with Markovian Dependency, *IEEE Trans. Information Theory*, 37, 1470–1475, 1991.

[5] P. Jacquet, and W. Szpankowski, Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combinatorial Theory Ser. A*, 66, 237–269, 1994.

[6] P. Jacquet and W. Szpankowski, Analytical depoissonization and its applications, *Theoretical Computer Science* **201** (1998), 1–62.

[7] S. Janson and W. Szpankowski, Analysis of an asymmetric leader election algorithm, *The Electronic Journal of Combinatorics* 6**4(1)**, R17 (1997), 1–16.

[8] P. Kirschenhofer, H. Prodinger, and W. Szpankowski, On the Variance of the External Path in a Symmetric Digital Trie *Discrete Applied Mathematics*, 25, 129–143, 1989.

[9] C. Knessl and W. Szpankowski, On the number of full levels in tries, *Random Structures and Algorithms* (2004), to appear

[10] D. E. Knuth, *Fundamental Algorithms*, 3rd ed. (Reading, Massachusetts: Addison Wesley, 1997).

[11] D. E. Knuth, *Selected Papers on Analysis of Algorithms*, CSLI, Stanford, 2000.

[12] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.

[13] S. Nilsson, *Radix Sorting & Searching*, PhD Thesis, Lund University, 1996.

[14] Y. Reznik, Some Results on Tries with Adaptive Branching, *Theoretical Computer Science*, 289, 1009-1026, 2002.

[15] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms* (Reading, Massachusetts: Addison-Wesley, 1996).

[16] V. Srinivasan and G. Varghese, Fast Address Lookups using Controlled Prefix Expansions, *ACM SIGMETRICS'98*, 1998.

[17] W. Szpankowski, On the Height of Digital Trees and Related Problems, *Algorithmica*, 6, 256–277, 1991.

[18] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences* (New York: Wiley, 2001).