Expected External Profile of PATRICIA Tries *

Abram Magner
Dept. Computer Science
Purdue University
W. Lafayette, IN 47907 U.S.A.
Email: anmagner@purdue.edu

Charles Knessl
Dept. Math. Stat. & Comp. Sci.
University of Illinois at Chicago
Chicago, IL 60607-7045 U.S.A.
Email: knessl@uic.edu

Wojciech Szpankowski
Dept. Computer Science
Purdue University
W. Lafayette, IN 47907 U.S.A.
Email: spa@cs.purdue.edu

Abstract

We consider PATRICIA tries on n random binary strings generated by a memoryless source with parameter $p \geq \frac{1}{2}$. For both the symmetric ($p = 1/2$) and asymmetric cases, we analyze asymptotics of the expected value of the external profile at level $k = k(n)$, defined to be the number of leaves at level k. We study three natural ranges of k with respect to n. For k bounded, the mean profile decays exponentially with respect to n. For k growing logarithmically with n, the parameter exhibits polynomial growth in n, with some periodic fluctuations. Finally, for $k = \Theta(n)$, we see super-exponential decay, again with periodic fluctuations. Our derivations rely on analytic techniques, including Mellin transforms, analytic depoissonization, and the saddle point method. To cover wider ranges of k and n and provide more intuitive insights, we also use methods of applied mathematics, including asymptotic matching and linearization.

Key Words: Digital trees, PATRICIA trie, tree profiles, analytic combinatorics, analysis of algorithms, recurrences, generating functions, poissonization, Mellin transform, saddle point method, matched asymptotics, linearization.

1 Introduction

A *digital tree* is a fundamental data structure on words in which the storage and retrieval of a word is based on its digits. Digital trees enjoy many important applications, including data compression and distributed hashing [12, 16]. There are several variations of digital trees, two of the most important being *tries* and *digital search trees*. Various parameters of random digital trees have been defined and studied extensively, including height, size, and fill-up level [14, 2]. Many of these can be rephrased in terms of external and internal *profiles*. The *external profile* of a digital tree on n strings at level k, denoted by $B_{n,k}$, is the number of leaves at distance k from the root. Study of profiles is motivated by the fact that distributional information about them implies information about many other parameters.

This paper completes the project of analyzing the expected external profile of digital trees under a Bernoulli source model; tries and digital search trees profiles were fully treated in [3, 13]. We are concerned here with a variant of tries called PATRICIA tries, which address an inefficiency in standard tries [11]. In particular, in a standard trie, if many strings share long prefixes, the result is a tree having many non-branching paths, which is a waste of space. In a PATRICIA trie, non-branching paths are *compressed*; that is, a non-branching path corresponding to symbols $x_1 \dots x_m$ is replaced by a single node whose parent edge is labeled with the string $x_1 \dots x_m$ (see Figure 1 for an illustration).

As the first important step toward a full characterization of PATRICIA tries, here we study the expected external profile $\mathbb{E}[B_{n,k}] = \mu_{n,k}$ of PATRICIA tries built from n strings generated by a memoryless source with probability of a "1" equal to $p \geq 1/2$ and probability of a "0" equal to $q := 1 - p$. The external profile is of particular mathematical interest in the case of PATRICIA

*This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, NSA Grants H98230-11-1-0184 and H98230-11-1-0141, and in addition NSF Grants DMS-0800568, and CCF-0830140.

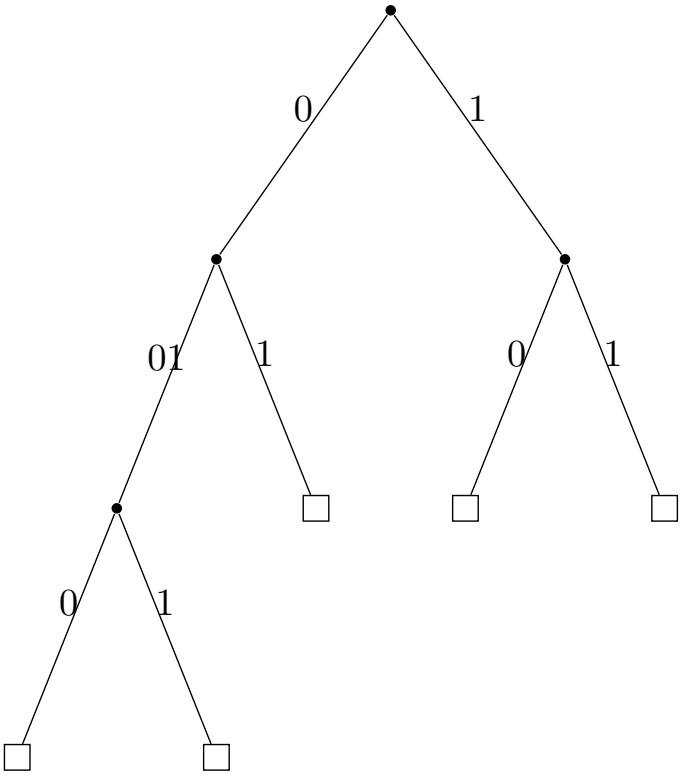

Figure 1: A PATRICIA trie on $n = 5$ strings ($s_1 = 0010\ldots$, $s_2 = 0011\ldots$, $s_3 = 01\ldots$, $s_4 = 10\ldots$, $s_5 = 11\ldots$). Note the path compression involved in the representation of s_1 and s_2. The external profile is given by $B_{5,0} = B_{5,1} = 0$, $B_{5,2} = 3$, $B_{5,3} = 2$.

tries, because it satisfies an unusual recurrence:

$$\mu_{n,k} = (p^n+q^n)\mu_{n,k}+\sum_{j=1}^{n-1}\binom{n}{j}p^j q^{n-j}(\mu_{j,k-1}+\mu_{n-j,k-1})$$

with appropriate initial conditions. The multiplicative factor and the incompleteness of the binomial sum are complications that do not arise in the analyses of tries and digital search trees (see [15]). This recurrence we solve asymptotically for various ranges of k and n. For k growing logarithmically with n, we solve it analytically by considering the Poisson transform, $\tilde{G}_k(z) := e^{-z}\sum_{n\geq 0}\mu_{n,k}\frac{z^n}{n!}$, of $\mu_{n,k}$. We shall find an expression for the Mellin transform $G_k^*(s)$, compute the inverse Mellin integral via the saddle point method, and apply analytic depoissonization to recover the asymptotics of $\mu_{n,k}$.

The peculiarities of the original recurrence are reflected in the form of the recurrence on $\tilde{G}_k(z)$:

$$\tilde{G}_k(z) = \tilde{G}_{k-1}(pz) + \tilde{G}_{k-1}(qz) + \tilde{f}_k(z),$$

where $\tilde{f}_k(z)$ is a function of $\tilde{G}_k(cz)$ with a complicated Mellin transform (see (3.7)), so that this functional equation cannot be solved explicitly. The primary difficulty in applying methods used for tries and digital search trees is the presence of the products $\tilde{G}_k(cz)e^{-(1-c)z}$ for $c = p, q$ in the definition of $\tilde{f}_k(z)$; this is not easily dealt with via standard Mellin functional identities, so $f_k^*(s)$ is implicitly given in terms of values of $\mu_{n,j}$. A similar difficulty arises in a solution to a problem posed by Knuth [6] and in the analysis of an asymmetric leader election algorithm [9]. Our problem has the additional complication that the recurrence (2.1) involves two variables. One of our main technical contributions is to tame the complexity of this recurrence, in particular showing that the Mellin transform $G_k^*(s)$ of $\tilde{G}_k(z)$ is expressible as the product of an entire function and the Euler gamma function $\Gamma(s+1)$ [1], such that some of the poles introduced by the Γ function are canceled by zeros of the entire function. We are thus able to show, via analytic techniques, that the expected profile in this range is of polynomial growth, with bounded oscillations.

For the same range ($k = O(\log n)$), we also give a more intuitive, though less mathematically precise, derivation via other methods. In particular, we apply an approximation similar in spirit to the saddle point method, directly to the recurrence (2.1).

As previously mentioned, we also solve the recurrence for several other ranges of k, in both the symmetric and asymmetric cases. This we do via methods of applied mathematics, including matched asymptotics and linearization. By these techniques, we show that for k bounded by a constant, the expected profile decays exponentially with n; for k growing logarithmically with n, it grows polynomially, with periodic fluctuations; and for $k = \Theta(n)$, it decays super-exponentially, again with periodic fluctuations.

The plan of the paper is as follows. In Section 2, we introduce some notation, give a precise formulation of the problem, present the main results in detail, and compare with results for other digital trees. In Section 3, we sketch the proofs of the main results.

2 Main Results

Here we give some notation that is used in the rest of the paper, present in detail the basic setup, and then give our main theorems and some of the intuition behind their proofs. We then discuss consequences and compare with similar results for other digital tree models.

2.1 Setup Throughout, the function $T(s)$ is given by

$$T(s) = p^{-s} + q^{-s}.$$

For any x, the fractional part of x, denoted by $\{x\}$, is given by

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

the function $\alpha(L)$ is given by

$$\alpha(L) = \alpha_L = \{\log_{1/p} L\},$$

and the constant Δ is given by

$$\Delta = \log(p/q) \geq 0.$$

All asymptotic notation is defined with $n \to \infty$ unless explicitly indicated otherwise. Define $B_{n,k}$ to be the random number of external nodes at level k of a PATRICIA trie over n independently generated strings, each an infinite sequence of i.i.d. Bernoulli random variables with probability p of taking the value "1" and $q = 1 - p$ of taking the value "0", with $p \geq q$. The fundamental recurrence for $\mu_{n,k} = \mathbb{E}[B_{n,k}]$ is

(2.1)
$$\mu_{n,k} = (p^n + q^n)\mu_{n,k} + \sum_{j=1}^{n-1} \binom{n}{j} p^j q^{n-j} (\mu_{j,k-1} + \mu_{n-j,k-1})$$

for $n \geq 2$ and $k \geq 1$. This recurrence arises from conditioning on the number of strings starting with "0". If $1 \leq j \leq n-1$ strings start with "0", then the expected external profile is a sum of contributions from the left subtree (a PATRICIA trie built on j strings) and from the right subtree (a PATRICIA trie built on $n - j$ strings). If, on the other hand, all strings start with the same symbol (which happens with probability $p^n + q^n$), then the path compression property applies, and the contribution is $\mu_{n,k}$.

The initial conditions are as follows:

$$\mu_{0,k} = 0$$

for all k,

$$\mu_{n,0} = \delta[n = 1], \quad \mu_{1,k} = \delta[k = 0],$$

and

$$\mu_{n,k} = 0$$

for $k \geq n$. The last condition, which, in the case of PATRICIA tries, arises from the path compression property, arises also in digital search tree profiles but not in those of standard tries.

The exponential generating function for $\mu_{n,k}$, defined to be

$$G_k(z) = \sum_{n \geq 0} \mu_{n,k} \frac{z^n}{n!}, \tag{2.2}$$

is then seen to satisfy the recurrence (for $k \geq 1$)

$$G_k(z) = e^{-qz} G_{k-1}(pz) + e^{-pz} G_{k-1}(qz) + f_k(z), \tag{2.3}$$

with initial condition $G_0(z) = z$ and where $f_k(z)$ is given by

$$f_k(z) = (G_k(pz) - G_{k-1}(pz)) + (G_k(qz) - G_{k-1}(qz)).$$

2.2 Asymmetric Case In this section, we present results for the asymmetric case ($p > q$), starting with the range $k = \Theta(\log n)$, for which we first give a result derived by analytic techniques. A sketch of the proof can be found in the last section.

THEOREM 2.1. (AVERAGE PROFILE FOR $k = \alpha \log n$) *Let $\epsilon > 0$ be independent of n and k, and let $\alpha \in \left(\frac{1}{\log(1/q)} + \epsilon, \frac{1}{\log(1/p)} - \epsilon\right)$. Then for $k = \alpha \log n$,*

$$\mathbb{E}[B_{n,k}] = H(\rho(\alpha), \log_{p/q}(p^k n)) \cdot \frac{n^{-\rho(\alpha)} T(\rho(\alpha))^k}{\sqrt{2\pi \kappa_*(\rho(\alpha)) k}} \left(1 + O(k^{-1/2})\right), \tag{2.4}$$

where

$$\rho(\alpha) = -\frac{1}{\log(p/q)} \log\left(\frac{\alpha \log(1/q) - 1}{1 - \alpha \log(1/p)}\right),$$
$$\kappa_*(\rho) = \frac{p^{-\rho} q^{-\rho} (\log(p/q))^2}{T(\rho)^2},$$

and $H(\rho, x)$ is a non-zero periodic function with period 1 given by

$$H(\rho, x) = \sum_{j \in \mathbb{Z}} A(\rho + it_j) \Gamma(\rho + 1 + it_j) e^{-2j\pi i x},$$

where $t_j = 2\pi j/\Delta$, and

$$A(s) = 1 + \sum_{j=1}^{\infty} T(s)^{-j} \sum_{n=j}^{\infty} T(-n)(\mu_{n,j} - \mu_{n,j-1}) \frac{\phi_n(s)}{n!}, \tag{2.5}$$

where $\phi_n(s) = \prod_{j=1}^{n-1}(s+j)$ for $n > 1$ and $\phi_n(s) = 1$ for $n \leq 1$.

We remark that the average profile given by the theorem can be written as

$$\mathbb{E}[B_{n,k}] = n^{\beta(\alpha)} K(\alpha, n),$$

where

$$\beta(\alpha) = -\rho(\alpha) + \alpha \log T(\rho(\alpha))$$

and $K(\alpha, n)$ is a slowly varying function with respect to n. This form will match the result of Theorem 2.2.

The next theorem presents results obtained via the method of matched asymptotics and other ideas from applied mathematics. The idea of asymptotic matching

is the following: suppose that we have two asymptotic expansions of $\mu_{n,k}$, each valid on some part of the domain of the problem (e.g., $k = O(1)$ and $k = \log_2 n + \xi$ in the symmetric case). If the domains of validity of the two expansions overlap, then the two should *match* in the intersection, which yields the matching condition to which we refer in the proofs. This condition allows us to determine constants and other information about our expansions. If, on the other hand, the two expansions do not match, then this implies that an intermediate scale, between the two under consideration, must be sought for a complete solution to the problem.

We include this derivation for two reasons: by this method we are able to cover a wider range of behaviors of k with respect to n, with the disadvantage of having to make some mild assumptions on the asymptotic form of $\mu_{n,k}$ (for example, the form

$$\mu_{n,k} \sim n^{\beta(\alpha)} H(\alpha, n)$$

assumed in the $k = \alpha \log n$ range is precisely an assumption that $\mu_{n,k}$ is of regular variation, in the sense of [5]). Furthermore, among the ranges analyzed is the one dealt with in Theorem 2.1; we present this new derivation as a more intuitive alternative, wherein we start with an application of a saddle point-like method directly to the recurrence (2.1).

THEOREM 2.2. (AVERAGE PROFILE FOR ALL RANGES) *Let $p > q$ and recall that $\Delta = \log(p/q)$.*

(i) For $k = O(1)$,

$$\mathbb{E}[B_{n,k}] \sim n q^k (1 - q^k)^{n-1}.$$

(ii) For $k = \alpha \log n$ with $\alpha \in \left(\frac{1}{\log(1/q)}, \frac{1}{\log(1/p)}\right)$,

$$\mathbb{E}[B_{n,k}] \sim n^{\beta(\alpha)} H(\alpha, n),$$

where

$$\beta(\alpha) = \frac{1}{\Delta}\left[\log\left(\frac{-(\alpha\log q + 1)}{1 + \alpha \log p}\right) + \alpha \log p \cdot \log\left(\frac{\alpha\Delta}{\alpha \log p + 1}\right) - \alpha \log q \cdot \log\left(\frac{\alpha\Delta}{-(\alpha \log q + 1)}\right)\right]$$

and $H(\alpha, n)$ is a slowly varying function with respect to n. This coincides with (2.4) of Theorem 2.1.

(iii) For $k = n - \ell$, with $\ell = O(1)$,

$$\mathbb{E}[B_{n,k}] \sim C_*(p) D_*(p) n! \cdot p^{k^2/2 + k/2} q^k \cdot \Psi(n - k),$$

where

$$C_*(p) = \prod_{j=2}^{\infty} (1 - p^j - q^j)^{-1},$$

$$D_*(p) = \prod_{j=2}^{\infty} \left(1 + \left(\frac{q}{p}\right)^{j-2}\right),$$

and

$$\Psi(\ell) = \frac{1}{2\pi i} \oint \frac{e^z}{z^\ell} \prod_{j=0}^{\infty} \left(\frac{1 - e^{-qp^j z}}{qp^j z}\right) \mathrm{d}z,$$

where the integral is taken around any counterclockwise contour encircling the origin.

For $\ell = n - k \to \infty$, $\Psi(\ell)$ is asymptotically equivalent to

$$\Psi(\ell) \sim \frac{1}{(\ell - 1)!} \ell^{\frac{\log q}{\log p} - \frac{1}{2}} \exp\left[-\frac{\log^2 \ell}{2 \log(1/p)}\right] \hat{\Psi}(\ell),$$

where $\hat{\Psi}(\ell)$ is the following bounded, periodic function (with $\alpha_\ell = \{\log_{1/p} \ell\}$):

$$\hat{\Psi}(\ell) = q^{\alpha_\ell} p^{-\alpha_\ell^2/2 - \alpha_\ell/2} \left(\frac{1 - e^{-qp^{-\alpha_\ell}}}{qp^{-\alpha_\ell}}\right) \cdot \prod_{J=1}^{\infty} \frac{(1 - e^{-qp^{J-\alpha_\ell}})(1 - e^{-qp^{-J-\alpha_\ell}})}{qp^{J-\alpha_\ell}}.$$

We comment that the analysis of the three scales here still leaves gaps in the asymptotics (that is, we have not covered all possible ranges). It is still necessary to consider cases where $\alpha = \frac{k}{\log n} \approx \frac{1}{\log(1/p)}$ and $\alpha \approx \frac{1}{\log(1/q)}$, since the expansion in (iii) cannot asymptotically match that in (ii) (or that in Theorem 2.1). Some preliminary results suggest that the appropriate transition scale is $n, k \to \infty$ with $k - \log_{1/p} n = O(1)$, and we will discuss it in depth in the full paper. Similarly, another expansion is needed for $\alpha \approx \frac{1}{\log(1/q)}$, which would connect the results in (i) and (ii).

2.3 Symmetric Case We now present results for the case $p = q = 1/2$.

For $k = O(1)$, it should be noted that the derived expression is different from the analogous one for the asymmetric case. In particular, the ratio of the two, when p and q are set to $1/2$, tends toward some constant not equal to 1. This occurs in the derivation as follows: for arbitrary $p \geq 1/2$ and $q \leq p$, the asymptotic formula for $\mu_{n,k}$ features two terms, the second of which is of lower order than the first when $p > q$ and of the same

order when $p = q$. The following example illustrates this phenomenon: consider $k = 1$. We can show that

$$\mu_{n,1} = \frac{n(pq^{n-1} + qp^{n-1})}{1 - p^n - q^n}.$$

In both the symmetric and asymmetric cases,

$$1 - p^n - q^n \sim 1,$$

so we can ignore the denominator. In the asymmetric case,

$$pq^{n-1} = o(qp^{n-1}),$$

so that

$$\mu_{n,1} \sim nqp^{n-1}.$$

In contrast, when $p = q = 1/2$,

$$pq^{n-1} = qp^{n-1} = 2^{-n};$$

that is, the two terms are of the same order, so that they both contribute to the leading term. Thus, we have

$$\mu_{n,1} \sim 2n2^{-n} = n2^{1-n},$$

which differs from nqp^{n-1} by a factor of 2. This phenomenon is the reason for the difference between the formulas in (i) of Theorem 2.2 and Theorem 2.3.

For the logarithmic range, we are able to glean more information than in the asymmetric case, because $\mu_{n,k}$ turns out to be asymptotically close to a product of n and a function that is periodic in $\log_2 n$, and we can then use matching conditions as $\xi \to -\infty$ to determine some information about the function's Fourier coefficients. The same phenomenon is not apparent in the asymmetric case as given in Theorem 2.2.

Finally, for the $k = n - \ell$, $\ell = O(1)$ range, we see nearly the same behavior for both the symmetric and asymmetric cases, and the derivation is essentially the same.

THEOREM 2.3. (AVERAGE PROFILE, SYMMETRIC CASE) *Let $p = q = 1/2$.*

(i) For $k = O(1)$ as $n \to \infty$,

$$\mathbb{E}[B_{n,k}] \sim \left(\frac{2^k - 1}{2^k}\right)^{n-1} n.$$

(ii) For $k = \log_2 n + \xi$ with $\xi = O(1)$,

$$\mathbb{E}[B_{n,k}] \sim n \left[C(\xi) + \sum_{j=-\infty, j\neq 0}^{\infty} C_j(\xi) e^{2\pi i j \log_2 n} \right],$$

where

$$C(\xi) \sim \exp(-2^{-\xi}), \quad \xi \to -\infty$$

and, for all $j \neq 0$, $C_j(\xi) = o(C(\xi))$ as $\xi \to -\infty$.

(iii) For $k = n - \ell$ with $\ell = O(1)$,

$$\mu_{n,k} \sim C_* n! 2^{-k^2/2 - k/2} \bar{k}_\ell,$$

where

$$C_* = \prod_{j=1}^{\infty} \frac{1}{1 - 2^{-j}}$$

and

$$\bar{k}_\ell = \frac{1}{2\pi i} \oint \frac{1}{z^\ell} \prod_{j=1}^{\infty} \left[\frac{2^j}{z} (e^{z2^{-j}} - 1) \right] \, \mathrm{d}z,$$

where the integral is taken over a contour encircling the origin.

For $\ell \to \infty$, the expression for $\bar{k}_\ell$ asymptotically simplifies to

$$\bar{k}_\ell \sim \frac{\ell^{3/2}}{\ell!} 2^{\alpha_\ell(\alpha_\ell+1)/2} \left(\frac{1 - e^{-2^{\alpha_\ell}}}{2^{\alpha_\ell}}\right) \exp\left(-\frac{\log^2 \ell}{2 \log 2}\right) \cdot \prod_{j=1}^{\infty} \frac{(1 - e^{-2^{\alpha_\ell + j}})(1 - e^{-2^{\alpha_\ell - j}})}{2^{\alpha_\ell - j}}.$$

As in the same range in the asymmetric case, factors involving α_ℓ yield oscillations that are periodic in $\log_2 \ell$.

We now briefly discuss some of the qualitative phenomena seen in the preceding results. For small k, in both the symmetric and asymmetric cases, the expected external profile exhibits roughly exponential decay in n. For the logarithmic ranges, we see polynomial growth, and it is clear in the symmetric case that there are fluctuations with period 1 in $\log_2 n$. The analysis leading to Theorem 2.2 does not show it, but similar fluctuations arise in the asymmetric case, as revealed by the analytic derivation. Finally, for k close to n, we see superexponential decay with an oscillating factor in both cases. In addition, we find in the asymmetric case that there are gaps between the first and second and the second and third ranges.

2.4 Comparison with Other Types of Digital Trees Here we compare the phenomena seen in our analysis with those observed in the analyses of other types of digital trees.

We start by comparing with tries. Analytically, they are somewhat similar, but with important differences. The saddle points of the integrand of the Mellin inversion are the same in both cases: the real-valued saddle point ρ is the same, and there are infinitely many regularly spaced saddle points on the imaginary line corresponding to ρ. This shared phenomenon is what gives rise to the oscillations in both cases in the

range of polynomial growth (discussed in more detail below). The singularities of $G_k^*(s)$, on the other hand, are different in the two cases. For regular tries, we see poles at $s = -2, -3, \ldots$, in contrast to the PATRICIA situation, where we see only poles at the integers less than or equal to $-k$. As a consequence, for $\alpha \in (\frac{1}{\log(1/q)} + \epsilon, \frac{1}{\log(1/p)} - \epsilon)$ for any constant $\epsilon > 0$, we see no effect of the poles on the asymptotics for PATRICIA tries, because the contour along which we compute the inverse Mellin transform has a real part which is contained in some bounded interval, while the poles of the integrand tend to $-\infty$ as k grows large. This is not the case for standard tries and results in more compact trees; for example, the height of the trie grows like $(2/H_2)\log n$ (H_2 is the second Rényi's entropy) while for DST and PATRICIA the growth is $1/\log(p^{-1})\log n$ (see [10]).

Qualitatively, in the asymmetric case, tries and PATRICIA tries are quite similar in the ranges that we have examined. In the small k range, we find that the two are asymptotically equivalent. For k in the logarithmic range, expected external profiles of both tries and PATRICIA tries exhibit polynomial growth with oscillations. Furthermore, the polynomials have the same order. Thus, the difference lies in the subpolynomial multiplicative factors. Finally, for $k = \Theta(n)$, expected profiles for both decay to 0, but the decay for PATRICIA tries is faster. Indeed, letting $\mu_{n,k}^{[T]}$ denote the expected external profile at level k for a standard trie on n strings,

$$\begin{aligned}\frac{\mu_{n,k}^{[T]}}{\mu_{n,k}} &\sim \frac{2pqn^2(p^2+q^2)^{k-1}}{n^k(n-k)^{1/2+\log q/\log p}p^{k^2/2+k/2}q^k} \\ &\quad \cdot \frac{1}{\exp\left[-\frac{\log^2(n-k)}{2\log(1/p)}\right]O(1)} \\ &\sim e^{\Theta(n^2)},\end{aligned}$$

which, for $k = \Theta(n)$, tends to ∞ because of the k^2 in the exponent of p in the denominator. Provided $n - k = \ell \to \infty$, oscillations appear in PATRICIA tries but are absent in standard tries.

Interestingly, in the symmetric case, standard tries and PATRICIA tries differ qualitatively: standard tries do not exhibit oscillations, to leading order, in the range of polynomial growth or in the range $k = \Theta(n)$. Meanwhile, our Theorem 2.3 shows that oscillations around $k = \log_2 n$ and $k = \Theta(n)$ do appear in PATRICIA tries.

Now we turn to digital search trees (DSTs), with which we compare in the logarithmic range in the asymmetric case. Analytically, PATRICIA tries are closer to DSTs than to standard tries. A vertical line of equally spaced saddle points also arises in the analysis of DSTs, and the location of the real-valued saddle point agrees with that in tries and PATRICIA tries, so that, again, oscillations arise in the region of polynomial growth. A difference arises in the location of singularities: in DSTs, there are no poles, owing to a phenomenon similar to one observed in our analysis: in both cases, $G_k^*(s)$ is shown to be asymptotically equal to a product of a Γ function and an entire function with zeros at certain negative integers. In the case of DSTs, all negative integer poles are canceled in this way.

As with tries both standard and PATRICIA, DSTs exhibit polynomial growth in the $k = \alpha \log n$ range, and an oscillating factor again arises due to the shared saddle point phenomenon. The polynomial order is the same as in the other two models.

In the symmetric case in the range $k = \log_2 n + \xi$, when $\xi \to \infty$, DST expected profiles exhibit periodic oscillations akin to those observed in PATRICIA profiles, but not, as mentioned earlier, in tries. The oscillations for ξ fixed that arise in PATRICIA tries are not seen in DSTs.

3 Proof Sketches

We now sketch the proofs of Theorems 2.1, 2.2, and 2.3. Since the most interesting phenomena arise when $k = \Theta(\log n)$ and, to a lesser extent, $k = \Theta(n)$, we discuss the corresponding derivations in greater detail than we do for $k = O(1)$.

3.1 Proof of Theorem 2.1 Our starting point is the Poisson transform $\tilde{G}_k(z) = e^{-z}G_k(z)$, which satisfies the recurrence

$$\tilde{G}_k(z) = \tilde{G}_{k-1}(pz) + \tilde{G}_{k-1}(qz) + \tilde{f}_k(z), \tag{3.6}$$

where

$$\begin{aligned}\tilde{f}_k(z) &= [\tilde{G}_k(pz) - \tilde{G}_{k-1}(pz)]e^{-qz} \\ &\quad + [\tilde{G}_k(qz) - \tilde{G}_{k-1}(qz)]e^{-pz},\end{aligned} \tag{3.7}$$

with initial condition $\tilde{G}_0(z) = ze^{-z}$. We then apply the Mellin transform

$$\int_0^\infty z^{s-1}\tilde{G}_k(z)\,\mathrm{d}z$$

to $\tilde{G}_k(z)$ to get a recurrence for $G_k^*(s)$. The initial condition derived from the path compression property implies that

$$\tilde{G}_k(z) = O(z^{k+1})$$

as $z \to 0$, and, by a standard argument by induction on increasing domains (see [7]), we show that, for any $\epsilon > 0$,

$$\tilde{G}_k(z) = O(z^{1+\epsilon})$$

as $z \to \infty$ in a cone containing the positive real axis, so that $G_k^*(s)$ is analytic at least in the strip $\Re(s) \in (-k-1,-1)$. The transformed function is seen to be of the form

$$G_k^*(s) = A_k(s)T(s)^k\Gamma(s+1),$$

where $A_k(s)$ and $T(s)^k$ are entire functions, with $A_k(s)$ implicitly defined in terms of various values of $\mu_{n,j}$:

$$A_k(s) = 1 + \sum_{j=1}^{k} T(s)^{-j} \sum_{n=j}^{\infty} T(-n)(\mu_{n,j} - \mu_{n,j-1})\frac{\phi_n(s)}{n!}.$$

The function $A_k(s)$ has the interesting property that, for $r \in \{1,2,\dots,k\}$, $A_k(-r) = 0$, so that the poles introduced by the Γ function at those points are canceled. The fact that $A_k(-r) = 0$ for $r \in \{2,\dots,k\}$ follows immediately from the previously established fact that $G_k^*(s)$ is analytic at these points, that $T(s)$ is never 0, and that $\Gamma(s+1)$ has poles at these points. The proof that $A_k(-1) = 0$ for all $k \geq 1$ is by induction, relying on the crucial property of $\phi_n(s)$ that its zeros are precisely $-1,-2,\dots,-(n-1)$: the base case, $k=1$, is handled easily by applying the initial conditions. For the inductive step, we write

$$\begin{aligned} A_k(-1) &= A_{k-1}(-1) \\ &\quad + T(-1)^{-k} \sum_{n=k}^{\infty} T(-n)(\mu_{n,k} - \mu_{n,k-1})\frac{\phi_n(-1)}{n!}. \end{aligned}$$

Then, by the inductive hypothesis, $A_{k-1}(-1) = 0$, and the sum vanishes because $\phi_n(-1) := \prod_{j=1}^{n-1}(-1+j) = 0$ for $n \geq 2$ (see (2.5)). This shows that $G_k^*(s)$ is meromorphically continuable to $\mathbb{C}$, with simple poles at integers less than or equal to $-k-1$.

The next step is to compute the inverse Mellin transform

$$\tilde{G}_k(z) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} G_k^*(s)z^{-s}\,\mathrm{d}s, \tag{3.8}$$

where $-k-1 < c < \infty$. We evaluate this integral via the saddle point method, defining $\rho = \rho(\alpha)$ to be the real solution to the equation

$$\frac{\mathrm{d}}{\mathrm{d}s}[k\log T(s) - s\log z] = 0.$$

Interestingly, the integrand has a unique saddle point on the real axis and infinitely many regularly spaced saddle points on the line $\Re(s) = \rho$. This leads to a fluctuating factor in $\tilde{G}_k(z)$. A fact about ρ should be noted here: as we vary α from $\frac{1}{\log(1/q)}$ to $\frac{1}{\log(1/p)}$, ρ goes from $-\infty$ to ∞, which corresponds to the boundaries for the range we consider.

Finally, we apply analytic depoissonization results to transfer asymptotics of $\tilde{G}_k(n)$ as $n \to \infty$ to asymptotics of $\mu_{n,k}$. Our tool of choice here is induction along increasing domains [7].

It remains to show that $H(\rho,x)$ is well defined for all $\rho, x \in \mathbb{R}$. That is, if we define $H_k(\rho,x)$ by the formula for $H(\rho,x)$ with A_k in place of A, then we need to show that

$$\lim_{k\to\infty} H_k(\rho,x)$$

exists and is equal to $H(\rho,x)$. To do this, we apply the *Dominated Convergence Theorem*. Thus, we need to show that $A_k(s)$ converges pointwise for all s, and then that the sum

$$\sum_{j\in\mathbb{Z}} A(s_j)\Gamma(s_j+1)e^{-2j\pi i x}$$

converges absolutely for all ρ, x (throughout, we define $s_j = \rho + it_j$). That is, we are choosing $A(s)$ as the dominating function, and we are showing that

$$\int |A(s)|\,\mathrm{d}\mu < \infty,$$

where the integral is with respect to the measure that puts mass $|\Gamma(s_j+1)|$ on each point s_j and mass 0 everywhere else.

To prove pointwise convergence of $A_k(s)$, we show absolute convergence of the sum whose mth term is

$$a_m = T(s)^{-m}\sum_{n=m}^{\infty}\eta_{n,m}\frac{\Gamma(n+s)}{\Gamma(n+1)},$$

where we define

$$\eta_{n,m} = T(-n)(\mu_{n,m} - \mu_{n,m-1}).$$

We can do this by showing that $|a_m|$ is upper bounded by the tail of a convergent geometric series. The idea is to split the sum into two parts:

$$\begin{aligned} &T(s)^{-m}\sum_{n=m}^{cm}\eta_{n,m}\frac{\Gamma(n+s)}{\Gamma(s+1)} \\ &+ T(s)^{-m}\sum_{n=cm+1}^{\infty}\eta_{n,m}\frac{\Gamma(n+s)}{\Gamma(s+1)} \end{aligned}$$

where c is a sufficiently large constant. The asymptotic formula for $\mu_{n,m}$ for m close to n happens to hold for $m \geq Cn$, for any constant C, so that, in the first sum, $|\mu_{n,m} - \mu_{n,m-1}| \leq c_1 e^{-c_2 m^2}$ for sufficiently large m (here, c_1 and c_2 are positive constants).

In the second sum, provided c is chosen appropriately, $|T(-n)T(s)^{-m}|$ decays exponentially with respect

to m and n. Combining this with the previous fact yields pointwise convergence of $A_k(s)$.

Showing that $\int |A(s)|\,\mathrm{d}\mu < \infty$ boils down to examining the behavior of the Γ function on vertical lines. Each term of the sum is given by

$$\begin{aligned}|a_j| &= |A(s_j)\Gamma(s_j+1)| \\ &\le |\Gamma(s_j+1)| \\ &+ \sum_{m\ge 1, n\ge m} |T(s_j)|^{-m}|\eta_{n,m}|\left|\frac{\Gamma(n+s_j)}{\Gamma(n+1)}\right|.\end{aligned}$$

Now, there exists a constant $0 < C < 1$ such that, for sufficiently large $|j|$ and all constant $x \in \mathbb{R}$,

$$\left|\frac{\Gamma(x+s_{j+\mathrm{sgn}(j)})}{\Gamma(x+s_j)}\right| \le C.$$

Also, $T(s_j) = T(s_{j'})$ for all $j, j' \in \mathbb{Z}$. Applying these two facts and the ratio test shows absolute convergence of the sum, which concludes the proof.

3.2 Proof of Theorem 2.2 The analysis for $k = O(1)$ is the simplest: we first derive asymptotic expressions for $k = 1, 2$, then use this to guess an asymptotic form for general k, in terms of some unknown function. We then determine this function using (2.1).

The range $k = \alpha\log(n)$ is more interesting, in terms of the methods used: we begin with the observation that the sum in (2.1) is very nearly the expected value of a function of a binomial random variable with parameters n and p. Furthermore, because of concentration of the binomial distribution, one expects that the dominant contribution to the sum comes from $j \approx np$. This phenomenon has been rigorously studied and exploited previously in, e.g., [4] and [8]. From this we get a dramatically simplified recurrence, of the form

$$\mu_{n,k} = \mu_{np,k-1} + \mu_{nq,k-1} + O(1).$$

In the next step, after applying the approximation, we assume the form

$$\mu_{n,k} = n^{\beta(\alpha)}H(\alpha, n),$$

with $H(\alpha, n)$ a slowly varying function with respect to n, and $\beta(\alpha)$ a function to be determined. The function $\beta(\alpha)$ is then determined by substitution into the simplified recurrence. In particular, we get

$$\begin{aligned}n^{\beta(\alpha)}H(\alpha,n) &= (np)^{\beta(f_p(\alpha))}H(f_p(\alpha), np) \\ &+ (nq)^{\beta(f_q(\alpha))}H(f_q(\alpha), nq),\end{aligned}$$

where

$$f_c(\alpha) = \frac{\alpha - \frac{1}{\log n}}{1 + \frac{\log c}{\log n}}.$$

Approximating the exponents using a Taylor expansion of β around α, then using the slow variation of H, yields, for n large, the equation

$$1 = (p^{\beta(\alpha)-\alpha\beta'(\alpha)} + q^{\beta(\alpha)-\alpha\beta'(\alpha)})e^{-\beta'(\alpha)}. \tag{3.9}$$

Differentiating both sides of (3.9), followed by some algebra, yields the Clairaut equation

$$\beta(\alpha) - \alpha\beta'(\alpha) = \rho(\alpha), \tag{3.10}$$

where $\rho(\alpha)$ is the real-valued saddle point occurring in the analytic derivation. This can be solved by letting $\beta(\alpha) = \alpha\gamma(\alpha)$, where γ is to be determined. Substitution of this expression into (3.10), followed by integration by parts, yields an expression for $\gamma(\alpha)$ that involves an integration constant, which we then determine by substitution of the derived expression for β into (3.9). We can prove that it coincides with (2.4) of Theorem 2.1: simple (but tedious) algebra shows that

$$\beta(\alpha) = -\rho(\alpha) + \alpha\log T(\rho(\alpha)).$$

The derivation for k close to n begins along lines similar to those for $k = O(1)$. That is, we explicitly derive expressions for $\mu_{n,n-1}, \mu_{n,n-2}$, and $\mu_{n,n-3}$. Then, from these, we posit an asymptotic form for general $\ell = n-k$, in terms of an unknown function $\xi(\ell)$. To find this unknown function, we use the original recurrence to derive a recurrence for $\xi(\ell)$, which is amenable to solution by generating functions and complex analytic techniques. From this we get an exact integral representation for $\xi(\ell)$ (cf. $\Psi(\ell)$ in Theorem 2.2), from which can be derived more explicit asymptotic expressions for $\ell \to \infty$.

3.3 Proof of Theorem 2.3 The derivations for the ranges $k = O(1)$ and $k = \Theta(n)$ are essentially the same in this case as in the asymmetric one. Thus, we briefly describe the derivation for $k = \log_2 n + \xi$.

We start as in the range $k = \Theta(\log n)$ in the asymmetric case, applying the saddle point-like approximation, which gives

$$\mu_{n,k} \approx 2\mu_{n/2,k-1}. \tag{3.11}$$

We then assume the asymptotic form

$$\mu_{n,k} \sim nF(\xi, n),$$

which we then substitute into (3.11). After some algebra, we find that F satisfies

$$F(\xi, n) = F\left(\xi, \frac{n}{2}\right),$$

so that F is periodic with period 1 in $\log_2 n$. Thus, it can be written as the Fourier series

$$F(\xi, n) = C(\xi) + \sum_{j=-\infty, j\neq 0}^{\infty} C_j(\xi) e^{2\pi i j \log_2 n}.$$

Matching this with the asymptotic expansion for $k = O(1)$ then yields explicitly the behavior of $C(\xi)$ as $\xi \to -\infty$, and a rougher estimate of the $C_j(\xi)$. The behavior of the coefficients as $\xi \to +\infty$ can also be obtained, by matching to the expansion for $k = \Theta(n)$ given in Theorem 2.3.

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] Luc Devroye. A note on the probabilistic analysis of patricia trees. *Random Struct. Algorithms*, 3(2):203–214, March 1992.

[3] Michael Drmota and Wojciech Szpankowski. The expected profile of digital search trees. *J. Comb. Theory Ser. A*, 118(7):1939–1965, October 2011.

[4] Philippe Flajolet. Singularity analysis and asymptotics of bernoulli sums. *Theoretical Computer Science*, 215:1–2, 1998.

[5] J. Galambos and E. Seneta. Regularly varying sequences. *Proceedings of the American Mathematical Society*, 41(1):pp. 110–116, 1973.

[6] Philippe Jacquet, Charles Knessl, and Wojciech Szpankowski. A note on a problem posed by D. E. Knuth on a satisfiability recurrence. *Combinatorics, Probability, and Computing*, 2014.

[7] Philippe Jacquet and Wojciech Szpankowski. Analytical depoissonization and its applications. *Theor. Comput. Sci.*, 201(1-2):1–62, July 1998.

[8] Philippe Jacquet and Wojciech Szpankowski. Entropy computations via analytic depoissonization. *IEEE Trans. Information Theory*, 45:1072–1081, 1998.

[9] Svante Janson and Wojciech Szpankowski. Analysis of an asymmetric leader election algorithm. *Electronic J. Combin*, 4:1–6, 1996.

[10] Charles Knessl and Wojciech Szpankowski. Limit laws for the height in patricia tries. *J. Algorithms*, 44(1):63–97, July 2002.

[11] Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2nd ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998.

[12] Moni Naor and Udi Wieder. Novel architectures for P2P applications: the continuous-discrete approach. In *Proceedings of the fifteenth annual ACM symposium on Parallel algorithms and architectures*, SPAA '03, pages 50–59, New York, NY, USA, 2003. ACM.

[13] G. Park, H. Hwang, P. Nicodème, and W. Szpankowski. Profiles of tries. *SIAM Journal on Computing*, 38(5):1821–1880, 2009.

[14] M. Regnier and P. Jacquet. New results on the size of tries. *Information Theory, IEEE Transactions on*, 35(1):203–205, 1989.

[15] Wojciech Szpankowski. Patricia tries again revisited. *J. ACM*, 37(4):691–711, October 1990.

[16] Wojciech Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, Inc., New York, NY, USA, 2001.