# On unary nodes in tries

Stephan Wagner

# *On unary nodes in tries*

Stephan Wagner[†]

*Department of Mathematical Sciences, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa*

The difference between ordinary tries and Patricia tries lies in the fact that all unary nodes are removed in the latter. Their average number is thus easily determined from earlier results on the size of tries/Patricia tries. In a well-known contention resolution algorithm, whose probabilistic model is essentially equivalent to tries, unary nodes correspond to repetitions, i.e., steps in the algorithm that do not resolve anything at all. In this paper, we take an individual's view on such repetitions: we consider the distribution of the number of repetitions a certain contender encounters in the course of the algorithm—which is equivalent to the number of unary nodes on the path from the root to a random string in a trie. We encounter an example of a sequence of distributions that does not actually converge to a limit distribution, but rather oscillates around a (discrete) limit distribution.

**Keywords:** tries, Patricia tries, contention trees, unary nodes, limit distribution

## 1 Introduction

Tries can certainly be said to belong to the most classical data structures, and different shape parameters have been studied extensively in the past under various models. Their main purpose is the storage and retrieval (hence the name *trie*, being part of the word *retrieval*, suggested by Fredkin [6]) of strings over a finite alphabet (bit strings, hexadecimal strings, words, DNA, etc.). Let us start with a brief definition.

A trie is built from a set of $n$ strings over a finite alphabet in such a way that the strings are stored in external nodes and that edges going out from an internal node correspond to different letters of the alphabet. Recursively, one can define tries as follows: if the set of strings has only one element, the corresponding trie only consists of a single external node containing this string; otherwise, the set of strings is split into groups according to their first letter; tries are built from each of these groups and attached to a common root. Figure 1 shows a simple example of a trie, built from the strings aabc…, acaa…, acab…, babc…, cbab…, cbbb…, cbca….

Tries were first proposed by de la Briandais [3] in the context of information processing. They play an important role in computer science due to their various applications to searching and sorting, fast retrieval [10], data compression [14], and others. The classical model for their analysis assumes that the set of strings is randomly generated by a memoryless source, and that the probabilities of the various letters are fixed. Under these assumptions, parameters such as the *size* (number of (internal) nodes) or the

[†]This material is based upon work supported financially by the National Research Foundation of South Africa under grant number 70560.

*external path length* (sum of the distances from the root to all external nodes) can be studied by means of techniques involving analytic poissonization/depoissonization and the Mellin transform (see for instance [1, 5, 8, 9, 12]). Besides expected values and variances, limit distributions have been determined as well. One of the first such distribution results, due to Jacquet and Régnier [7], states, for example, that the distribution of the (random) size of a binary trie tends to a normal distribution.

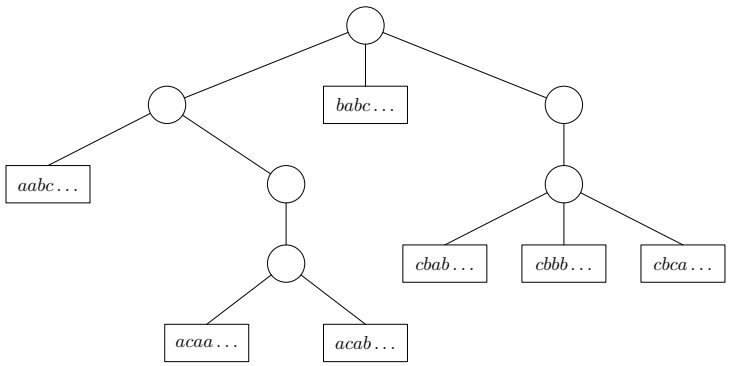


**Fig. 1:** An example of a trie

There are also more general (and thus also more realistic) source models that have been studied, for example involving finite Markov chains (see for instance [1]). For the sake of simplicity, we will stick to the case in which the $m$ letters of the alphabet in use are assigned positive probabilities $p_1, p_2, \ldots, p_m$ and all letters of a random string are independently generated, even though all results should extend to more general probabilistic models. In this simple case, the average size of a trie generated by $n$ strings is given by

$$\frac{1}{-\sum_{j=1}^{m} p_j \log(p_j)} \cdot n + nQ_1(n) + o(n),$$

see [1]. Here $Q_1(n)$ is an oscillating function (possibly 0) of small amplitude compared to the first term. This is a typical phenomenon in the analysis of tries. Note the occurrence of the *entropy* associated with the probabilities $p_1, p_2, \ldots, p_m$ in the denominator.

A variant of tries is known as *Patricia tries*; the only difference lies in the fact that unary nodes (with only one child) are compressed, thus making the data structure slightly more efficient. See Figure 2 for the Patricia trie generated from the aforementioned set of strings. Parameters of Patricia tries are somewhat more intricate to study, see Bourdon's article [1] for an extensive study of the most important shape parameters *size* and *external path length*. The average size is given by a formula similar to that for tries:

$$\frac{-\sum_{j=1}^{m}(1-p_j)\log(1-p_j)}{-\sum_{j=1}^{m} p_j \log(p_j)} \cdot n + nQ_2(n) + o(n).$$

It is worth noting that the size of Patricia tries is deterministic in the binary case, since all internal nodes are binary nodes in this case, which implies that their number is exactly $n-1$. In the general $m$-ary case, this phenomenon does not occur any longer.

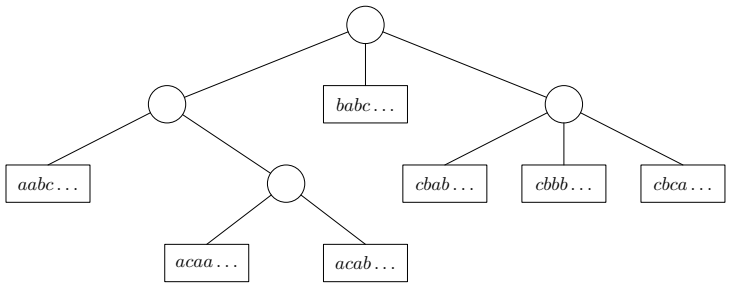


**Fig. 2:** The Patricia trie associated with the trie in Figure 1

Let us finally mention another aspect of tries related to a contention resolution scheme. Suppose that $n$ submitters contend for access to a broadcasting channel. An algorithm to resolve this situation is the contention tree algorithm due to Capetanakis [2], Tsybakov and Mikhailov [16]: for some given degree $m$, we let $m$ broadcasting slots be grouped into a frame. Now each transmitter selects one of these slot at random, with probabilities $p_1, p_2, \ldots, p_m$ (typically chosen to be equal, but there are variants where unequal probabilities are more efficient). If a slot is picked by only one contender, transmission is successful; otherwise, if a collision occurs at a certain slot, a new frame is opened for all contenders who picked this slot, and the procedure is applied recursively until all collisions are resolved. It is plain to see that this approach leads to a tree structure that is equivalent to the aforementioned description of tries. Figure 3 shows the contention tree that corresponds to the trie of Figure 1—the numbers in the slots indicate the number of contenders that choose the respective slot. It is possible in this procedure that no

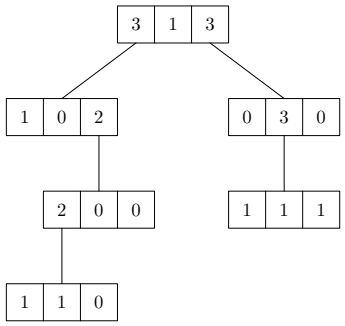


**Fig. 3:** The contention tree corresponding to the trie in Figure 1

resolution is achieved in a certain step at all, because all contenders pick the same slot of a frame. Such repetitions correspond exactly to unary nodes in tries, which constitute the difference between tries and Patricia tries. They occur surprisingly frequently, their average number being linear in $n$. Combining results on the size of tries and Patricia tries, one obtains that the average number of repetitions, given the probabilities $p_1, p_2, \ldots, p_m$ and the number $n$ of strings, is asymptotically

$$\frac{1 + \sum_{j=1}^{m} (1 - p_j) \log(1 - p_j)}{-\sum_{j=1}^{m} p_j \log(p_j)} \cdot n + nQ_3(n) + o(n). \tag{1}$$

Using the techniques of Jacquet and Régnier [7], one can also prove a central limit theorem for this quantity, the limit distribution being Gaussian. We will not give the details in this paper—the reader is referred to [8] and [15, Chapter 10] for very general results on analytic depoissonization.

Since repetitions are so numerous, it usually makes a difference in the study of trie parameters whether they are taken into account or not (in other words, whether tries or Patricia tries are considered): the size and the external path length are but two examples; let us mention the *number of $k$-cousins* as another example. We call a set of $k$ strings $k$-cousins if they are stored in a common subtrie rooted at some internal node, and no other strings occur in this subtrie. For instance, in the binary case, 2-cousins are exactly pairs of two strings sharing a common parent node. The number of $k$-cousins was studied in a recent paper of Mahmoud and Ward [11], motivated by the fact that measures for the similarity of strings play an important role in applications. Under the usual model of a memoryless source with given probabilities $p_1, p_2, \ldots, p_m$, the mean number of $k$-cousins was determined by Mahmoud and Ward to be

$$\frac{1-\sum_{j=1}^m p_j^k}{-k(k-1)\sum_{j=1}^m p_j \log(p_j)} \cdot n + nQ_4(n) + o(n).$$

The corresponding parameter for contention trees would be the number of frames with exactly $k$ contenders (henceforth called $k$-frames). In this context, however, it makes more sense to take repetitions into account (which is not done for $k$-cousins; there might be more than one possible root for the common subtrie). This, however, does not pose major difficulties: the simplest way to derive a result that takes repetitions into account is to notice that once a $k$-frame is reached, the number of repetitions to follow before the number of contenders is reduced follows a geometric distribution with parameter $1-\sum_{j=1}^m p_j^k$ (the probability that the group of $k$ contenders is split into several non-empty subgroups). Hence we obtain a formula for the mean number of $k$-frames that is even slightly simpler since the average number of repetitions cancels exactly with the numerator:

$$\frac{1}{-k(k-1)\sum_{j=1}^m p_j \log(p_j)} \cdot n + nQ_5(n) + o(n).$$

In other words, $k$-frames roughly make for a fraction of $\frac{1}{k(k-1)}$ of all frames.

In the following, we take an individual's view on repetitions: how often can a certain contender expect to encounter the frustrating situation that nothing is resolved at all, in the sense that all other contenders in the same frame choose the exact same slot? In the setting of tries generated by strings, one can formulate this in a more positive way: suppose that a trie is used to store a dictionary, and that a word is entered. An autocomplete tool suggests the following letter to be entered whenever it is unique (among words stored in the dictionary). Then the parameter we consider corresponds to the number of letters that are suggested by this tool if a random word from the dictionary is entered (which is exactly the number of unary nodes on the path from the root to the node where the word is stored).

In view of the fact that the total number of repetitions is of linear asymptotic order in $n$ and that repetitions are most likely to occur if the number of contenders is small, it is heuristically clear that the expected number of repetitions encountered by an individual is bounded. In the following, we show that the distribution of this number either converges to a discrete limit distribution or oscillates around one. As can also be seen from the above results, such oscillation phenomena are typical for tries and related structures.

## 2 The average number of repetitions encountered by an individual

Let $a_n(u)$ be the probability generating function for the number of repetitions that a certain (fixed) contender encounters (equivalently, the number of unary nodes on the path from the root to a random external node, as mentioned above), given the total number $n$ of contenders. Then the following equation is satisfied for $n > 1$:

$$a_n(u) = \sum_{k=0}^{n-1} \binom{n-1}{k} \sum_{j=1}^{m} p_j^{k+1} q_j^{n-k-1} a_{k+1}(u) + (u-1) \sum_{j=1}^{m} p_j^n a_n(u),$$

where we set $q_j = 1 - p_j$. Note that $\binom{n-1}{k} \sum_{j=1}^{m} p_j^{k+1} q_j^{n-k-1}$ is the probability that exactly $k$ fellow contenders choose the same slot, from which this equation follows easily. In particular, the sum $\sum_{j=1}^{m} p_j^n$ represents the probability that a repetition occurs, i.e., all contenders pick the same slot. Furthermore, it is clear that the initial value is $a_1(u) = 1$. Now consider the exponential generating function

$$A(x,u) = \sum_{n=1}^{\infty} \frac{a_n(u)}{n!} x^n.$$

The above equation can now be translated to an equation for $A(x,u)$ as follows:

$$\begin{aligned} A(x,u) &= \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} \frac{k+1}{n(n-k-1)!} \sum_{j=1}^{m} p_j^{k+1} q_j^{n-k-1} \frac{a_{k+1}(u)}{(k+1)!} x^n + (u-1) \sum_{n=2}^{\infty} \sum_{j=1}^{m} p_j^n \frac{a_n(u)}{n!} x^n \\ &= \sum_{k=0}^{\infty} \sum_{n=k+1}^{\infty} \frac{k+1}{n(n-k-1)!} \sum_{j=1}^{m} \left(\frac{p_j}{q_j}\right)^{k+1} \frac{a_{k+1}(u)}{(k+1)!} (q_j x)^n + (u-1) \sum_{j=1}^{m} A(p_j x, u) \\ &\quad - (u-1)x. \end{aligned}$$

This can be simplified by differentiating with respect to $x$, which merely amounts to a shift in the coefficients of $A$:

$$\begin{aligned} A_x(x,u) &= \sum_{k=0}^{\infty} \frac{k+1}{x} \sum_{j=1}^{m} (p_j x)^{k+1} \frac{a_{k+1}(u)}{(k+1)!} \sum_{n=k+1}^{\infty} \frac{(q_j x)^{n-k-1}}{(n-k-1)!} + (u-1) \sum_{j=1}^{m} p_j A_x(p_j x, u) - (u-1) \\ &= \sum_{k=0}^{\infty} \frac{k+1}{x} \sum_{j=1}^{m} (p_j x)^{k+1} \frac{a_{k+1}(u)}{(k+1)!} e^{q_j x} + (u-1) \sum_{j=1}^{m} p_j A_x(p_j x, u) - (u-1) \\ &= \sum_{j=1}^{m} p_j e^{q_j x} \sum_{k=0}^{\infty} (k+1) \cdot \frac{a_{k+1}(u)}{(k+1)!} (p_j x)^k + (u-1) \sum_{j=1}^{m} p_j A_x(p_j x, u) - (u-1) \\ &= \sum_{j=1}^{m} p_j e^{q_j x} A_x(p_j x, u) + (u-1) \sum_{j=1}^{m} p_j A_x(p_j x, u) - (u-1) \\ &= \sum_{j=1}^{m} p_j (e^{q_j x} + u - 1) A_x(p_j x, u) - (u-1) \end{aligned}$$

For $u = 1$, one has $a_n(u) = 1$ for all $n \geq 1$ and thus $A_x(x, 1) = e^x$, which indeed satisfies the functional equation. Now consider the function $B(x, u) = e^{-x} A_x(x, u)$, which is the Poisson transform of the sequence $a_n(u)$ (shifted). Then one obtains

$$B(x, u) = \sum_{j=1}^{m} p_j (1 + (u-1)e^{-q_j x}) B(p_j x, u) - (u-1)e^{-x}. \tag{2}$$

Let us consider an interesting special case first: if $m = 2$ and $p_1 = p_2 = \frac{1}{2}$, one obtains

$$B(x, u) = (1 + (u-1)e^{-x/2}) B(x/2, u) - (u-1)e^{-x},$$

which has the explicit solution $B(x, u) = \frac{1}{2-u} + \frac{1-u}{2-u} e^{-x}$, so that $A_x(x, u) = \frac{1}{2-u} e^x + \frac{1-u}{2-u}$ and thus $a_n(u) = \frac{1}{2-u}$ for all $n > 1$. Hence we have the following result:

**Proposition 1** *In the unbiased binary case ($m = 2$, $p = q = \frac{1}{2}$), the distribution of the number of repetitions encountered by a certain contender is geometric: with probability $2^{-1-\ell}$, it is equal to $\ell$. This is independent of the number of contenders ($> 1$).*

This seems to be the only special case in which the distribution does not depend on the number of contenders. The simple geometric law does not persist in the more general case either, not even asymptotically. Before we look at general distributions, let us first study the mean; we differentiate with respect to $u$ and plug in $u = 1$ to obtain

$$B_u(x, u) = \sum_{j=1}^{m} p_j (1 + (u-1)e^{-q_j x}) B_u(p_j x, u) + \sum_{j=1}^{m} p_j e^{-q_j x} B(p_j x, u) - e^{-x}$$

and thus

$$\begin{aligned} B_u(x, 1) &= \sum_{j=1}^{m} p_j B_u(p_j x, 1) + \sum_{j=1}^{m} p_j e^{-q_j x} B(p_j x, 1) - e^{-x} \\ &= \sum_{j=1}^{m} p_j B_u(p_j x, 1) + \sum_{j=1}^{m} p_j e^{-q_j x} - e^{-x}. \end{aligned}$$

In order to proceed, we apply a general depoissonization theorem:

**Theorem 2 ([15, Theorem 10.5])** *Let $G(z) = \sum_{n=0}^{\infty} \frac{g_n}{n!} z^n$ be the Poisson transform of a sequence $g_n$, and assume that it satisfies a functional equation of the form*

$$G(z) = \sum_{j=1}^{m} \gamma_j(z) G(p_j z) + t(z),$$

*where $\sum_{j=1}^{m} p_j = 1$. Suppose further that $G(z)$ is an entire function, and that the following conditions hold for some $\beta$, $0 < \theta < \frac{\pi}{2}$, $0 < \eta < 1$, $B > 0$ and all $z$ with $|z| > \xi$:*

- *For* $z \in S_\theta = \{z : |\arg(z)| < \theta\}$,

$$\sum_{j=1}^{m} |\gamma_j(z)| p_j^\beta \le 1 - \eta$$

*and*

$$|t(z)| \le B|z|^\beta.$$

- *For* $z \notin S_\theta$ *and some* $\alpha < 1$,

$$|\gamma_j(z)| e^{p_j \Re(z)} \le \frac{1}{3} e^{\alpha p_j |z|}$$

*for all* $j$ *and*

$$|t(z)| e^{\Re(z)} \le \frac{1}{3} e^{\alpha |z|}.$$

*Then*

$$g_n = G(n) + O(n^{\beta-1}).$$

See [15] (the theorem is only stated in the special case $m = 2$ there; the generalization is straightforward) and [8]. It is easy to check that the conditions are satisfied for the Poisson transform $B_u(z, 1)$, where $\beta$ can be chosen to be any positive real number. Hence we only have to consider the behavior of $B_u(n, 1)$ as $n \to \infty$. To this end, we use the standard technique for this type of functional equation, which is the Mellin transform: if $B^*(s)$ denotes the Mellin transform of $B_u(z, 1)$, then the functional equation translates to

$$B^*(s) = \sum_{j=1}^{m} p_j^{1-s} B^*(s) + \left( \sum_{j=1}^{m} p_j q_j^{-s} - 1 \right) \Gamma(s)$$

or

$$B^*(s) = \frac{\sum_{j=1}^{m} p_j q_j^{-s} - 1}{1 - \sum_{j=1}^{m} p_j^{1-s}} \Gamma(s)$$

for $-1 < s < 0$. Now we apply the Mellin inversion formula

$$B_u(n, 1) = \frac{1}{2\pi i} \int_{-1/2-i\infty}^{1/2+i\infty} B^*(s) n^{-s}\, ds$$

and shift the path of integration to the right, collecting residues at all the singularities (see [4] for a formal treatment of this procedure). Since $B^*(s)$ has poles at all solutions of the equation

$$\sum_{j=1}^{m} p_j^{1-s} = 1, \tag{3}$$

one needs information on the location of these solutions. This is given by the following result:

**Proposition 3 ([11])** *The roots of the characteristic equation* (3) *fall in the vertical strip* $0 \le \Re(s) \le \Delta$ *for some* $\Delta$ *that depends on* $p_1, p_2, \ldots, p_m$*; furthermore, if the plane is cut into horizontal slices of height* $\frac{2\pi}{|\log \min_j p_j|}$*, then each of these slices contains exactly one root.*

In view of this result and Theorem 2, the aforementioned Mellin transform technique yields

$$\frac{-\sum_{j=1}^m p_j \log q_j}{-\sum_{j=1}^m p_j \log p_j} + \sum_{\rho \neq 0} \frac{\left(\sum_{j=1}^m p_j q_j^{-\rho} - 1\right) \Gamma(\rho) n^{-\rho}}{-\sum_{j=1}^m p_j^{1-\rho} \log p_j} + o(n^{-\Delta}) + O(n^{\beta-1})$$

for the average number of repetitions encountered by an individual, where the sum is taken over all roots $\rho$ of the characteristic equation. Since $\Gamma(\rho)$ decreases exponentially as the imaginary part of $\rho$ goes to infinity, the sum converges. Now there are essentially two cases:

- If there are further characteristic roots (other than 0) on the vertical line $\Re(s) = 0$, then the contributions of all such roots sum to a periodic function in $\log n$, whose Fourier series can be determined explicitly. This is the *periodic case* in the terminology of [1]. This occurs if the numbers $\log p_j$ are rational multiples of each other. In particular, in the uniform case ($p_1 = p_2 = \ldots = p_m = \frac{1}{m}$), the characteristic equation (3) reduces to $m^s = 1$, whose solutions are $\rho_j = \frac{2\pi i j}{\log m}$, $j \in \mathbb{Z}$. Hence one obtains the Fourier series

  $$1 - \frac{\log(m-1)}{\log m} + \sum_{j \in \mathbb{Z}, j \neq 0} \frac{\exp(-2\pi i j \log(m-1)/\log m) - 1}{\log m} \Gamma\left(\frac{2\pi i j}{\log m}\right) \exp\left(\frac{-2\pi i j \log n}{\log m}\right).$$

  Note in particular that all terms in the sum vanish for $m = 2$ (in this case, the Mellin transform $B^*(s)$ is simply $\Gamma(s)$). See Figure 4 for a plot of this function in the case $m = 3$.

- If there are no such roots, then the contributions of all characteristic roots other than 0 only add to a term of order $o(1)$.

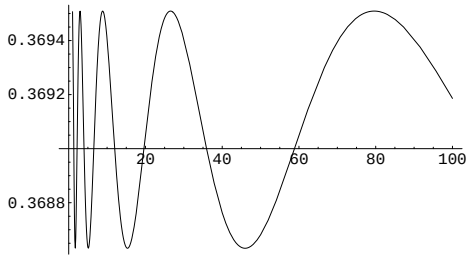


**Fig. 4:** A plot of the oscillating function in the case $m = 3$, $p_1 = p_2 = p_3 = \frac{1}{3}$

Let us collect all these results in a single theorem:

**Theorem 4** *As the number of contenders goes to infinity, the expected number of repetitions encountered by an individual is given by*

$$\frac{-\sum_{j=1}^m p_j \log q_j}{-\sum_{j=1}^m p_j \log p_j} + Q_6(n) + o(1),$$

*where $Q_6(n)$ is a small oscillating function (possibly 0). In particular, it oscillates around $1 - \frac{\log(m-1)}{\log m}$ in the unbiased case $p_1 = p_2 = \ldots = p_m = \frac{1}{m}$.*

Let us compare this to equation (1), which gives the average total number of repetitions. Of course, this average must be less than $n$ times the average given in Theorem 4 for reasons of double-counting (which is also the reason why Theorem 4 cannot be obtained directly from the results on tries and Patricia tries mentioned in the introduction). Roughly speaking, the quotient

$$\frac{-\sum_{j=1}^{m} p_j \log q_j}{1 + \sum q_j \log q_j}$$

(average total number of repetitions, counted with multiplicity, divided by the average total number of repetitions), gives the average number of contenders involved in a repetition. From this interpretation, it follows that the quotient must be $\geq 2$, which of course can also be shown directly. In the unbiased case, the quotient tends to 2 as $m \to \infty$.

# 3 Limit distribution

Let us now consider the distribution in more detail. Again, oscillation phenomena arise, in particular in the periodic case. In this case, we observe the interesting phenomenon that there is no limit distribution, but that the distribution of the number of repetitions oscillates around a certain distribution. Since the mean is $O(1)$, one naturally expects a discrete distribution (as opposed to the continuous distributions for size and depth, see [7, 13]), which is indeed the case. However, except for the very special case $m = 2$ (Proposition 1), it is none of the classical distributions and can only be described somewhat implicitly. Our aim is the following theorem:

**Theorem 5** *Let $R_n$ be the (random) number of repetitions encountered by a certain contender in our contention resolution algorithm; equivalently, $R_n$ is the number of unary nodes on the path from the root to a random external node in a trie. Then the following cases occur:*

- *In the aperiodic case (i.e., there are no characteristic roots on the vertical line $\Re(s) = 0$ other than $s = 0$), or if $m = 2$ and $p_1 = p_2 = \frac{1}{2}$, $R_n$ tends to a discrete limit distribution.*
- *In the periodic case (further characteristic roots on the vertical line $\Re(s) = 0$), except for $m = 2$, $p_1 = p_2 = \frac{1}{2}$, the distribution of $R_n$ oscillates around a fixed distribution, in the sense that*

$$\mathbb{P}(R_n = k) = L_k + P_k(n) + o(1)$$

*for all $k \geq 0$, where $L_k$ is a constant and $P_k(n)$ is a function that is periodic in $\log n$.*

**Proof:** The case $m = 2$, $p_1 = p_2 = \frac{1}{2}$ is already covered by Proposition 1. Now let us consider the Poisson transform $B(x, u)$ again. We have

$$r_{n,k} = \mathbb{P}(R_n = k) = (n-1)![u^k x^{n-1}] e^x B(x, u).$$

Now recall the functional equation (2):

$$B(x, u) = \sum_{j=1}^{m} p_j (1 + (u-1)e^{-q_j x}) B(p_j x, u) - (u-1)e^{-x}.$$

We consider the coefficient of $u^k$, which we denote by $C_k(x)$:

$$C_k(x) = e^{-x} \sum_{n=1}^{\infty} \frac{r_{n,k}}{(n-1)!} x^{n-1} = [u^k]B(x,u).$$

Now start with $C_0(x)$; from the functional equation for $B(x,u)$, we obtain

$$C_0(x) = \sum_{j=1}^{m} p_j (1 - e^{-q_j x}) C_0(p_j x) + e^{-x}.$$

Rewriting this as

$$C_0(x) = \sum_{j=1}^{m} p_j C_0(p_j x) + e^{-x} \left( 1 - \sum_{j=1}^{m} p_j \sum_{n=1}^{\infty} \frac{r_{n,0}}{(n-1)!} (p_j x)^{n-1} \right),$$

we see that the structure of the functional equation is the same as in the analysis of the mean, and that the conditions of Theorem 2 are satisfied (note for this that $0 \leq r_{n,0} \leq 1$, so that the inner sum (over $n$) can be estimated by $e^{p_j|x|}$). This also shows that the Mellin transform of the expression

$$e^{-x} \left( 1 - \sum_{j=1}^{m} p_j \sum_{n=1}^{\infty} \frac{r_{n,0}}{(n-1)!} (p_j x)^{n-1} \right)$$

converges for $\Re(s) > -1$ (note that we also have a zero at $x = 0$) and represents an entire function there. Hence our Mellin transform technique shows that $C_0(x)$ tends to a constant as $x \to \infty$ in the aperiodic case, and oscillates around a constant in the periodic case, which implies the same behavior for $r_{n,0}$. See Figure 5 for a comparison of the two cases ($m = 2$, $p_1 = \frac{1}{3}$, $p_2 = \frac{2}{3}$ vs. $m = 3$, $p_1 = p_2 = p_3 = \frac{1}{3}$).

For $k \geq 1$, one can proceed in exactly the same way; one simply notes that

$$C_1(x) = \sum_{j=1}^{m} p_j (1 - e^{-q_j x}) C_1(p_j x) + \sum_{j=1}^{m} p_j e^{-q_j x} C_0(p_j x) - e^{-x}$$

and

$$C_k(x) = \sum_{j=1}^{m} p_j (1 - e^{-q_j x}) C_k(p_j x) + \sum_{j=1}^{m} p_j e^{-q_j x} C_{k-1}(p_j x)$$

for $k > 1$, which one rewrites as

$$C_k(x) = \sum_{j=1}^{m} p_j C_k(p_j x) + \sum_{j=1}^{m} p_j e^{-q_j x} (C_{k-1}(p_j x) - C_k(p_j x)).$$

Once again, the Mellin transform of the last summand can be shown to converge for $\Re(s) > -1$, and one obtains the theorem for $k \geq 1$ in the same way as for $k = 0$. □

**Remark 1** *The distributions in Theorem 5 are only given implicitly in view of the fact that the Mellin transforms of functions defined in terms of the functions $C_k(x)$ occur. It seems difficult to obtain more explicit expressions, but numerical values for the probabilities can be computed.*

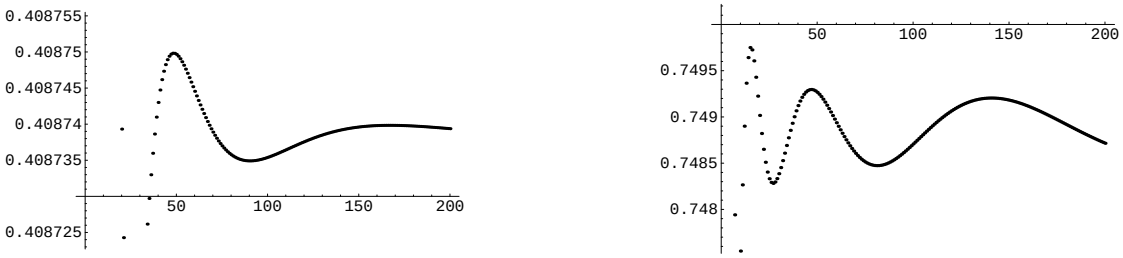


**Fig. 5:** Plots of $r_{n,0}$ in the cases $m = 2$, $p_1 = \frac{1}{3}$, $p_2 = \frac{2}{3}$ and $m = 3$, $p_1 = p_2 = p_3 = \frac{1}{3}$

# 4 Autocomplete revisited

Let us finally consider the autocomplete tool mentioned in Section 1 again: a trie is used to store a dictionary of words; as a word is entered letter by letter, the tool aims to automatically detect the next letter. If the following letter is not unique (unary node), one of the branches is selected at random, with probability proportionate to the number of words stored in the respective branch: if $k$ of the words are stored in the subtrie rooted at the current node, and $\ell$ of them in a certain branch, then the autocomplete tool selects the letter corresponding to this branch with probability $\frac{\ell}{k}$.

We analyze the average number of correctly suggested letters by this algorithm: if this number is denoted by $s_n$ for a total number of $n$ words, then we have the recursion

$$s_n = \sum_{k=1}^{n} \binom{n-1}{k-1} \sum_{j=1}^{m} p_j^k q_j^{n-k} \left( \frac{k}{n} + s_k \right)$$

for $n > 1$, with $s_1 = 0$. If we consider the (shifted) Poisson transform again, i.e.,

$$S(x) = e^{-x} \sum_{n=0}^{\infty} \frac{s_{n+1}}{n!} x^n,$$

then the methods of Section 2 yield, after some manipulations,

$$S(x) = \sum_{j=1}^{m} p_j S(p_j x) + \sum_{j=1}^{m} p_j^2 (1 - e^{-x}) + \sum_{j=1}^{m} p_j q_j \cdot \frac{1 - (x+1)e^{-x}}{x}.$$

Once again, the conditions of Theorem 2 are satisfied, and the Mellin transform of $S(x)$ is given by

$$S^*(s) = -\left( 1 - \sum_{j=1}^{m} p_j^{1-s} \right)^{-1} \left( \sum_{j=1}^{m} p_j^2 \Gamma(s) + \sum_{j=1}^{m} p_j q_j \frac{\Gamma(s+1)}{s-1} \right).$$

This Mellin transform has poles at 0 (double), 1 and all characteristic roots. The double pole at 0 gives rise to a logarithmic term that dominates the fluctuation, so that one ends up with the following result:

**Theorem 6** *The expected number of successes (correctly suggested letters) of the autocomplete tool explained above is given by*

$$\frac{\sum_{j=1}^m p_j^2}{-\sum_{j=1}^m p_j \log p_j} \log n + K + Q_7(n) + o(1),$$

*where $K$ is a constant and $Q_7(n)$ is a small oscillating function (possibly $0$).*

If one compares this to the expected depth of a randomly selected word, which is (see [1])

$$\frac{1}{-\sum_{j=1}^m p_j \log p_j} \log n + L + Q_8(n) + o(1),$$

one finds that an average proportion of $\sum_{j=1}^m p_j^2$ of the letters on the way to the desired node is correctly suggested by the autocomplete tool. Note that this is precisely the coincidence probability (the probability that two randomly generated words have the same first letter and are thus stored in the same branch).

## Acknowledgment

I would like to thank Alois Panholzer for valuable discussions and for his encouragement.

## References

[1] J. Bourdon. Size and path length of Patricia tries: dynamical sources context. *Random Structures Algorithms*, 19(3-4):289–315, 2001. Analysis of algorithms (Krynica Morska, 2000).

[2] J. I. Capetanakis. Tree algorithms for packet broadcast channels. *IEEE Trans. Inform. Theory*, 25(5):505–515, 1979.

[3] R. da la Briandais. File searching using variable length keys. In *Proceedings of the Western Joint Computer Conference*, volume 15, pages 295–298. Spartan Books, New York, 1959.

[4] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: harmonic sums. *Theoret. Comput. Sci.*, 144(1-2):3–58, 1995. Special volume on mathematical analysis of algorithms.

[5] P. Flajolet and P. Jacquet. Analytic models for tree communication protocols. In A. R. Odoni, L. Bianco, and G. Szegö, editors, *Flow Control of Congested Networks*, volume 38, pages 223–234. Springer-Verlag, 1987.

[6] E. Fredkin. Trie memory. *Commun. ACM*, 3(9):490–499, 1960.

[7] P. Jacquet and M. Régnier. Trie partitioning process: limiting distributions. In *CAAP '86 (Nice, 1986)*, volume 214 of *Lecture Notes in Comput. Sci.*, pages 196–210. Springer, Berlin, 1986.

[8] P. Jacquet and W. Szpankowski. Analytical de-Poissonization and its applications. *Theoret. Comput. Sci.*, 201(1-2):1–62, 1998.

[9] P. Kirschenhofer, H. Prodinger, and W. Szpankowski. On the variance of the external path length in a symmetric digital trie. *Discrete Appl. Math.*, 25(1-2):129–143, 1989. Combinatorics and complexity (Chicago, IL, 1987).

[10] D. E. Knuth. *The art of computer programming. Volume 3*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1973.

[11] H. M. Mahmoud and M. D. Ward. Average-case analysis of cousins in $m$-ary tries. *J. Appl. Probab.*, 45(3):888–900, 2008.

[12] G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski. Profile of tries. In *LATIN 2008: Theoretical informatics*, volume 4957 of *Lecture Notes in Comput. Sci.*, pages 1–11. Springer, Berlin, 2008.

[13] B. Rais, P. Jacquet, and W. Szpankowski. Limiting distribution for the depth in PATRICIA tries. *SIAM J. Discrete Math.*, 6(2):197–213, 1993.

[14] R. Sedgewick. *Algorithms*. Addison-Wesley Series in Computer Science. Addison-Wesley Publishing Company Advanced Book Program, Reading, MA, 1983.

[15] W. Szpankowski. *Average case analysis of algorithms on sequences*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2001.

[16] B. S. Tsybakov and V. A. Mikhaĭlov. Free synchronous packet access in a broadcast channel with feedback. *Problems Inform. Transmission*, 14(4):32–59, 1978.