# Fringe Analysis Revisited

Ricardo A. Baeza-Yates

Depto. de Ciencias de la Computación

Universidad de Chile

Blanco Encalada 2120

Santiago, Chile\*

#### Abstract

Fringe analysis is a technique used to study the average behavior of search trees. In this paper we survey the main results regarding this technique, and we improve a previous asymptotic theorem. At the same time we present new developments and applications of the theory which allow improvements in several bounds on the behavior of search trees. Our examples cover binary search trees, AVL trees, 2-3 trees, and B-trees.

Categories and Subject Descriptors: F.2.2 [Analysis of Algorithms and Problem Complexity]: Nonnumerical Algorithms and Problems – computations on discrete structures; sorting and searching; E.1 [Data Structures]; trees.

#### Contents

1	Introduction	2
2	The Theory of Fringe Analysis	4
3	Weakly Closed Collections	9
4	Including the Level Information	11
5	Fringe Analysis, Markov Chains, and Urn Processes	13

<sup>\*</sup>This work was partially funded by Research Grant FONDECYT 93-0765. e-mail: rbaeza@dcc.uchile.cl

#### 1 Introduction

Search trees are one of the most used data structures to manage dynamic data sets. To study their average-case performance, there exist several analytical techniques. Classical ones are generating functions coupled with differential equations, which use complex analysis to obtain asymptotical results. These techniques are covered in many textbooks like [Knu73, GK81, VF90, Mah92]. In this survey we present another technique, called *fringe analysis*, which is not covered by standard textbooks, and is mainly used for the average-case analysis of balanced search trees.

Fringe analysis was formally introduced by Yao in 1974 [Yao74, Yao78] and was also discovered independently by Nakamura and Mizoguchi [NM78]. However, the first fringe type analysis was done by Knuth [Knu73, Solution to problem 6.2.4.10, pages 679-680]. Fringe analysis is a method used to analyze search trees that considers only the bottom part or *fringe* of the tree. From the behavior of the subtrees on the fringe, it is possible to obtain bounds on most complexity measures of the complete tree and some exact results.

Classical fringe analysis considers only insertions, and the model used is that the n! possible permutations of the n keys used as input are equally likely. A search tree built under this model is called a random tree. This is equivalent to saying that the n-th insertion has the same probability of falling in any of the n + 1 leaves of the tree.

This technique has been applied to almost all search trees and related problems:

- variants of binary search trees [BY91, PM85, Pob93],
- 2-3 trees [Yao78, Bro79b, EZG<sup>+</sup>82, BYP85, VD76, LW79],
- AVL trees [Bro79a, Meh82, GZ82, BYGZ92],
- multiway search trees and variants [CIOW81, BY87b, CP88, BYC92],
- B-trees [Yao78, NM78, EZG<sup>+</sup>82, Wri85], its variants [BY89a, BYL89, Fre79, BY89b, CK86, BY90b, BY90a], and related operations (concurrency for example),
- symmetric Binary B-trees [ZOG85],
- 1-2 trees [OW80, OS80, Oli81],
- red-black trees [GS78],
- generalized search trees [MP84, CG87],

- bounded disorder files [Mat90, BY94], and
- external sorting [CGMP91].

The typical measures obtained are bounds in the expected number of nodes, several probabilities of events happening during an insertion, expected space utilization, etc [GBY91]. Also it is possible to obtain bounds for the height of the tree [Ziv82, EZG<sup>+</sup>82, BY85] or probabilities related to concurrent access to the tree [KW80, EZG<sup>+</sup>82, BYP85]. Table 1 shows some upper and lower bounds for the fraction of height balanced nodes and the probability of a rotation in AVL trees, and the fraction of internal nodes (with respect to all the n keys stored) and the probability of a split in 2-3 trees and B-trees of order m (internal nodes can hold between m and 2m keys). We also include the storage utilization for the case of 2-3 and B-trees. Some of these results require that n be large.

Measure $(n \to \infty)$	AVL-trees	2-3 trees	B-trees
Fraction of nodes	[0.5637, 0.7799]	[0.7377, 0.7543]	$\frac{1}{2m\ln 2} + O(m^{-2})$
Probability of rotation/split	[0.3784, 0.7261]	$[0.7212, 0.5585 + 0.03308 \log_2(n+1)]$	$\left[\frac{1}{2(m+1)\ln 2}, \ \frac{1}{m}\right]$
Storage utilization	ĺ,	[0.6629, 0.6778]	$\ln 2 + O(m^{-1})$

Table 1: Main results obtained with fringe analysis.

Some attempts to include deletions into the analysis had been made for AVL-trees [Meh82], and B-trees [Miz79, QK80, BY87a, JS89]. However, the random model does not apply after deletions, and a different model must be used.

In this review we give an unifying view to several variations of fringe analysis, and we present the complete solution for a fringe analysis type recurrence. Some theorems and proofs are included to keep the text self-contained and/or because they are contained in papers that are difficult to obtain. Fringe type recurrences can be applied to the analysis of any class of search trees, ranging from binary search trees to B-trees. The asymptotic solution shows a fast convergence of the process to the steady state. The connection between fringe analysis and urn processes shows also that the variance tends to 0 with high probability.

### 2 The Theory of Fringe Analysis

The theory of fringe analysis was formalized in Eisenbarth  $et\ al.\ [EZG^+82]$ . The fringe of the tree is defined in terms of a finite collection C of trees. A collection C is closed if the effect of an insertion only affects the subtree of the fringe in which it is performed, and produces one or more members of the same collection.



Figure 1: 2-3 tree fringe collection of height one.

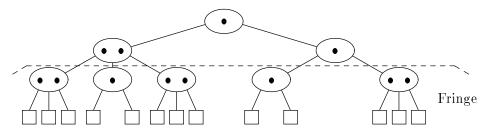


Figure 2: A 2-3 tree and its fringe of height one.

**Example 1.** Figure 1 shows a tree collection that defines a fringe of height one in a 2-3 tree, while figure 2 shows a 2-3 tree and the fringe corresponding to this collection. The composition of the fringe can be described by the number of subtrees of each type [Yao78], or by the probability that a randomly chosen leaf belongs to each member in C [EZG<sup>+</sup>82]. This process defines a Markov chain [EZG<sup>+</sup>82]. Let n be the number of keys in the tree,  $A_i(n)$  be the expected number of subtrees of type i in the fringe,  $L_i$  the number of leaves in type i, and

$$P_i(n) = \frac{A_i(n)L_i}{n+1}$$

be the probability of a leaf belonging to a subtree of type i in a random 2-3 tree with n keys.

An insertion in a type I subtree produces a type II subtree, while an insertion in a type II subtree produces two type I subtrees. So,

$$A_1(n) = A_1(n-1) - P_1(n-1) + 2P_2(n-1)$$
 and

$$A_2(n) = A_2(n-1) - P_2(n-1) + P_1(n-1)$$
.

Using only  $P_i(n)$  and matrix notation we have

$$\vec{P}(n) = \left(\mathbf{I} + \frac{1}{n+1}\mathbf{H}\right)\vec{P}(n-1)$$
,

where  $\mathbf{I}$  is the identity matrix and  $\mathbf{H}$  is called the *transition* matrix, given by

$$\mathbf{H} = \left[ \begin{array}{cc} -3 & 4 \\ 3 & -4 \end{array} \right] .$$

By replacing the condition  $P_1(n) + P_2(n) = 1$  in the above recurrence, we obtain only one recurrence in  $P_1(n)$ , namely

$$P_1(n) = P_1(n-1) + \frac{1}{n+1}(4-7P_1(n-1)), \quad P_1(1) = 1.$$

Note that if  $P_1(n-1) = 4/7$  then  $P_1(n)$  remains constant. Iterating the recurrence we see that  $P_1(6) = 4/7$ , giving  $\vec{P}(n) = [4/7, 3/7]^T$  for  $n \ge 6$  (a formal solution is given later).

In general,  $\mathbf{H} = \mathbf{H_2} - \mathbf{H_1}$  where  $\mathbf{H_2}$  represents the transitions from one type to another type, that is,  $h_{2ij}$  is the the probability that an insertion in type j produces one or more subtrees of type i times the number of leaves of type i created; and  $\mathbf{H_1} = \operatorname{diag}(L_i) + \mathbf{I}$  represents the leaves lost by each type after an insertion plus one. It is not difficult to see that  $\operatorname{det}(\mathbf{H}) = 0$ .

A fringe analysis is *connected* [EZG<sup>+</sup>82], if there exist  $\mathbf{H}_i$  such that  $det(\mathbf{H}_i) \neq 0$ , where  $\mathbf{H}_i$  is the matrix  $\mathbf{H}$  with the *i*-th row and *i*-th column deleted.

The asymptotic solution to a connected fringe analysis, as shown later, is given by

$$\vec{P}(n) = \vec{x} + O(n^{Re(\lambda_2)} \log^{m-1} n) ,$$

where  $\vec{x}$  is the solution to  $\mathbf{H}\vec{x}=0$ , normalized such that  $\sum x_i=1$ , and  $\lambda_2$  is the second largest eigenvalue in absolute value with  $Re(\lambda_2)<0$  and multiplicity m (the largest eigenvalue is always 0 with multiplicity 1) [EZG<sup>+</sup>82, BYG90].

In many cases, it is possible to use the structure of  $\mathbf{H}$  to simplify the solution of the system of equations [BYP85]. In the following paragraphs we give the exact solution of the recurrence for all n, and its asymptotic solution.

We start from the matrix recurrence equation

$$\vec{P}(n) = \left(\mathbf{I} + \frac{1}{n+1}\mathbf{H}\right)\vec{P}(n-1)$$
.

Let m be the dimension of the recurrence. Because the rank of  $\mathbf{H}$  is m-1, we include the condition that  $\sum_{i} P_i(n) = 1$ , to obtain the following equation  $(\vec{P} \text{ now has one component less})$ 

$$\vec{P}(n) = \left(\mathbf{I} + \frac{1}{n+1}\mathbf{T}\right)\vec{P}(n-1) + \frac{1}{n+1}\vec{F} ,$$

where  $\mathbf{T} = \mathbf{H}_{(m-1)\times(m-1)} - \mathbf{H}'$  where  $\mathbf{H}_{(m-1)\times(m-1)}$  denotes the principal minor of  $\mathbf{H}$ ,  $\vec{F}$  is the last column of  $\mathbf{H}$  up to the (m-1)-th row, and  $\mathbf{H}'$  is a  $(m-1)\times(m-1)$  matrix where every column is equal to  $\vec{F}$ . In the sequel of this paper  $\mathbf{X}^{\underline{n}} = (\mathbf{X} - (n-1)\mathbf{I})\cdots(\mathbf{X} - \mathbf{I})\mathbf{X}$  denotes descendent factorials over matrices.

Theorem 2.1 The solution for a connected fringe analysis of a closed collection of trees is given by

$$\vec{P}(n) = -\mathbf{T}^{-1}\vec{F} + \frac{(-1)^{n+1}}{(n+1)!}(-\mathbf{T} - \mathbf{I})^{\frac{n+1}{2}}\vec{C}$$
,

with  $\vec{C}$  obtained from the initial condition  $\vec{P}(n_0) = [1,0,0,...,0]^T$ , where  $n_0$  is the number of elements in the smallest subtree type of the fringe collection.

**Proof:** Introducing the generating function (see [Knu69, page 86])

$$\vec{P}(z) = \sum_{n>0} \vec{P}(n)z^n ,$$

in the matrix recurrence, we obtain the following first order linear differential equation

$$\frac{d\vec{P}(z)}{dz} = \left(\frac{2z-1}{z(1-z)}\mathbf{I} + \frac{1}{1-z}\mathbf{T}\right)\vec{P}(z) + \frac{1}{z(1-z)^2}\vec{F} ,$$

whose solution is

$$\vec{P}(z) = \frac{1}{z} e^{-(\mathbf{T} + \mathbf{I}) \ln(1-z)} \vec{C} - \frac{1}{z(1-z)} \mathbf{T}^{-1} \vec{F} ,$$

where  $\vec{C}$  is obtained from the initial condition.

For  $n \geq n_0$ , we have

$$\vec{P}(n) = [z^n]\vec{P}(z) = -\mathbf{T}^{-1}\vec{F} + \sum_{k>0} (-\mathbf{T} - \mathbf{I})^k [z^{n+1}] \frac{\ln^k (1-z)}{k!} \vec{C} ,$$

where  $[z^n]P(z)$  denotes the coefficient in  $z^n$  of P(z). But

$$[z^n] \frac{\ln^k (1-z)}{k!} = \frac{(-1)^n}{n!} \mathcal{S}_n^{(k)} ,$$

for  $n \geq k$  where  $\mathcal{S}_n^{(k)}$  denotes Stirling numbers of the first kind [Knu69, page 65]. Therefore, for  $n \geq n_0$ 

$$\vec{P}(n) = -\mathbf{T}^{-1}\vec{F} + \frac{(-1)^{n+1}}{(n+1)!} \sum_{k=0}^{n+1} \mathcal{S}_{n+1}^{(k)} (-\mathbf{T} - \mathbf{I})^k \vec{C} ,$$

or [Knu69, page 65]

$$\vec{P}(n) = -\mathbf{T}^{-1}\vec{F} + \frac{(-1)^{n+1}}{(n+1)!}(-\mathbf{T} - \mathbf{I})^{\frac{n+1}{2}}\vec{C}$$
.

A preliminary version of this theorem was presented in [BYG90].

**Example 2.** In Example 1, we have  $\vec{P}(1) = [1,0]$ , T = -7, F = 4 and C = 1/35, obtaining

$$P(z) = \frac{4}{7z(1-z)} + \frac{(1-z)^6}{35z} ,$$

or

$$P_1(n) = \frac{4}{7} - \frac{(-1)^n}{35} \binom{6}{n+1}$$
.

Note that the second term is 0 for n > 5.

The next step is to obtain the asymptotic behavior of the solution. Let  $\mathbf{R} = -\mathbf{T} - \mathbf{I}$ . We want the asymptotic value of

$$\vec{\epsilon}(n) = \frac{(-1)^{n+1}}{(n+1)!} \sum_{k=0}^{n+1} \mathcal{S}_{n+1}^{(k)} \mathbf{R}^k \vec{C}$$

Theorem 2.2 Asymptotically, every component of  $\vec{\epsilon}(n)$  is

$$O(n^{Re(\lambda_1)} \log^{m-1} n)$$
,

where  $\lambda_1$  is the eigenvalue of T with largest real component, and m is its multiplicity. Note that  $\lambda_1$  is equal to the second largest eigenvalue of H and that  $Re(\lambda_1) < 0$ .

**Proof:** We decompose **R** in its upper normal Jordan form [Gan59, Vol. 1, pages 152, 200-202]

$$\mathbf{R} = \mathbf{P} \ \mathbf{J} \ \mathbf{P}^{-1}$$

with J being an upper-diagonal matrix obtained from the eigenvalues of R. Then, we obtain

$$\vec{\epsilon}(n) = \mathbf{P} \frac{(-1)^{n+1}}{(n+1)!} \sum_{k=0}^{n+1} S_{n+1}^{(k)} \mathbf{J}^k \mathbf{P}^{-1} \vec{C}$$
.

Therefore, we have summations of the form

$$M_j = \frac{(-1)^{n+1}}{(n+1)!} \sum_{k=0}^{n+1} \mathcal{S}_{n+1}^{(k)} \binom{k}{j} \lambda^{k-j} .$$

By using a well known combinatorial relation [Knu69, page 65, equation 40] we have

$$M_0 = (-1)^{n+1} \binom{\lambda}{n+1} .$$

From Eisenbarth *et al.* [EZG<sup>+</sup>82] we know that  $Re(\lambda) > 0$  for  $\lambda$  an eigenvalue of  $-\mathbf{T}$ . If  $\lambda$  is an integer we have  $M_0 = 0$  for  $n \geq \lambda$ . If  $\lambda$  is not an integer (that is,  $\lambda$  is real or complex), we have [Knu69, page 15]

$$M_0 = (-1)^{n+1} \binom{\lambda}{n+1} = \binom{n-\lambda-1}{-\lambda-1} = \frac{\Gamma(n-\lambda)}{\Gamma(-\lambda)\Gamma(n+1)}$$
.

Therefore, asymptotically, we have

$$M_0 = \frac{n^{-(\lambda+1)}}{\Gamma(-\lambda)} + O(n^{-\lambda-2})$$

(see [AS72, page 253] for details about the  $\Gamma$  function).

Analogously, we obtain

$$M_j = ((\Psi(n-\lambda) - \Psi(-\lambda))^j + O(1/n)) M_0 = O(M_0 \log^j n),$$

where  $\Psi(x) = \frac{d}{dx}(\ln \Gamma(x)) = \ln n + O(1)$  [AS72, page 259].

But if  $\lambda$  is an eigenvalue of  $\mathbf{T}$ , then  $-(\lambda + 1)$  is an eigenvalue of  $-\mathbf{T} - \mathbf{I}$ . Then, we can state that every component of  $\vec{\epsilon}(n)$  is

$$O(n^{Re(\lambda_1)} \log^{m-1} n)$$

where  $\lambda_1$  is the eigenvalue of **T** with largest real component, and m is its multiplicity.

In all the analyses that appear in the literature, the multiplicity of the second eigenvalue is 1. However, in general this may be not true. For example, the following transition matrix has  $\lambda_2 = \lambda_3 = -3$ , and

$$\mathbf{H} = \begin{bmatrix} -1 & 1 & 2 \\ 1 & -2 & 1 \\ 0 & 1 & -3 \end{bmatrix}$$

with the steady state solution  $\vec{P} = [5/9, 1/3, 1/9]^T$ .

**Example 3.** For the second order analysis of 2-3 trees [Yao78, EZG<sup>+</sup>82] (a seven type collection), the second eigenvalue of **H** is -6.55 + 6.25i, and then, the order of each component in  $\vec{\epsilon}(n)$  is proportional to

$$99.01 \cos(6.25 \ln n) n^{-6.55} + O(n^{-7.55})$$
.

Note that the periodic term has logarithmic period (complex part of  $\lambda_2$ ).

In many cases it might be difficult to obtain the second eigenvalue exactly, if the order of the matrix is bigger than 4. It might even be difficult to obtain it numerically for a large matrix. In those cases, we can use theorems that bound the eigenvalues of a matrix (for example, Gerschgorin's theorem [Wil65, Chapter 2, S13]). In our example, from the **T** matrix we can obtain  $Re(\lambda_1) \leq -2$ , which is far from -6.55, but good enough for asymptotic purposes. Another application of these bounding techniques can be found in [BYG89].

### 3 Weakly Closed Collections

For some classes of balanced trees, an event on the fringe may depend on events happening outside the fringe. In that case unknown probabilities are introduced to handle this problem. For example, in AVL trees, the composition of the fringe may change due to rotations involving nodes outside the fringe [Meh82, GZ82, BYGZ92].

The following theorem extends Theorem 2.1 to the case of having unknown probabilities [BYGZ92].

Theorem 3.1 Let  $t_n^i$  be m unknown probabilities. The solution to a recurrence of the form

$$\vec{P}(n) = (\mathbf{I} + \frac{1}{n+1}\mathbf{T})\vec{P}(n-1) + \frac{1}{n+1}\vec{F} + \frac{1}{n+1}\sum_{i=1}^{m} t_{n-1}^{i}\vec{G}_{i} ,$$

where  $\vec{F}$  and  $\vec{G}_i$  (i = 1..m) are constant vectors, is

$$\vec{P}(n) = -\mathbf{T}^{-1}\vec{F} + \frac{(-1)^{n+1}}{(n+1)!}\mathbf{R}^{\frac{n+1}{2}}\vec{C} + \sum_{i=1}^{m} \sum_{k>n_0}^{n-1} (-1)^{n-k} t_k^i \sum_{j=0}^{n-k-1} \frac{\mathbf{R}^j \mathbf{T}^{\frac{n-k-j-1}{2}}}{(n-j+1)(n-k-j-1)!j!} \vec{G}_i$$

where  $\mathbf{R} = -\mathbf{T} - \mathbf{I}$ , and  $\vec{C}$  is obtained from the initial condition  $\vec{P}(n_0) = [1, 0, 0, ..., 0]^T$ , where  $n_0$  is the number of elements in the smallest subtree type of the fringe collection.

If  $t_n^i \leq t^i$  for all  $n \geq n_1$ , we have

$$\vec{P}(n) \le -\mathbf{T}^{-1} \left( \vec{F} + \sum_{i=1}^{m} t^{i} \vec{G}_{i} \right) + O(1/n)$$

for some  $n_1 > n_0$ . A similar relation holds if  $t_n^i \geq t^i$ .

**Example 4.** The smallest weakly AVL tree collection involves three types and one unknown probability [GZ82, BYGZ92] (figure 3). This collection is *non-ambiguous*, that is, a subtree is a type I subtree only if does not belong to a type II or a type III subtree. Otherwise, it is possible to show that this collection is closed if it is an ambiguous collection [GZ82].

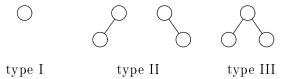


Figure 3: Weakly closed AVL tree collection.

The composition of the fringe is changed due to a rotation outside the fringe if an insertion happens in one of the leaves shown in figure 4 with dashed lines.

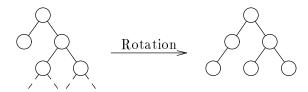


Figure 4: Insertions that depend on an unknown probability.

The recurrence equation for this case is [BYGZ92]

$$\vec{P}(n) = \left(\mathbf{I} + \frac{1}{n+1} \begin{bmatrix} -3 & 0 & 2\\ 3 & -4 & 3\\ 0 & 3 & -5 \end{bmatrix}\right) \vec{P}(n-1) + \frac{t_{n-1}}{n+1} \begin{bmatrix} -4\\ 0\\ 4 \end{bmatrix}.$$

From Brown's result [Bro79b] we have that  $P_2(n) = 3/7$  and  $P_1(n) + P_3(n) = 4/7$  for n > 5, reducing the problem to a recurrence in only one variable [BYGZ92]:

$$P_1(n) = P_1(n-1)\left(1 - \frac{5}{n+1}\right) + \frac{8}{7(n+1)} - \frac{4t_{n-1}}{n+1}$$

with  $P_1(6) = 0$ . The solution is, for n > 5,

$$P_1(n) = \frac{8}{35} - 4 \sum_{j>4}^{n-1} \frac{(j+1)^{\frac{3}{2}}}{(n+1)^{\frac{5}{2}}} t_j.$$

If  $t_n$  converges for large n, we obtain

$$P_1(n) = \frac{8}{35} - \frac{4t_{\infty}}{5} + O(1/n)$$
.

Note that with a bigger fringe collection, we can obtain  $t_n$ . However, other unknown probabilities appear [GZ82, BYGZ92]. If we can prove that there is a fringe collection without unknown probabilities, capturing the subtrees of  $t_n$ , or a fringe collection in where  $t_n$  does not depend in other probability, that would imply that  $t_n$  converges. However, it seems that this does not happen, and there is always an event above the fringe collection that modifies the fringe. Therefore, it may be possible that  $t_n$  oscillates around an average value.

### 4 Including the Level Information

As in Knuth's analysis [Knu73, problem 6.2.4.10], it is possible to include information about the level of a leaf. This is formalized in Poblete and Munro to study locally balanced search trees [PM85], and has been used to study binary search trees [BY91, CG87] and multiway trees [BY87b, CP88, BYC92].

In this case, we define  $P_i(n, k)$  as the average number of leaves belonging to subtrees of type i at level k in a tree with n elements.

Again, the insertion process defines a set of recurrences. Introducing the generating function  $P_i(n,y) = \sum_k P_i(n,k) y^k$  the following matrix recurrence is obtained:

$$\vec{P}(n,y) = \left(\mathbf{I} + \frac{1}{n+1}\mathbf{H}(y)\right)\vec{P}(n-1,y)$$

with the initial condition  $\vec{P}(n_0, y) = \vec{F}(y)$ .

As before, the solution is

$$\vec{P}(n,y) = \frac{(-1)^{n+1}}{(n+1)!} (-\mathbf{H}(y) - \mathbf{I})^{\frac{n+1}{2}} \vec{D}(y)$$

with  $\vec{D}(y)$  obtained from the initial condition.

Note that  $\vec{P}(n,1)$  gives the probability of finding a leaf of each type. The matrix  $\mathbf{H} = \mathbf{H}(1)$  has 0 as maximal eigenvalue with multiplicity 1, and all the others have negative real part [EZG<sup>+</sup>82, PM85]. Then,  $\vec{P}(n,1)$  has the same asymptotic solution as  $\vec{P}(n)$ .

More interesting is how to extract information from  $\vec{P}(n,y)$  about the external path length of the tree. Let  $C_{n,k}$  be the probability that k comparisons are needed for an unsuccessful search in a tree with n keys. Clearly,

$$C_{n,k} = \sum_{i} P_i(n,k) .$$

Introducing the generating function

$$C_n(y) = \sum_k C_{n,k} y^k ,$$

we have

$$C_n(y) = \vec{P}(n,y) \cdot \vec{1}$$

where  $\cdot$  denotes the dot product of two vectors and  $\vec{1}$  is a vector with all its elements equal to 1.

The expected number of comparisons for an unsuccessful search is

$$\overline{C}_n = \sum_k k \ C_{n,k} = \frac{\mathrm{d}}{\mathrm{dy}}(C_n(y)) |_{y=1} = C'_n(1) \ .$$

Following Theorem 2.1, we have that the higher order term is given by  $\lambda_1(1)$ , with multiplicity 1. Then, ignoring lower order terms

$$\vec{P}(n,y) = \frac{\Gamma(n-\lambda_1(y)+1)}{\Gamma(-\lambda_1(y))(n+1)!}\vec{x}(y)$$

where  $\vec{x}(y)$  is the solution of  $\mathbf{H}(y)\vec{x}(y)$ , such that  $\sum_i x_i(1) = 1$ . Then

$$\vec{C}_n(y) = \frac{\Gamma(n - \lambda_1(y) + 1)}{\Gamma(-\lambda_1(y))(n+1)!}.$$

Taking the derivative with respect to y, and evaluating in y = 1 we obtain

$$\overline{C}_n = \lambda_1'(1) \ H_{n+1} + O(1) \sim \lambda_1'(1) \ln n$$

where  $H_n = \sum_{i=1}^n 1/i$  denotes the harmonic numbers. This result was obtained in a different way by Poblete and Munro [PM85]. Recently, Poblete [Pob93] extended the result shown above to a general multidimensional case. That is, given a random variable  $\vec{X}(n)$ , with generating function

$$\vec{X}(n,y) = \vec{c}(y)\vec{A}(n,y)$$

where  $\vec{A}(n,y)$  is the generating function of the number of subtrees of each type on the fringe, and  $\vec{c}(y)$  is a nonzero row vector of polynomials in y with non-negative coefficients. Then, if  $\mathbf{H}(y)$  is the transition matrix of the insertion process, and  $\mathbf{H}(y)$  has a dominant eigenvalue  $\lambda_1(y)$  at y=1, and  $\lambda_1(y) \neq -1, -2, \cdots$  when  $y \to 1$ , we have

$$\overline{X}(n) = \lambda_1'(1)\Psi(n + \lambda_1(1)) + O(1) .$$

A similar result holds for the variance, and it is shown that the probability distribution of  $\vec{X}(n)$  converges weakly to the normal distribution.



Figure 5: Local balancing in a binary search tree

**Example 5.** The simplest local balance restructuring in a binary search tree is to perform a rotation whenever a son is appended to a node without a brother [PM85]. This case is depicted in figure 5, where a rotation is performed if an insertion falls on one of the shaded leaves.

There are two types of subtrees (see figure 6) that define three types of leaves, which we denote A, B and C.

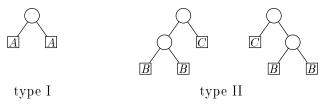


Figure 6: Type of subtrees

Noting that the number of leaves of type C at level k is half the number of leaves of type B at level k+1, we can express the problem only as a function of leaves of type A and B. The matrix recurrence in the generating functions is then [PM85]  $(P_1(n, y) = A(n, y), P_2(n, y) = B(n, y))$ 

$$\vec{P}(n,y) = \left(\mathbf{I} + \frac{1}{n+1} \begin{bmatrix} -3 & 6 \\ 2y & 4 \end{bmatrix}\right) \vec{P}(n-1,y)$$

with initial condition  $\vec{P}(1, y) = [y, 0]^T$ .

The eigenvalues of  $\mathbf{H}(y)$  are  $-7/2 \pm \sqrt{1+48y}/2$ . Computing the derivative and evaluating in y=1, we have

$$\overline{C}_n = \frac{12}{7} H_{n+1} + O(1) \sim 1.188 \log_2 n$$
.

This should be compared with  $1.386 \log_2 n$  for random binary search trees and  $\log_2 n$  for completely balanced binary trees.

## 5 Fringe Analysis, Markov Chains, and Urn Processes

Although it is not explicit, a fringe analysis recurrence is equivalent to a Markov chain, and the final expected probability vector can be considered as the steady state solution of a finite Markov chain

without absorbing states [KS83]. For that reason, it is not surprising to find that the probability distributions involved converge to normal laws. On the same vein, fringe analysis was connected to urn processes, when Bagchi and Pal [BP85] proved that the two type collection of Example 1 can be described as an urn process which converges to a normal distribution with variance converging to 0 as n approaches infinity.

Aldous et al. [AFP88] formalized the relation between fringe analysis and urn processes. This relation holds provided that the matrix  $\mathbf{H_2}$  is irreducible. This is the case for any reasonable insertion algorithm, due to the cyclic nature of the process, and is equivalent to the notion of connected fringe analysis.

They state the following theorem (in a slightly different way)

Theorem **5.1** In a connected fringe analysis,  $\vec{P}(n) \to \vec{x}$  almost surely as  $n \to \infty$ , where  $\vec{x}$  is the solution of

$$\mathbf{H}\vec{x} = 0$$

normalized such that  $\sum_i x_i = 1$ .

That is, all previous bounds obtained for the expected value of a measure depending linearly on  $\vec{x}$ , are also valid for the measure itself as a random variable. Further results on this line are given in [Gou89].

**Example 6.** From the above theorem, we have now that  $E(1/x) \to 1/E(x)$  as  $n \to \infty$  for a random variable x. This fact allows us to improve the bounds for the expected storage utilization in B-trees [EZG<sup>+</sup>82]. So, for a B-tree of order m, we can now state

$$\ln 2 - \frac{2\ln 2 - 1}{4m} + O(m^{-2}) \le \overline{U} \le \ln 2 + \frac{1}{4m} + O(m^{-2}).$$

The approximate value of  $2 \ln 2 - 1$  is 0.3863.

Using an overflow technique in the last level, that splits a node only when itself and its brothers are full, we obtain

$$\overline{U} = 1 - \left(\frac{3}{\ln 2} - 2\right) \frac{1}{4m} + O(m^{-2}).$$

Other improvements of this kind are given in [Pal90].

### Acknowledgments

We thank the helpful comments of the referees and the editor.

#### References

- [AFP88] D. Aldous, B. Flannery, and J.L. Palacios. Two applications of urn processes: The fringe analysis of search trees and the simulation of quasi-stationary distributions of Markov chains. Probability in the Enginnering and Informational Sciences, 2:293–307, 1988.
- [AS72] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1972.
- [BP85] A. Bagchi and A.K. Pal. Asymptotic normality in the generalized Polya-Eggenberger urn model, with an application to computer data structures. SIAM J. of Algebraic and Discrete Methods, 6:394-405, 1985.
- [Bro79a] M.R. Brown. A partial analysis of random height-balanced trees. SIAM J on Computing, 8(1):33-41, Feb 1979.
- [Bro79b] M.R. Brown. Some observations on random 2-3 trees. *Inf. Proc. Letters*, 9(2):57–59, Aug 1979.
- [BY85] R.A. Baeza-Yates. Analysis of algorithms in search trees. Master's thesis, Dept. of Computer Science, University of Chile, Santiago, Chile, January 1985. (in Spanish).
- [BY87a] R.A. Baeza-Yates. Analyzing deletions in B-trees. Dept. of Computer Science, Univ. of Waterloo (unpublished manuscript), 1987.
- [BY87b] R.A. Baeza-Yates. Some average measures in m-ary search trees. Inf. Proc. Letters, 25:375–381, July 1987.
- [BY89a] R.A. Baeza-Yates. Expected behaviour of B<sup>+</sup>-trees under random insertions. Acta Informatica, 26(5):439–472, 1989.
- [BY89b] R.A. Baeza-Yates. Modeling splits in file structures. *Acta Informatica*, 26(4):349–362, 1989.
- [BY90a] R.A. Baeza-Yates. An adaptive overflow technique for the B-tree. In F. Bancilhon, C. Thanos, and D. Tsichritzis, editors, Extending Data Base Technology Conference (EDBT 90), pages 16-28, Venice, March 1990. Springer Verlag Lecture Notes in Computer Science 416.

- [BY90b] R.A. Baeza-Yates. A storage allocation algorithm suitable for file structures. *Information Systems*, 15(5):515–521, 1990.
- [BY91] R. Baeza-Yates. Height balance distribution of search trees. *Information Processing Letters*, 39(6):317–324, 1991.
- [BY94] R.A. Baeza-Yates. Analysis of bounded disorder. In B. Rovan I. Privara and P. Ruzicka, editors, 19th MFCS'94, LNCS 841, pages 233–244, Kosice, Slovakia, August 1994. Springer Verlag.
- [BYC92] R.A. Baeza-Yates and W. Cunto. Unbalanced multiway trees improved by partial expansions. *Acta Informatica*, 29(5):443-460, 1992.
- [BYG89] R. Baeza-Yates and G.H. Gonnet. Efficient text searching of regular expressions. In G. Ausiello, M. Dezani-Ciancaglini, and S. Ronchi Della Rocca, editors, ICALP'89, Lecture Notes in Computer Science 372, pages 46-62, Stresa, Italy, July 1989. Springer-Verlag.
- [BYG90] R.A. Baeza-Yates and G.H. Gonnet. Average case analysis of algorithms using matrix recurrences. In 2nd International Conference on Computing and Information, ICCI'90, pages 47–51, Niagara Falls, Canada, May 1990.
- [BYGZ92] R. Baeza-Yates, G.H. Gonnet, and N. Ziviani. Improved bounds for the expected behaviour of AVL trees. *BIT*, 32(2):297–315, 1992.
- [BYL89] R.A. Baeza-Yates and P-Å. Larson. Performance of B<sup>+</sup>-trees with partial expansions.

  IEEE Trans. on Knowledge and Data Engineering, 1:248–257, June 1989.
- [BYP85] R.A. Baeza-Yates and P.V. Poblete. Reduction of the transition matrix of a fringe analysis and its application to the analysis of 2-3 trees. In 5th International Conference of the Chilean Computer Science Society, pages 56–82, Santiago, Chile, 1985. (English version to appear in Foundations of Computer Science).
- [CG87] Walter Cunto and J.L. Gascon. Improving time and space efficiency in generalized binary search trees. *Acta Informatica*, 24:583–594, 1987.
- [CGMP91] Walter Cunto, Gaston H. Gonnet, J. Ian Munro, and P.V. Poblete. Fringe analysis for extquick: An in situ distributive external sorting algorithm. *Information and Computation*, 92(2):141–160, Jun 1991.

- [CIOW81] K. Culik II, Thomas Ottmann, and Derick Wood. Dense multiway trees. *ACM TODS*, 6(3):486–512, Sep 1981.
- [CK86] J-H. Chu and Gary D. Knott. An analysis of B-trees and their variants. Technical Report CS-TR-1737, University of Maryland, Nov 1986.
- [CP88] W. Cunto and P. Poblete. Transforming unbalanced multiway trees into a practical external data structure. *Acta Informatica*, 26(3):193–211, 1988.
- [EZG<sup>+</sup>82] B. Eisenbarth, N. Ziviani, Gaston H. Gonnet, Kurt Mehlhorn, and Derick Wood. The theory of fringe analysis and its application to 2-3 trees and B-trees. *Information and Control*, 55(1):125–174, Oct 1982.
- [Fre79] G.N. Frederickson. Improving storage utilization in balanced trees. In *Allerton Conference*, pages 255–264, Monticello, IL, 1979.
- [Gan59] F.R. Gantmacher. The Theory of Matrices (2 Vols). Chelsea Publishing Company, New York, 1959.
- [GBY91] G.H. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures In Pascal and C.* Addison-Wesley, Wokingham, UK, 1991. (second edition).
- [GK81] D. Greene and D. Knuth. Mathematics for the Analysis of Algorithms (2nd ed.). Birkhauser, Boston, 1981.
- [Gou89] R. Gouet. A martingale approach to strong convergence in a generalized Pólya-Eggenberger urn model. Statistics & Probability Letters, 8:225–228, Aug 1989.
- [GS78] L.J. Guibas and Robert Sedgewick. A dichromatic framework for balanced trees. In FOCS, volume 19, pages 8–21, Ann Arbor MI, Oct 1978.
- [GZ82] G.H. Gonnet and N. Ziviani. Expected behaviour analysis of AVL trees. Technical Report CS-82-18, Department of Computer Science, University of Waterloo, 1982.
- [JS89] T. Johnson and D. Shasha. Utilization of B-trees with inserts, deletes and modifies. In *PODS'89*, pages 235–246, Philadelphia, March 1989.
- [Knu69] D.E. Knuth. The Art of Computer Programming: Fundamental Algorithms, volume 1. Addison-Wesley, Reading, Mass., 1969.

- [Knu73] D.E. Knuth. The Art of Computer Programming: Sorting and Searching, volume 3. Addison-Wesley, Reading, Mass., 1973.
- [KS83] J.G. Kemeny and J.L. Snell. Finite Markov Chains. Springer-Verlag, New York, 1983.
- [KW80] Y.S. Kwong and D. Wood. Approaches to Concurrency in B-Trees, pages 402–413.
  Springer-Verlag, 1980.
- [LW79] J.A. Larson and W.E. Walden. Comparing insertion schemes used to update 3-2 trees. Inform. Systems, 4:127–136, 1979.
- [Mah92] Hosam Mahmoud. Evolution of Random Search Trees. John Wiley, New York, 1992.
- [Mat90] G. Matsliach. Performance analysis of file organizations that use multi-bucket data leaves. *Information Processing Letters*, 36:301–310, Dec 1990.
- [Meh82] Kurt Mehlhorn. A partial analysis of height-balanced trees under random insertions and deletions. SIAM J on Computing, 11(4):748-760, Nov 1982.
- [Miz79] T. Mizoguchi. On required space for random split trees. In *Allerton Conference*, pages 265–273, Monticello, IL, 1979.
- [MP84] J. Ian Munro and Patricio V. Poblete. Fault tolerance and storage reduction in binary search trees. *Information and Control*, 62(2-3):210–218, Aug 1984.
- [NM78] T. Nakamura and T. Mizoguchi. An analysis of storage utilization factor in block split data structuring scheme. In *VLDB*, volume 4, pages 489–495, Berlin, Sep 1978.
- [Oli81] H.J. Olivie. On random son-trees. Int. J Computer Math, 9:287–303, 1981.
- [OS80] Thomas Ottmann and W. Stucky. Higher order analysis of random 1-2 brother trees. BIT, 20(3):302-314, 1980.
- [OW80] Thomas Ottmann and Derick Wood. 1-2 brother trees or AVL trees revisited. Computer Journal, 23(3):248-255, Aug 1980.
- [Pal90] J.L. Palacios. Pointwise storage performance of m-ary trees, B-trees and generalized binary search trees. New Jersey Institute of Technology, 1990.
- [PM85] P.V. Poblete and J.I. Munro. The analysis of a fringe heuristic for binary search trees.

  Journal of Algorithms, 6:336–350, 1985.

- [Pob93] P.V. Poblete. The analysis of heuristics for search trees. *Acta Informatica*, 30(3):233–248, 1993.
- [QK80] Konrad H. Quitzow and Manfred R. Klopprogge. Space utilization and access path length in B-trees. *Inform. Systems*, 5:7–16, 1980.
- [VD76] J.R. Van Doren. An asymptotic analysis of minimum order B-trees. Dept. of Computing and Information Sciences, Oklahoma State University, Feb 1976.
- [VF90] J.S. Vitter and P. Flajolet. Average-case analysis of algorithms and data structures. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science (volume A)*, chapter 9, pages 431–524. Elsevier and MIT Press, Amsterdam/Cambridge, 1990.
- [Wil65] J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, London, 1965.
- [Wri85] W.E. Wright. Some average performance measures for the B-tree. *Acta Informatica*, 21:541–557, 1985.
- [Yao74] A.C-C. Yao. On random 3-2 trees. Technical Report UIUCDCS-R-74-679, Departament of Computer Science, University of Illinois at Urbana, Oct 1974.
- [Yao78] A.C-C. Yao. On random 2-3 trees. Acta Informatica, 9(2):159-170, 1978.
- [Ziv82] N. Ziviani. The Fringe Analysis of Search Trees. PhD thesis, Department of Computer Science, University of Waterloo, 1982.
- [ZOG85] N. Ziviani, H.J. Olivie, and Gaston H. Gonnet. The analysis of an improved symmetric binary B-tree algorithm. *Computer Journal*, 28(4):417–425, Aug 1985.