BIT 32 (1992), 297–315.

IMPROVED BOUNDS FOR THE EXPECTED BEHAVIOUR OF AVL TREES

RICARDO BAEZA-YATES, GASTON H. GONNET and NIVIO ZIVIANI

Depto. de Cs. de la Computación, Universidad de Chile, Santiago, Casilla 2777, Chile.
Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.
Depto. de Ciencia de Computação, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

Abstract.

In this paper we improve previous bounds on expected measures of AVL trees by using fringe analysis. A new way of handling larger tree collections that are not closed is presented. An inherent difficulty posed by the transformations necessary to keep the AVL tree balanced makes its analysis difficult when using fringe analysis methods. We derive a technique to cope with this difficulty obtaining the exact solution for fringe parameters even when unknown probabilities are involved. We show that the probability of a rotation in an insertion is between 0.37 and 0.73 (and seems to be less than 0.56), that the fraction of balanced nodes is between 0.56 and 0.78, and that the expected number of comparisons in a search seems to be at most 12% more than in the complete balanced tree.

CR categories: E.1, F.2.2, G.2.1.

1. Introduction.

Balanced tree structures are efficient ways of storing information. They provide an excellent solution for the dictionary data structure problem. For N elements the operations *find*, *insert*, and *delete* can be done in $O(\log N)$ units of time. The most popular, for main memory, are AVL trees (also called Height Balanced trees).

AVL trees were introduced by Adel'son-Vel'skii and Landis in 1962 [1]. A binary search tree is AVL if the height of the subtrees at each node differ by at most one. A balance field in each node can indicate this with two bits: $+1$, higher right subtree; 0, equal heights; -1, higher left subtree.

The first valuable attempt to analyze a balanced search tree in the average case was performed by Yao [2]. In this work Yao presented a method which he used to

The work of the first author was also supported by the Institute for Computer Research of the University of Waterloo, the second author by a Natural Sciences and Engineering Research Council of Canada Grant No. A-3353, and the third by a Brazilian Coordenação do Aperfeiçoamento de Pessoal de Nível Superior Contract No. 4799/77 and by the University of Waterloo.

Received January 1991. Revised December 1991.

obtain a partial analysis of 2–3 trees and B-trees. The method used by Yao was later used by Brown [3] to obtain a partial analysis of AVL trees. In his analysis Brown considered the collection of AVL subtrees with three or less leaves and called it the fringe of the AVL tree. By analyzing the fringe of large AVL trees Brown was able to derive bounds on the expected number of balanced nodes in the whole tree.

An improvement on Brown's results for AVL trees was obtained by Mehlhorn [4], through the study of 1–2 brother trees [5]. The main technical contribution of Mehlhorn's paper is a method for analyzing the behaviour of 1–2 brother tree schemes where the rebalancing operations require knowledge about the brother of a node. Using the close relationship between 1–2 brother trees and AVL trees [6], Mehlhorn was able to improve the bounds on the expected number of balanced nodes in AVL trees. Also, Mehlhorn [7] presented a fringe analysis of AVL trees under random insertions and deletions.

Consider an AVL tree T with N keys and consequently $N + 1$ external nodes. These N keys divide all possible key values into $N + 1$ intervals. An insertion into T is said to be a *random insertion* if it has an equal probability of being in any of the $N + 1$ intervals defined above. A *random AVL tree* with N keys is an AVL tree constructed by making N successive random insertions into an initially empty tree. In this paper we assume that all trees are random trees.

We now define certain complexity measures:

- Let $\bar{b}(N)$ be the expected number of balanced nodes in an AVL tree after the random insertion of N keys into an initially empty tree;
- Let $\bar{r}(N)$ be the expected number of rotations required during the insertion of the $(N + 1)$st key into a random AVL tree with N keys;
- Let $\bar{C}(N)$ be the expected number of comparisons in a successful search in an AVL tree with N keys;
- Let $\bar{f}(N)$ be the expected number of nodes in the fringe of an AVL tree after the random insertion of N keys into an initially empty tree;
- Let $\bar{h}(N)$ be the expected height of a random AVL tree with N elements;
- Let $\bar{c}(N)$ be the expected number of balance changes in an AVL tree during the N random insertions into an initially empty tree.

Table 1 shows the summary of our main results related to AVL trees.

2. Fringe analysis.

In this section we survey briefly fringe analysis, and we formalize fringe properties to deal with AVL-tree collections. We do so by following an example related to AVL trees.

Let us define a tree collection as a finite collection $\mathscr{C} = \{T_1, \ldots, T_m\}$ of trees. The collection of AVL trees with three leaves or fewer forms a tree collection, as shown in Figure 1.

Table 1. *Tree collection of AVL trees with three leaves or fewer.*

Tree Collection Size	Kind	$\bar{f}(N)$	$\bar{r}(N)$	$\bar{b}(N)/N$	$\bar{c}(N)/N$
2 [3]	closed	$0.57N$		[0.47, 0.86]	
3 (Ours)	closed ambiguous	$0.66N$	[0.29, 0.86]	[0.51, $0.86N$]	[1.43, 2.34]
3	weakly-closed ambiguous	[$0{,}66N$, $0.69N$]	[0.29, 0.77] [8]	[0.51, 0.81] [4]	[1.47, 2.25] [8]
4 (Ours)	weakly-closed ambiguous	[$0.747N$, $0.753N$]	[0.37, 0.73]	[0.56, 0.78]	[1.59, 2.15]

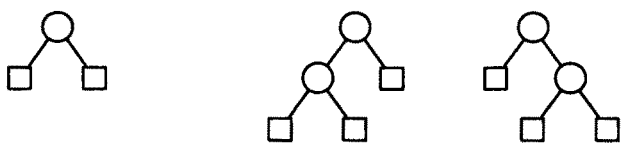

Fig. 1. Tree collection of AVL trees with three leaves or fewer.

The fringe of a tree consists of one or more subtrees that are isomorphic to members of a tree collection $\mathscr{C}$. Typically, the fringe will contain all subtrees that meet this definition; for example the fringe of an AVL tree that corresponds to the tree collection of Figure 1 is obtained by deleting all nodes at a distance greater than 2 from the leaves. Figure 2 shows an instance of an AVL tree with eleven keys in which the fringe that corresponds to the tree collection of Figure 1 is delimited by a dashed line.

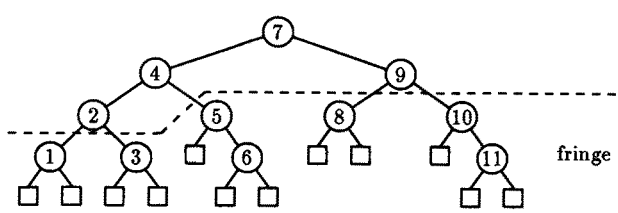

Fig. 2. An AVL tree and its fringe.

The composition of the fringe can be described in several ways. One possible way is to consider the probability that a randomly chosen leaf of the tree belongs to each

of the members of the corresponding tree collection [9]. In other words, the probability p is

$$(1) \qquad p_i(N) = \frac{\textit{Expected number of leaves of type } i \textit{ in an } N\textit{-key tree}}{N+1}.$$

We now introduce some concepts about the fringes of search trees.

DEFINITION. A tree collection $\mathscr{C} = \{T_1, \ldots, T_m\}$ is *weakly-closed* if for all $j \in [1, \ldots, m]$ an insertion into T_j always leads to one or more T_i, $i \in [1, \ldots, m]$.

DEFINITION. A tree collection $\mathscr{C}$ is *closed* when (i) $\mathscr{C}$ is weakly-closed and (ii) the effect of an insertion on the composition of the fringe is determined only by the subtree of the fringe where the insertion is performed.

The tree collection of Figure 1 is an example of a closed tree collection (proved by Brown [3]). On the other hand the collection of AVL trees with more than 2 and fewer than 6 leaves (see Figure 7) is not closed. This is because an insertion into a type 2 tree of Figure 10, when the type 2 tree is part of the fringe of an AVL tree, may cause a rotation higher in the tree, and the composition of the fringe depends on this rotation at the higher level. Figure 3 shows an instance of an AVL tree where an insertion into a type 2 tree does not lead to a type 3 tree as expected.

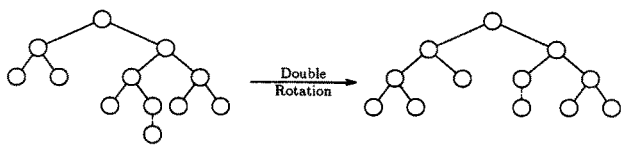

Fig. 3. Example of an insertion that unexpectedly changes the fringe of an AVL tree. (Dashed edge shows the point of insertion).

DEFINITION. A tree collection $\mathscr{C}$ is *ambiguous* when a tree in $\mathscr{C}$ appears as a subtree of another tree in $\mathscr{C}$. Figure 4 shows an AVL tree collection that is ambiguous, since a tree of type 1 is a subtree of trees of type 3.

DEFINITION. A fringe defined by a tree collection $\mathscr{C}$ and a set of transitions between members of $\mathscr{C}$ is *ambiguous* if it is not possible to determine which subtrees belongs to the fringe.

DEFINITION. A tree collection $\mathscr{C}$ is *open* if it is not weakly-closed.

The transitions between trees of a tree collection can be used to model the insertion process. In an insertion of a key into the type 1 tree shown in Figure 1 two

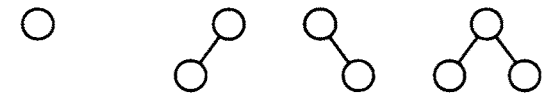

Fig. 4. Tree collection of AVL trees with more than 1 and less then 5 leaves. (leaves not shown).

leaves of type 1 are lost and three leaves of type 2 are obtained. In an insertion of a key into the type 2 tree three leaves of the type 2 are lost and four leaves of the type 1 tree are obtained.

To improve the results obtained by Brown we need larger tree collections. A new tree collection for AVL trees composed of trees with four leaves or fewer is shown in Figure 4. In Brown's paper, the fringe is defined statically; from the tree itself we can decide how to break it up into the different subtree types without knowing the sequence of insertions that built it. For this collection, the fringe is defined dynamically; that is, every insertion, in addition to building a new tree, defines the fringe of the new tree as a function of the fringe of the old tree and the transformations caused by the insertion.

THEOREM 2.1. *The AVL tree collection shown in Figure* 4 *is closed.*

PROOF. Similar to the proof by Brown [3] ■.

The results obtained with this fringe are better than Brown's results, and some are as good as Mehlhorn's results. They are shown in Table 1.

Consider the tree collection of AVL trees shown in Figure 1. It is possible to see by studying the transitions of each type of subtree that

$$\vec{P}(N) = \left(I + \frac{H}{N+1}\right)\vec{P}(N-1) \tag{2}$$

where $\vec{P}(N) = [p_1(N), p_2(N)]$, $H = \begin{bmatrix} -3 & 4 \\ 3 & -4 \end{bmatrix}$ is the transition matrix, and I is the identity matrix. In general, the same recurrence type holds for a tree collection with m types.

Formal details of the development of this theory can be found in Ziviani [10], or in Eisenbarth, Ziviani, Gonnet, Mehlhorn and Wood [9]. The following theorem [9, 11] is basic for deriving our new results.

THEOREM 2.2. *Let H be the $m \times m$ transition matrix of a connected fringe analysis problem, as in Eq.* 2. *Let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of H, then*

$$\lambda_1 = 0 > \operatorname{Re}\lambda_2 \geq \operatorname{Re}\lambda_3 \geq \ldots \geq \operatorname{Re}\lambda_m.$$

Let t be the multiplicity of λ_2. Then for every vector $\vec{P}(N)$

$$\vec{P}(N) = \vec{P}(\infty) + O(\log^t N\, N^{\operatorname{Re}\lambda_2})$$

where $\vec{P}(N)$ is defined by Eq. 1, and $\vec{P}(\infty)$ is the unique solution of the system of equations

$$H\vec{P}(\infty) = 0 \quad \textit{and} \quad \sum_i p_i(\infty) = 1.$$

Let $A_i(N)$ be the expected number of trees of type i in a random search tree with N keys. Let L_i be the number of leaves of the type i tree. Note that Eq. 1 can be written as

$$p_i(N) = \frac{A_i(N)L_i}{N+1}. \tag{3}$$

3. Basic relations.

In this section we show how to relate the fringe of an AVL tree to the defined measures. To simplify notation, $p_i(N)$ is written as p_i throughout the remainder of this paper.

LEMMA 3.1. *The expected number of rotations in a random AVL tree with N keys during the $(N+1)$st insertion is bounded above by*

(*i*) $$\bar{r}(N) = 1 - Pr\{\textit{no rotation}\}$$

and (*ii*) $\bar{r}(N) \leq \bar{r}(N)$ *in the fringe + rotations outside the fringe.*

PROOF. For case (i) it is known that the maximum number of rotations per insertion in an AVL tree is 1. For case (ii) $\bar{r}(N)$ must be less than or equal to the number of rotations per insertion in the fringe plus all possible rotations per insertion that may occur outside the fringe. ■

LEMMA 3.2. *The expected number of nodes in the fringe of an AVL tree with N keys that corresponds to a tree collection $\mathscr{C} = T_1, \ldots, T_m$ is*

$$\bar{f}(N) = \sum_{i=1}^{m} \frac{p_i(L_i - 1)}{L_i}(N+1).$$

PROOF. Each subtree type has $L_i - 1$ nodes. Then, $\bar{f}(N) = \sum_{i=1}^{m}(L_i - 1)A_i(N)$. Using Eq. 3 we have the desired result. ■.

LEMMA 3.3. *The expected number of balanced nodes in a random AVL tree with N keys is bounded above by*

(*i*) $$\bar{b}(N) = N - \bar{u}(N) \quad \textit{for} \quad N \geq 1$$

and (*ii*) $$\bar{b}(N) \leq \bar{b}(N) \textit{ in the fringe} + [N - \bar{f}(N)] \quad \textit{for} \quad N \geq 1,$$

where $\bar{u}$ *is the expected number of unbalanced nodes.*

PROOF. For case (i) $\bar{b}(N) + \bar{u}(N) = N$. For case (ii) $\bar{b}(N)$ must be less than or equal to the number of balanced nodes in the fringe plus all nodes outside the fringe. ■

THEOREM 3.1. *The expected number of balance changes during N random insertions in an initially empty AVL tree is bounded by*

$$N(2 + \bar{r}(N)) - \bar{b}(N) - 1.44 \log_2 N + O(1) \leq \bar{c}(N)$$
$$\leq N(2 + \bar{r}(N)) - \bar{b}(N) - \log_2 N + O(1).$$

PROOF. Mehlhorn and Tsakalidis [8] show that

$$\bar{c}(N) = \bar{u}(N) + R(N) + N - \bar{h}(N),$$

where $R(N) = \sum_{i=1}^{N} \bar{r}(i)$. Using this definition, Lemma 3.5, and the fact [12] that

$$\log_2 (N) \leq \bar{h}(N) \leq 1.44 \log_2 (N),$$

we get the desired result. ■

4. Weakly-closed AVL tree collections.

If the effect of an insertion on the composition of the fringe is determined not only by the subtree of the fringe where the insertion is performed, but by some other transformation that may happen outside the fringe, then the tree collection is weakly-closed (Definition 2). We will show that the tree collection of AVL trees with four or less leaves shown in Figure 4 is not closed if the fringe is not ambiguous.

LEMMA 4.1. *If the trees shown in Figure 4 from the fringe of a random AVL tree with N keys and* $N \to \infty$, *then an insertion into a leaf of a type 3 subtree* (*i*) *decreases by one the number of type 3 subtrees and increases by one the number of type 1 and type 2 subtrees; or* (*ii*) *decreases by one the number of type 1 subtrees and increases by one the number of type 2 subtrees.*

PROOF. By Mehlhorn [4]. ■.

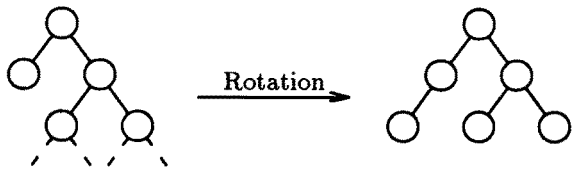

Fig. 5. AVL transformation with changes (symmetric transformations occur). (Dashed edges show the point of insertion).

Lemma 4.1 tells us that any AVL tree collection that contains type 3 shown in Figure 4 is not closed, if the fringe is not ambiguous, (i.e. it is weakly-closed). In fact it is not difficult to show that every AVL tree type that contains more than one internal node and has its root node balanced suffers from the same type of misbehaviour that occurs with type 3 (i.e. consider the AVL tree with six nodes). Consequently an AVL tree collection that contains a tree type with the root node balanced and has more than two types is weakly-closed, if the fringe is not ambiguous (cf. Def. 2).

We know from Lemma 4.1 that an insertion into the type 3 tree shown in Figure 4, when it belongs to a fringe of an AVL tree with N keys, produces a transition that is not well defined: the transition depends on the unknown probability t_N which also depends on N. First of all let us give a more precise meaning to t_N. Let I be the expected number of leaves in an AVL tree with N keys such that an insertion in one of the I leaves causes the transformations shown in Figure 5. Thus

$$t_N = \frac{I}{N+1}.$$

Although the probability t_N is unknown it cannot assume arbitrary values between 0 and 1.

LEMMA 4.2. *The probability* t_N *is bounded by* $0 \leq t_N \leq \min(2p_1(N), p_3(N))$.

PROOF. The number of type 1 subtrees ($A_1(N)$) must be greater than or equal to the number of type 3 subtrees that have a brother of type 1 ($t_N(N+1)/4$). This is because not all type 1 subtrees are brothers of type 3 subtrees and in the worst case all type 3 subtrees have a brother of type 1. This gives $t_N \leq 2p_1(N)$. Also, the t_N leaves are a subset of p_3 leaves. ■

The probability t_N models an absolute event. On the other hand, Mehlhorn [4] used a conditional probability of the brother of a type 1 subtree being a type 3 subtree. The main problem is that he used $A_3(N)/A_1(N)$ to bound this probability. However, in general, $E[x/y] \neq E[x]/E[y]$ where x and y are random variables, and $E[x]$ is the expected value of x. Our approach avoids this problem.

Table 2 (page 308) shows the exact value of t_N for $N \leq 20$.

From the results of Lemma 4.1 we can examine the insertion process and obtain

$$(4)\qquad \vec{P}(N) = I + \frac{1}{N+1}\begin{bmatrix} -3 & 0 & 2 \\ 3 & -4 & 3 \\ 0 & 3 & -5 \end{bmatrix} \vec{P}(N-1) + \frac{t_{N-1}}{N+1}\begin{bmatrix} -4 \\ 0 \\ 4 \end{bmatrix}$$

$$= \left(I + \frac{H}{N+1}\right)\vec{P}(N-1) + \frac{t_{N-1}}{N+1}\vec{U}$$

where t_N depend on N. Figure 6 shows how the value of the third column of H and the value of $\vec{U}$ were obtained.

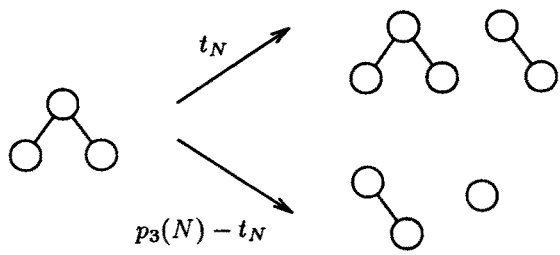

Fig. 6. Transitions for type 3.

The following theorem shows that the bounds for p_1 and p_3 are independent of the convergence of t_N.

THEOREM 4.1. *The solution of Eq.* 4 *is bounded by*

$$\tfrac{8}{91} \le p_1 \le \tfrac{8}{35}, p_2 = \tfrac{3}{7}, \quad \tfrac{12}{35} \le p_3 \le \tfrac{44}{91}$$

for $N \ge 12$ *with the condition* $p_1 + p_3 = \frac{4}{7}$.

COROLLARY 4.1. $t_N \le 16/91$ *applying Lemma* 4.2.

PROOF. The number of subtrees with two leaves is 2/7 and with three leaves 1/7 for $N \ge 6$ [3]. This implies

$$(5)\qquad \frac{p_1}{L_1} + 2\frac{p_3}{L_3} = \tfrac{2}{7} \quad \text{and} \quad p_2 = \tfrac{3}{7}$$

for $N \ge 6$. Introducing this in Eq. 4, we obtain for $N \ge 6$

$$p_1(N) = p_1(N-1)\left(1 - \frac{5}{N+1}\right) + \frac{8}{7(N+1)} - \frac{4}{N+1}t_{N-1}$$

and

$$p_3(N) = p_3(N-1)\left(1 - \frac{5}{N+1}\right) + \frac{12}{7(N+1)} + \frac{4}{N+1} t_{N-1}.$$

Using Lemma 4.2 $p_1(N)$ is bounded by

$$(6) \qquad p_1(N-1)\left(1 - \frac{13}{N+1}\right) + \frac{8}{7(N+1)} \leq p_1(N) \leq p_1(N-1)\left(1 - \frac{5}{N+1}\right) + \frac{8}{7(N+1)}.$$

The solution of a recurrence of the form

$$x_N = x_{N-1}\left(1 - \frac{a}{N+1}\right) + \frac{b}{N+1}$$

is $x_N = b/a$ for $N \geq (a-1)$ if a is an integer.

Solving Eq. 6 gives $8/91 \leq p_1 \leq 8/35$ for $N \geq 12$. In the same way, using Lemma 4.2 and relations 5 we obtain $12/35 \leq p_3 \leq 44/91$ for $N \geq 12$. ■

The results for this tree collection are shown in Table 1, and are similar to Mehlhorn [4] and Mehlhorn and Tsakalidis [8] results (although using a different formulation). The lower bounds are the minimum values and the upper bounds the maximum values to hold for any value of p_1 and p_3 in the given ranges, with the constraint that $p_1 + p_3 = 4/7$.

5. Larger weakly-closed AVL tree collections.

In Section 4 we showed that any AVL tree collection that contains a tree type with its root node balanced and with more than two types is weakly-closed. This happens because every AVL tree type that contains more than one internal node and has its root node balanced suffers from the same type of misbehaviour that occurs with type 3 of Figure 4, as described in Lemma 4.1.

It is easy to prove a lemma similar to Lemma 4.1 for the tree collection shown in Figure 7. The only difference in the proof of such a lemma is that now the trees shown in Figure 8 add another unknown probability that we call s_N, and is divided in two cases, $s1_N$ and $s2_N$. In these cases the number of type 3 trees decreases by one and the number of type 1 and type 2 trees increases by one. The recurrence relation corresponding to the tree collection shown in Figure 7 involves the two unknown probabilities t_N and s_N, as follows

$$(7) \quad \vec{P}(N) = \left(I + \frac{1}{N+1}\begin{bmatrix} -3 & 0 & 0 & 6/5 \\ 3 & -4 & 4 & 12/5 \\ 0 & 4 & -5 & 12/5 \\ 0 & 0 & 5 & -6 \end{bmatrix}\right) \vec{P}(N-1) + \frac{1}{N+1}\begin{bmatrix} -2 & 2 \\ 3 & 3 \\ 4 & 0 \\ -5 & -5 \end{bmatrix}\begin{bmatrix} t_{N-1} \\ s_{N-1} \end{bmatrix}$$

where t_N and s_N depend on N. Figure 9 shows the transitions in which t_N and s_N appear.

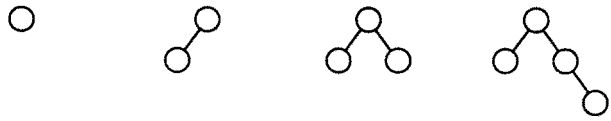

Fig. 7. Tree collection of AVL trees with five or less leaves.
(Leaves not shown).

For trees of size $N \leq 20$ we are able to obtain s_N exactly. Table 2 shows the values for $s1_N$ (cases (a) and (b) of Fig. 8), $s2_N$ (case (c) of Fig. 8) and $s_N = s1_N + s2_N$.

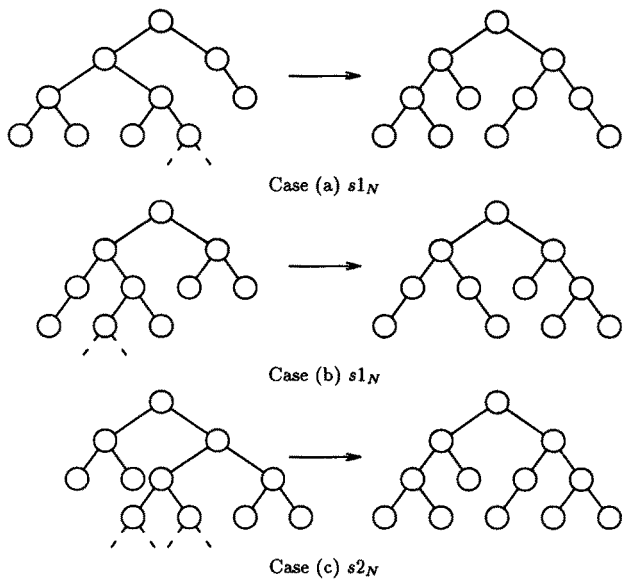

Fig. 8. Transformations in the fringe with changes (symmetric transformations occur).
(Dashed edges shows the points of insertion).

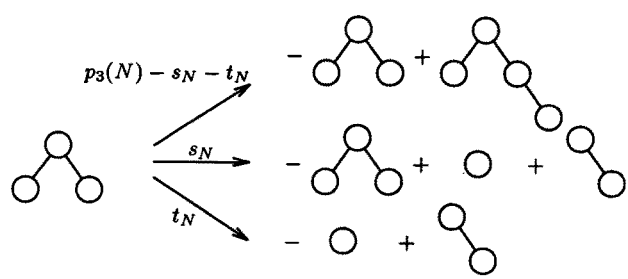

Fig. 9. Transitions for type 2 that involves t_N and s_N.

Table 2. *Exact values for t_N and s_N.*

N	t_N	N	t_N	N	$s1_N$	$s2_N$	s_N
5	0.4	13	0.09376	10	0.045455	0.	0.045455
6	0.	14	0.08948	11	0.	0.038961	0.038961
7	0.	15	0.08914	12–14	0.	0.	0.
8	0.09524	16	0.09074	15	0.003207	0.	0.003207
9	0.11429	17	0.09299	16	0.005758	0.001983	0.007741
10	0.10390	18	0.09490	17	0.007114	0.003889	0.011003
11	0.11039	19	0.09613	18	0.007560	0.005117	0.012677
12	0.10739	20	0.09672	19	0.007428	0.005705	0.013133
				20	0.007752	0.005830	0.013582

LEMMA 5.1. *We have $t_N = 2p_1(N)$.*

PROOF. A type 1 subtree is always brother of a type 3 subtree, since otherwise we would have a type 3 or a type 4 subtree. Then, the number of type 3 subtrees with a type 1 subtree as a brother is $A_1(N)$. ■

Now the relation between this collection to Brown's collection [3] gives us the following equations:

$$\frac{p_3}{3} + \frac{p_4}{5} = \tfrac{1}{7} \quad \text{and} \quad p_1 + p_3 + \tfrac{2}{5}p_4 = \tfrac{4}{7}. \tag{8}$$

Using these equations and Lemma 5.1, the new matrix recurrence relation is reduced for $N \geq 6$ to

$$\vec{P}(N) = \left(I + \frac{1}{N+1}\begin{bmatrix} -7 & -2 \\ 9 & -8 \end{bmatrix}\right)\vec{P}(N-1) + \frac{1}{N+1}\left(\begin{bmatrix} 6/7 \\ 12/7 \end{bmatrix} + s_{N-1}\begin{bmatrix} 2 \\ 3 \end{bmatrix}\right) \tag{9}$$

with $\vec{P}(N) = [p_1(N), p_2(N)]$, and $p_3(N)$ and $p_4(N)$ are obtained from Eqs. 8.

LEMMA 5.2. *s_N is bounded by*

$$0 \leq s_N \leq \frac{p_3(N) - 2p_1(N)}{3} = \tfrac{4}{21} - p_1(N) - \tfrac{2}{15}p_4(N) = \tfrac{2}{21} + \tfrac{2}{5}p_2(N) - p_1(N).$$

COROLLARY 5.1. $0 \leq s_N \leq \frac{4}{21}$.

PROOF. In the worst case all type 3 subtrees without a type 1 subtree as a brother belong to trees associated with s_N. Using Lemma 5.1 this number is $A_3(N) - A_1(N)$. If all are associated with $s1_N$ we have $2(A_3(N) - A_1(N))/2$ leaves and if all are associated with $s2_N$ we have $4(A_3(N) - A_1(N))/3$ leaves. Using relations 8 we obtain the last equalities. ■

Although we have again reduced the problem to an unknown probability, s_N, how do we solve recurrence 9 in general? The next theorem, one of the most important of our results, gives the exact solution to this type of recurrences.

THEOREM 5.1. *Let t_N^i be m unknown probabilities. The solution to a recurrence of the form*

$$\vec{P}(N) = \left(I + \frac{1}{N+1}H\right)\vec{P}(N-1) + \frac{1}{N+1}\vec{F} + \frac{1}{N+1}\sum_{i=1}^{m} t_{N-1}^i \vec{D}_i,$$

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

$$\vec{P}(N) = -H^{-1}\vec{F} + \frac{(-1)^{N+1}}{(N+1)!}R^{\underline{N+1}}\vec{D}$$

$$+ \sum_{i=1}^{m}\sum_{k \geq N_0}^{N-1}(-1)^{N-k}t_k^i \sum_{j=0}^{n-k-1} \frac{R^j H^{\underline{N-k-j-1}}}{(N-j+1)(N-k-j-1)!j!}\vec{G}_i$$

where $R = -H - I$, and $H^{\underline{n}} = (H-(n-1)I)\ldots(H-I)H$ denotes descendent factorials over matrices, and with $\vec{D}$ obtained from the initial condition

$$\vec{P}(N_0) = [1, 0, 0, \ldots, 0]^T,$$

where N_0 is the number of elements in the smallest subtree type of the fringe collection.

COROLLARY 5.2. *If $t_N^i \leq t^i$ for all $N \geq N_1$, then we have*

$$\vec{P}(N) \leq -H^{-1}\left(\vec{F} + \sum_{i=1}^{m} t^i \vec{G}_i\right) + O(1/N)$$

for some $N_1 > N_0$. A similar relation holds if $t_N^i \geq t^i$.

PROOF. Introducing the generating function

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

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

$$\frac{d\vec{P}(z)}{dz} = \left(\frac{2z-1}{z(1-z)}I + \frac{1}{1-z}H\right)\vec{P}(z) + \frac{1}{z(1-z)^2}\vec{F} + \frac{1}{1-z}\sum_{i=1}^{m} t^i(z)\vec{G}_i,$$

where $t^i(z)$ is the generating function associated to t_N^i. The solution of the previous equation is (this can be checked by simple substitution)

$$\vec{P}(z) = \frac{1}{z}e^{R\ln(1-z)}\vec{D} - \frac{1}{z(1-z)}H^{-1}\vec{F} - \sum_{i=1}^{m}\frac{1}{z}e^{R\ln(1-z)}\int z t^i(z) e^{H\ln(1-z)}dz\,\vec{G}_i$$

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

For $N \geq N_0$, we have

$$\vec{P}(N) = [z^N]\vec{P}(z) = -H^{-1}\vec{F} + \sum_{k \geq 0} R^k [z^{N+1}] \frac{\ln^k(1-z)}{k!} \vec{D} - \sum_{i=1}^{m} [z^N]\vec{U}_i(z)$$

where $[z^n]P(z)$ denotes the coefficient in z^n of $P(z)$, and

$$\vec{U}_i(z) = \frac{1}{z} e^{R \ln(1-z)} \int z t^i(z) e^{H \ln(1-z)} dz\, \vec{G}_i$$

$$= \frac{1}{z} e^{R \ln(1-z)} \sum k \geq N_0 \sum_{j \geq 0} \int z t_k^i z^{k+j+1} H^j \frac{\ln^j(1-z)}{j!} dz\, \vec{G}_i.$$

But

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

for $n \geq k$ where $\mathscr{S}_n^{(k)}$ denotes Stirling numbers of the first kind [13]. Using the previous formula and that

$$\sum_{k=0}^{n+1} \mathscr{S}_{n+1}^{(k)} H^k = H^{\underline{n+1}}$$

we obtain the desired result. This simplifies if t_i^k is bounded by a constant, giving the corollary. ■

Applying this solution to the recurrence for the collection of Section 4 we obtain, for $N > 5$,

$$p_1(N) = \tfrac{8}{35} - 4 \sum_{j>4}^{N-1} \frac{(j+1)^4}{(N+1)^3} t_j$$

Using Lemma 5.2 and Corollary 5.2 (choosing $N_1 = 20$) we can bound the set of probabilities p_i. Boot-strapping this result in Lemma 5.2 we obtain the following improved bounds for t_N and s_N.

LEMMA 5.3. *The probabilities t_N and s_N are bounded by*

$$\tfrac{24}{259} \approx 0.09266 \leq t_N \leq \tfrac{1168}{9583} \approx 0.1218 \quad \textit{and} \quad 0 \leq s_N \leq \tfrac{4}{37} \approx 0.1081.$$

The previous lemma gives the following bounds: $12/259 \leq p_1 \leq 584/9583$, $69/259 \leq p_2 \leq 3099/9583$, $108/259 \leq p_3 \leq 4220/9583$, and $240/1369 \leq p_4 \leq 10/37$. The next lemma gives the relation between the current tree collection and a larger tree collection that is used for the final results.

LEMMA 5.4. *The probabilities, q_1, of the tree collection shown in Figure* 10 *are*

$$q_1 = p_2, \quad q_2 = p_3 - 2p_1, \quad q_3 = p_4, \quad \textit{and} \quad q_4 = 3p_1.$$

PROOF. Using Lemma 5.1. ■

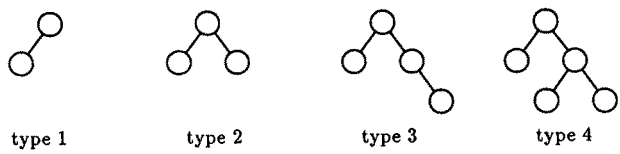

Fig. 10. Tree collection of AVL trees with more than 2 and less than 7 leaves. (Leaves not shown).

The following results are for the collection of Figure 10 using Lemma 5.3 and the bounds on the p_i's.

THEOREM 5.2. *The expected number of rotations in a random AVL tree with N keys during the* $(N+1)$*st insertion is bounded by*

$$2\frac{q_1}{L_1} + 2\frac{q_3}{L_3} + 4\frac{q_4}{L_4} + s_N \leq \bar{r}(N) \leq 2\frac{q_1}{L_1} + 2\frac{q_3}{L_3} + 4\frac{q_4}{L_4} + q_2.$$

COROLLARY 5.3. $\frac{14}{37} \approx 0.3784 \leq \bar{r}(N) \leq \frac{994}{1369} \approx 0.7261$ *for* $N \geq 20$.

PROOF. The expression above can be obtained by observing Figure 10 and using Lemma 3.1. ■

CONJECTURE 5.1. *The expected number of comparisons in a successful search is upper bounded by*

$$\bar{C}_N \leq \tfrac{60}{37} H_N + O(1) \approx 1.117 \log_2 N,$$

where $H_N = \sum_{i=1}^{N} 1/i$ *is the Nth harmonic number and* $\bar{r}(N) \leq 0.5573$.

PROOF. For each input of N elements (permutation) we model the tree process as building a random binary search tree (BST) and then applying a transformation to the corresponding AVL tree. This transformation is based in each rotation performed in the AVL tree. Each rotation in the fringe decreases the internal path length by at least one (in fact, the path length is decreased by two only in the rotation of the case (a) of Figure 8). We assume that the effect of all rotations above the fringe does not increase the internal path length (this is not true in general for a single rotation). This seems to be true, but we were not able to prove it. If a fringe rotation was done in the ith insertion, that rotation decreases the internal path length of i elements. Then

$$C_N^{AVL} \leq C_N^{BST} - \sum_{i=1}^{N-1} \frac{Rot_i}{i}$$

where Rot_i is 0 or 1 depending on if for that input there was a fringe rotation in the ith insertion. Taking the expected value over all inputs, we have

$$\bar{C}_N^{AVL} \leq \bar{C}_N^{BST} - \sum_{i=1}^{N-1} \frac{\bar{r}(i)}{i}.$$

It is known that $\bar{C}_N = 2H_N + O(1)$ for a random binary search tree. In the ith insertion, a fringe rotation happens with probability 14/37 for $i > N_1$ (previous theorem). Thus, $\bar{C}_N \leq 2H_N - \frac{14}{37}\sum_{i=1}^{N-1} 1/i + O(1)$ which gives the desired result. This conjecture also holds for unsuccessful searches (C'_N), by using the relation $C_N = (1 + 1/N)C'_N - 1$. Because a complete balanced tree has $C'_N = \log_2 N + O(1)$, $\bar{r}(N)$ is at most $2 - 1/\ln(2)$. ■

LEMMA 5.5. *The expected number of nodes in the fringe of an AVL tree with N keys corresponding to the tree collection of Figure* 10 *is* $[\frac{7159}{9583} \approx 0.7471](N + 1) \leq \bar{f}(N) \leq [\frac{195}{259} \approx 0.7529](N + 1)$ *for* $N \geq 20$.

PROOF. The above expression can be obtained by observing Figure 10 and by using Lemma 3.1. ■

From the previous lemma, taking a balanced tree and the tallest tree outside the fringe, and adding the average height of the fringe, we obtain the following bounds

$$\log_2(N) + 0.3413 \leq \bar{h}(N) \leq 1.44 \log_2 N - 0.4464.$$

LEMMA 5.6. *The expected number of unbalanced nodes outside the fringe defined by the current tree collection, of a random AVL tree with N keys, is at least* $(s_N/4)(N + 1)$.

PROOF. The above expression is obtained as follows: a tree associated with s_N always has one unbalanced node (the root) outside the fringe. In the worst case all these trees belongs to case (c) of Fig. 8. ■

LEMMA 5.7. *The expected number of balanced nodes outside the fringe defined by the current tree collection, of a random AVL trees with N keys, is at least*

$$\left(\frac{q_1}{L_1} + \frac{q_2}{L_2} - \frac{q_3}{L_3} - \frac{q_4}{L_4}\right)\frac{(N + 1)}{3}.$$

PROOF. Type 1 or 2 subtrees may be brothered with type 3 or type 4 subtrees. In that case we have $A_1(N) + A_2(N)$ subtrees of height 3. In the worst case for balanced nodes, the $A_3(N) + A_4(N) - A_1(N) - A_2(N)$ remaining subtrees of height 2 are brothers of the subtrees of height 3, i.e. $A_1(N) + A_2(N)$ pairs. Any 3 of the others $(A_3(N) + A_4(N) - A_1(N) - A_2(N))$ must generate at least one balanced node (see Figure 11). ■

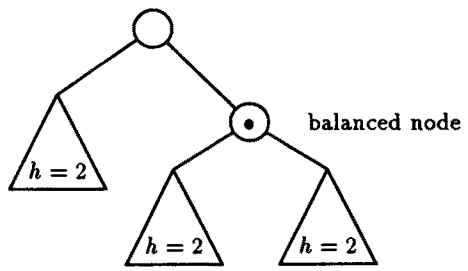

Fig. 11. Balanced nodes outside the fringe.

THEOREM 5.3. *The expected number of balanced nodes in a random AVL tree with N keys is bounded by*

$$\left(\frac{4q_1}{3L_1} + \frac{10q_2}{3L_2} + \frac{5q_3}{3L_3} + \frac{11q_4}{3L_4}\right)(N+1) \leq \bar{b}(N)$$

$$\leq N - \left(\frac{q_1}{L_1} + 2\frac{q_3}{L_3} + \frac{q_4}{L_4} + \frac{s_N}{4}\right)(N+1)$$

COROLLARY 5.4. $(\frac{146}{259} \approx 0.5637) \leq \dfrac{\bar{b}(N)}{N} \leq (\frac{202}{259} \approx 0.7799)$ *for* $N \geq 20$.

PROOF. The left hand side of the above expression is obtained by observing Figure 10, Lemma 5.7 and by using Eq. 3. The right hand side is obtained by using Lemmas 3.3, 5.5, and 5.6. ■

THEOREM 5.4. *The expected number of balance changes during N random insertions in an initial empty AVL tree is bounded by*

$$\tfrac{414}{259}N + o(N) \approx 1.5985N \leq \bar{c}(N) \leq \tfrac{61718}{28479}N + o(N) \approx 2.1468N$$

for $N \geq 20$.

PROOF. Similar to Theorem 3.1. ■

Experimental results show that $C'_N \approx 1.0176 \log_2 N + 0.0513$ [14], $\bar{r}(N) \approx 0.47$ [14], $\bar{b}(N) \approx 0.68N$ [12], and $\bar{c}(N) \approx 1.78N$ [8].

Table 3 shows simulation results for larger trees for t_N and s_N, obtained with a 95% confidence interval. For example, from Table 3 the value of $s2_N$ seems to converge to 12/2500 when N is large, but we are not able to prove it. Moreover $s2_N$ may oscillate smoothly, in such a way that simulations cannot detect (e.g. consider $s2_N = 12/2500 + \cos(\ln N)/10000$.)

Table 3. *Simulation results for t_N and s_N.*

Tree Size	Number of Trees	t_N	s_N (percent)		
			Total	$s1_N$	$s2_N$
20	50000	0.0968 ± 0.0011	1.367 ± 0.036	0.789 ± 0.023	0.00578 ± 0.00029
50	10000	0.0964 ± 0.0016	1.062 ± 0.047	0.565 ± 0.028	0.496 ± 0.038
100	5000	0.0961 ± 0.0016	1.094 ± 0.048	0.605 ± 0.030	0.490 ± 0.039
500	5000	0.09524 ± 0.00070	1.110 ± 0.022	0.614 ± 0.013	0.496 ± 0.018
1000	2000	0.09559 ± 0.00078	1.090 ± 0.023	0.593 ± 0.015	0.497 ± 0.019
2500	2000	0.09611 ± 0.00048	1.086 ± 0.015	0.6114 ± 0.0096	0.474 ± 0.012
5000	1000	0.09603 ± 0.00048	1.097 ± 0.015	0.6029 ± 0.0095	0.494 ± 0.012
10000	1000	0.09565 ± 0.00035	1.091 ± 0.010	0.6050 ± 0.0066	0.4859 ± 0.0085
15000	500	0.09582 ± 0.00043	1.098 ± 0.013	0.6078 ± 0.0077	0.490 ± 0.010
20000	500	0.09550 ± 0.00037	1.101 ± 0.011	0.6019 ± 0.0064	0.4986 ± 0.0087

LEMMA 5.8. *If* $\lim_{N\to\infty} t_N = t$ *and* $\lim_{N\to\infty} s_N = s$ *exist, then* $t = \frac{24}{259} + \frac{10}{37}s$.

PROOF. Replacing s_{N-1} by s in Eq. 9, solving it, and using Lemma 5.1 and Theorem 2.2. ■

We conjecture that t_N and s_N converge. In fact, the simulation results agree with Lemma 5.8. For example, if we assume that s_N for large N is approximately 0.011, then using Lemma 5.8 the value for t_N is 0.0956. The simulation value is 0.0955.

6. Conclusions.

We have formalized the concept of weakly-closed and ambiguous tree collections, in relation to AVL-trees. We present a new closed AVL tree collection that allows to obtain almost all the results obtained by Mehlhorn using a weakly-closed collection [4]. In Section 4 we model weakly-closed tree collections, giving the exact solution to the new fringe analysis recurrence in Section 5. Our last tree collection improves all previous known results for the expected case of AVL-trees. By using this exact solution it is possible to analyze larger tree collections. In fact, a tree collection with subtrees from height 2 to 4 includes 10 types and at least 8 unknown probabilities.

Like AVL trees, weight-balanced trees are balanced by single and double rotations [12, Sec. 6.2.3]. For this reason only small tree collections of weight-balanced trees are closed. For large tree collections we find the same type of difficulties shown for AVL trees. Consequently, the technique presented for the analysis of AVL trees is also suitable for the analysis of weight-balanced trees.

Acknowledgements.

We wish to acknowledge the helpful suggestions from Patricio Problete and the anonymous referees.

REFERENCES

1. G. M. Adel'son-Vel'skii and E. M. Landis. *An algorithm for the organization of information.* Doklady Akademia Nauk SSSR, 146 (2): 263–266, 1962. English translation in Soviet Math. Doklady 3, 1962, 1259–1263.
2. A. C.-C. Yao. *On random 2–3 trees.* Acta Informatica, 9 (2): 159–170, 1978.
3. M. R. Brown. *A partial analysis of random height-balanced trees.* SIAM J on Computing, 8 (1): 33–41, Feb 1979.
4. K. Mehlhorn. *A partial analysis of height-balanced trees.* Technical Report A79/13, Universität des Saarlandes, Saarbrücken, West Germany, 1979.
5. Th. Ottmann and H. W. Six. *Eine neue Klasse von ausgeglichenen Binärbäumen.* Angewandte Informatik, 9: 395–400, 1976.
6. Thomas Ottmann and Derick Wood. 1–2 *brother trees or AVL trees revisited.* Computer Journal, 23 (3): 248–255, Aug 1980.
7. Kurt Mehlhorn. *A partial analysis of height-balanced trees under random insertions and deletions.* SIAM J on Computing, 11 (4): 748–760, Nov 1982.
8. Kurt Mehlhorn and A. Tsakalidis. *An amortized analysis of insertions into AVL-trees.* SIAM J on Computing, 15 (1): 22–33, Feb 1986.
9. 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.
10. N. Ziviani. *The Fringe Analysis of Search Trees.* PhD thesis, Department of Computer Science, University of Waterloo, 1982.
11. 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. Also as Technical Report CS-89-16, Dept. of Computer Science, U. of Waterloo, 1989.
12. D. E. Knuth. *The Art of Computer Programing: Sorting and Searching,* volume 3. Addison-Wesley, Reading, Mass., 1973.
13. D. E. Knuth. *The Art of Computer Programming: Fundamental Algorithms,* volume 1. Addison-Wesley, Reading, Mass., 1969.
14. N. Ziviani and F. W. Tompa. *A look at symmetric binary B-trees.* Infor, 20 (2): 65–81, May 1982.