# On the average size of the intersection of binary trees

R. Baeza-Yates
R. Casas
J. Díaz
C. Martinez

Report LSI-89-23

**Abstract:** The probability distribution for binary search trees could be considered a more realistic distribution to do statistics on trees than the general uniform distribution. The average analysis of algorithms for binary search trees, yields very different results from those obtained under the uniform distribution. Moreover, the analysis itself is a lot more complex. In this work we carry out this analysis for the computation of the average size of the intersection of two binary trees. The development of this analysis involves Bessel functions which appear in the solutions of partial differential equations, and the result is an average size of $O(n^{2\sqrt{2}-2}/\sqrt{\log\ n})$, contrasting with the unrealistic $O(1)$ which appears as solution when considering a uniform distribution.

**Resum:** Per a calcular cost mitjà d'operacions o algorismes sobre arbres, la distribució de probabilitat per als arbres binaris de recerca es pot considerar com una distribució més realista que la distribució uniforme utilitzada normalment. En aquest treball, realitzem la computació de la talla mitjana de la intersecció de dos arbres binaris. El desenvolupament d'aquesta anàlisi involucra funcions de Bessel que apareixen com a resultat d'una equació diferencial. El resultat es una talla mitjana de $O(n^{2\sqrt{2}-2}/\sqrt{\log\ n})$, la qual cosa contrasta amb el $O(1)$ que apareix com a solució, quan es considera una distribució uniforme.

# On the Average Size of the Intersection of Binary Trees[*]

R. Baeza-Yates[†] R.Casas[‡] J.Díaz[‡] C.Martínez[‡]

November 17, 1989

**Abstract**

The probability distribution for binary search trees could be considered a more realistic distribution to do statistics on trees than the general uniform distribution. The average analysis of algorithms for binary search trees, yields very different results from those obtained under the uniform distribution. Moreover, the analysis itself is a lot more complex. In this work we carry out this analysis for the computation of the average size of the intersection of two binary trees. The development of this analysis involves Bessel functions which appear in the solutions of partial differential equations, and the result is an average size of $O(n^{2\sqrt{2}-2}/\sqrt{\log\, n})$, contrasting with the unrealistic $O(1)$ which appears as solution when considering a uniform distribution.

## 1 Introduction

Most results on average analysis of algorithms on trees have been done by considering a uniform distribution over all the trees with the same number of nodes. As most of the binary trees of a given size are very skew, this distribution has the drawback that the average statistics obtained often are not the ones expected by "common sense".

Moreover there have been some statistics on trees which consider other types of distributions. In particular, a great amount of work has been done on statistics for Binary Search Trees. Most of this work relates to the average

[*] This research was supported by the ESPRIT BRA Program of the EC under contract no. 3075,project ALCOM.

[†] Department Computer Science, University of Waterloo, Canada

[‡] Departament de Llenguatges i Sistemes, Universitat Politècnica Catalunya, Barcelona

analysis of algorithms associated with the manipulation of this particular data structure [Knu73].

Recall that binary search trees are binary trees whose nodes are labeled in a certain way. The model underlying this random tree model is that of random permutation. Each permutation on the symmetric group of size $n$ is taken with a uniform probability of $1/n!$, and this uniform distribution induces a nonuniform probability distribution on the set $\mathcal{B}_n$ of all binary trees with $n$ internal nodes [Knu73,Fla88]. This distribution, which from now on we shall denote as *bst-distribution*, can also be used as the underlying distribution to do statistics on non-labeled binary trees. Devroye proved that the average height of binary trees under the bst-distribution is asymptotically $O(\log\ n)$ [Dev86]. This result marks a difference with the average height of binary trees under the uniform distribution model, which tends to $O(\sqrt{n})$ [FO82].

We believe that there are two reasons for using the bst-distribution as the canonical distribution to do statistics on binary trees. On the one hand, the bst-distribution tends to assign higher probability of existence to the more balanced binary trees, which in many cases makes the distribution look more realistic than the uniform distribution. On the other hand, we think there are reasonable chances that it is feasible to compute the average complexity of many algorithms without getting bogged down at the difficulty of the mathematical computations, as the present work indicates.

The computation of the average size of the intersection of binary trees, appears in a natural way in the analysis of a number of algorithms; for example in processes involving tree matching [FS87] or unification [CDS89]. In fact the intersection of binary trees taken as operation between binary trees is exactly the kernel of the algorithm of Shuffle of two trees described in [CKS89]. As a matter of fact, the average time complexity of the Shuffle of two binary trees coincides with twice the average size of the intersection of the two trees.

We have chosen to study the average size of the intersection of two binary trees, because of its simplicity. Our first consideration has been to investigate the kind of analysis which we shall deal with, when doing statistics under the bst-distribution. That is the reason why we selected a problem which yields a very elementary development and solution, when considering the uniform distribution . (Under the uniform distribution, the average intersection of two trees is always the constant 2, regardless of the size of the trees [CKS89]). The use of the bst-distribution introduces a partial differential equation which has a solution in terms of Bessel functions,

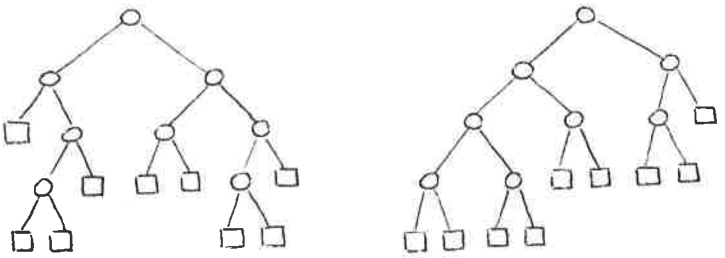

Figure 1: Binary trees $T_1$ (left) and $T_2$ (right)

which seems to correspond to the structure induced by this distribution.

## 2 A recursive definition of the bst-distribution

We begin by introducing a new recursive way to look at the bst-distribution. Unless otherwise stated, we shall use the definitions and notation as in [Knu69]

Given a binary tree $T$ let us denote by $T^r$ and $T^l$ respectively the right and left subtrees of the root of $T$. Let us define the following probability distribution over the set of binary trees with $n$ nodes,

$$p(T) = \begin{cases} 1 & \text{if } T = \square \\ \frac{p(T^l) \cdot p(T^r)}{1+|T^l|+|T^r|} & \text{otherwise} \end{cases}$$

where $p$ reads probability, $|T|$ denotes the size (number of internal nodes) of the tree $T$, and $\square$ denotes the leaf of a binary tree.

For example, given the binary tree $T_1$ described in Fig. 1, its probability is $p(T_1) = \frac{1}{384}$, while for the tree $T_2$ in the same figure, the probability is $p(T_2) = \frac{1}{240}$.

It is easy to verify that this probability corresponds to the frequency of all the binary search trees whith the same shape $T$.

The recursive manner in which we express this probability distribution, is very handly to simplify a lot some of the classical proofs about average behavior of binary search. The interested reader could convince himself by proving, using the propossed formulation, some well known facts about the expected behavior of random binary search trees.

In a natural way, we can extend the above definition to pairs of binary trees in the following way, given binary trees $T_1$ and $T_2$,

$$p(T_1, T_2) = \frac{p(T_1) \cdot p(T_2)}{1 + |T_1| + |T_2|}$$

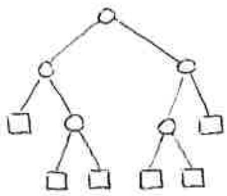

Figure 2: Intersection of trees in figure 1

Notice that for all $n \geq 0$, this definition satisfies the condition

$$\sum_{|T_1|+|T_2|=n} p(T_1, T_2) = 1$$

## 3 Average Size of the Intersection of Two Trees

Let $\mathcal{B}$ denote the set of all binary trees. Given trees $T_1, T_2 \in \mathcal{B}$ we wish to compute the average size of the intersection of the two trees, where the intersection of $T_1$ and $T_2$ is given by:

```
Intersection $(T_1, T_2)$
    If $T_1$ or $T_2$ is $\square$ then $(T_1 \cap T_2) = \square$
    else $(T_1 \cap T_2) = (T_1^l \cap T_2^l) \overset{\circ}{\frown} (T_1^r \cap T_2^r)$
```

where $(T_1 \cap T_2)$ denotes the intersection of trees $T_1$ and $T_2$.

Figure 2, shows the result of the intersection of the trees in Figure 1.

We shall define the size of the intersection of trees $T_1$ and $T_2$ by

$$s(T_1, T_2) = \begin{cases} 0 & \text{if } \square \in \{T_1, T_2\} \\ 1 + s(T_1^l, T_2^l) + s(T_1^r, T_2^r) & \text{otherwise} \end{cases}$$

We wish to compute the average value of $s(T_1, T_2)$ over all the pairs $(T_1, T_2)$ with $|T_1| + |T_2| = n$. Let $\tilde{s}(n)$ denote this average value, then we get

$$\tilde{s}(n) = \sum_{|T_1|+|T_2|=n} s(T_1, T_2) \cdot p(T_1, T_2)$$

Following the standard techniques [FV87,GJ83] let us define the following generating function:

$$S(z) = \sum_{(T_1,T_2)\in\mathcal{B}^2} s(T_1, T_2) \cdot p(T_1, T_2) \cdot z^{|T_1|+|T_2|} \tag{1}$$

We have to evaluate

$$\bar{s}(n) = [z^n]S(z) \tag{2}$$

where $[z^n]S(z)$ denotes the $n^{th}$ coefficient in the expansion of $S(z)$. For this let us define another generating function of two variables

$$S(x,y) = \sum_{(T_1,T_2)\in\mathcal{B}^2} s(T_1,T_2)p(T_1)p(T_2)x^{|T_1|}y^{|T_2|} \tag{3}$$

It follows that

$$S(z) = \frac{1}{z}\int_0^z S(t,t)dt \tag{4}$$

We use the following descomposition of the cartesian product of binary trees

$$\mathcal{B}^2 = (\square,\square) + \square \times (\mathcal{B}-\square) + (\mathcal{B}-\square)\times\square + (\mathcal{B}-\square)^2 \tag{5}$$

From equation (3) and using (5) we get the following hyperbolic partial differential equation

$$\frac{\partial^2 S(x,y)}{\partial x\partial y} = \frac{1}{(1-x)^2(1-y)^2} + \frac{2S(x,y)}{(1-x)(1-y)} \tag{6}$$

subject to the boundary conditions: for all $x$ and $y$, $S(x,0)=0$ and $S(0,y)=0$. These boundary conditions are given by the intersection of a tree and a leaf and the intersection of a leaf and a tree respectively.

Equation (6) can be rewritten using

$$S(x,y) = \Psi(x,y) - \frac{1}{(1-x)(1-y)} \tag{7}$$

where $\Psi(x,y)$ satisfies the homogeneous equation

$$\frac{\partial^2\Psi}{\partial x\partial y} = \frac{2\Psi}{(1-x)(1-y)}$$

with boundary conditions $\Psi(x,0) = \frac{1}{1-x}$ and $\Psi(0,y) = \frac{1}{1-y}$.

Making the change of variables

$$\begin{cases} X = -\sqrt{2}\ln(1-x) \\ Y = -\sqrt{2}\ln(1-y) \end{cases}$$

and making $G(X,Y)$ be $\Psi(1-e^{-X/\sqrt{2}}, 1-e^{-Y/\sqrt{2}})$, we finally obtain the hyperbolic differential equation

$$\frac{\partial^2 G}{\partial X\partial Y} = G \tag{8}$$

subject to boundary conditions $G(X,0) = e^{X/\sqrt{2}}$, $G(0,Y) = e^{Y/\sqrt{2}}$.

This system can be solved by the method of Riemann (Chapter 5 of [Cop75]) to yield

$$\begin{aligned} G(X,Y) &= \frac{1}{\sqrt{2}} \int_0^X e^{t/\sqrt{2}} J_0\left(2i\sqrt{(X-t)Y}\right) dt + \\ &\quad \frac{1}{\sqrt{2}} \int_0^Y e^{t/\sqrt{2}} J_0\left(2i\sqrt{(Y-t)X}\right) dt + J_0(2i\sqrt{XY}) \end{aligned} \tag{9}$$

where $J_0$ denotes the Bessel function of the first kind of order 0.

We are interested in obtaining asymptotics to the $[z^n]S(z)$, but by (4) we know that $[z^n]S(z) = \frac{1}{n+1} \cdot [z^n]S(z,z)$, which together with (7) gives

$$\tilde{s}(n) = \frac{1}{n+1}[z^n]\Psi(z,z) - 1 \tag{10}$$

To obtain an asymptotic value for $[z^n]\Psi(z,z)$ we need the following result,

**Lemma 1**

$$[z^n]\Psi(z,z) \approx c_1 \cdot [z^n]J_0(-2\sqrt{2} \cdot i \cdot \ln(1-z))$$

*where $\approx$ stands for asymptotical equivalence, and $c_1 = 3 + 2\sqrt{2}$*

The highly technical proof of this lemma is given in the appendix at the end of the paper.

On the other hand, using the Laplace method for integrals (see for example chapter 4 of [DB58] ) we derive the asymptotics for the $n^{th}$ coefficient in the expanssion of the Bessel function,

$$[z^n]J_0(-2\sqrt{2} \cdot i \cdot \ln(1-z)) \approx c_2 \cdot \frac{n^{2\sqrt{2}-1}}{\sqrt{\log n}} \cdot \left(1 + O(\frac{1}{\log n})\right) \tag{11}$$

where the value of constant $c_2$ is given by

$$c_2 = \frac{\sqrt{\log e}}{2^{5/4}\sqrt{\pi} \cdot \Gamma(2\sqrt{2})} = 0.0910284$$

Lemma 1 together with (10) and (11) gives the following result,

**Theorem 1** *Under the bst-distribution, the average size of the intersection of two trees behaves asymptotically as*

$$\tilde{s}(n) = c \cdot \frac{n^{2\sqrt{2}-2}}{\sqrt{\log n}} \cdot \left(1 + O(\frac{1}{\log n})\right)$$

*with* $c = c_1 c_2 = 0.530552\cdots$

Again, it should be emphasized, that under the usual uniform distribution, the average size of the intersection of two trees is the constant 2, which is a quite different result from the one we just obtained.

As said in the introduction, it also follows from Theorem 1 that under the bst-distribution, the average complexity of the algorithm of Shuffle of two trees is also

$$2c \cdot \frac{n^{2\sqrt{2}-2}}{\sqrt{\log n}} \cdot \left(1 + O(\frac{1}{\log n})\right)$$

while with the uniform distribution the average complexity of the Shuffle algorithm is $4(1 + O(\frac{1}{n}))$ [CKS89].

## 4 Conclusions

It seems that the apparition of a hyperbolic partial differential equation as the one in section 3, depends directly on the definition of probability distribution given in section 2, and it is rather independent of the nature of the problem under consideration. Current work by the authors seems to confirm this hypothesis. For instance, when considering other simple algorithms, like the equality of trees, the same methodology works and it also yields a hyperbolic differential equation. The present paper could be considered as a first formulation of the kind of framework inhereht to the statistics on trees under the bst-distribution.

**Acknowledgment**

We thank Josep Grané for pointing us to the solution of equation (**6**); Carles Simó for his advice in obtainning Lemma 1 and Philippe Flajolet for many interesting suggestions.

# A Appendix: Proof of Lemma 1

Let $G(Z,Z) = A(Z) + J_0(2iZ)$ with

$$A(Z) = \sqrt{2}\int_0^Z e^{t/\sqrt{2}} J_0\left(2i\sqrt{(Z-t)Z}\right) dt$$

Let us recall the series expansion

$$J_0(x) = \sum_{k\geq 0} \frac{(-1)^k}{(k!)^2}\left(\frac{x}{2}\right)^{2k}$$

then

$$A(Z) = \sqrt{2}\int_0^Z e^{t/\sqrt{2}}\left(\sum_{k\geq 0}\frac{(-1)^k}{(k!)^2}\cdot(-(Z-t)Z)^k\right) dt = \sqrt{2}\cdot\sum_{k\geq 0}\frac{Z^k}{(k!)^2}\cdot\Phi_k(Z)$$

where

$$\Phi_k(Z) = \int_0^Z e^{t/\sqrt{2}}(Z-t)^k dt = (\sqrt{2})^{k+1}\cdot k!\cdot\sum_{j>k}\frac{1}{j!}\left(\frac{Z}{\sqrt{2}}\right)^j$$

so we get

$$A(Z) = 2\sum_{k\geq 0}\frac{(Z\sqrt{2})^k}{k!}\left(\sum_{j>k}\frac{\left(\frac{Z}{\sqrt{2}}\right)^j}{j!}\right)$$

Let us consider the coefficient $a_n = [Z^n]A(Z)$, we can distinguish three different cases;

if $n = 0$ then $a_0 = 0$,

if $n = 2p+1$, then

$$a_{2p+1} = \frac{2}{(\sqrt{2})^{2p+1}(2p+1)!}\cdot\sum_{k=0}^{p}\binom{2p+1}{k}2^k$$

if $n = 2p$, then

$$a_{2p} = \frac{2}{2^p(2p)!}\cdot\sum_{k=0}^{p-1}\binom{2p}{k}2^k$$

so we can conclude that

$$\begin{aligned} a_{2p+1} &= c'_p\cdot\frac{2}{(p!)^2}\cdot\frac{\sqrt{2}}{p+1} \\ a_{2p} &= c''_p\cdot\frac{2}{(p!)^2} \end{aligned}$$

and it is straightforward to prove that $c'_p$ and $c''_p$ tend to 1 as $p$ tends to $\infty$. So we conclude that $G(Z, Z)$ is asymptotically equivalent to

$$3J_0(2iZ) + \sqrt{2} \cdot \frac{d}{dZ} J_0(2iZ)$$

and we get the statement of the lemma, by making the change of variable $Z = -\sqrt{2}\ln(1-z)$.

## References

[CDS89] R. Casas, J. Díaz, and J. M. Steyaert. Average-case analysis of Robinson's unification algorithm with two different variables. *Information Processing Letters*, 31:227–232, June 1989.

[CKS89] C. Choppy, S. Kaplan, and M. Soria. Complexity analysis of term-rewriting systems. *Theoretical Computer Science*, 67(2):261–282, 1989.

[Cop75] E.T. Copson. *Partial Differential Equations*. Cambridge University Press, Cambridge, 1975.

[DB58] N.G. De Bruijn. *Asymptotic Methods in Analysis*. Dover, New York, 1958.

[Dev86] L. Devroye. A note on the height of binary search trees. *J.ACM*, 33(3):489–498, July 1986.

[Fla88] P. Flajolet. Random tree models in the analysis of algorithms. In *Performance'87*, pages 171–187, 1988.

[FO82] P. Flajolet and A. Odlyzko. The average height of binary trees and other simple trees. *JCSS*, 25(2):171–213, Oct 1982.

[FS87] P. Flajolet and J.M. Steyaert. A complexity calculus for recursive tree algorithms. *Mathematical Systems Theory*, 19:301–331, 1987.

[FV87] P. Flajolet and J.S. Vitter. *Average-case Analysis of Algorithms and Data Structures*. Technical Report 718, INRIA, Aug 1987.

[GJ83] I. Goulden and D. Jackson. *Combinatorial Enumerations*. Wiley, New York, 1983.

[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.