Search Trees

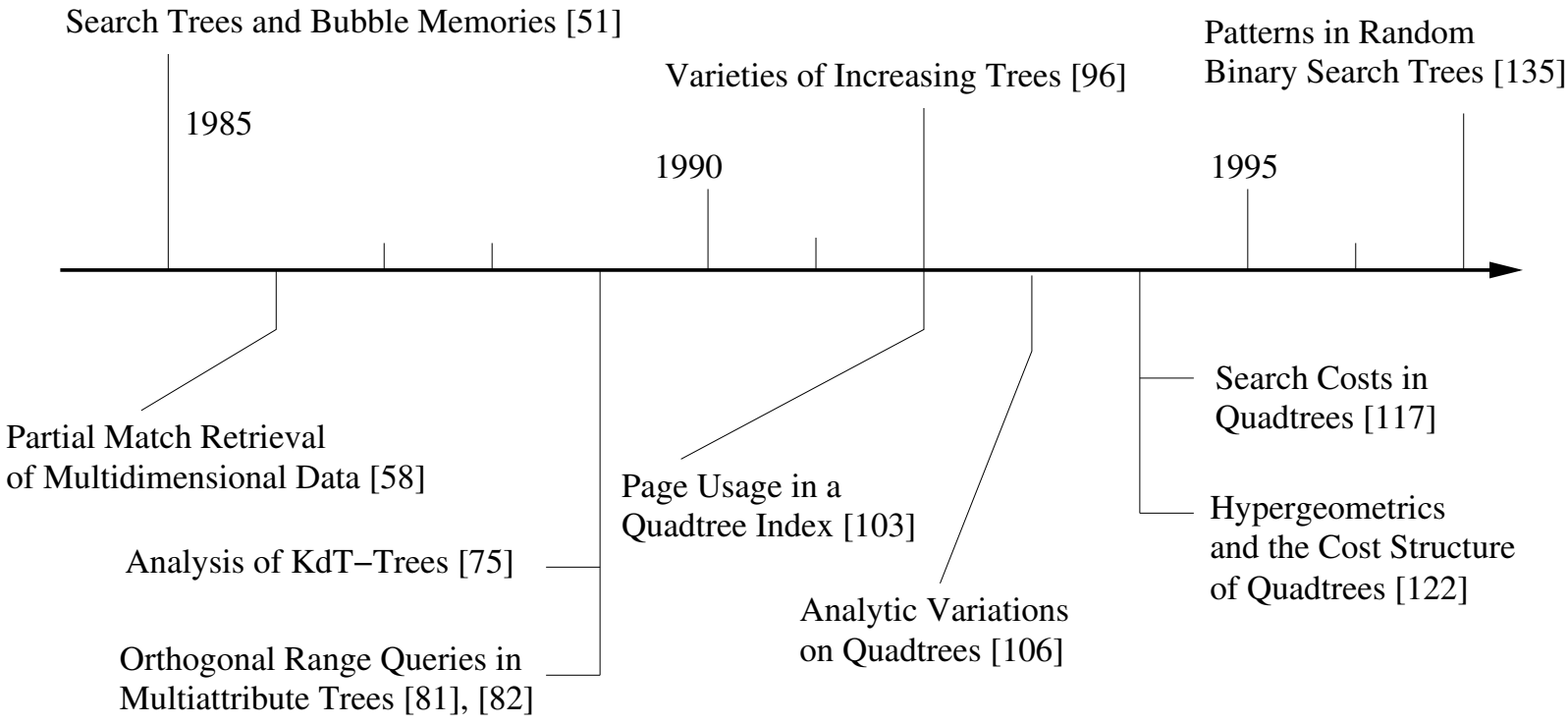
Conrado Martínez

Univ. Politècnica de Catalunya, Spain

Philippe Flajolet and Analytic Combinatorics

Paris, December 2011

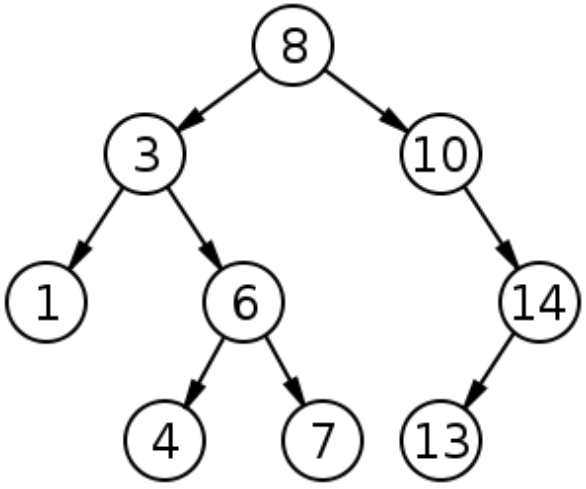
Timeline


Search Trees and Bubble Memories [51]
1985
Varieties of Increasing Trees [96]
1990
Patterns in Random
Binary Search Trees [135]
1995
Partial Match Retrieval
of Multidimensional Data [58]
Analysis of KdT−Trees [75]
Orthogonal Range Queries in
Multiattribute Trees [81], [82]
Page Usage in a
Quadtree Index [103]
Analytic Variations
on Quadtrees [106]
Search Costs in
Quadtrees [117]
Hypergeometrics
and the Cost Structure
of Quadtrees [122]

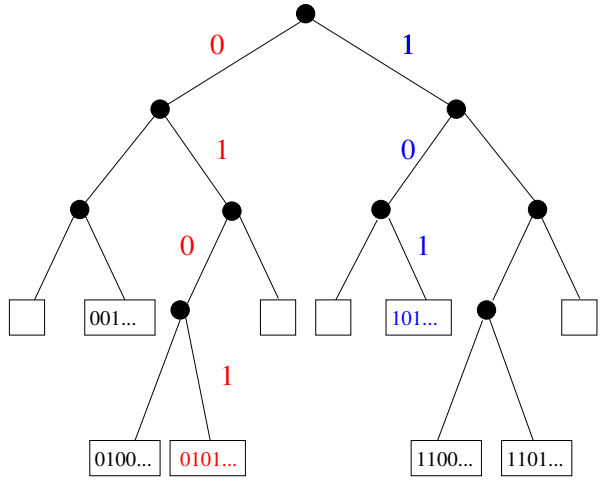
Search Trees

Search trees are fundamental data structures in all areas of Computer Science

1. Data-driven search trees, e.g., binary search trees
2. Space-driven search trees, e.g., tries



A binary search tree (BST)

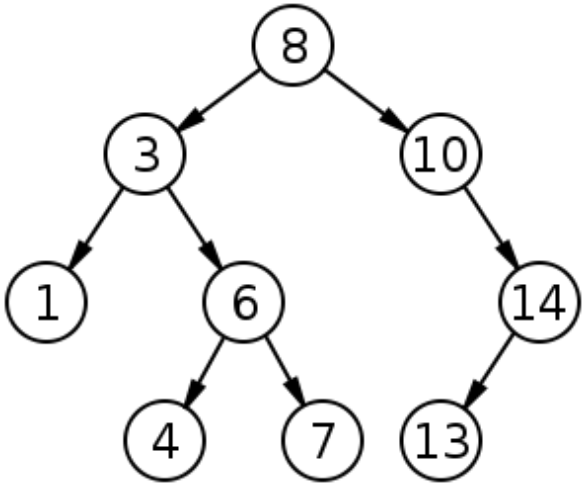


A binary trie

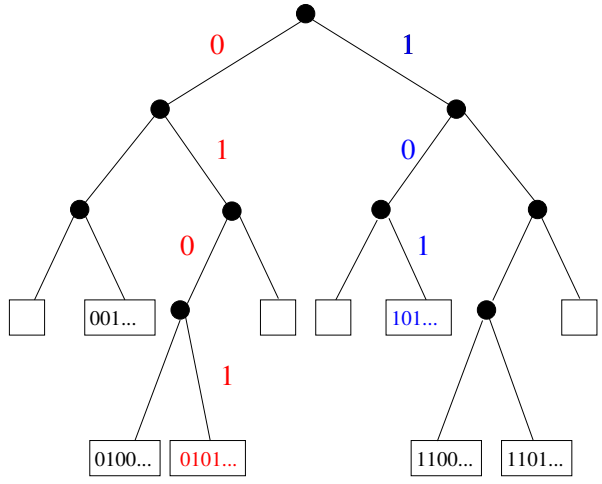
Search Trees

Search trees are fundamental data structures in all areas of Computer Science

1. Data-driven search trees, e.g., binary search trees
2. Space-driven search trees, e.g., tries



A binary search tree (BST)



A binary trie

Search Trees

In this talk we briefly review Philippe Flajolet's works on data-driven search trees:

1. Binary Search Trees and Relatives: [51], [96], [135]
2. Multidimensional Search Trees
   - Kd-trees: [58], [75]
   - Multiattribute trees: [81], [82]
   - Quadtrees: [103], [106], [117], [122]

Search Trees

In this talk we briefly review Philippe Flajolet's works on data-driven search trees:

1. Binary Search Trees and Relatives: [51], [96], [135]
2. Multidimensional Search Trees
   - Kd-trees: [58], [75]
   - Multiattribute trees: [81], [82]
   - Quadtrees: [103], [106], [117], [122]

Search Trees

In this talk we briefly review Philippe Flajolet's works on data-driven search trees:

1. Binary Search Trees and Relatives: [51], [96], [135]
2. Multidimensional Search Trees
   - Kd-trees: [58], [75]
   - Multiattribute trees: [81], [82]
   - Quadtrees: [103], [106], [117], [122]

Search Trees

In this talk we briefly review Philippe Flajolet's works on data-driven search trees:

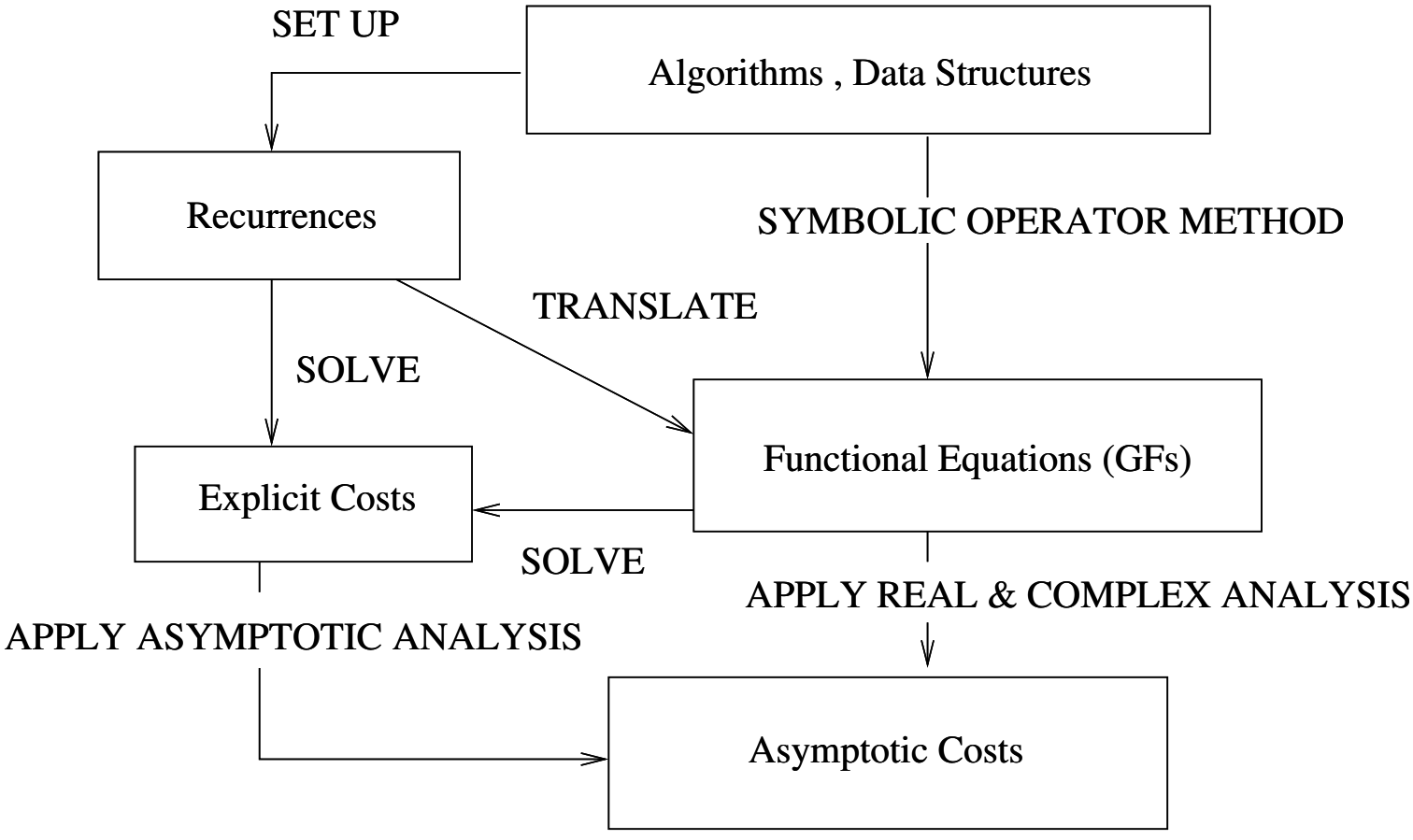
1. Binary Search Trees and Relatives: [51], [96], [135]
2. Multidimensional Search Trees
   - Kd-trees: [58], [75]
   - Multiattribute trees: [81], [82]
   - Quadtrees: [103], [106], [117], [122]

Search Trees

In this talk we briefly review Philippe Flajolet's works on data-driven search trees:

1. Binary Search Trees and Relatives: [51], [96], [135]
2. Multidimensional Search Trees
   - Kd-trees: [58], [75]
   - Multiattribute trees: [81], [82]
   - Quadtrees: [103], [106], [117], [122]

AofA in the 80s and early 90s



Search Trees and Bubble Memories

SEARCH TREES AND BUBBLE MEMORIES (*)

by Philippe FLAJOLET [1], Thomas OTTMANN [2] and Derick WOOD [3]

Communicated by J. BERSTEL

Abstract. – *We consider the storage of binary search trees in major-minor loop configurations of bubble memories. This leads, under reasonable assumptions, to the investigation of two cost measures for binary search trees, free search cost FCOST, and root-reset search cost RCOST. We analyze the average case behaviour of both cost measures and characterize their associated minimal cost trees. The average case average case analyses are themselves of interest since they are examples of the application of a recently developed methodology.*

T. Ottmann

D. Wood

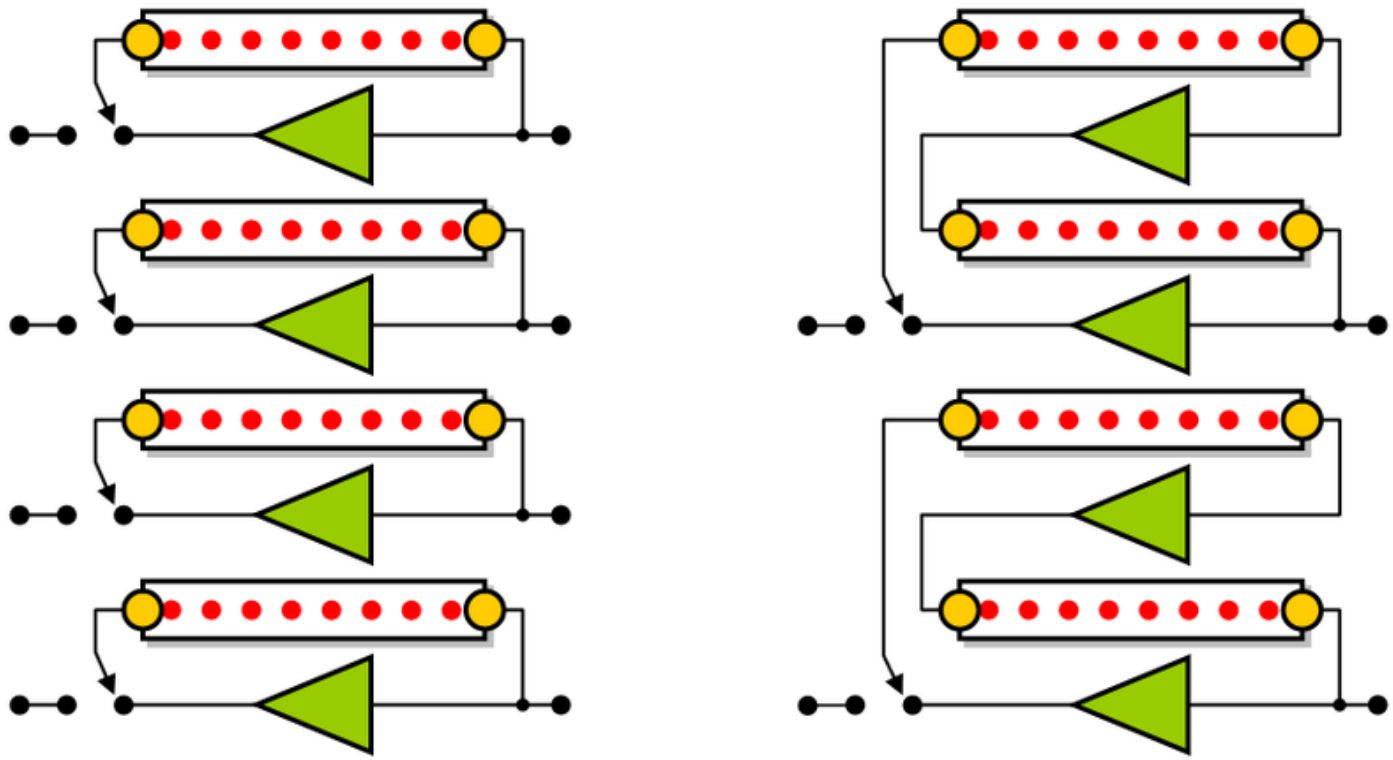
Philippe Flajolet, Thomas Ottmaan, and Derick Wood.
Search trees and bubble memories.
RAIRO. Informatique Théorique (Theoretical Informatics), 19:137–164, 1985.

Search Trees and Bubble Memories

Bubble memories emerged during the 70s as a very promising technology for permanent storage.
The improvements of magnetic disk technologies in the early 80s signaled the decline of the bubble memory.

Search Trees and Bubble Memories

One important problem at the time was how to deploy data structures in bubble memories in the most efficient way—taking advantage of some of its features and avoiding some of the pitfalls.



Search Trees and Bubble Memories

The most efficient way to store BSTs in bubble memories was a circular double-linked lists.
Two search strategies were contemplated giving rise to two different measures of cost:

$$\text{FCOST}(T) = \sum_{u,v \text{ in } T} \text{dist}(u,v),$$

$$\text{RCOST}(T) = \sum_{u,\ v \text{ in } T} \text{rdist}(u,v),$$

with $\text{rdist}(u,v) = \text{dist}(u,v)$ if v is a descendant of u in T, or
$\text{rdist}(u,v) = \text{dist}(\text{root},v) +$
path length of unsuccessful search of v in the subtree rooted at u

Search Trees and Bubble Memories

- The paper is fine early example of the symbolic method that Flajolet so firmly contributed to establish as one of the pillars of Analytic Combinatorics.
- It also anticipates the singularity analysis which would consolidate Philippe's international recognition and preminence, and so profoundly shaped the area.
- In many aspects, it is a very representative paper using the emerging "technologies" of the eighties, focusing in average-case complexity.

Search Trees and Bubble Memories

Example

Recursive definition:

$$p(\circ(t_1, t_2)) = p(t_1) + p(t_2) + |t_1| + |t_2|$$
$$d(\circ(t_1, t_2)) = d(t_1) + d(t_2) + 2(p(t_1) + |t_1|)(|t_2| + 1) + 2(p(t_2) + |t_2|)(|t_1| + 1)$$

Functional equations:

$$T(z) = \frac{1}{1-z}, \quad S(z) = zT'(z) = \frac{z}{(1-z)^2}$$
$$P(z) = 2\int_0^z (P(z) + S(z))T(z)dz$$
$$D(z) = 2\int_0^z D(z)T(z)dz + 4\int_0^z (P(z) + S(z))(T(z) + S(z)))T(z)dz$$

Varieties of Increasing Trees

VARIETIES OF INCREASING TREES

François Bergeron
LACIM
Université du Québec à Montréal
Case Postale 8888, Succursale A
Montréal, Québec H3C3P8
Canada

Philippe Flajolet
Algorithms Project
INRIA Rocquencourt
78153 Le Chesnay
France

Bruno Salvy
Algorithms Project
INRIA Rocquencourt
78153 Le Chesnay
France

Abstract

An increasing tree is a labelled rooted tree in which labels along any branch from the root go in increasing order. Under various guises, such trees have surfaced as tree representations of permutations, as data structures in computer science, and as probabilistic models in diverse applications.

We present a unified generating function approach to the enumeration of parameters on such trees. The counting generating functions for several basic parameters are shown to be related to a simple ordinary differential equation,

$$\frac{d}{dz}Y(z) = \phi(Y(z)),$$

which is non linear and autonomous.

Singularity analysis applied to the intervening generating functions then permits to analyze asymptotically a number of parameters of the trees, like: root degree, number of leaves, path length, and level of nodes. In this way it is found that various models share common features: path length is $O(n \log n)$, the distribution of node levels and number of leaves are asymptotically normal, etc.

F. Bergeron

B. Salvy

François Bergeron, Philippe Flajolet, and Bruno Salvy.
Varieties of increasing trees.
In Jean-Claude Raoult, editor, *Proceedings of the 17th Colloquium on Trees in Algebra and Programming (CAAP '92)*, volume 581 of *Lecture Notes in Computer Science*, pages 24–48, Berlin/Heidelberg, 1992. Springer.

Varieties of Increasing Trees

- An increasing tree is a labelled rooted tree such that the labels in the path from the root to any node in the tree are increasing. Increasing trees might be plane or non-plane.
- In *Varieties*, Bergeron, Flajolet and Salvy consider "simple families of increasing trees" (both plane and non-plane), developing an unified framework for their enumeration and the analysis of some of their fundamental parameters, e.g., the total path length

Varieties of Increasing Trees

- The paper extends to increasing trees the programme in the famous papers by Meir and Moon on simple families of trees
- Binary increasing trees (a.k.a. heap ordered trees, tournaments) are isomorphic to binary search trees
- Recursive trees is another important family failing within this framework

Varieties of Increasing Trees

- s_k = number of "symbols" of arity $k = |\mathcal{S}_k|$, $s_0 > 0$, $s_k > 0$ for some $k \geq 2$,

$$\mathcal{T} = \mathcal{S}_0 + \mathcal{S}_1 \times^{\square} \mathcal{T} + \mathcal{S}_2 \times^{\square} \mathcal{T} \times \mathcal{T} + \cdots$$

  $A \times^{\square} B$ denotes the boxed product of A and B: the smallest label must be attached to an atom in the A-component

- Plane trees: $\phi(u) = \sum_{k \geq 0} s_k u^k$
- Non-plane trees: $\phi(u) = \sum_{k \geq 0} s_k u^k / k!$
- Enumeration: $Y_n = n![z^n]Y(z) = \#$ increasing trees of size n

$$\frac{dY}{dz} = z\phi(Y(z))$$

Varieties of Increasing Trees

- For instance, the number Y_n of increasing trees of size n is, for polynomial varieties (ϕ is a polynomial of degree d)

$$Y_n \sim n! K \rho^{-n} n^{-(d-2)/(d-1)}$$

where $\rho = \int_0^\infty \frac{dt}{\phi(t)}$, K is a constant depending on the family

- The paper also analyzed inductive parameters, i.e., those that can be described by the following recursion:

$$c(t) = c(t_1) + \cdots c(t_r) + f(|t|)$$

for a tree t with subtrees $t_1, \ldots, t_r$

AofA in the 90s

- Singularity analysis coming of age —it is very well developed and understood, and becomes the building block for many advances in the 90s
- Pioneering works, e.g., Flajolet & Soria (1990) on Gaussian limiting distributions for the number of components in random combinatorial structures
- Hwang's quasi-power theorem in the mid 90s provides a powerful technique to easily establish Gaussian limiting distributions and rates of convergence to the CLT
- Increasing interest in limiting distributions; average complexity and variance are not the end of the picture

Patterns in Random Binary Search Trees

X. Gourdon

Patterns in Random Binary Search Trees

Philippe Flajolet,[1] Xavier Gourdon,[1] Conrado Martínez[2]
[1]*Algorithms Project, INRIA-Rocquencourt, F-78153 Le Chesnay, France*
[2]*Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Pau Gargallo, 5, E-08028 Barcelona, Spain*

Received 11 October 1996; accepted 30 January 1997

ABSTRACT: In a randomly grown binary search tree (BST) of size n, any fixed pattern occurs with a frequency that is on average proportional to n. Deviations from the average case are highly unlikely and well quantified by a Gaussian law. Trees with forbidden patterns occur with an exponentially small probability that is characterized in terms of Bessel functions. The results obtained extend to BSTs a type of property otherwise known for strings and combinatorial tree models. They apply to paged trees or to quicksort with halting on short subfiles. As a consequence, various pointer saving strategies for maintaining trees obeying the random BST model can be precisely quantified. The methods used are based on analytic models, especially bivariate generating function subjected to singularity perturbation asymptotics. *Random Struct. Alg.*, **11**, 223–244 (1997)

Key Words: binary search tree; limit distribution; pattern; random tree; singularity analysis

Patterns in Random Binary Search Trees

- This paper was the outcome of my collaboration with Xavier and Philippe on a few problems that I left open in my PhD
- The goal was to analyze parameters such as the number of occurrences of a given pattern u in a random binary search trees, much in the same way Flajolet and Steyaert did with simple families of trees

Patterns in Random Binary Search Trees

- Ocurrences of patterns in random BSTs, like in random strings, trees, permutations and many other combinatorial structures fail under the very general and ubiquitous Borges' paradigm: the number of occurences of a pattern u (or patterns in a finite collection) has a Gaussian limiting distribution
- All boils down to the study of

$$\frac{\partial}{\partial z}F(z,y) = F^2(z,y) + (y-1)\lambda(u)|u|z^{|u|-1},$$

with $\lambda(u)$ the probability to obtain a BST with shape u

- The change $F(z,y) = -w_z(z,y)/w(z,y)$ leads to a Riccati DE and the solution is

$$w(z,y) = A_m(z) - zB_m(z)$$

with $m = |u|$ and A_m, B_m cylinder (Bessel) functions of order $-1/(m+1)$ and $1/(m+1)$ resp.; singularity pertubation methods yield the various results about the limiting distribution

Patterns in Random Binary Search Trees

- The paper also studied the probability that a random BST does not contain occurrences of a pattern or very few (fixed k occurrences $\to$ Poisson), local limit laws for the number of occurrences, and the number K_n of distinct subtrees in a random BST

$$K_n \leq 4 \ln 2 \frac{n}{\ln n} \left(1 + O\left(\frac{\log \log n}{\log n}\right)\right)$$

A lower bound $\Omega(n/\log n)$ was proved by Devroye

Multidimensional Search Trees

1. K-dimensional Search Trees (K-d Trees)
2. Multiattribute Trees
3. Quadtrees

K-dimensional Search Trees

W. Cunto

C. Puech

Philippe Flajolet and Claude Puech.

Tree structures for partial match retrieval.
In *Proceedings of the 24th Annual Symposium on Foundations of Computer Science*, pages 282–288. IEEE Computer Society Press, 1983.
Subsumed by [PF058]

Philippe Flajolet and Claude Puech.

Partial match retrieval of multidimensional data.
Journal of the ACM, 33:371–407, 1986.

Walter Cunto, Gustavo Lau, and Philippe Flajolet.

Analysis of kdt-trees: kd-trees improved by local reorganisations.
In Frank Dehne, Jörg-Rüdiger Sack, and Nicola Santoro, editors, *Proceedings of the 1989 Workshop on Algorithms and Data Structures (WADS '89)*, volume 382 of *Lecture Notes in Computer Science*, pages 24–38. Springer, Berlin/Heidelberg, 1989.

K-dimensional Search Trees

W. Cunto

C. Puech

Philippe Flajolet and Claude Puech.

Tree structures for partial match retrieval.
In *Proceedings of the 24th Annual Symposium on Foundations of Computer Science*, pages 282–288. IEEE Computer Society Press, 1983.
Subsumed by [PF058]

Philippe Flajolet and Claude Puech.

Partial match retrieval of multidimensional data.
Journal of the ACM, 33:371–407, 1986.

Walter Cunto, Gustavo Lau, and Philippe Flajolet.

Analysis of kdt-trees: kd-trees improved by local reorganisations.
In Frank Dehne, Jörg-Rüdiger Sack, and Nicola Santoro, editors, *Proceedings of the 1989 Workshop on Algorithms and Data Structures (WADS '89)*, volume 382 of *Lecture Notes in Computer Science*, pages 24–38. Springer, Berlin/Heidelberg, 1989.

Partial match retrieval of multidimensional data

Partial Match Retrieval of Multidimensional Data

PHILIPPE FLAJOLET

INRIA, Rocquencourt, France

AND

CLAUDE PUECH

Université de Paris-Sud, Orsay, France and Ecole Normale Supérieure, Montrouge, France

Abstract. A precise analysis of partial match retrieval of multidimensional data is presented. The structures considered here are multidimensional search trees (k-d-trees) and digital tries (k-d-tries), as well as structures designed for efficient retrieval of information stored on external devices. The methods used include a detailed study of a differential system around a regular singular point in conjunction with suitable contour integration techniques for the analysis of k-d-trees, and properties of the Mellin integral transform for k-d-tries and extendible cell algorithms.

Categories and Subject Descriptors: F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems—*sorting and searching*; G.2.1 [**Discrete Mathematics**]: Combinatorics—*counting problems*; *generating functions*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process*

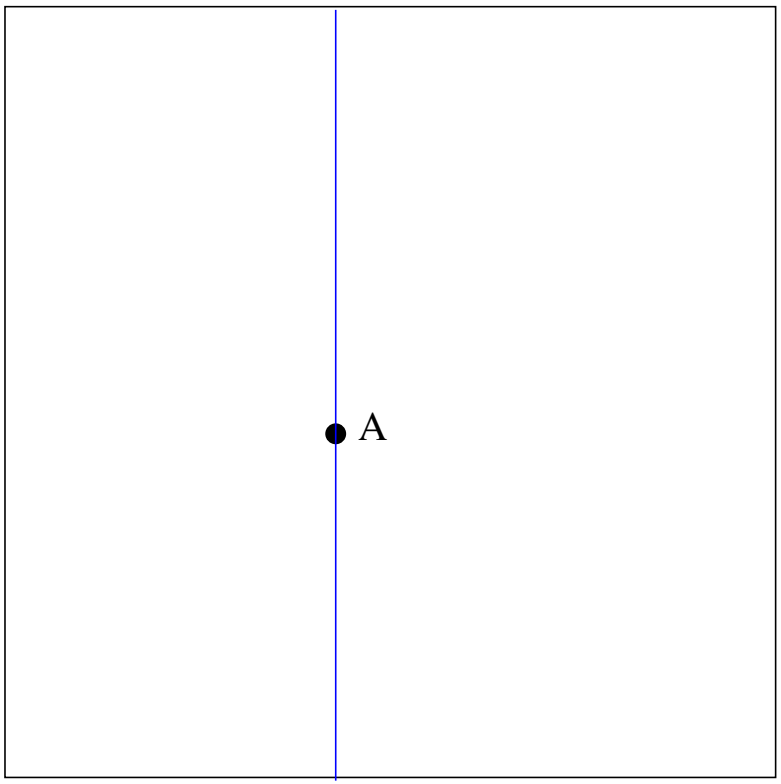
General Terms: Algorithms, Performance

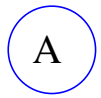
Additional Key Words and Phrases: Analysis of algorithms, data structures, multidimensional search, partial match, trees

Partial match retrieval of multidimensional data

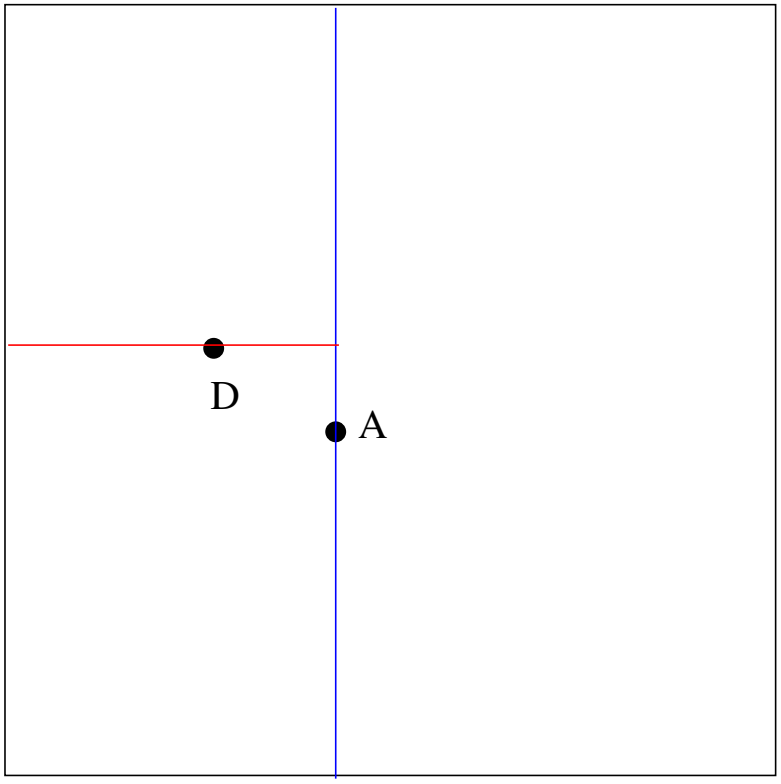
- [PF058] constitutes a landmark in the analysis of multidimensional data structures. Prior to the work of Flajolet and Puech, the analysis of partial matches, orthogonal range search, etc. assumed that, on average, the performance of the index (kd-tree, quadtree, ...) would be as if it were perfectly balanced—and this turned out to lead to wrong conclusions!
- The paper studied the performance of partial matches in kd-trees, kd-tries and grid files; the analysis showed that digital methods (=space-driven) outperform the comparison-based trees (=data-driven)

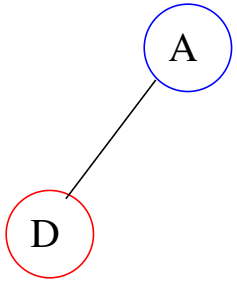
Partial match retrieval of multidimensional data



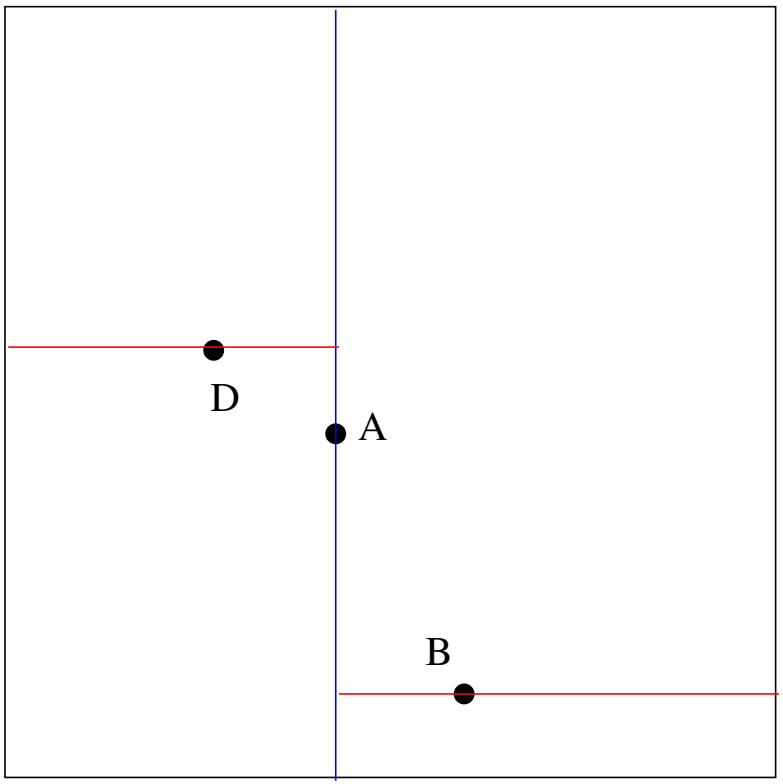


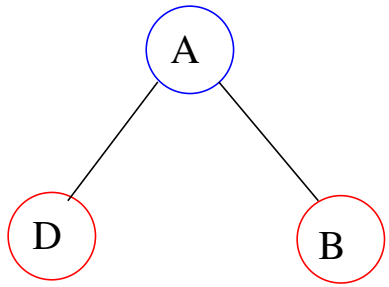
Partial match retrieval of multidimensional data



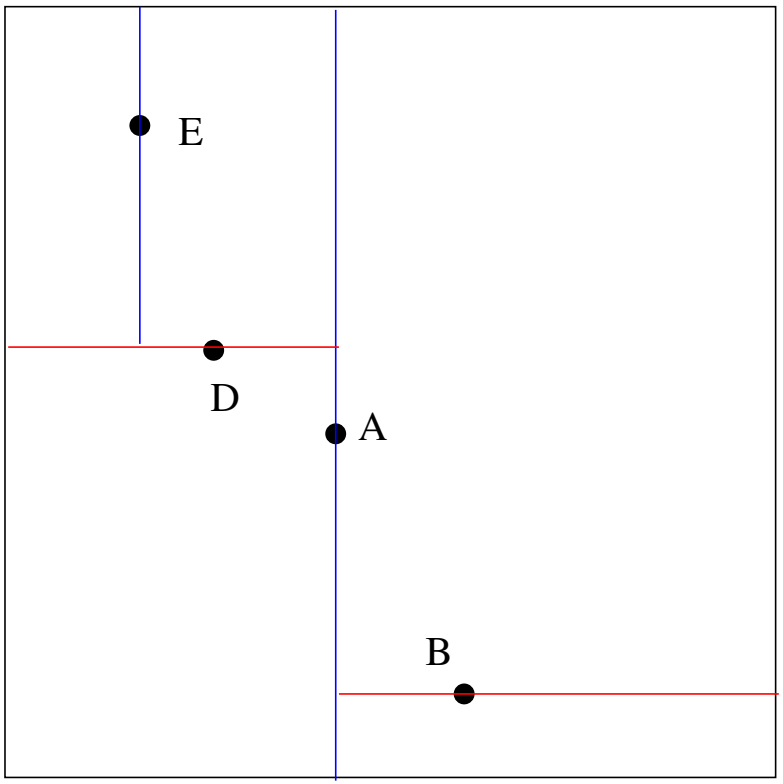


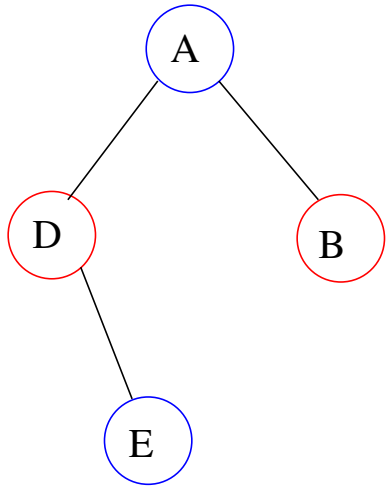
Partial match retrieval of multidimensional data



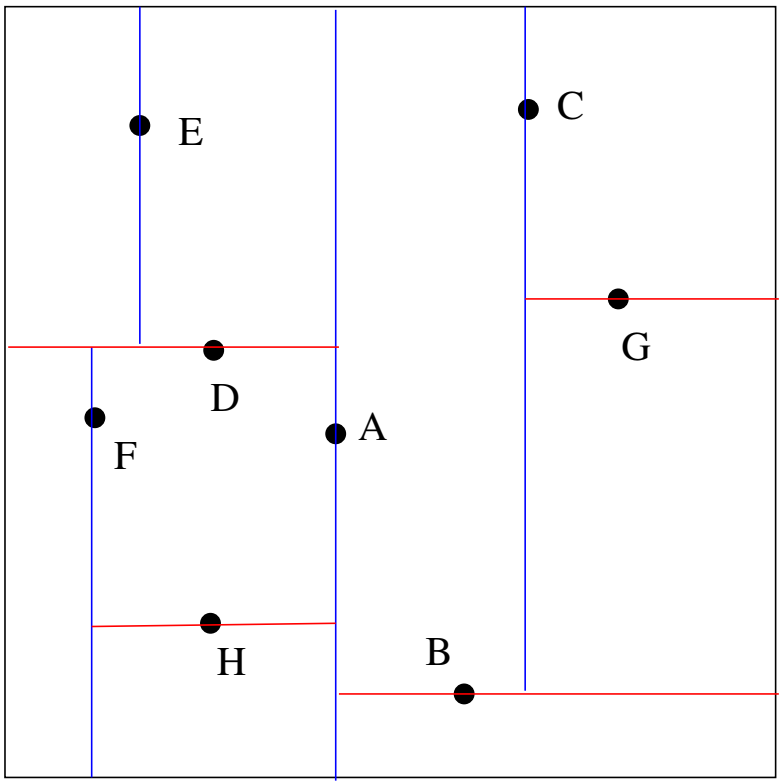


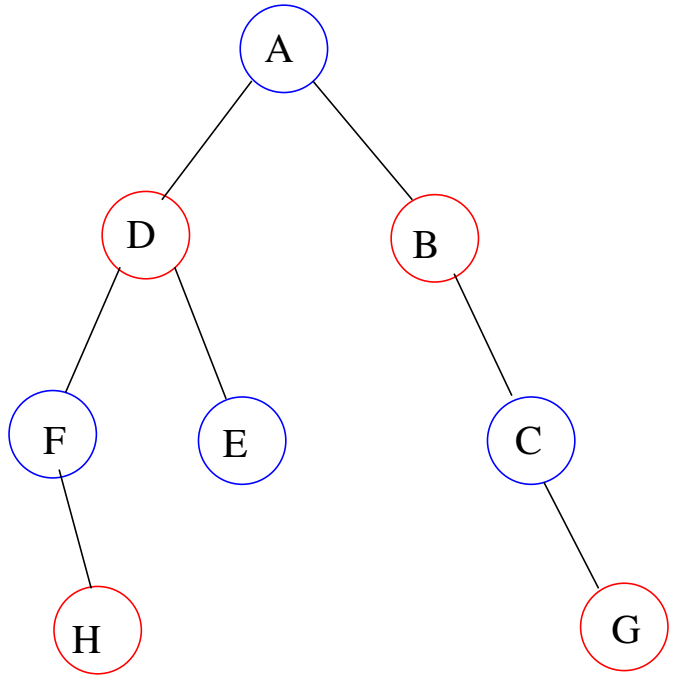
Partial match retrieval of multidimensional data





Partial match retrieval of multidimensional data





Partial match retrieval of multidimensional data

- The goal of a partial match query $q = (q_0, q_1, \ldots, q_{k-1})$ with $q_i \in (0, 1)$ or $q_i = *$ is to retrieve all records x in the kd-tree such that $x_i = q_i$ for all i such that $q_i \neq *$; when $q_i = *$ we say that the coordinate is not specified
- Partial match queries are assumed random, i.e., the specified q_i are independently and randomly drawn from $(0, 1)$
- A key observation is that the shape of a random kd-tree is the same as that of a random BST, i.e., any node has probability $1/n$ to be the root and subtrees are random independent kd-trees

Partial match retrieval of multidimensional data

- For the analysis, it is crucial to deal with query patterns u: binary strings that indicate which of the s coordinates are specified (S) and which of the remaining $k-s$ are not $(*)$
- The symbolic method + the BST model immediately yield

$$\frac{dc_{*\cdot v}(z)}{dz} = 2\frac{c_v(z)}{1-z} + \frac{1}{(1-z)^2},$$

$$\frac{d^2}{dz^2}\left(zc_{S\cdot v}(z)\right) = 2\frac{1}{1-z}\frac{d}{dz}(zc_v(z)) + \frac{2}{(1-z)^3},$$

where $c_u(z) = \sum_{n\geq 0}\mathbb{E}[P_{n,u}]\,z^n$, GF of expected cost of a partial match with query pattern u

Partial match retrieval of multidimensional data

- The problem can be cast as the solution of a system of linear differential equations
- The paper is a good example of singularity analysis at its best: we need no explicit solution to get asymptotic estimates for the coefficients (here expected cost of the partial match)
- The characteristic polynomial of the matrix gives the exponent α, and a delicate analysis allows to derive a full asymptotic expansion of $d_u(z) = (zc_u(z))'$ around the singularity $z = 1$

Partial match retrieval of multidimensional data

- The last part —a "preview" of what would be the systematized and fully developed singularity analysis and transfer lemmas— yields the asymptotics for the expected costs

$$\mathbb{E}[P_{n,u}] = \beta_u n^{\alpha(s/k)}(1 + o(1)),$$

where β_u is a constant depending on the query pattern and α is the unique positive solution of

$$(\alpha + 2)^s(\alpha + 1)^{k-s} = 2^k$$

Partial match retrieval of multidimensional data

- These results very quite surprising, as they showed that the often conjectured expected cost $\Theta(n^{1-s/k})$ is wrong: $\alpha(s/k) > 1 - s/k$
- The paper does not end here: the analysis is also carried out in kd-tries (the multidimensional analogue of binary tries) and grid-files to show that partial match in these digital data structures is more efficient on the average: $\mathbb{E}[P_n] = \Theta(n^{1-s/k})$
- An explicitly given periodic fluctuation multiplies the main order term in the expected cost of partial matches in both kd-tries and grid files: Mellin transform and residue computations are the key technologies involved here

Analysis of kdt-trees

- In [PF075], the analysis of partial match retrieval is extended to kdt-trees; in kdt-trees all fringe subtrees of size $\leq 2t+1$ are locally balanced
- The average cost of partial matches is

$$\mathbb{E}[P_n] = \beta n^{\alpha_t(s/k)}(1+o(1)),$$

with β a constant depending on the specific pattern of search and t, and $\alpha = \alpha_t(x)$ the unique positive solution of

$$((\alpha+t+1)^{\overline{t+1}})^{1-x}((\alpha+t+2)^{\overline{t+1}})^{x} = (t+2)^{\overline{t+1}},$$

$$x^{\overline{k}} = x(x+1)\cdots(x+k-1)$$

- The paper also was the first to give $\mathbb{V}[P_n] = \Theta(n^{2\alpha})$

N.B. P_n is the cost of an idealized partial match in which the query specified values are randomly picked at each subtree of the recursion, hence results about variance must be taken with a grain of salt

Multiattribute Trees

D. Gardy

Danièle Gardy, Philippe Flajolet, and Claude Puech.

On the performance of orthogonal range queries in multiattribute and doubly chained trees.
In Frank Dehne, Jörg-Rüdiger Sack, and Nicola Santoro, editors, *Proceedings of the 1989 Workshop on Algorithms and Data Structures (WADS '89)*, volume 382 of *Lecture Notes in Computer Science*, pages 218–229. Springer, Berlin/Heidelberg, 1989.

Danièle Gardy, Philippe Flajolet, and Claude Puech.

Average cost of orthogonal range queries in multiattribute trees.
Information Systems, 14:341–350, 1989.
Paper [GFP89b] is closely related.

Orthogonal Range Queries in Multiattribute Trees

- Multiattribute trees are a sort of "hybrid" data structure, closer to digital search methods than to comparison based
- Each record in the collection is a k tuple $R = (r_1, \ldots, r_k)$ where each r_i belongs to a finite domain ($\equiv$ alphabet) D_i
- The multiattribute tree is basically a trie, and double chained trees are an implementation equivalent to list-tries

Orthogonal Range Queries in Multiattribute Trees

- Orthogonal range queries are modeled as k ranges, one per domain, with the probability $p_i(m)$ that the ith range of the query is of size m
- The probability distribution for records stems from the assumption that each attribute is independently chosen, with probability $1/|D_i|$ for the value of the ith attribute

Orthogonal Range Queries in Multiattribute Trees

- The results of the paper rely heavily in symbolic methods; here is a typical result: the expected cost of a random query in a multiattribute of size n is

$$C_n = 1 + \sum_{j=1}^{k-1} \overline{m_1 m_2} \cdots \overline{m_j} \left(1 - \frac{\binom{d-d_{>j}}{n}}{\binom{d}{n}} \right)$$

$$d_i = |D_i|, d_{>j} = d_{j+1} d_{j+2} \cdots d_k, d := d_{>0}$$

Orthogonal Range Queries in Multiattribute Trees

- Paginated variants and pruned variants are also investigated; and [PF082] also gives formulas for simple and partial match queries (corollaries)
- The journal paper [PF082] explores the asymptotic behavior under two possible scenarios: 1) n fixed with respect the size of the domains, 2) n grows as the size of the domains grows
- The conference paper [PF081] hinted at an interesting question which was not further explored in [PF082]: the effect of the ordering of the attributes in the cost of orthogonal range query

Quadtrees




G. Gonnet G. Labelle L. Laforest J.-M. Robson

Philippe Flajolet, Gaston Gonnet, Claude Puech, and John Michael Robson.

The analysis of multidimensional searching in quad–trees.
In *Proceedings of the Second Annual ACM–SIAM Symposium on Discrete Algorithms (SODA '91)*, pages 100–109, 1991.
Subsumed by [PF106].

Mamoru Hoshi and Philippe Flajolet.

Page usage in a quadtree index.
BIT, 32:384–402, 1992.

Philippe Flajolet, Gaston Gonnet, Claude Puech, and John Michael Robson.

Analytic variations on quadtrees.
Algorithmica, 10:473–500, 1993.

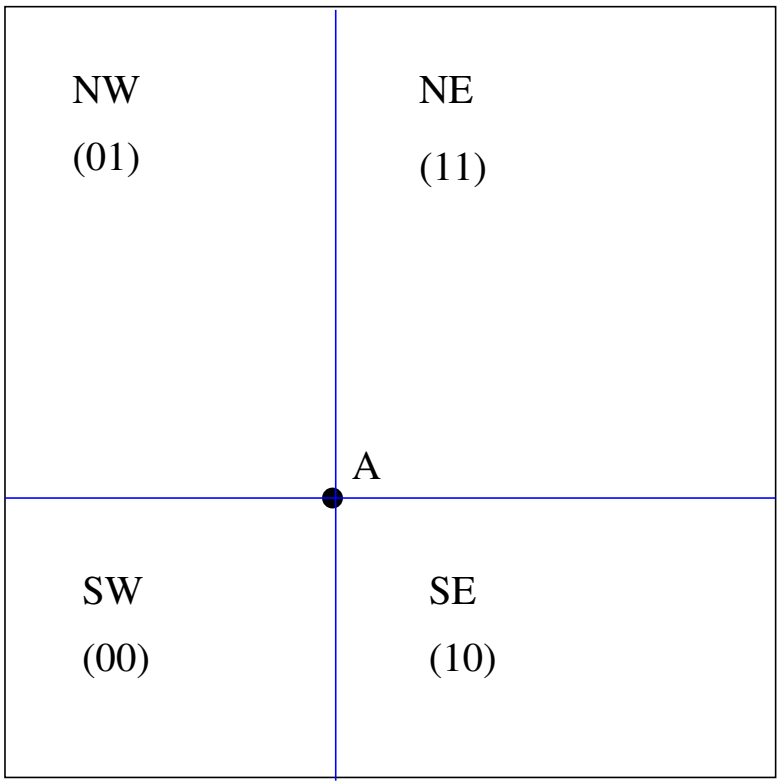
Philippe Flajolet and Thomas Lafforgue.

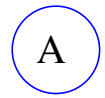
Search costs in quadtrees and singularity perturbation asymptotics.
Discrete & Computational Geometry, 12:151–175, 1994.

Philippe Flajolet, Gilbert Labelle, Louise Laforest, and Bruno Salvy.

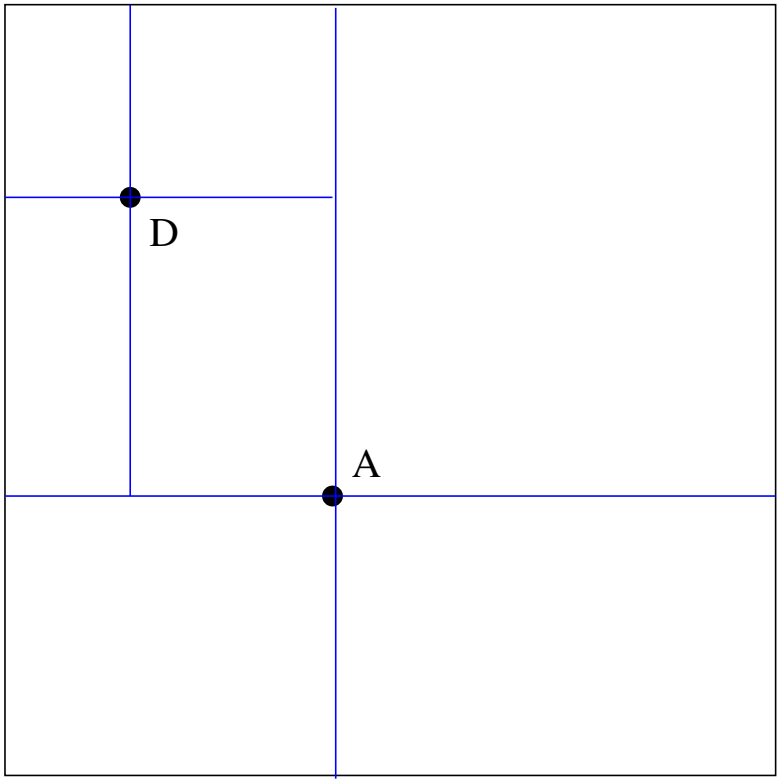
Hypergeometrics and the cost structure of quadtrees.
Random Structures & Algorithms, 7:117–144, 1995.

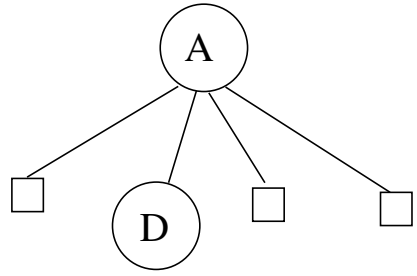
Quadtrees



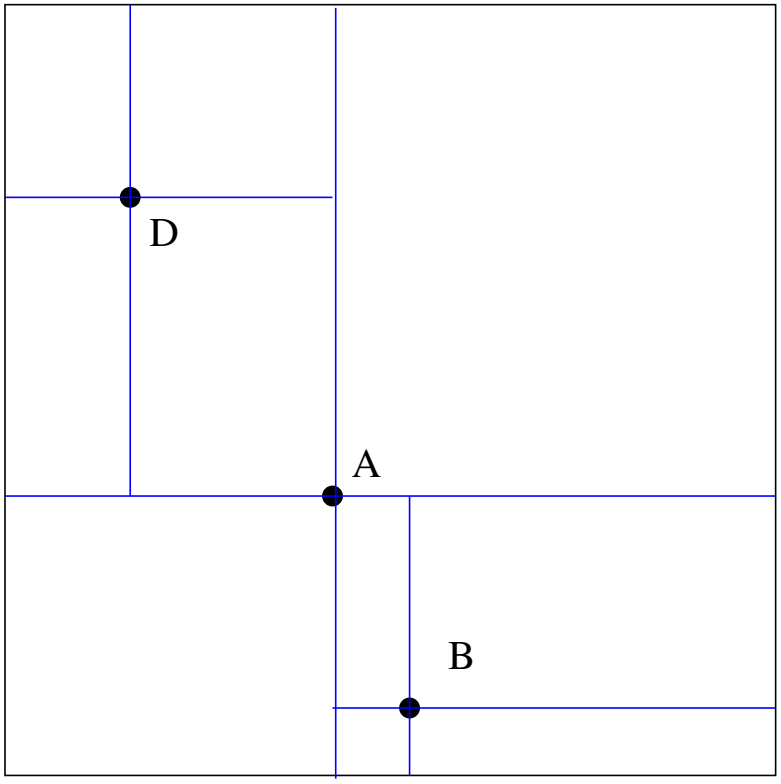


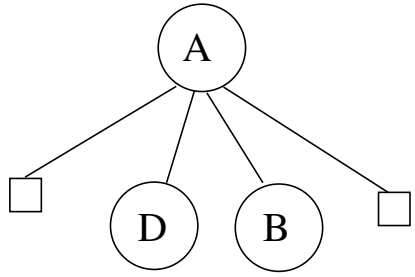
Quadtrees



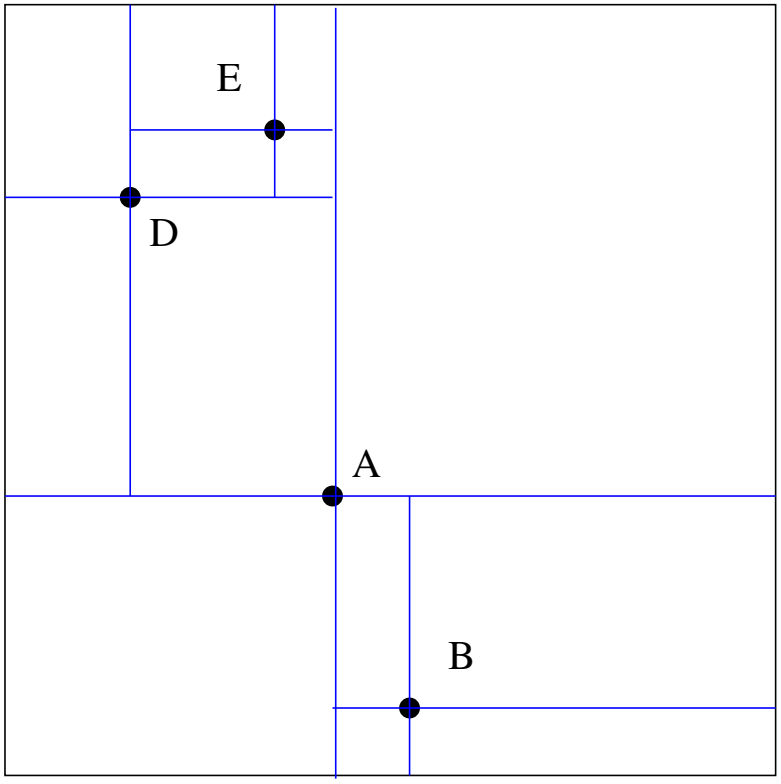


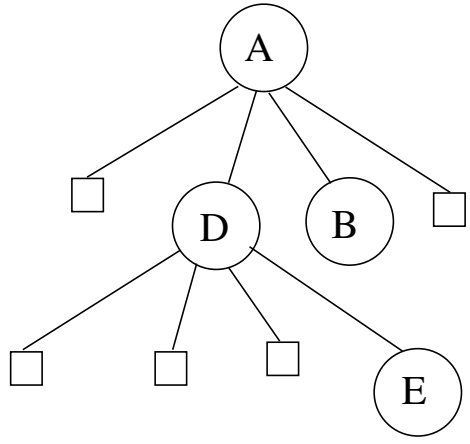
Quadtrees



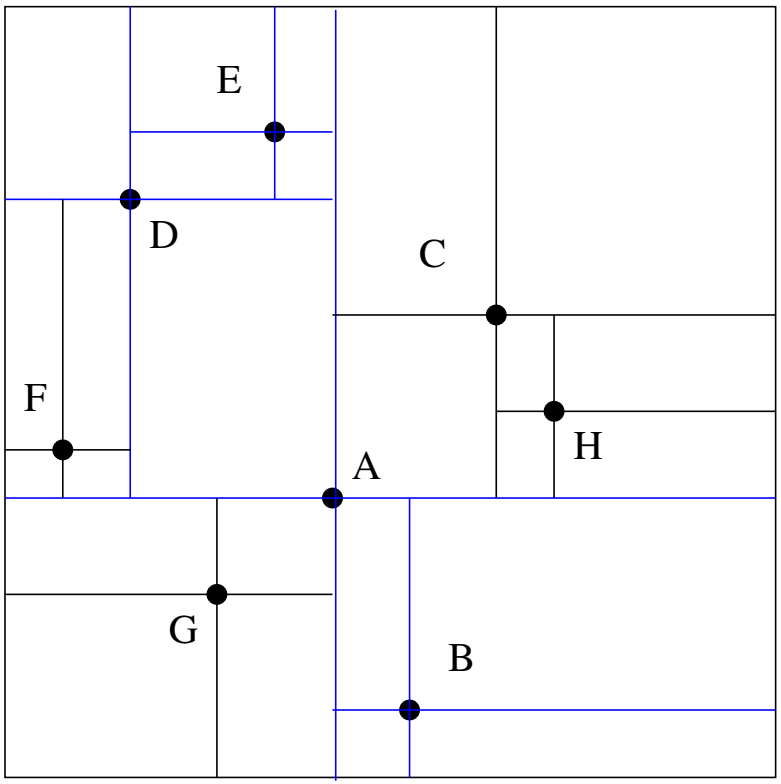


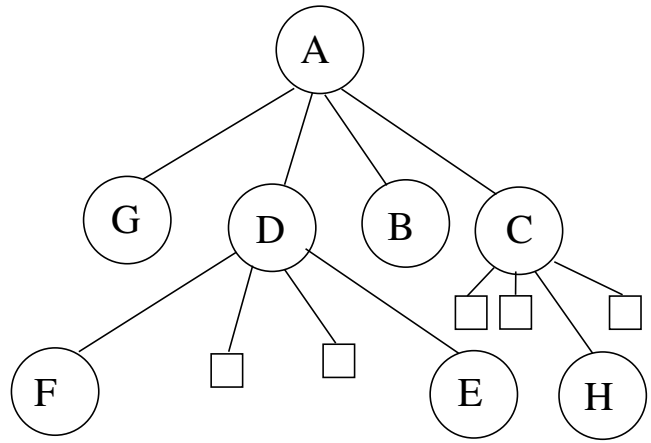
Quadtrees





Quadtrees





Page Usage in Quadtree Indexes

- [PF103] considers the filling ratio in a bucketed 2d quadtree; when the number of points in a region is $\leq B$, we do not subdivide any further and the elements are stored in a bucket; Hoshi and Flajolet showed that this filling ratio (# of items per bucket) is about 33%; some other questions are also addressed like the number of occurrences of a fixed pattern in a random quadtree —we would consider similar questions on random BSTs a few years later in [PF135]
- The techniques involved are now fairly standard, in particular, clever applications of singularity analysis to obtain the expected number of buckets ($\sim \gamma_B n$) or to analyze the asymptotic behavior of the filling ratios ($\gamma_B \to 3/B$)

Page Usage in Quadtree Indexes

- This paper was one early example were the use of Maple for the analysis of algorithms was crucial, as indicated in many points in the paper and

  > *"The authors would like to acknowledge the constant help and support of the Maple system that might well have been a coauthor of the paper . . ."*

- Philippe has been interested and made many theoretical and practical contributions to Computer Algebra; and he was also a very proficient Maple user

Analytic Variations on Quadtrees

- *Analytic Variations on Quadtrees* is the first rigorous and most complete analysis of the performance of several important operations on quadtrees until the date of its publication; there was a preliminary version in SODA 91 [PF093]
- The paper covers the analysis of the expected performance of two fundamental operations on quadtrees of dimension d: exact search and partial match

Analytic Variations on Quadtrees

- The general framework laid out in the paper is that the expected costs of interest can be cast as

$$C_n = t_n + \sum_{0 \leq k < n} \pi_{n,k} C_k,$$

  for suitably chosen t_n and splitting probabilities $\pi_{n,k}$
- Under the model of randomly and independently drawn points in $[0,1]^d$ this entails, for instance for $d = 2$, $t_n = n$ and $\pi_{n,k} = \frac{4}{n}(H_n - H_k)$ for path length and $t_n = 1$ and $\pi_{n,k} = \frac{4(n-k)}{n(n+1)}$ for partial match
- Similar, but more complicated forms for the $\pi_{n,k}$ can be found for larger dimensions

Analytic Variations on Quadtrees

- The divide-and-conquer recurrences giving expected costs can be translated into linear (integro)differential systems of equations
- For $d = 2$ the corresponding ODEs can be easily and explicitly solved yielding exact and asymptotic expressions for the expected cost of a search C_n and a partial match P_n

$$C_n = H_n - \frac{n+1}{6n} + \frac{H_n}{3n} = \ln n + O(1)$$

$$P_n = \gamma n^{\alpha},$$

$$\alpha = (\sqrt{17} - 3)/2 \approx 0.56, \quad \gamma = \frac{\Gamma(2\alpha + 2)}{2\Gamma^3(\alpha + 1)} \approx 1.595$$

- For larger dimensions the systems where not explicitly solved, but singularity analysis can be carried out nevertheless, much in the same way it was done for the analysis of partial match in k-d trees

Analytic Variations on Quadtrees

- The fundamental theory behind this is that the singularities of the solution to the system arise from the singularities of the coefficients
- A particular important case is when the coefficient matrix is meromorphic and the singularity is a simple pole $\rightarrow$ regular singularity
- The solutions to the homogeneous system are then of the form

$$(1-z)^{-\alpha} \sum_{n \geq 0} c_n (1-z)^n,$$

  with α the roots of a polynomial (the indicial equation)
- When the difference of two roots is integers then a second family of solutions involing a $\log(1-z)$ enters teh picture
- The particular solutions to the inhomogeneous system are obtained via the variation of constants method
- This path provides the sought answers, in particular, $C_n = \frac{2}{d} \ln n + O(1)$ and $P_n = \gamma_d n^{\alpha(s/d)}$, the exponent α being the same as for partial match in k-d trees

Search Costs in Quadtrees and Singularity Perturbation Asymptotics

- The paper [117] focuses in the costs of random successful searches C_n and random unsuccessful searches D_n (depth of insertion) in random quadtrees of size n
- Flajolet and Lafforgue show that suitably normalized versions of both RVs weakly converge to normal distributions
- They also show uniform exponential tails for the probability of large deviations of D_n and a local limit theorem also exists

Search Costs in Quadtrees and Singularity Perturbation Asymptotics

- The fundamental methodological contribution is the singularity perturbation method; under suitable conditions, if we have an expansion $F(z,u) \sim c(u)(1-z)^{-b(u)}$ around the dominant singularity $z=1$, uniform in u, then $[z^n]F(z,u) = p_n(u) \sim c(u)\frac{n^{b(u)-1}}{\Gamma(b(u))}$, directly leading to a Gaussian limit distribution
- The main difficulty was indeed the singular expansion of $F(z,u)$ (in turn a solution to a system of linear differential equations) uniform in u; this involves a delicate analysis of the behavior of the expansion for u near 1, i.e., the "perturbation" around $u=1$
- In the particular case of quadtrees of dimension $d \geq 3$, this necessitated separately studying the cases of odd dimension (easy) and of even dimension (integer differences between the eigenvalues of the system when $u=1$, no such difference when $u \neq 1 \rightarrow$ "different" expansions of $F(z,u)$!)

Hypergeometrics and the Cost Structure of Quadtrees

- [PF122] is a sort of recap of the previous papers, with a general framework to investigate additive parameters in quadtrees (that includes internal path length, storage occupation, ...)
- The key idea is to work with the Euler transform of the corresponding GFs

$$f^*(z) = \mathcal{E} f(z) = \frac{1}{1-z} f\left(\frac{z}{z-1}\right)$$

Hypergeometrics and the Cost Structure of Quadtrees

- Under such transformation, the diff. equation satisfied by $f^*(z)$ is particularly simple and solvable in terms of generalized hypergeometric functions, allowing explicit computation of the expected values

$$f_n^* = [\![n]\!]! \sum_{2 \le j \le n} \frac{t_j^* - t_{j-1}^*}{[\![j]\!]!}$$

$$[\![n]\!]! = \prod_{j=3}^{n} \left(1 - \frac{2^d}{j^d}\right), \quad [\![2]\!]! = 1$$

$$f_n = \sum_k \binom{n}{k} (-1)^k f_k^*, \qquad t_n^* = \sum_k \binom{n}{k} (-1)^k t_k,$$

Hypergeometrics and the Cost Structure of Quadtrees

- The asymptotic behavior of f_n can be estimated through the Mellin-Lindelöf integral representation of $f^*(z)$; collecting the residues σ in a strip $a < \Re(s) < n_0$

$$f^*(z) = \sum_{n \geq n_0} (-1)^n \phi(n) z^n = -\sum_{\sigma} \operatorname{Res}\left(\phi(s) z^s \frac{\pi}{\sin \pi s}\right) + O(t^a),$$

$$z \to \infty$$

- Or via singularity analysis of the expansion of $f(z)$ around $z = 1$; for additive costs, $f(z)$ can be explicitly expressed in terms of hypergeometric functions

Epilogue

This survey spans several papers (10-13) published over ∼15 years, from 1983 to 1997, an exciting period for AofA

- Singularity analysis
- Techniques to prove limit distributions + AofA proves the normal distribution is normal :)
- Birth of AofA: Dagstuhl 1993
- ...

During that period Philippe made many fundamental methodological contributions, but he always kept interest and made significant contributions on applications, harmonically blending theory and practice. Philippe gave many examples on how applications (for instance, relevant parameters in different families of search trees) could be analyzed with scientific rigor and unprecedented precision and insight, and developed beautiful but purposeful mathematical techniques in the way