Internationales Begegnungs- und Forschungszentrum Für Informatik

Schloß Dagstuhl

Seminar Report

Average Case Analysis of Algorithms

July 7 - 11, 1997

Context. During the week of July 7–11, 1997, the Third International Seminar on "Average-Case Analysis of Algorithms" was held at Schloß Dagstuhl. This Seminar followed two earlier meetings on average-case analysis of algorithms held in Schloß Dagstuhl in July 1993 and 1995[1].

Analysis of algorithms aims at a precise prediction of the expected performance of algorithms under well-defined randomness models of data. What is sought is a precise description of the average-case behaviour of algorithms, often a very meaningful practical measure of complexity as it permits us to quantify precisely implied constants, as well as compare and finely tune implementations. In Knuth's words:

> "People who analyze algorithms have double happiness. First of all they experience the sheer beauty of elegant mathematical patterns that surround elegant computational procedures. Then they receive a practical payoff when their theories make it possible to get other jobs done more quickly and more economically."

The field of average-case analysis of algorithms is closely related to that of random structures and their quantitative properties. Two major classes of methods prevail here, namely *analytic methods* and *probabilistic methods*. Probabilistic methods are especially effective in the case of problems of high complexity: for instance, in combinatorial optimization, models like random graphs shed light on the behaviour of exponential time algorithms, typical examples being graph colouring or hamiltonian paths. The Dagstuhl seminar was to a large extent devoted to an in-depth investigation of analytic methods that are based on generating functions and asymptotic analysis. Such methods are especially effective in the case of decomposable combinatorial structures and they go well with modularity in data structures as well as with algorithms of low polynomial complexity. Boundaries are however not rigid: on the one hand, some polynomial time algorithms are being analysed by probabilistic methods; on the other hand, analytic methods have recently proved capable of yielding probability distributions estimates as well as attacking random graph problems.

The field of analysis of algorithms draws its problems from computer science and its methods from mathematics. Its relevance to practical computer science is clear. Quantitative analyses percolate into classical treatises

[1] See Seminar No. 9328, Report No. 68 (Ph. Flajolet, R. Kemp, H. Prodinger), and Seminar No. 9527, Report No. 119 (Ph. Flajolet, R. Kemp, H. Prodinger, R. Sedgewick).

and profoundly help us shape up efficient implementations; see the books by Knuth (*The Art of Computer Programming*) and Sedgewick (*Algorithms*), each printed at about 200,000 copies. Many complexity-sensitive applications, like massive indexing of data or communication in local and large area networks clearly benefit from analytic results through such books or through direct channels. Instances are the recent use of tries and hashing in the design of ultra highspeed routers or the "ternary tree protocol" (normalized as IEEE 802.14) for internet access on cable networks.

The Dahstuhl seminar series has played an important rôle in structuring the analysis of algorithms community. (Although the keyword "analysis of algorithms" is present in most theoretical computer science conferences, this community did not have a "home" prior to the three Dagstuhl seminars devoted to the subject.) First, a special effort has been devoted to making the results of these seminars widely available through the publication of special issues of journals, namely:

— Special issue on Mathematical Analysis of Algorithms, H. Prodinger and W. Szpankowski editors. In *Theoretical Computer Science*, volume 144 (1–2), 1995, 322 pages.

— Special issue on Average-Case Analysis of Algorithms, P. Flajolet and W. Szpankowski editors. In *Random Structures and Algorithms*, volume 10 (1–2), 1997, 303 pages.

— Special issue on Average-Case Analysis of Algorithms, H. Prodinger and W. Szpankowski editors. In *Algorithmica*, volume in preparation, 1999.

Second, the Dagstuhl series of meetings have been instrumental in shaping up the community. An outcome is the plan elaborated at Dagstuhl in July 1997 to set up a collection of home pages on the world-wide web, at

`http://www-rocq.inria.fr/algo/AofA/index.html`

and to initiate a regular series of meetings. In particular, continuations of the series are scheduled in July 1998 in Princeton (under the auspices of the DIMACS special year on massive data sets), and in Barcelona in 1999.

Overview. The twenty-eight abstracts in this report illustrate the vitality of the subject. They range from methodological to applied, covering such diverse problems as string matching and computational biology, hashing, tree data structures, selection problems in statistics, data compression and information-theory, adaptive data structures and learning, real-time and systems programming, as well as relations with computer algebra.

General methodology is often approached under a perspective that combines probabilistic intuition and analytic insight, and it is discussed in several talks. Urn models underly a great many applications, like hashing, allocations, or learning (Gardy); Brownian motion can be put to good use in the analysis of tree parameters (Louchard), while many properties of random trees in discrete mathematics are unified by a general theory that is distinctively different from branching processes (Devroye). Alternating sums, present in several algorithms, can be treated systematically by complex analysis. Many saddle point computations that arise in combinatorics can be performed automatically by a computer algebra system (Salvy), though hard problems arise in asymptotic enumerations when analysing constrained set partitions (Odlyzko). The classical framework of "analytic combinatorics" is capable of characterizing limit distributions of a great many parameters of combinatorial structures (Flajolet).

Patterns in strings are of interest for information retrieval, indexing, and also computational biology. We have by now a fair understanding of pattern occurrences in random strings (Régnier). Formal languages also lead to interesting enumerative problems (Nebel), some of which are related to random generation (Liebehenschel), where a computer algebra system like Maple offers an interesting functionality. Patterns in strings are also closely related to information theory and to digital tries that we know how to analyse under a rich variety of models (Vallée). Many problems in information theory, most notably compression, can in fact be now subjected to analytic methods, resulting in the emergence of "analytic information theory" (Szpankowski).

Trees are the data structure *par excellence*. The combinatorial models still pose intriguing questions. Very recent solutions to the problem of the width —equivalently the queue size in a traversal— have been found (Gittenberger), while balanced trees can be analysed under this model (Kemp). The search tree model is of direct impact on data structuring. We now see that height can be attacked by analytic methods (Drmota). This model also applies to quicksort and quickselect, where multiple selection (Mahmoud),

optimal sampling (Martínez), and locality of search (Prodinger) can be precisely analysed. Sorting poses problems, some similar but others quite different when one takes into account measures of presortedness (Hwang) or fast approximate sorting by networks (Sedgewick).

Hashing algorithms constitute direct access methods with a very high service rate under favorable probabilistic conditions. In critical applications (*e.g.*, routers), it is often crucial to quantify precisely the "risk" of long delays: a definitive solution to the variance analysis of the linear probing strategy was presented for the first time at the seminar (Poblete, Viola).

Finally, several talks illustrate the diversity of topics where analysis of algorithms may be applied. Instances are computational geometry (Golin), self-organizing search (Fill), reservation policies in communication systems (Coffman), communication protocols (Jacquet), and clock synchronization in real-time systems (Schmid).

The Organizers,
P. Flajolet (Paris),
R. Kemp (Frankfurt am Main),
H. Mahmoud (Washington),
H. Prodinger (Wien)

Participants

Rafael Casas, Barcelona
Ed Coffman, Murray Hill
Luc Devroye, Montreal
Michael Drmota, Wien
James A. Fill, Baltimore
Philippe Flajolet, Le Chesnay
Danièle Gardy, Versailles
Bernhard Gittenberger, Wien
Mordecai Golin, Hong Kong
Hsien-Kuei Hwang, Taiwan
Philippe Jacquet, Le Chesnay
Rainer Kemp, Frankfurt
Peter Kirschenhofer, Leoben
Jens Liebehenschel, Frankfurt
Guy Louchard, Bruxelles
Hosam M. Mahmoud, Washington
Conrado Martínez, Barcelona
Markus Nebel, Frankfurt
Andrew M. Odlyzko, Florham Park
Patricio V. Poblete, Santiago
Helmut Prodinger, Wien
Mireille Régnier, Le Chesnay
Bruno Salvy, Le Chesnay
Ulrich Schmid, Wien
Robert Sedgewick, Princeton
Robert Smythe, Washington
Jean-Marc Steyaert, Palaiseau
Wojciech Szpankowski, W. Lafayette
Brigitte Vallée, Caen
Alfredo Viola, Montevideo

Contents

Markus Nebel
On the Average Complexity of the Membership Problem for a Generalized Dyck Language
Andrew M. Odlyzko
On Set Partitions with Unequal Block Sizes
Patricio V. Poblete
Transform Methods for the Analysis of Algorithms
Helmut Prodinger
On the Analysis of Ascendants and Descendant for Various Families of Trees
Mireille Régnier
Some Limit Theorems in Pattern Matching
Bruno Salvy
Saddle-Point Method and Computer Algebra
Ulrich Schmid
Challenges in Interval-Based Clock Synchronization
Robert Sedgewick
Open Problems in the Analysis of Sorting and Searching Algorithms
Robert Smythe
Probabilistic Analysis of String-Matching Algorithms
Jean-Marc Steyaert, Palaiseau
Wojciech Szpankowski
Analytical Information Theory: Analysis of Lempel-Ziv scheme
Brigitte Vallée
Dynamical Systems and Average-Case Analysis of General Tries
Alfredo Viola
Recent Results in the Analysis of Linear Probing Hashing Algorithms

Abstracts

Reservation Probabilities

by Ed Coffman

Requests to use a resource arrive in a rate-λ Poisson stream, each request being specified by an arrival time T, advance notice A, and duration R of use. A request is accepted if and only if $[T + A, T + A + R]$ is disjoint from all currently reserved intervals. The three parameters form independent i.i.d. sequences. What is the long-run fraction of time that the resource is reserved? This defines the *reservation probability*.
We present results for three special cases: *slotted time*- reservation intervals have durations $R = 1$ and start at integer times, *short notice* - the support of A is entirely to the left of the support of R, and *bimodal advance notice* - requests are either immediate or have a fixed large advance notice. Finally, we give a limit law for $R = 1$ and $A \sim U(0, a)$ as $a \to \infty$. The derivation of the limiting reservation probability $p(\lambda)$ analyzes a generalization of the classical parking problem to obtain the "incomplete Renyi constant"

$$p(\lambda) = \int_0^\lambda e^{-2\int_0^v \frac{1\Gamma e^\Gamma x}{x} dx} dv.$$

Towards Universal Limit Laws for Depths and Heights of Random Trees

by Luc Devroye

We formulate a general model for random trees that uses the distribution of n balls over an infinite binary tree. The model includes as special cases the binary search tree, the m-ary search tree, the fringe-balanced search trees, the quadtree, the simplex tree, the trie and the digital search tree. If D_n is the depth of a random ball (node), we show that $D_n/\log n \to c$ in probability for some constant c, and $(D_n - c\log n)/\sqrt{\log n} \to \mathcal{N}(0, \sigma^2)$ in distribution for $\sigma > 0$, where $\mathcal{N}$ denotes the normal law.

A Retarded Differential Equation and the Height of Binary Search Trees

by Michael Drmota

In 1986 Luc Devroye showed that the expected value of the height of binary search trees of size n is asymptotically $\mathbb{E}[H_n] \sim c \log n$ where $c > 2$ is the solution of $\left(\frac{2e}{c}\right)^c = e$. In this talk a different approach to this result is presented.
Let $y_k(x)$, $(h \geq 0, x \geq 0)$ be recursively defined by $y_0(x) = 1$ and by $y'_{h+1}(x) = y_h(x)^2$, where $y_h(0) = 1$. Then

$$\sum_{n\geq 0} \mathbb{E}[H_n] x^n = \sum_{h\geq 0} \left(\frac{1}{1-x} - y_h(x) \right).$$

It is shown that

$$\sum_{h\geq 0} \left(\frac{1}{1-x} - y_h(x) \right) = \frac{c}{1-x} \log \frac{1}{1-x} \mathcal{O}\left(\frac{\left(\frac{1}{1-x}\right)^{1/2}}{1-x} \right), \; x \to 1- \; .$$

Hence by Karamata's Tauberian theorem you again get $\mathbb{E}[H_n] \sim c \log n$.
The main ingredients of the proof is a (approximate) solution of the retarded differential equation

$$\Phi'(x) = -\frac{1}{\alpha^2} \Phi\left(\frac{x}{\alpha}\right)^2, \; \Phi(0) = 1, \; \alpha = e^{1/c}.$$

Setting $\tilde{y}_h(x) := \alpha^h \Phi(\alpha^h(1-x))$ you have $\tilde{y}'_{h+1} = y_h(x)^2$ and you can approximate $y_h(x)$.

Recent Development in the Probabilistic Analysis of Self-Organizing Search

by Fames A. Fill

We derive upper and lower bounds on the total variation distance to stationarity for the distribution of search cost under the move-to-front (MTF) rule for self-organizing lists with i.i.d. record requests. These enable us to obtain sharp rates of convergence for serveral standard examples of weights,

including Zipf's law and geometric weights, as the length of the list becomes large. The upper bound also shows that a number of moves of the order of the length of the list is uniformly sufficient for near-stationarity over all choices of weights. Concerning the stationary search cost distribution itself, we use a representation obtained by considering the continuized (i.e. Poissonized) MTF Markov chain to derive, for of the each standard examples, the asymptotic distribution for long lists.
In her Ph.D. dissertation, Josefin Bodell has recently shown how some of these results can be extended to accomodate (1) a second user whose record request probabilities may differ from those of the controller of the list but whose requests do not alter the list's order, (2) time-dependent request probabilities, and (3) "cold starts". We discuss these results, together with recent large deviation results of Predray Jelenković for the stationary distribution of search cost for infinite lists with summable weights.
Finally, we briefly discuss ongoing work concerning a variant of MTF in which at each step a requested subset of records is moved to the front, and interesting extensions there of.

The Ubiquitous Gaussion Law in Analytic Combinatorics
by Philippe Flajolet

Recent years see a convergence between analytic methods and probabilistic methods in the analysis of algorithms, of data structures, and of classical combinatorial structures. As a consequence, there are now many basic algorithms whose behaviour is understood not only in the average case, but also in distribution. The goal of this talk is to examine reasons for which the Gaussion distribution surfaces so often.
Start with discrete combinatorial structures that are "decomposable" in a suitable technical sense. Examples include strings, words, trees, search trees, many classes of special graphs, permutations, etc. It is then possible to set up functional equations for bivariate generating functions that keep track of important structural parameters. In many cases, singularities smoothly move in the complex plane, and this induces a "quasi-powers" sheme that leads to Gaussian laws.
Applications include paging and patterns of binary search trees, patterns in strings, planar geometric configurations, order patterns in random permuta-

tions, serach cost in point quadtrees, etc.
The whole approach fits within the framework of analytic combinatorics.

Some Applications of Urn Models to the Analysis of Algorithms
by Danièle Gardy

Urn models are frequently used in the analysis of algorithms to describe random allocation phenomena.
Basically, an urn model is simply a sequence of urns, into which are thrown balls according to certain rules; after this operation the urns are in a random configuration, and the focus is on a parameter such as the number of (non) empty urns, or of urns containing a specified number of balls.
Roughly speaking, the class of models we present in this talk can be partitioned into three parts:

- the "static" models : A specified number of balls is thrown; what is the final configuration?
- the "waiting time" models : The interest is on the number of balls that have to be thrown before obtaining for the first time a specified configuration;
- the "dynamic" models : How does the configuration evolve when the balls are thrown at different times, i.e. can we characterize the resulting random process?

When some property of independence is satisfied, static problems are often well described by generating functions, and general results are available. The generating function approach can also give results in the waiting-time problems, but seems to fail on dynamic problems, which probably require more probabilistic tools.

The Number of Nodes of Given Degree in Random Trees
by Bernhard Gittenberger

Let T_n denote the set of unlabeled unrooted trees of size n and let $k \geq 1$ be given. By assuming that every tree of T_n is equally likely it is shown that

the limiting distribution of the number of nodes of degree k is normal with mean value $\sim \mu_k n$ and variance $\sim \sigma_k^2 n$ with positive constants μ_k and σ_k. Besides, the asymptotic behaviour of μ_k and σ_k for $k \to \infty$ as well as the corresponding multivariate distributions are derived. Furthermore, similar results can be proved for plane trees, for labeled trees, and for forests.

Probabilistic Divide-and-Conquer Recurrences for Geometric Random Variables
by Mordecai Golin

Let S be a set of points in the plane. A triangulation of S is a maximal set of non-crossing edges (connecting vertices of S). The weight of a triangulation is the sum of the length of the edges in the triangulation. A Minimum Weight Triangulation is a triangulation with minimum weight among all triangulations of S. We let $MWT(S)$ represent this weight.
Let S_n be a set of n points chosen i.i.d. from the uniform distribution over a unit square in the plane. The quantity we analyze is $MWT(S_n)$. We prove that $\exists c > 0$ such that

$$\frac{MWT(S_n)}{\sqrt{n}} \to 0$$

where the convergence is in expectation and probability. The basic technique used is to prove that $MWT(S_n)$ satisfies a divide-and-conquer type inequality and a "continuity" condition. Any quantity satisfying such inequalities and conditions will then be shown to "converge". The methods demonstrated can be used to analyze many other geometric quantities as well.

Presorting Algorithms: An Average-Case Point of View
by Hsien-Kuei Hwang

We introduce the concept of presorting algorithms, quantifying and evaluating the performance of such algorithms with the average reduction of the number of inversions. Stages of well known algorithms such as quicksort are evaluated in such a framework and shown to cause a meaning drop in the inversion statistic. The expected value, variance and generating function for the decrease in number of inversions are computed.

Analytic Information Theory in Service of Queueing with Aggregated Exponential ON/OFF Sources
by Philippe Jayquet

We use tools developed for the analysis of algorithms (Mellin transform, generating functions) and apply them to traffic and performance analysis of communication networks. We focus on the exponential ON/OFF source model. We show that the aggregation of an infinite number of independent ON/OFF sources satisfying some "profile" conditions, generates traffic with long range dependences and self-similarities. These latter phenomena have been recently depicted in real network traffics. It is shown that such models induce queue size and network access delays with polynomial tail. We model the network as a single queue with a single exponential server. The analysis borrows from Mellin transform, singularity analysis and asymptotic expansion theory.

On the Expected Number of Nodes at Level k in 0-Balanced Trees
by Rainer Kemp

An ordered tree with height n and m leaves is called 0-balanced if all leaves have the same level. We compute the average number of nodes (with specified degree) appearing in a given level in a 0-balanced ordered tree as well as in a 0-balanced t-ary ordered tree.
With respect to the former class we shall show that the average rate of increase of nodes amounts to $\rho = \frac{m-1}{n}$ passing from one level to the next one. The same fact holds for nodes with a degree one at a large level. The average number of nodes with a degree two and that one with a degree greater than two tends to ρ and zero for large levels, respectively.
The class of 0-balanced t-ary trees corresponds to the set of all code trees associated with all n-block codes with m code words over a given alphabet with cardinality t. In that case, we shall show that all nodes with maximal degree t are concentrated at levels smaller than $\log_t(m)$ and all nodes with degree one appear at levels greater than $\log_t(m)$, on the average.

Some Remarks on Alternating Sums and Mittag-Leffler's Theorem
by Peter Kirschenhofer

Alternating sums of the type $\sum_k \binom{n}{k}(-1)^k f(k)$ frequently arise with the average case analysis of algorithms. The two main approaches in the asymptotic treatment of such sums are Mellin transforms and Rice's integrals. In a recent paper in TCS 1995 Flajolet and Sedgewick have given an overview on the Rice's integral method thereby showing that for certain types of meromorphic functions $f(z)$ explicit formulae that exhibit well the asymptotic growth may be gained, too.
We focus on the aim of deriving such explicit formulae but start instead of taking Rice's integrals from the Mittag-Leffler partial fraction decomposition of $f(z)$. The method is illustrated for examples like the analysis of approximate counting or the internal path length of digital search trees.

Ranking and Unranking of Lexicographically Ordered Words: An Average-Case Analysis
by Jens Liebehenschel

We consider all words of length n of a formal language. If these words are arranged according to the lexicographical order, ranking means to determine the position of a word of the language. Unranking is the inverse operation of ranking. For a given formal language we compute the length of the minimal prefix of a word to be read to determine its position on the average, if the word is read from left to right. The length of the minimal prefix to be read only depends on the language itself, not on the ranking algorithm. After having derived a general expression, we demonstrate the result by discussing selected applications.

Random Trees and Brownian Excursion Local Times
by Guy Louchard

Random trees and Brownian Excursions (BE) are closely related. The BE is the limiting process for the depth-first traversal of a random rooted planar tree (hence the relation between the number of nodes at some level end the BE local time), also the generation size of a Galton-Watson process (condi-

tioned on the total progess) weakly converges to the B. E. local time.
In a joint work with B. Gittenberger, we have analyzed the B. E. multi-dimensional local time density with two different methods. In this talk we present a method based on Kec's formula for Brownian Motion path integrals and on a technique due to Bluny. We obtain the density as an integral in the complex plane depending on the product of Bessel functions.

Probabilistic Analysis of Multiple Quick Select
by Hosam M. Mahmoud

We investigate the distribution of the number of comparisons made by Multiple Quick Select (a variant of Quick Sort for finding order statistics). By convergence in the Wasserstein metric space, we show that a limit distribution exists for a suitably normalized version of the number of comparisons. We characterize the limiting distribution by an inductive convolution and find its variance. We show that the limiting distribution is smooth and prove that it has a continuous density with unbounded support.

Optimal Sampling Strategies in Quicksort and Quickselect
by Conrado Martínez

Sampling strategies for Quicksort and Quickselect amount to select a sample of size $s = 2k + j$ out of the n elements to be sorted (to get involved in the selection process in the case of quickselect) and find the median of the sample to use it as the pivot for a partitioning stage. It is well known that this sampling strategy reduces the probability of uneven partitions and improves the average number of comparisons to be made.
In this talk, I show that the best ("optimal") results are obtained when $k = \alpha \cdot \sqrt{n} + o(\sqrt{n})$, i.e. the size of the sample grows with n, taking into account the comparisons need to partition, the exchanges made during partition and also the comparisons and exchanges made to select the median of the sample. The constant α can be analytically computed in terms of the relative costs of comparisons and swaps, and the cost of the median-finding algorithm.
This is joint work with Salvador Roura,from UPC, Barcelona.

On the Average Complexity of the Membership Problem for a Generalized Dyck Language
by Markus Nebel

A general concept for analyzing the average complexity of the membership problem for any formal language is used in order to examine a generalization of the Dyck language. Our investigation is motivated by the fact that the Dyck language has a distinguished behaviour concerning that parameter. Surprisingly, that behaviour is lost even by small variations without utilising any opposite controls, e.g. adapting probabilities. This observation supports the significance of the Dyck language in computer science.

On Set Partitions with Unequal Block Sizes
by Andrew M. Odlyzko

The asymptotic behaviour of a_n, the number of set partitions of an n-element set into blocks of distinct sizes was detemined by Arnold Knopfmacher, Boris Pittel, Bruce Richmond, Dudley Stark, George Szekery, Nick Wormold and the speaker. The behaviour a_n is more complicated than is typical for set partition problems. Although there is a simple generating function, so that

$$a_n = n![z^n] \prod_{k=1}^{\infty} \left(1 + \frac{z^k}{k!}\right),$$

the usual analytic methods for estimating coefficients fail, so elementary tools are used to obtain most of the results.

Transform Methods for the Analysis of Algorithms
by Patricio V. Poblete

We study the properties of three mathematical transforms, that can be used to analyze data structures such as hash tables and skip lists. They are the Poisson Transform, defined as

$$\tilde{a}_m(x) = \mathcal{P}[a_{m,n}; x] = e^{-mx} \sum_{n \geq 0} \frac{(mx)^n}{n!} a_{m,n}$$

the Diagonal Poisson Transform:

$$\grave{a}_c(x) = \mathcal{D}_c[a_n, x] = (1-x)\sum_{n\geq 0} e^{-(n+c)x}\frac{((n+c)x)^n}{n!}a_n,$$

and the Binomial Transform:

$$\hat{a} = \mathcal{B}_s a_n = \sum_n (-1)^n \binom{s}{n} a_n.$$

We also discuss connections between the two Poisson transforms and a "depoissonization" theorem.
This talk is based on joint work with J. Ian Munro, Alfredo Viola and Tom Papadakis.

On the Analysis of Ascendants and Descendants for Various Families of Trees
by Helmut Prodinger

The number of ascendants of a particular node in a tree is the length of the chain from the root to the node, whereas the number of descendants is the size of the subtree rooted at this node.
For binary search trees and locally balanced binary search trees (a la Poblete/Munro) these parameters are analyzed via generating functions in 3 variables. Connections to Quicksort (standard and median-of-three) are pointed out.
The instance of (Catalan) binary trees is considered, and the enumeration of the nodes might by obtained by preorder, inorder or postorder traversal. Explicit Formulae are obtained for expectations and variances in all instances.
Multiple Quickselect is also considered, namely the "grand averages" when searching for p random elements. For the number of recursive calls (passes) and comparisions explicit formulae (expectation, variance) are obtained.
These results were obtained partially with coauthors Kirschenhofer, Martínez, and Panholzer.

Some Limit Theorems in Pattern Matching
by Mireille Régnier

We present a few probabilistic theorems that are suitable for pattern matching analysis. Notably, the subadditive ergodic theorem allows to get a. s. convergence or linearity results. We also remark that many pattern matching problems are expressed in terms of m-independent random variables; results on asymptotic normality follow. We discuss various applications such as: string searching algorithms, occurences of set of patterns as well as their waiting time, longest common subsequence, statistical distance, ...

Saddle-Point Method and Computer Algebra
by Bruno Salvy

We describe new algorithms for asymptotic expansion of exp-log or inverse exp-log functions which overcome the problem of indefinite cancellation. An application to the asymptotics of the average number of parts in a set partition and its variance are given.

Challenges in Interval-Based Clock Synchronization
by Ulrich Schmid

We survey some of our results on interval-based clock synchronization in distributed fault-tolerant systems. Unlike interval synchronization approaches, our technique provides approximately synchronized clocks maintaining both precision and accuracy w. r. t. external time. This is accomplished by means of a time representation relying on intervals that capture external time, providing accuracy information encoded in interval lengths. Our - quite delicate - worst case analysis utilizes a novel interval-based framework for establishing precision and accuracy bounds subject to a fairly detailed system model. Apart from individual clock rate and transmission delay bounds, our system model incorporates non-standrad features like clock granularity and broad cost latencies as well. Some experimental results, however, indicated that the obtained worst case bounds are almost meaningless for describing the average behaviour. Moreover, those bounds do not properly reflect the fact that interval-based clock synchronization algorithms - unlike traditional ones

- benefit from non-worst case settings. This poses the challange of conducting a proper average-case analysis of precision and, in particular, accuracy of interval-based clock synchronization. Trying to do this, however, is not at all trivial: Any meaningful analysis must be based upon a realistic - but still manageable - probabilistic system model for clocks, networks and faults. Moreover, some problems of its own right like j-th order statistics with $j = j(n)$ appear in the analysis of convergence functions. We cannot report on results of a meaningful analysis yet, but can nevertheless claim that fault-tolerant distributed algorithms for partially non-synchronous systems are certainly a source of new problems for the analysis of algorithms.

Open Problems in the Analysis of Sorting and Searching Algorithms
by Robert Sedgewick

We survey the $\sim$ 60 open problems described in Knuth's volume 3 (those rated 46-50) and discuss the $\sim$ 20 that have been solved in the 25 years since publication of the book.
Three open problems are discussed in detail:

(i) Average-case analysis of shellsort or one of its variants.

(ii) Are balanced trees, such as AVL, 2-3, or red-black trees, asymptotically optimal? (of height $\log n$ with coefficient 1).

(iii) Development of sorting networks that are substantially better than Batcher's networks for practical problem sizes.

These problems have been open for 30 years or more, despite many attempts to solve them.

Probabilistic Analysis of String-Matching Algorithms
by Robert Smythe

Two problems are considered: (i) finding all occurrences of a pattern of length m in a text of length n, as n becomes large; (ii) finding the first k occurrences of such a pattern.
For problem (i), we analyze the Boyer-Moore-Horspool algorithm with a text from a finite alphabet, either with i.i.d. characters or a text string forming a Markov chain. If the Markov chain is irreducible aperiodic, we get in either case an explicit evaluation of the average number of comparisons as $n \to \infty$. A central limit for the number of comparisons is also shown to hold. For the original Boyer-Moore algorithm, the case is more difficult: for simple examples we can compute the average number of comparisons, but the problem rapidly becomes computationally infeasible. A central limit result holds for this case also.
For problem (ii), we investigate the approximation of the waiting time to the kth occurence of the pattern in an i.i.d. text. If the pattern has low "self-similarity", we show that Poisson approximation is very accurate for large values of n.

Analytical Information Theory: Analysis of Lempel-Ziv Scheme
by Wojciech Szpankowski

We argue for the need of "analytical information theorey" that solves problems of information theory using analytical techniques. As an illustation we presented a detailed analysis of Lempel-Ziv scheme. We show that a simplified version of the problem can be reduced to the analysis of digital search trees (DST). We analyze such trees in a Markovian model, that is, when sequences are generated by a Markov chain. Using generating functions, poissonization, Mellin transform and depoissonization we obtain our results. Finally, we sketch how to extend these results to Lempel-Ziv model.
Joint work with P. Jaquest and J. Tang.

Dynamical Systems and Average-Case Analysis of General Tries
by Brigitte Vallée

Digital trees or tries are a versatile data structure that implements "dictionary" operations on sets of words (namly, insert, delete and query), as well as set-theoretic operations like set union or set intersection. It can also be used for text searching, interval search and partial match retrieval. In a recent paper, Bentley and Sedgewick even show that a carefully design implementation can be more efficient than hashing while offering considerably wider functionality.
The cost of these operations is mainly measured by three parameters of the tree structure: heigth, number of internal nodes and external path length. This talk is devoted to the average case analysis of these three basic parameters. Here, data items are (infinite) words that are produced by a common mechanism, called a source in information theory contexts. The dependence on the initial input distribution has already been considered by Devroye in the case of the simplest source, the Bernoulli source. We treat here the case of a general dynamic source and a general initial input distribution, which we call a *probabilistic dynamic source*.
The following results are established:

- The height of a trie has expectation $\sim B \log n$ and its limit probability distribution is of the doubly exponential type with sharp tail properties.
- The average search time is $\sim A \log n$.
- The average size of a trie is, up to possible small fluctuations, well approximated by a quantity that is $\approx An$

There, the constants A and B are well-characterized quantities that are expressible in terms of entropy and coincidence probability.

Recent Results in the Analysis of Linear Probing Hashing Algorithms
by Alfredo Viola

In this talk we survey recent results in problems related with Linear Probing Hashing.

We start with the analysis of the variance of the cost of a successful search for a random element when the Last-Come-First-Served heuristic is used. This problem leads us to develop a new mathematical transform that we call Diagonal Poisson Transform.
We follow with the analysis of Linear Probing Hashing with buckets. We show how mathematical transforms, singularity analysis and mathematical software (we used Maple) play a key role in the derivation of our results.
Finally we present an analysis of the variance of the total displacement of the keys.
This talk is based on joint work with Patricio Poblete and Ian Munro.

At the end of the seminar there was an *open problem session.* At the URL *http://pauillac.inria.fr/algo/AofA/index.html* you can find a column dedicated to the problems presented there (and others) and the discussion on them.