Internationales Begegnungs- und
Forschungszentrum Für Informatik

Schloß Dagstuhl

Seminar Report 9527

Average Case Analysis of Algorithms

July 3 - 7, 1995

Overview

Analysis of algorithms aims at a precise prediction of the expected performance of algorithms under well-defined randomness models of data. The classical aspects are well covered by Knuth's *magnum opus* who treated, already more than twenty years ago, many aspects of fundamental algorithms, semi-numerical algorithms, or sorting and searching. In this, and other domains, what is sought is a precise description of the average-case behaviour of algorithms, often a very meaningful practical measure of complexity.

The field is undergoing tangible changes. We see now the emergence of combinatorial and asymptotic methods that permit us to classify data models into broad categories that are susceptible of a unified treatment. This has two important consequences for the analysis of algorithms: it becomes possible to predict average-case behaviour under more *complex* data models (for instance, nonuniform models and even Markovian dependencies); at the same time it becomes possible to analyse much more structurally *complex* algorithms since we have a much higher level grasp on the average-case analysis process.

In this perspective, generating functions together with their complex-geometric properties have often lead to a better understanding of what goes on. We witness here the emergence of general combinatorial and analytic schemas that leads to a structuring of what may be called "analytic combinatorics", and this has in turn strong implications for the analysis of algorithms itself. A further consequence is that we now see more and more possibilities of characterizing not only averages of computation costs but also variances, and even distributions.

Several talks in this report are devoted to the general methodology just outlined. Such a simple problem as "how many times must one shuffle a deck of cards before it looks almost random?" is clearly related to sorting and searching and the analysis requires deep methods of algebraic combinatorics (Bergeron). The related problem of the longest increasing sequence in a random permutation —the answer is $\sim 2\sqrt{n}$— nicely illustrates the interplay between algebraic methods and the analysis of generating functions (Odlyzko). It also bears relations to Zeilberger's recent "holonomic" paradigm in which many of these and similar analysis can be cast.

On a more analytic register, the diagonal Poisson transform (Viola) and the depoissonization lemma (Jacquet, Szpankowski) are tools of a general nature that are strongly tied with the analysis of hashing and digital trees or data compression. The study of largest components in composite structures is amenable to a general treatment (Gourdon) via a classification into general schemas that are also effective in the extraction of limit distributions (Soria, Drmota). We now know from such works that seemingly diverse statistical problems like cycles in permutations, components in functional graphs, or irreducible factors in polynomials over finite fields, obey common laws. This in turn has direct consequences regarding sorting (*in situ* permutation), computer algebra (polynomial factorization), or cryptography (functional graphs).

Trees have often been said to be the single most important discrete structure of computer science, and not unnaturally a description of application problems should

start with them. First there is the purely combinatorial model of uniform random trees (Gittenberger) that has been used repeatedly for the analysis of trees in parsing or symbolic manipulation; an interesting instance is the register allocation problem in code generation (Prodinger), which strangely enough is also related to several problems in the physical sciences.

Next, there come trees as a data structure related to order informations. We now have a very precise understanding of quickselect, a basic algorithm for finding order statistics, even in its distributional aspects (Mahmoud) or in its refined median-of-three version (Kirschenhofer). The classical heap structure can also be analyzed in the average case revealing some intricate fractal behaviour that contrasts with a classical Gaussian distribution (Steyaert). Quadtrees are an important data structure for geometrical data and they can be now fully analyzed on average as well as in distribution (Flajolet). Such typical analyses would barely have been possible a decade ago.

Trees based on digital information are essential to textual data processing, data compression, and they surface as an analytic model of many situations in distributed computing like leader election or communication protocols. In their basic version, they can be analysed under a variable key length model (Nebel). Bucket trees that are useful in a paging environment are attacked by a sophisticated use of the Mellin transform technology (Hubalek). Last, the methods developed in this context extend well to skip lists, an attractive randomized alternative to balanced trees (Martinez).

Mathematical analysis of algorithms is also gradually extending to more recent or to less classical areas of computer science. Circuits are an instance, where the depth of a random circuit proves to be much smaller on average than in the worst-case (Diaz). Gate matrix design may even benefit from methods of random graph theory (Karoński). Generating function methods prove operational in the reliability analysis of a cellular network (Sipala). Finally, ideas that stem from the analysis of Shellsort lead to the design of highly regular practical sorting networks that appear to sort almost surely (Sedgewick).

Other new applications of analysis of algorithms concern: patterns in strings like in DNA sequences (Régnier), computing with faulty processors (Louchard), parallel simulations (Greenberg), random and exhaustive generation (Kemp), simulation (Robson), computer algebra (Gonnet), occupancy problems (Gardy) and parallel scheduling (Wright).

This meeting is the second of its kind. The first one (Dagstuhl report #68) has given rise to a healthy 315 pages special issue of the journal *Theoretical Computer Science* (vol. 144, number 1-2). The present meeting is coupled to a special issue of the journal *Random Structures and Algorithms*, due to appear in 1997.

The Organizers
Philippe Flajolet
Rainer Kemp
Helmut Prodinger
Robert Sedgewick

Participants

François Bergeron, Montreal
Joseph Díaz, Barcelona
Michael Drmota, Wien
Philippe Flajolet, Le Chesnay
Danièle Gardy, Versailles
Bernhard Gittenberger, Wien
Massimiliano Goldwurm, Milano
Gaston Gonnet, Zürich
Xavier Gourdon, Le Chesnay
Albert Greenberg, Murray Hill
Walter J. Gutjahr, Wien
Friedrich Hubalek, Wien
Philippe Jacquet, Le Chesnay
Michal Karoński, Poznań
Rainer Kemp, Frankfurt am Main
Peter Kirschenhofer, Wien
Guy Louchard, Bruxelles
Hosam M. Mahmoud, Washington
Conrado Martínez, Barcelona
Markus E. Nebel, Frankfurt am Main
Andrew M. Odlyzko, Murray Hill
Helmut Prodinger, Wien
Mireille Régnier, Le Chesnay
John M. Robson, Bordeaux
Robert Sedgewick, Princeton
Paolo Sipala, Trieste
Michèle Soria, Paris
Jean-Marc Steyaert, Palaiseau
Wojciech Szpankowski, W. Lafayette
Uwe Trier, Frankfurt am Main
Alfredo Viola, Waterloo
Paul E. Wright, Murray Hill

Contents

Markus E. Nebel
Digital Search Trees with Keys of Variable Length

Andrew M. Odlyzko
Increasing Subsequences in Random Permutations

Helmut Prodinger
Solution of a Problem of Yekutieli and Mandelbrot

Mireille Régnier
Frequency of a Pattern Occurence in a (DNA) Sequence

John M. Robson
Fast Simulation of Random Trees

Robert Sedgewick
Shellsort

Paolo Sipala
Reliability of a Cellular Network

Michèle Soria
Special Limit Distributions

Jean-Marc Steyaert
On the Number of Heaps and the Cost of Heap Construction

Alfredo Viola
The Diagonal Poisson Transform

Paul E. Wright
The Ratio of the Extreme to the Sum in a Random Sequence with Applications

Abstracts

An Algebraic Approach to the Analysis of Mixing and Sorting Algorithms

by François Bergeron

Let $\pi : S_n \to [0,1]$ be a probability distribution on the symmetric group. A mixing rule is defined to be the element $(\sum_{\sigma \in S_n} \pi_\sigma \cdot \sigma)$ of the group algebra $\mathbb{R}[S_n]$ of S_n. To analyze the behaviour of related mixings $(\sum \pi_\sigma \cdot \sigma)^k$, we can often find a "semisimple" commutative subalgebra $\mathbb{H}$ of $\mathbb{R}[S_n]$. This is to say that $\mathbb{H}$ admits a basis of "idempotent" $e_1, e_2, \ldots, e_n$ such that $e_i e_j = 0$, if $i \neq j$. If $(\sum \pi_\sigma \cdot \sigma) \in \mathbb{H}$ we can write this element as a linear combination of the e_i's:

$$(\sum \pi_\sigma \cdot \sigma) = \sum \hat{\pi}_i e_i,$$

which can be thought of as the "Fourier transform" of $(\sum \pi_\sigma \cdot \sigma)$. It is clear that $(\sum \pi_\sigma \cdot \sigma)^k = \sum \hat{\pi}_i^k e_i$ and if we can compute the inverse Fourier transform, we obtain explicit expressions for the probability distribution $(\sum \pi_\sigma \sigma)^k$. We illustrate these steps with iterates of transpositions, shuffles and similar problems for other contexts than that of the symmetric group algebra.

Average Height of Uniform Increasing Monotone Circuits

by Joseph Díaz

We prove that, under the UC distribution, the average height of a randomly generated circuit with N nodes (gates+inputs) is $\leq (e \log 4 + O(1)) \log_2 N$.

(Joint work with M. J. Serra, P. Spirakis and T. Tsuki)

Combinatorial Constructions and Limiting Distributions: Predecessors in Random Mappings

by Michael Drmota

Let $\mathcal{F}_n$ be the set of random mappings $\varphi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ (such that every mapping is equally likely). For $x \in \{1, 2, \ldots, n\}$ the elements of $\bigcup_{k \geq 0} \varphi^{-k}(\{x\})$ are called the "predecessors" of x. Furthermore, let N_r denote the random variable which counts the number of points $x \in \{1, 2, \ldots, n\}$ with exactly r predecessors. It is shown how to identify the limiting distribution of N_r as $n \to \infty$.

If $r = r(n) = o(n^{\frac{2}{3}})$ then the limiting distribution is Gaussian, if $r \sim cn^{\frac{2}{3}}$ then it is Poisson, and in the remaining case $rn^{-\frac{2}{3}} \to \infty$ it is degenerate. Furthermore, N_r is a Poisson approximation if $r \to \infty$.

Analytic Variations on Quadtrees
by Philippe Flajolet

Quadtrees are a data structure suitable for searching multidimensional data. In this talk, we show how the average case analysis of quadtrees leads to differential equations of the so-called "holonomic" type. Use of the classical theory of Fuchsian singularities in conjunction with singularity analysis permits to analyze asymptotically expectations (and sometimes even limiting distributions) for parameters like: path length, number of leaves, number of pages in index quadtrees, partial match retrieval, and the like.
(Joint work with G. Gonnet, C. Puech, J. M. Robson, M. Hoshi, T. Lafforgue, G. Labelle, L. Laforest and B. Salvy)

Analysis of an Occupancy Problem in the Static and Dynamic Cases
by Danièle Gardy

We consider the following problem: allocate at random n balls among d urns; analyze the number of urns with at least one ball. The case of urns that can receive an unbounded number of balls for a known total number of balls, has been studied extensively in the literature.

We first present an extension to the case of urns with finite capacity. Then we study what happens in both urn models when we can add or delete balls according to some rules.

We prove that the number of balls behaves as a Gaussian Markov process, and the number of non-empty urns as a Gaussian process. Our results allow us to study the influence of parameters such that the ratio *number of urns/time of observation*, the number of balls at the end of the observation interval, the capacity of a bounded urn, or the rules for insertion or deletion.

The Contour and Profile of Random Trees

by Bernhard Gittenberger

We study four stochastic processes describing the contour and the profile of simply generated random trees: the contour is described by the traverse process which is the process of the node heights during pre-order traversal of the tree and the contour process constructed from the leaf heights of the tree. The processes associated to the profile are the number of nodes and the number of leaves at a given height. Using multivariate generating functions and singularity analysis we obtain the following results: if we scale the above processes in a suitable way, then the contour processes converge weakly to Brownian excursion and the profile processes to its local time.

Average Case Analysis of Membership Problems for Trace Languages

by Massimiliano Goldwurm

In this talk we present a short survey on the analysis of algorithms for membership problems for trace languages. These languages can be defined as subsets of free partially commutative monoids and their properties are studied as an extension of the traditional theory of formal languages.

A *concurrent alphabet* is a pair $< \Sigma, \mathfrak{I} >$ where Σ is a finite alphabet and $\mathfrak{I}$ is a symmetric and irreflexive relation over Σ. Let $\mathbb{M}(\Sigma, \mathfrak{I})$ be the quotient monoid $\Sigma^* / \approx_{\mathfrak{I}}^*$, where $\approx_{\mathfrak{I}}^*$ is the reflexive and transitive closure of the relation $\approx_{\mathfrak{I}}$ defined as follows: $\forall x, y \in \Sigma^*$ and for all $a, b \in \Sigma$, $xaby \approx_{\mathfrak{I}} xbay$ if and only if $(a, b) \in \mathfrak{I}$. A *trace* $t = [x]_{\mathfrak{I}}$ is any element of $\mathbb{M}(\Sigma, \mathfrak{I})$, while a *trace language* T is any subset of $\mathbb{M}(\Sigma, \mathfrak{I})$. The class of *rational trace languages* on the monoid $\mathbb{M}(\Sigma, \mathfrak{I})$ is defined as the class of languages T such that $T = \{[x] | x \in L\}$, where $L \subseteq \Sigma^*$ is a regular language. It can be proved that the above definition describes the smallest class of trace languages containing all finite sets and closed with respect to the operation of union, product, and star [...]. Analogously, we define the class of context-free trace languages as the class of subsets $T \subseteq \mathbb{M}(\Sigma, \mathfrak{I})$ such that $T = \{[x] | x \in L\}$, where $L \subseteq \Sigma^*$ is a context-free language [...]. The *membership problem* for a trace language $T \subseteq \mathbb{M}(\Sigma, \mathfrak{I})$ is defined as the problem of verifying, for an input $x \in \Sigma^*$, whether $[x]$ belongs to T.

An important notion that plays a key role in the analysis of membership problems is the notion of prefix of a trace. Given two traces $p, t \in \mathbb{M}(\Sigma, \mathfrak{I})$, we say that p is a *prefix* of t, if $t = p \cdot q$ for some $q \in \mathbb{M}(\Sigma, \mathfrak{I})$; similarly, q is called *suffix* of t. Note that if p is a prefix of t then there exist $x, y \in \Sigma^*$ such that $t = [x]$, $p = [y]$ and y is a (string) prefix of x. Observe that, while the number of prefixes of a word x is $|x| + 1$, in the case of a trace $t \in \mathbb{M}(\Sigma, \mathfrak{I})$, the number of prefixes also depends on the concurrent alphabet and the symbols of Σ occuring in t.

In the talk we present some algorithms for recognition of rational and context-free

trace languages. It turns out that the time and space complexity of these procedures are related to the number of prefixes of the input trace. In particular the mean number of prefixes of a trace of length n is proportional to the average time computation required by the algorithm for recognition of rational trace languages. The moments and the variance of the number of prefixes of a random trace of length n can be evaluated by using a sort of bijective argument that reduces the problem to determining the number of words of given length in certain regular languages. For some concurrent alphabets, using a different approach, it is also possible to give the asymptotic distribution of the number of prefixes of a trace of length n.

Random Problems Derived from Computer Algebra
by Gaston Gonnet

We present some open problems in the analysis of algorithms which arise from computer algebra. These problems, as opposed to the most common ones being analyzed these days, are real problems, and their solutions would be extremely useful to the area. This probably requires some explanation. CA algorithms are often heuristic algorithms or polyalgorithms, often subjectable to a large amount of tuning. These algorithms often need to be tested quite extensively too. For these two reasons, it becomes very relevant to be able to generate random problems with particular properties. Generating these random CA problems is the main underlying topic of this presentation. We present 4 open problems.

(1) Generate random polynomials (all polynomials with integer coefficients) $p(x, y)$ and $q(x, y)$, so that the solutions of $p(x, y) = q(x, y) = 0$, are all rational (or involving algebraic numbers of low degree). In other words, the resultant of p, q factors into all linear (or small degree) factors. (Note: for random p, q, the $\text{res}(p, q, x)$ with exponentially high probability will not factor, and hence all the solutions of the system of equations involve a high degree algebraic number).

(2) Are there easy, sparse, classes of polynomials $p(x, y, z, \ldots)$, $q(x, y, z, \ldots)$ so that when $|p| = \Omega(1)$, $|q| = \Omega(1)$ then $|p * q| = O(|p| + |q|)$? $|p|$ denotes the number of terms of $p(x, y, z, \ldots)$. (Note: normally, $|p * q| = O(|p| * |q|)$).

(3) Let $A * x = b$ be a sparse linear system of equations. Let the entries of A be all different symbols (no simplification possible). A is $n \times n$.

  (3.a) What is the number of random non-zero entries so that $\det(A) \neq 0$ with probability $1/2$? (Result should be an asymptotic expansion in n to at least $O(1)$, it is known that the result is $O(n * log(n))$).

  (3.b) For this number of non-zero terms, when $\det(A) \neq 0$, how many factors does $\det(A)$ have? (Note: this will simulate quite well linear systems with structure and with simple solutions).

(4) Let $p(x, y, z, \ldots)$ be a polynomial with small integer coefficients (i.e. random between $-m \ldots m$). What are the probabilities of:

(4.a) $p = p1 * p2$ (non-trivial factors)

(4.b) $p = p1(p2(x, y, z, \ldots))$ (non-trivial composition)

(4.c) $p = p1 * p2 * p3 \ldots$ (all linear factors)

Largest Component in Random Combinatorial Structures

by Xavier Gourdon

The talk presents a general analytic framework dedicated to the evaluation of the size of the largest component in random combinatorial structures. It applies to composite combinatorial structures of the form $\Phi(\mathcal{P}) \approx \varphi$; the meaning is that φ is formed by substituting atoms of Φ by $\mathcal{P}$-structures. It translates into generating functions as $C(z) = F(P(z))$.

Our framework applies when the generating functions involved are algebraic - logarithmic near their dominant singularity. Three cases appear, depending on which function, $F(w)$ or $P(z)$, dictates the singularity of $F(P(z))$. In the subcritical case, the singularity is only dictated by $P(z)$ and leads to a discrete law limit for $n - L_n$, where L_n is the size of the largest component (roughly, the size of the largest component is nearly the total size of the structure). In the critical case (both $F(w)$ and $P(z)$ dictate the singularity of $F(P(z))$), we prove a central limit theorem $\lim_{n\to\infty} Pr(L_n \leq \frac{n}{\lambda}) = f(\lambda)$ for all fixed $\lambda > 1$ (the size of the largest component is proportional to the size of the structure). In the supercritical case (only $F(w)$ dictates the singularity of $F(P(z))$), we prove a double exponential law limit, leading to a distribution concentrated near $\log n$.

Three typical examples corresponding to each of the three cases are: size of the largest subtree of a Catalan tree (subcritical), size of the largest Cayley tree in a random mapping (critical), size of the largest summand in composition (supercritical).

Parallel Simulations

by Albert Greenberg

Discrete event simulation is the most general and widely used method for investigating the behaviour of large, complex computer and communication systems. Unfortunately, such simulations often require very large computer memories and very long runs. This has motivated a great deal of research in parallel and distributed simulation methods.

In the first part of this talk, I consider a very simple stochastic "event-coupling" model, which describes the degree of parallelism available in the large systems we want to simulate. By analysing an approximating system of ordinary differential equations, we bound from above and below the length of the critical path in an associated event dependancy graph. The good news is that the critical path length grows very slowly with the system parameters. This analysis helps to motivate and explain the performance of an efficient optimistic simulation method, known as "synchronous relaxation". An application of synchronous relaxation to the simulation of large circuit-switched networks is described. Implemented on a 16K processor (masPar) this simulation is about 30 times faster than an optimized serial counterpart running on a workstation.

In the second part of my talk, I consider a very simple stochastic "asynchronous-updates" model, which describes the behaviour of conservative parallel simulations. Again via the analysis of a system of differential equations, we characterize the performance of the simulation in terms of a few system parameters. This helps to motivate and explain an efficient conservative method, which we have applied to simulate wireless, cellular networks. By exploiting a new idea of "slackness" in these simulations, our masPar program runs about 120 times faster than the optimized workstation counterpart.

Finally, a series of experiments is presented describing how the communication structure (i.e., the allowed patterns of communication between subsystems) of the system being simulated effects the speed of optimistic and conservative methods.

Mellin Convolutions in the Analysis of Bucket Digital Trees

by Friedrich Hubalek

We show that the variance of some parameters of bucket digital search trees is asymptotically $(C + \delta(\log_2 N))N$ (under the symmetric Bernoulli model). We use a method by Flajolet and Richmond, but handle the pecularities of variance-calculations, namely the binomial convolution, with a combination of exponential and ordinary generating functions and with the (repeated) application of Mellin's convolution integral. This approach has the advantage that the symmetry of the problem is preserved and we need no deep transformation formulae to show that the variance is rather small. Furthermore, this method relies on asymptotic properties rather the explicit representations of generating functions, hence we hope it is a first step towards a general result on the smallness of the variance of trie-parameters.

Depoissonization Lemma and its Appliations
by Philippe Jacquet and Wojciech Szpankowski

Often the Poisson model is easier to analyze than the original model (i.e. the Bernoulli model). In the analytical analysis of algorithms, this means that the bivariate Poisson generating function satisfies simpler functional-differential equations than the original model. The question is how to extract original results from the one obtained in the Poisson model. This is called "depoissonization". In our talk we consider the case where one is able only to derive an asymptotical approximation for the Poisson model. Then, a care careful analysis is required to get Bernoulli model results.

Average Case Analysis of the Gate Matrix Layout Problem
by Michal Karoński

We present an application of a new model of random graphs - random intersection graphs - to gate matrix circuit design. In this model two vertices are adjacent provided their assigned sets intersect. We explore the evolution of random intersection graphs by studying thresholds for appearance and disappearance of small induced subgraphs. In particular we are interested in determining thresholds for which an intersection graph, representing a particular gate matrix layout (GML), is an interval graph. One can show that in such a case the GML optimization problem is easily solvable.
(Joint work with E. R. Scheinerman and K. B. Singer)

Prefixes of Formal Languages:
Their Relation to the Analysis of Particular Algorithms
by Rainer Kemp

We discuss an interrelation between prefixes of formal languages and the average case analysis of algorithms generating combinatorial objects lexicographically or solving the membership problem.

In the first part, we consider a parameter defined on a given formal language $\mathfrak{L} \subseteq T^*$ (=prefixes of a fixed length in the set $\mathfrak{L} \cap T^n$) whose expected value is appropriate to measure the average running time of an algorithm generating the words in the language lexicographically, even in the case that no information about the algorithm itself is being available. This observation implies a **complete** average case analysis, including higher moments about the origin and the cumulative distribution function. We demonstrate our results by discussing various concrete applications, such as the generation of words in a given regular language, the generation of subsets of a given set,

the generation of Dyckwords, the generation of t-ary ordered trees according to Ruskey and Zaks, the generation of ordered trees with bounded height and the generation of various classes of 0-balanced ordered trees.

The second part of this talk is devoted to the average case analysis solving the membership problem for a formal language $\mathfrak{L} \subseteq T^*$. Scanning a given word $w \in \mathfrak{L} \cap T^n$ from left to right, this word can be rejected if we find a prefix of minimal length which cannot be extended to a word of the language. The average running time of such an algorithm can be expressed by the number of prefixes of a certain length in $\mathfrak{L} \cap T^n$. We show that the asymptotical behaviour of the average running time required to solve the membership problem for regular languages has the form $\varphi_1(n)n + \varphi_2(n),\ n \to \infty$, where φ_1, φ_2 are bounded periodic functions. In the case of the Dyck language, the asymptotic behaviour is given by $4\pi^{-\frac{1}{2}} n^{\frac{1}{2}} - 2,\ n \to \infty$. Higher moments and further distribution results are computed, too.

Analysis of Hoare's FIND-Algorithm with Median-of-Three Partition
by Peter Kirschenhofer

Hoare's FIND algorithm can be used to select the j-th element out of a file of n elements. It bears a remarkable similarity to Quicksort; in each pass of the algorithm, a pivot element is used to split the file into two subfiles, and recursively, the algorithm proceeds with the subfile that contains the sought element. As in Quicksort, different strategies for selecting the pivot element are reasonable. In this talk, we consider the median-of-three version, where the pivot element is chosen as the median of a random sample of three elements. We give explicit formulae for both the average number of passes and comparisons, when any relative ordering of the n elements in the file is equally likely. We also indicate how higher moments of these parameters can be derived.
(Joint work with C. Martínez and H. Prodinger)

Finding the Maximum with Error Probabilities: A Sequential Analysis
by Guy Louchard

Assume that n players are represented by n reals, uniformly distributed over the unit interval. We assume that the error probability of a comparison between two players depends linearly on the distance between the players. Using a sequential analysis approach, we present an algorithm to estimate the maximum ξ of the players with an error less than ε.

Mean cost, variance and central moment generating functions are analyzed.

Limit Distributions for some Sorting Algorithms
by Hosam M. Mahmoud

We demonstrate that QUICKSELECT, a one-sided version of QUICKSORT suitable for finding order statistics, has an infinitely divisible limit distribution which is explicitly characterized. QUICKSELECT (also known as FIND) can easily be adapted to find several order statistics simultaneously. This version of the algorithm is called MULTIPLE QUICKSELECT. We present average case analysis for MULTIPLE QUICKSELECT.

Tree-growing search strategies are discussed and a simple sufficient condition for normality is given. Thus most practical implementations of insertion sort have asymptotically normal behaviour.

Statistics over Sequences of Geometric Random Variables
by Conrado Martínez

Skip lists are probabilistic data structures that allow to dynamically maintain a set of n items (Pugh, 1990). Their analysis heavily relies on the properties of the underlying sequences of i.i.d. geometric random variables. For instance, the performance of search in skip lists is directly related to the maximum of n i.i.d. geometric random variables (height of skip list) and the number of right-to-left maxima in a sequence of geometric r.v. (horizontal path length). The so-called optimized search algorithm for skip list - a search stategy that avoids redundant key comparisons - was discussed in more detail; in particular, the kind of the mathematical problems arising in such analyses and the techniques available to solve them.

Digital Search Trees with Keys of Variable Length
by Markus E. Nebel

We consider Digital Search Trees with keys of variable length, where it might happen that one key is a prefix of another. Since this situation is not practicable for the traditional insertion algorithm, a modification of this method is introduced. This modification is based on the idea not to store a key at the same position for the whole lifetime of the tree. An average case analysis of the number of position exchangements is performed under some kind of "worst case model".

Increasing Subsequences in Random Permutations
by Andrew M. Odlyzko

Let L_n denote the length of the largest increasing subsequence of a permutation of $\{1, \ldots, n\}$. The distribution of L_n has been studied intensively for a long time, and it is known the $\mathbf{E}(L_n) \sim 2\sqrt{n}$, as $n \to \infty$, and that there is a sharp transition, with almost all permutations in S_n having $L_n \geq c_n$ for some constant $c_n \sim 2\sqrt{n}$, as $n \to \infty$, and with very few permutations having $L_n > c_n + n^{\frac{1}{3}}$, for example. However, little is known about this transition zone. This talk presented a new approach to this problem, based on a generating function of I. Gessel. This approach was developed by B. Poonen, H. Widom, H. Wilf and the speaker. So far this approach has only provided information about the tails of the distribution of L_n. One of the generating functions that occur in the analysis also occur in the physics literature, in the sudy of quantum gravity, and one of the main open questions is whether the methods used by the physicists can be made more rigorous.

Solution of a Problem of Yekutieli and Mandelbrot
by Helmut Prodinger

The register function (or Horton-Strahler number) of a binary tree is defined recursively as follows: Leaves get the number 0, and, if a left subtree has number a and a right subtree has number b, the whole tree gets the maximum of a and b, or, if $a = b$, the value $a + 1$. Yekutieli and Mandelbrot asked the following question: if the tree has register function p, how many maximal subtrees of register function $p - 1$ are there? Experiments indicate that the average value of this parameter oscillates between 3 and 4.

Using generating functions, Mellin transforms and singularity analysis, a precise version of the above claim was shown, the solution involving Catalan's constant $1 - \frac{1}{3^2} + \frac{1}{5^2} - \frac{1}{7^2} + \cdots$.

Some extensions are also sketched.

Frequency of a Pattern Occurrence in a (DNA) Sequence
by Mireille Régnier

Our goal is to assess the limiting distribution of the frequency of the occurrences of a given pattern H in a random text T of length n. We study both the so-called Bernoulli model and Markovian model. We prove that the number of pattern occurrences (overlapping copies are counted separately) tends to a normal distribution, and we derive

explicit and asymptotic formulas for the mean and the variance of the number of occurrences. During the course of the derivation we compute the probability of exactly i occurrences of H in the text T. We derive the generating function of this probability, and using an analytical technique we derive in a uniform manner all results announced above. Applications of these results range from wireless communications to approximate pattern matching, molecular biology, games, codes and stock market analysis. These findings are of particular interest to molecular biology problems such as finding patterns with unexpected (high or low) frequencies (the so-called contrast words) and gene recognition.

(Joint work with W. Szpankowski)

Fast Simulation of Random Trees

by John M. Robson

I consider the height h_n of randomly constructed binary search trees of n nodes. To estimate the variance of this height, a number of such trees are constructed. By constructing an array giving the number of external nodes at each level, the construction time is reduced to $O(\log^4 n)$. The results suggest that the variance remains bounded and small for large n. Attempting to prove this has yielded the result that, for any monotonic unbounded increasing function f, $\liminf_{n\to\infty} Pr[h_n > \mathbf{E}[h_n] + f(n)] = 0$.

Shellsort

by Robert Sedgewick

During the past twenty-five years the Shellsort algorithm has been studied in some detail, generally from the perspective of worst case performance. Upper bounds of the form $O(N^{1+\frac{\varepsilon}{\sqrt{\log M}}})$ have been developed, and it has been shown that this aymptotic form is best possible. No interesting case (from a practical point of view) has been successfully analyzed in the average case, and a number of variants of the algorithm are worthy of study. Simulations show one simple variant to have good average case performance, including correspondance to a $\log N$ depth probabilistic sorting network: for a decreasing sequence $d_i = \lfloor \alpha^i \rfloor$, $i = \ldots, 1$, with $\alpha < 1.3$, perform all compare-exchanges $a[x] := a[x+d]$, then all compare-exchanges $a[x+d] := a[x+2d]$. An average case analysis of this variant of Shellsort may be possible, based on the observation (suggested by simulations) that the distribution $D(k) = Pr\{a[x] = x + k\}$ (when sorting a permutation) is approximately normal.

Reliability of a Cellular Network

by Paolo Sipala

A linear cellular array consists of processing elements v_i, i integer; each element v_i is connected to the elements v_{i+s}, for all $s \in S$, where S is a given set of connection spans. Each element is operational with probability p, and fails (independantly) with probability $q = 1 - p$. We study the probability that an array remains connected in spite of element failures. This evaluation is obtained using a finite-dimensional Markov chain, from which a reliability generating function is derived.

Special Limit Distributions

by Michèle Soria

In many analytic schemes arising in combinatorics, limiting distributions can be seen as a direct reflection of structural characteristics of the underlying combinatorial constructions, and analytic properties of the involved generating functions.

We consider functional schemes $P(u, z) = F(uC(z))$, corresponding to the combinatorial construction of substitution, and are interested in the limiting distribution of the number of components in a random structure of large size, namely

$$\lim_{n\to\infty} \frac{[u^k]F(u) \cdot [z^n]C^k(z)}{[z^n]F(C(z))},$$

for $k = \mu_n + \chi\sigma_n$, $\chi = O(1)$.

Three complementary cases can be distinguished, according to the singular behaviour of $F(C(z))$ being dictated by the dominant singularity of $C(z)$ – subcritical case –, or $F(z)$ – supercritical case –, or both – critical case –.

In the subcritical case, an F' limiting law is to be expected: e.g. derivative of Geometric for F corresponding to sequence construction, or derivative of Poisson for set construction.

The supercritical case leads to Gaussian limiting distributions. In the critical case, when $F(z)$ is an alg-log function and $C(z)$ is algebraic with a leading term in $(1 - z)^\lambda$, $0 < \lambda < 1$, the limiting distribution has an hypergeometric representation for λ rational (e.g Rayleigh or Maxwell laws for $\lambda = \frac{1}{2}$).

On the Number of Heaps and the Cost of Heap Construction

by Jean-Marc Steyaert

Heaps are a well known data structure which allows an efficient sorting algorithm. Surprisingly their combinatorial properties are still partially worked out - exact summation formulae have been stated, but most asymptotic behaviours are still unknown. This is largely due to the chaotic look of most parameters. We present a number of general asymptotic results which give insight on the difficulties encountered when dealing with the asymptotics of the number of heaps of a given size and the cost of heap construction. In particular, we exhibit the influence of arithmetic functions in this chaotic behaviour. We also show that the distribution function of the cost of heap construction by Floyd's algorithm is asymptotically normal.

(Joint work with Hwang Hsien Kuei)

The Diagonal Poisson Transform

by Alfredo Viola

In this talk we present a new mathematical transform, called the "Diagonal Poisson Transform", and some of its applications. This transform can be seen as a variation of the Poisson transform, that is used to study hashing algorithms. We present several important properties of it, and its use to analyze some hashing algorithms, to solve some general classes of recurrences, to find some generalizations of Abel sums, and to find inverse relations.

(Joint work with Patricio Poblete and Ian Munro)

The Ratio of the Extreme to the Sum in a Random Sequence with Applications

by Paul E. Wright

If $X_1, X_2, \ldots, X_n$ is a sequence of non-negative independent random variables with common distribution function $F(t)$, we write M_n for the maximum of the sequence and S_n for its sum. The ratio variate $R_n = M_n/S_n$ is a quantity arising in the analysis of process speedup and the performance of scheduling tasks in parallel. O'Brien (1980) showed that $R_n \to 0$ almost surely as $n \to \infty$ if and only if $\mathbf{E}[X_1] < \infty$. Since $\{R_n\}$ is a uniformly bounded sequence, it follows that $\mathbf{E}[X_1] < \infty$ implies $\mathbf{E}[R_n] \to 0$ as $n \to \infty$.

Here we show that, provided either (i) $\mathbf{E}[X_1^2] < \infty$ or that (ii) $1-F(t)$ is a regularly

varying function with index $p < -1$, it follows that

$$\mathbf{E}[R_n] = \frac{\mathbf{E}[M_n]}{\mathbf{E}[S_n]}[1 + o(1)], \quad (n \to \infty).$$

Since the asymptotics of $\mathbf{E}[M_n]$ is often readily calculated, this provides a useful estimate for the most significant behaviour of the ratio R_n in expectation. We apply this result to multiprocessor scheduling policies, obtain rates of convergence of list scheduling policies to optimality in expectation (the ratio of the makespan of a fixed policy to that of the optimal list schedule) and to the behaviour of sample statistics.