DIMACS

Probabilistic Analysis of Algorithms

May 11–14, 1997

Organizers:

Professor Alan Frieze
Carnegie Mellon University

Professor Mike Molloy
University of Toronto

Contents

Limit Theorems in Geometric Probability

Joseph E. Yukich
Dept. of Mathematics
Lehigh University
Bethlehem, PA 18015-3174
`jey0@LEHIGH.edu`

Abstract

We prove a.s. limit theorems for the lengths of graphs of problems arising in Geometric Probability. We describe the asymptotics of the lengths of graphs of the following problems:

1. classical Euclidean optimization problems (TSP, MST, and minimal matching)
2. geometric location problems,
3. minimal triangulations,
4. bi-partite matching,
5. k nearest neighbors graph,
6. Voronoi and Delaunay tessellations.

While some of these problems have been extensively investigated, different problems have often required different techniques. We show that these problems have intrinsic similarities and that they may be successfully studied using a single approach.

Improving Ajtai's Connection

Jin-Yi Cai
Dept of Computer Science
SUNY Buffalo
`cai@cs.buffalo.edu`

Abstract

Recently Ajtai established a connection between the worst case and the average case complexity of the shortest lattice vector problem. Based on this work Ajtai and Dowk have proposed a cryptographic protocol whose security is provably equivalent, up to a large polynomial factor, to a well known lattice problem, in the worst case. (Compare this with the RSA and the factoring problem.)

In provable terms, the Ajtai result is still far from being efficient for cryptographic purposes. It is also intellectually challenging to investigate the true extent to which the average case and the worst case complexity can be reduced to each other. We present improvements to this work, reducing the exponent in the polynomial factor in the connection from around 8 to around 3, making the prospect of cryptographic and other applications more realistic.

Joint work with Ajay Nerurkar.

Locality-preserving Hashing in Higher Dimensions

Santosh Vempala
School of Computer Science
Carnegie-Mellon University
5000 Forbes Avenue
Pittsburgh PA 15213
`svempala@cs.cmu.edu`

Abstract

Hashing is an excellent technique for data storage and look-up. We consider locality-preserving hashing, where, in addition to the usual requirement of low expected bucket size, neighboring points are mapped to neighboring points. In a recent paper, Linial and Sasson showed that for points from the integer line, there exists a family of hash functions with these properties. We present locality-preserving hash families where the domain is a d-dimensional cube and show similar guarantees. We mention some lower bounds which indicate that our results are tight upto constant factors.

A Nonreversible Markov Chain Sampling Method

Persi Diaconis
Cornell & Harvard Universities
Maths and ORIE
`http://www.math.harvard.edu/~diaconis`

Radford Neal
Statistics & Computer Science
University of Toronto
`radford@cs.toronto.edu`

and

Susan Holmes
Department of Bioemtry
Cornell University
`sph11@cornell.edu`

Abstract

We introduce a nonreversible version of the Metropolis algorithm and show how it can improve on the diffusive behavior of the classical algorithm. The running time of some versions is analysed using both probabilistic techniques and an explicit diagonalisation. Numerical examples suggest real improvements over Metropolis in some cases.

Redundancy of the Lempel-Ziv Incremental Parsing Rule

Serap Savar
Bell Labs
Murray Hill, NJ 07974
savari@research.bell-labs.com

Abstract

The Lempel-Ziv codes are universal variable-to-fixed length codes that have become virtually standard in practical lossless data compression. For any given output string from a unifilar, Markov source, we use renewal theory to upper bound the difference between the number of binary digits needed by the Lempel-Ziv incremental parsing rule (LZ '78) to encode the string and the self-information of the string. We use this result to demonstrate that for unifilar, Markov sources, the redundancy of encoding the first n letters of the source output with LZ '78 is $O(1/(\ln n))$, and we upper bound the exact form of convergence. We conclude by considering the relationship between the code length and the empirical entropy associated with a string.

Probabilistic Analysis of String-matching Algorithms

R. T. Smythe
George Washington University
Washington, D.C.
`smythe@gwis2.circ.gwu.edu`

Abstract

We investigate the probabilistic behavior of the Boyer-Moore (BM) string-matching algorithm and a variant of this algorithm, the Boyer-Moore-Horspool algorithm (BMH). We consider the problem of finding all occurrences of a pattern of length m in a text of length n, as n becomes large.

For the case of a texgt consisting a i.i.d. characters from a (non-uniform) finite alphabet, there is an explicit evaluation of the average number of comparisons as n gets large for BMH, as well as a central limit theorem for the number of comparisons. The case of BM is more difficult: for simple examples we can compute the average number of comparisons, but the problem rapidly becomes computationally infeasible. A central limit holds for this case also.

When the text is governed by a finite irreducible Markov chain rather than an i.i.d. sequence, some results for the BMH case can still be deduced, using the theory of Markov renewal processes.

Algorithm analysis via diffusion processes

René Schott
CRIN-CNRS
Universitté Henri Poincaré
Nancy, France
`Rene.Schott@loria.fr`

Abstract

We will show trough several examples (dynamic algorithms, distributed algorithms, exhaustion of shared memory, data structures maxima, etc...) how to use diffusion processes in order to characterize time and space complexities. Brownian motion, Brownian excursion, Brownian motion with reflecting and absorbing boundaries, Ornstein-Uhlenbeck processes, stochastic integrals are among the diffusion processes which appear in the analysis. Mean values, variances and limiting ditributions of the costs are explicitely calculated.

Randomized Splay Trees

Martin Fürer
Department of Computer Science & Engineering
Pennsylvania State University
220 Pond Laboratory
`furer@cse.psu.edu`

Abstract

In order to compare the performance of different methods to handle binary search trees, we need to investigate the constant factors in the running times. We use the following simple measure. Traversing an edge to access an item costs one time unit, while a single rotation costs R time units. (R might have a value of about 3 for a reasonable implementation.)

With this measure the basic splaying algorithm of Sleator and Tarjan runs with an amortized time bound of $3(1 + R)\log n + O(1)$, while their simple splaying algorithm only uses $2(1 + R)\log n + O(1)$ amortized steps per access (log's are to the base 2).

We study several randomized versions of splaying algorithms using a moderate amount of randomness (O(1) coin tosses per access). The aim is to decrease the constant factor $2(1 + R)$. We actually don't improve the constant factor in the worst case. But the expected time is improved, while the factor for the (worst case) amortized time stays the same. The expectation is taken over all possible coin tosses for a worst case sequence of access paths. Furthermore, we observe a trade-off between expected and amortized time. If the amortized time per access and per insertion are only required to be $O(\log n)$, then the expected time per access can be improved more drastically.

Finally, a generalized global splaying method is introduced. It allows even the amortized time to decrease, if the selection of a new parent is used as a basic step. A single rotation consists of 3 such basic steps. Interesting open problems concern the dependencies between the Dynamic Optimality Conjectures for the various splaying methods.

Randomized Allocation Processes

Artur Czumaj
Heinz Nixdorf Institute &
Department of Mathematics & Computer Science
University of Paderborn
Fuerstenallee 11
Paderborn, Germany D-33102
`email artur@uni-paderborn.de`

Abstract

We investigate various processes allocating balls into bins that arise in applications in dynamic resource allocation and on-line load balancing. We consider the scenario when m balls that arrive sequentially are to be allocated into n bins on-line and without using a global controller.

Traditionally, the main aim of allocation processes is to place the balls into bins to minimize the maximum load in bins. However, in many applications it is equally important to minimize the number of random choices performed (the allocation time). We study adaptive allocation schemes that achieve optimal tradeoffs between the maximum load, the maximum allocation time, and the average allocation time.

We investigate allocation processes that may reallocate the balls. We provide a tight analysis of the maximum load of on-line processes that during placing a new ball may reassign the balls within up to d randomly chosen bins.

We extend our analysis to infinite processes, in which in each step a random ball is removed and a new ball is placed according to our scheduling rule. In particular, we provide a tight estimation of the time needed for the infinite process to be in the state very close to its equilibrium.

Finally we give a tight analysis of the maximum load of the off-line process in which each ball may be placed into one of d randomly chosen bins. We apply this result to competitive analysis of on-line load balancing processes.

(Joint work with Volker Stemann)

Tree-Growing Search Strategies

Hosam M. Mahmoud
Department of Statistics
The George Washington University
Washington, D.C. 20052
`hosam@gwis2.circ.gwu.edu`

Abstract

Using the concept of a "tree-growing" search strategy, we show that most practical insertion sort algorithms are "normal," i.e. the number of comparisons they use (suitably normed) is asymptotically distributed as a Gaussian random variable. A sufficient condition for asymptotic normality is presented. The condition specifies the relation between the variance of the number of comparisons and the rate of growth in height of the sequence of trees that the search strategy "grows."

Inferring big divergent trees from realistic length biomolecular sequences

Tandy Warnow
Department of Computer & Information Science
University of Pennsylvania
Philadelphia PA 19104-6389
`tandy@central.cis.upenn.edu`

Abstract

The construction of evolutionary trees is a fundamental problem in biology. All current polynomial time methods (to our knowledge) are unlikely to recover the topology of the true tree from sequences of realistic lengths (bounded, perhaps, by 10,000 nucleotides) for large sets of widely divergent sequences, even if such methods are known to be able to reconstruct the correct topology given long enough sequences. Therefore, all current polynomial time methods may be incapable of addressing difficult and important data sets such as the Tree of Life. We address this problem and present a new polynomial time algorithm for reconstructing evolutionary trees called the Short Quartets Method. Our method computes topologies of subtrees induced by "short" quartets, which avoid the problematic influence of pairs of widely divergent sequences. We reconstruct a tree on the entire set of sequences from these inferred subtrees when they are compatible. We describe a simple implementation of this method based upon applying the four-point method to derived distances, and study its convergence rate for inferring trees under the general Markov model of evolution. We show that for each fixed range of the mutation probabilities associated to the model tree, our method has a convergence rate that is at worst polynomial, while it is polylogarithmic for typical trees. By comparison, all previous distance-based methods (including Neighbor-Joining, perhaps the most popular method in use, and more recent algorithms developed by the theoretical computer science community) have convergence rates that generally grow superpolynomially. Consequently, our method, even implemented using distances, should achieve accuracy on significantly shorter sequences than these other distance-based methods.

This is joint work with Peter Erdos (Hungarian National Academy of Sciences), Mike Steel (Univeresity of Canterbury, Christchurch New Zealand), and Laszlo Szekely (University of South Carolina).

Discrete Tomography

Larry Shepp
Rutgers University
Bushch Campus Piscataway, NJ 08854
`shepp@stat.rutgers.edu`

Abstract

The discrete Radon transform of a finite subset, S, of the integer lattice Z^2 or Z^3 is the number, $\nu_S = \nu_S(L)$, of points in $L \cap S$, for any straight line, L. Using a transmission electron microscope, one can measure $\nu_S(L)$ for any line L in a main direction of the lattice. It is desired to reconstruct S from the measured values of ν_S.

Practical limitations dictate that only a small number (like 4) of directions can be specified. Despite this severe limitation, despite the NP hardness of the inverse problem, and despite the provable non-uniqueness of the inverse problem, useful reconstructions of S can be obtained.

Bin Packing in a Random World

E. G. Coffman, Jr.
Bell Labs
Lucent Technologies
700 Mountain Ave. Rm. 2C-381
Murray Hill, NJ 07974
egc@research.bell-labs.com

Abstract

In this talk, we focus on the probabilistic analysis of bin packing algorithms; the algorithms are mostly of greedy type and the analysis is mostly asymptotic. The discussion centers on relatively new directions, interval packing and dynamic packing being two such. In the former, a bin is still taken as the unit interval $[0, 1]$, but the items to be packed are taken as random subintervals of $[0, 1]$, thus adding item position to the usual item size of classical bin packing; in valid packings, no two items can overlap. In dynamic variants of both interval and classical bin packing, items come and go at random times. Expected wasted space under simple greedy algorithms like First Fit and Best Fit remains a prime objective, but in the more demanding, dynamic formulations, a smaller objective is often necessary and sufficient conditions for stability and for the existence of finite expected values. Massive change in services and technology in the communications arena has created a surge of new applications. We describe some and show where they yield new problems or just a new terminology in which to express old problems.

Average Case Analyses of First Fit and Random Fit Bin Packing

Michael Mitzenmacher
Digital Systems Research Center
130 Lytton Avenue
Palo Alto, CA 94301
`michaelm@pa.dec.com`

Abstract

We prove that the First Fit bin packing algorithm is stable under the input distribution $Uk-2,k$, settling an open question from the recent survey by Coffman, Garey, and Johnson. Our proof generalizes the multi-dimensional Markov chain analysis used by Kenyon, Rabani, and Sinclair to prove that Best Fit is stable under this distribution. Our proof is motivated by an analysis of Random Fit, a simpler packing algorithm related to First Fit, which we also show is stable under this distribution.

Edge coloring in polynomial time, on average

Ljubomir Perkovic
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
`ljuca+@cs.cmu.edu`

Abstract

I present an algorithm that optimally colors the edges of a simple graph. I then show that the average running time of the algorithm on a randomly chosen graph on n vertices is polynomial in n, assuming a uniform distribution on graphs with n vertices.

Joint work with B. Reed.

Algorithmic and Combinatorial Aspects of the Connectedness of Random Graphs

Nicola Galli[1] *and* **Klaus Simon**
Institute for Theoretical Computer Science
ETH-Zentrum
CH-8092 Zürich
{galli,simon}@inf.ethz.ch

Abstract

Markov Chains are a very adequate tool to study random graphs. In particular, we can interpret many problems as disintegration-like or Yule-like processes. The main purpose of this talk is to analyze the connectedness of a graph as a disintegration-like process.

First we consider the number of isolated vertices in a random graph $G_{n,p}$. Let $P_{n,\ell}$ be the probability that ℓ of the n vertices are isolated and $q = 1 - p$, then $P_{n,\ell}$ satisfies

$$P_{n,\ell} = q^{n-1} \cdot (P_{n-1,\ell-1} - P_{n-1,\ell}) + \sum_{i=0}^{n-1-\ell} \binom{\ell+i}{i} \cdot p^i \cdot q^\ell \cdot P_{n-1,\ell+i},$$

where $P_{0,0} = 1$ and n represents a discrete time. Although this formula looks intractable, we show that it implies a much simpler recursion for the generating function

$$g_n(z) \stackrel{\text{def}}{=} P_{n,0} + z \cdot P_{n,1} + \cdots P_{n,n} \cdot z^n = g_{n-1}(1 + q(z-1)) + q^{n-1} \cdot (z-1) \cdot g_{n-1}(z).$$

This can be solved by traditional methods yielding exact closed forms for the distributions and the moments.

Next, we extend the previous example to the number of isolated connected substructures (cliques, trees, etc). From the analysis of the multivariate generating function of the corresponding joint probability distribution, we derive exact forms, asymptotic approximations, and results on their convergence rate. In particular, we show that the multivariate case does not essentially differ from the univariate one. Further, these formulæ explain very simply the relationship between the probabilistic and the combinatorial aspects of this problem.

Finally, we analyze the Breath-First-Search (BFS) algorithm in a similar way and we demonstrate the equivalence between the connectedness of a graph and the number of edges considered during the BFS. More precisely, for the case that an edge exists with probability

$$p = \frac{c \cdot \log n}{n},$$

where $c > 1$, we find that the expected number of considered edges is given by

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

which imply the threshold $p = \frac{\log n}{n}$ for the connectedness (that is, we give an algorithmic proof for this problem). Moreover, this interpretation of the processes explains in a simple manner why the number of isolated nodes and the connectedness of a graph have the same threshold behavior: They are both disintegration processes.

Further, we derive similar results for Depth-First-Search and we discuss the trade-off between the number of reached nodes and the considered edges, which can be used to find a strategy that reduces the costs of a search.

The Analysis of a List-Colouring Algorithm on a Random Graph

Dimitris Achlioptas *and* **Michael Molloy**
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada M5S 1A1
{optas, molloy}@cs.toronto.edu

Abstract

We introduce a natural k-coloring algorithm and analyze its performance on random graphs with constant expected degree c ($G_{n,p=c/n}$). For $k = 3$ our results imply that almost all graphs with n nodes and $1.923n$ edges are 3-colorable. This improves the lower bound on the threshold for random 3-colorability significantly and completely settles a long-standing open question of Bollobás. We also provide a tight asymptotic analysis of the algorithm. We show that for all $k \geq 3$, if $c \leq k \ln k - 3/2k$ then the algorithm almost surely succeeds, while for any $\epsilon > 0$, and k sufficiently large, if $c \geq (1 + \epsilon)k \ln k$ then the algorithm almost surely fails. The analysis is based on the use of differential equations to approximate the mean path of certain Markov chains.

Average-case complexity of shortest-paths problems in the vertex-potential model

Volker Priebe
Max-Planck-Institut fuer Informatik
Im Stadtwald
Saarbruecken, Germany D-66123
`priebe@mpi-sb.mpg.de`

Abstract

Average-case complexity of shortest-paths problems in the vertex-potential model.

We study the average-case complexity of shortest-paths problems in the vertex-potential model. The vertex-potential model is a family of probability distributions on complete directed graphs with *arbitrary* real edge lengths but without negative cycles. We show that on a graph with n vertices and with respect to this model, the single-source shortest-paths problem can be solved in $O(n^2)$ expected time, and the all-pairs shortest-paths problem can be solved in $O(n^2 \log n)$ expected time.

This is joint work with Colin Cooper, Alan Frieze, and Kurt Mehlhorn.

Computer Simulation Demonstration

Achieving 100% of Capacity in Mobile Cellular Networks Under Uniform Random Distribution of Mobile Units

(Demonstrated 1:15 pm, Wednesday, May 14th, Room C217)

David W. Matula
Southern Methodist University
Dallas, Texas 75275
matula@seas.smu.edu

Abstract

This simulation demonstrates the evolution of mobile unit frequency channel assignment in an $n \times n$ triangular lattice with f frequency channels per lattice point as demand evolves from 0 to 100% ($= fn^2$ mobile units in network).

For $f = 1, 2$, the evolution is seen to behave much as chemical polymer gellation - small blocked regions/larger blocked regions/growth of an "infinite" blocked region encompassing a greater and greater portion of the region.

For $f = 6, 7$, a paradigm shift has occurred. The evolution shows almost no blockage (e.g. $< 1\%$), until demand reaches $100 - \epsilon)\%$ of fn^2, at which time there is a "big bang" chaotic transformation to almost all blockage (e.g. $> 99\%$ of region blocked).

A related theoretical result and conjecture is stated as follows:

Balls & Urns on a Lattice

An $n \times n$ triangular lattice is embedded on a torus with an urn at each of the n^2 lattice points. A total of kn^2 balls are thrown into the triangles uniformly at random. Each ball may be placed in any of the three urns at its triangular boundary. The goal is to assign the balls to the urns to balance the number in each urn. Let j be the difference between the least and most in the n^2 urns according to the assignment that minimizes j.

The simulation supports the conjecture (some parts proved) that if $k = \Omega(\log n)$, then the probability of a perfect assignment ($j = 0$) approaches unity for sufficinetly large n.

Average-case analysis of algorithms for polynomials over finite fields

Daniel Panario
University of Toronto
Department of Computer Science
Sanford Fleming Building
10 King's College Road
Toronto, Ontario, Canada M5S 1A4
daniel@cs.toronto.edu

Abstract

In this talk, we focus on how analytic combinatorics can be used in algebraic problems that deal with polynomials over finite fields. The kind of problems we are interested in are the factorization of polynomials, the irreducibility test of polynomials, and so on (all over finite fields).

First, we briefly introduce generating functions and asymptotic analysis. Then, we discuss in some detail the average-case analysis of some algorithms for factoring polynomials over finite fields. Finally, as time permits, we comment on other related problems.

Partially based on joint works with Philippe Flajolet, Joachim von zur Gathen, Shuhong Gao, and Xavier Gourdon.

Average Complexity: Traps and Escapes

Leonid Levin
Boston University
CAS Computer Science
111 Cummington Street
Bonston, MA 02215
`lnd@cs.bu.edu`

Abstract

Average case analysis always seemed more relevant than the worst case. But its many pitfalls kept theorists on a safe distance for a long time. In the last decade we gained some experience in robust and safe navigation among the many subtleties of the minefield which I will try to share.

Average Case Complexity

Jie Wang
Department of Mathematical Sciences
University of North Carolina at Greensboro
Greensboro, NC 27412
`wangjie@uncg.edu`

Abstract

This talk is the second part of the same topic following Levin's talk. In this talk, we survey average-case NP-complete problems and some of the techniques of obtaining average-case completeness results.

The Analysis of Linear Probing Hashing with Buckets

Alfredo Viola
Pedeciba Informatica
Casilla de Correo 16120
Distrito 6
Montevideo, Uruguay
`viola@fing.edu.uy`

and

Patricio Poblete
Department of Computer Science
University of Chile
Casilla 2777
Santiago, Chile
`ppoblete@dec.uchile.cl`

Abstract

We present the first exact analysis of a linear probing hashing scheme with buckets of size b. From the generating function for the Robin Hood heuristic we obtain exact expressions for the cost of successful searches. For a full table, with the help of Singularity Analysis, we find the asymptotic expansion of this cost up to $O((bm)^{-1})$. We conclude with a new approach to study certain recurrences that involve truncated exponentials. A new family of numbers that satisfies a recurrence resembling that of the Bernoulli numbers is introduced. These numbers may prove helpful in studying recurrences involving truncated generating functions.

So, our analysis solves the open problem presented by D. Knuth in question 6.4.56 [M48] in the volume 3 of the collection "The Art of Computer Programming".

Data Sensitive Analysis and a Posteriori Performance Guarantees

Mark Jerrum
University of Edinburgh
currently visiting DIMACS
`jerrum@lunar.rutgers.edu`

Abstract

Worst-case analyses of the computational complexity of problems or algorithms often result in bounds that are unduly pessimistic when considered in the light of computational experience, the exponential worst-case running time of the simplex algorithm being a particularly striking example. One response to this empirical observation has been to turn to average case analyses. While it sometimes leads us into to some interesting mathematics, average case analysis can be criticised on both practical and philosophical grounds: the probability distribution on problem instances is usually unknown, and even the assumption of its existence is likely to be invalid.

An alternative approach is to study algorithms for which a useful performance guarantee is not known a priori, but becomes available as a side effect of the computation. This talk will concentrate on examples of this approach in sampling and learning.