D I M A C S

Center for Discrete Mathematics and Theoretical Computer Science

DIMACS SPECIAL YEAR ON MASSIVE DATA SETS: 1998 ANNUAL REPORT

DIMACS Center • Rutgers University • 96 Frelinghuysen Road • Piscataway, NJ 08854-8018 • USA
TEL: 732-445-5928 • FAX: 732-445-5932 • EMAIL: `center@dimacs.rutgers.edu`
WWW: `http://dimacs.rutgers.edu/` • TELNET: `telnet info.rutgers.edu 90`

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs - Research, Bell Labs, NEC Research, and Telcordia Technologies (formerly Bellcore)

Overview:

The DIMACS special year on massive data sets began with prespecial year activities in 1996-97 and is continuing through 1999.

Organizers:

- Robert Sedgewick, Princeton University (chair)
- Jon Bentley, Bell Labs
- Martin Farach-Colton, Rutgers University
- Joan Feigenbaum, AT&T Labs - Research
- Helene Kulsrud, IDA-CCR, Princeton

Sponsors and Collaborating Institutions:

The National Science Foundation
The New Jersey Commission on Science and Technology
AT&T Labs - Research
Bell Labs
Bellcore
Institute for Defense Analyses/Center for Communications Research (IDA/CCR)
NEC Research Institute
Rutgers University
Princeton University

Theme of Program:

The emergence of the computer as an essential tool in scientific research and an essential ingredient in commercial systems has led to the proliferation of massive amounts of data: Data from astrophysical models, genetic sequencing, geographical information systems, world-wide ecological monitoring, weather prediction, telecommunications applications, biological models for medical applications, government information systems. Researchers in all of these applied areas currently face daunting computational problems in organizing and extracting useful information from these massive data sets. DIMACS' special year is seeking to stimulate new research to address these problems. We are seeking to learn common properties among diverse applications, identify fundamental computational problems that transcend particular applications, develop and analyze new practical algorithms, and establish partnerships with researchers in other areas, with the goal of developing solutions for specific scientific and commercial problems.

Workshops, Miniworkshops and Special Events, Feb. 1998 - Nov. 1998

- **MiniWorkshop on Mathematical Methods for High Performance Data Mining**
  - Dates: April 27-28, 1998
  - Location: Princeton University
  - Attendance: 52
  - Organizers: Helene Kulsrud, Institute for Defense Analyses;
    Robert Grossman, University of Illinois

The goal of the workshop was to survey some of the successful methods for Data Mining in Massive Data Sets. Unlike most of the other workshops in this series, applications were not restricted to a single field. Instead, in depth descriptions of five methods and the application areas to which they have been successfully applied were presented. This seemed particularly useful at this time when many of the well known methods are being extended to apply to applications different from those for which they were originally developed. There were several talks on Latent Semantic Indexing using Singular Value Decomposition (SVD), a method originally developed for problems in document retrieval. Michael Littman gave a general introduction to the method and described some exciting new applications such as scoring standard essay tests and modeling human lexical acquisition. Randall Dahlberg presented a new application to cross-language identification and emphasized the need for clustering techniques to identify the results. A significant difficulty in the use of SVD is the determination of the exact number of singular values to use in a given problem. Theoretical research on this problem is being carried out by Santosh Vempala and his colleagues and was presented by Vempala. As a result of Vempala's presentation, Littman is attempting to include these new methods into his suite of SVD programs.

Robert Grossman led off the talks on the use of tree based classifiers. Due to the sensitivity of these methods to the exact training data, Grossman emphasized the new ideas for combining regression trees - bagging, boosting and attribute averaging. He touched on the problems of distributed data and the use of parallelism. Work on the application of regression trees to problems in the medical diagnosis field were cited as good examples of the need for these new methods. Shannon gave a more complete description of his experiences in using these combining methods to improve the applicability of regression trees. Volinsky showed the use of Bayesian Classifiers to solve similar problems. On the theory side, Jorma Rissanen presented some new ideas for the design of classification trees using minimum description length and discussed some interesting problems when the minimum code length is one.

The status of the use of Association rules was covered by two papers directed at the use of computer parallelism. These studies are extremely significant since we will need all the parallelism we can muster to apply these methods to the petabyte data sets. Mohammed Zaki described an expansion of the "A Priori" method to work in a parallel computer. Beginning with "A Priori" assumptions, Vipin Kumar and his group have developed a hypergraph method which creates clusters of objects and allows for parallel computation, and produces better results than some standard clustering techniques such as Auto-class or K-means.

The current status of the statistical method called Acquaintance was presented by Claudia Pearce.

The talks of both Kumar and Volinsky indicated another area of progress. These presentations illustrated the use of combined methods and may indicate a trend to deal with the difficult aspects of individual methods. Volinsky begins with trees generated with uncertainty and applies Bayesian model averaging to produce better probabilities in the models.

The conference concluded with a startling and important paper presented by Michael Jordan of MIT. This work attempts to unify the entire field of graphical methods and set it on firm and consistent mathematical foundation. The graphical model formalism is an elegant combination of graph and probability theory that allows realistic and flexible models to be formed. In particular the problem of dense graphs, usually impractical to solve, is given a new framework. This paper encouraged the listeners to think about new methods.

- **Astrophysics and Algorithms: A DIMACS Workshop on Massive Astronomical Data Sets**
  - Dates: May 6-8, 1998
  - Location: Princeton University
  - Attendance: 72
  - Organizers: Michael Vogeley, Princeton University; David Spergel, Princeton University; Don Sun, Bell Labs; Tony Tyson, Bell Labs

This workshop was motivated by the "data avalanche" that astronomers now face, as several observational projects have produced or promise to yield Terabyte-class data sets. Examples include sky surveys in optical and infrared wavelengths, microwave background anisotropy satellite experiments, helioseismology data, gravitational radiation detection experiments, and results from N-body/hydrodynamical simulations. Processing of these data and subsequent statistical analyses clearly require advanced computational methods. Issues of scalability in database design and analysis algorithms that are well-known in, e.g., the telecommunications industry and defense establishment are now being grappled with by astronomers. A principal goal of this workshop was to foster interaction between the Astronomy community and Discrete Mathematics and Theoretical Computer Science communities. The participants were an unusual mix of mathematicians, astronomers, and computer scientists. The majority of the speakers were astronomers who are involved in collecting, archiving, and analyzing large astronomical data sets along with notable contributions from researchers in discrete mathematics and computer science.

The workshop began with keynote addresses by leaders in the respective fields of Astrophysics and Algorithms. Princeton University provost and astrophysicist Jeremiah Ostriker traced the evolution of computing in astrophysics from his own experiences as a graduate student student at Yerkes Observatory to his state-of-the-art N-body/hydrodynamical simulations on a 64-processor SGI Origin 2000. Robert Sedgewick noted the differing nature of computing in Astronomy and Theoretical Computer Science.

One session addressed analysis of large optical surveys. David Bennett described analysis of an imaging survey (500GByte) to look for evidence of MACHOS (Massive Compact Halo Objects) that might comprise some of the so-called "dark matter" in the universe. David Monet detailed the effort to catalog objects detected on photographic plates (object detections now number in excess of 6 billion, stored on 581 CD-ROMs) and the challenge of archiving and analyzing data products that range from "G-rated" (object catalogs) to "X-rated" (the raw pixel database). Barry Lasker discussed all-sky image and catalog databases that have also been constructed from photographic plates, which are now on-line for use by the astronomical community. Lasker's talk highlighted database issues such as cross-matching, classification, and sky partitioning.

A second session moved to astronomical surveys at longer wavelength. Carol Lonsdale described the data processing and archiving of results from the 2 Micron All Sky Survey (2MASS) project which collects 19 Gbyte of data per night. Rick White summarized status, results, and analysis algorithms used for the FIRST radio survey, the results of which are also available to the astronomical community from a WWW page. The NASA/IPAC Extragalactic Database, which archives all published extragalactic object detections and serves this catalog over the Internet, was reviewed by Barry Madore. Returning to analysis of optical data, Steve Odewahn discussed his use of neural network methods for galaxy classification.

A third session focused on database methods and computing technology. Alex Szalay presented an overview of the Sloan Digital Sky Survey (SDSS) that is now beginning to image most of the Northern sky in five optical bandpasses and will also obtain spectroscopy of one million galaxies. The raw data for this sky survey will comprise more than 40TByte, reduced to 1TByte of processed data. Laney Kulsrud reviewed mathematical methods for data mining in massive data sets, particularly algorithms that scale to Petabyte-class data sets. Tom Prince proclaimed that "10 Terabytes is not a large data set" and charted the growth of computing technology, both in the past and predicted for the near future. Much discussion ensued about the need for people (and/or compilers) who know how to take advantage of the benefits of parallel computing.

A fourth session was devoted to issues of data analysis. Sarbani Basu spoke about methods for analyzing helioseismology data, in particular mathematical methods applied to invert the spectrum of modes in the Sun. William Press's talk "How to Connect the Dots, FAST, in a Noisy World" focussed on the problem of curve fitting with noisy, sparse, irregularly sampled data with inhomogeneous errors, and promised code to do so from his WWW site. George Djorgovski reviewed scientific results from analysis of catalogs of objects found from digitization of optical plates from the second Palomar Sky Survey, emphasizing AI methods to classify the objects. Andreas Brieden discussed the mathematical problem of deriving approximate relations for properties of convex hulls in n-dimensional spaces. Sergei Shandarin presented a set of statistics to describe the shape of compact surfaces that are useful for description of the large-scale distribution of matter in the universe.

A fifth session was devoted to problems of analysis of the anisotropy of the Cosmic Microwave Background radiation (CMB). Francois Bouchet reviewed plans and status for future satellite missions and our current knowledge of anisotropy of the CMB, and described the computational challenge of reducing the one hundred billion data points to ten or so fundamental physical quantities. Regarding the latter, Bouchet pointed out the need for matrix inversion methods that are significantly faster than order n-cubed. Max Tegmark elaborated on mathematical methods for making maps of the CMB sky from such satellite data and described lossless methods for power spectrum estimation from these data. Dan Rockmore detailed methods for fast spherical transforms that compute in order n-log(n) time. His experience with different compilers/platforms led to discussion of analysis of algorithms viz. the dependence of performance on memory access and other factors. Julian Borrill discussed fast methods to bound the likelihood function of cosmological models that predict the observed power spectrum of the CMB anisotropy.

A final session dealt with classification problems and multiscale image analysis. Jon Hakkila described attempts to classify gamma-ray bursts, the fundamental properties of which are poorly understood, using data mining techniques and articial intelligence classification methods. Fionn Murtagh and Albert Bijaoui both spoke about multiresolution methods for image analysis in Astronomy. Renyue Cen discussed the computational challenges of computing and analyzing cosmological simulations that yield a Petabyte of stored data. Cen also described a novel method for out-of-core Fast Fourier Transforms.

Astronomers at this workshop were pleased to find that, when compared with some other fields, Astronomy does not yet have a massive data set problem and that many of the challenges faced by astronomers have been addressed in other contexts. The key to successful astronomical data analysis is clearly to collaborate with researchers in computer science and in fields where scalable algorithms have been developed. Although the computer scientists faced the daunting task of decoding much astronomical jargon, once this hurdle was mounted many productive discussions followed regarding how existing technology could be applied to Astronomy problems. Several speakers from the Astronomy side suggested problems that might be of interest for, e.g., a CS graduate student.

A common theme of this workshop was that Astronomy faces a paradigm shift, following which many researchers will spend their time data mining a "Digital Sky" compiled from the vast array of multi-wavelength sky surveys now underway, in contrast to individual efforts at a single telescope (though certainly that mode will continue to provide results at the cutting edge of research). Clearly, collaboration with the DIMACS community will be of increasing importance.

- **Workshop on External Memory Algorithms and Visualization**
  - Dates: May 20-22, 1998
  - Location: Rutgers University
  - Attendance: 54
  - Organizers: James Abello, AT&T Labs; A.Birilis, AT&T Labs; Steve Mahaney, NSF; Yossi Matias, Bell Labs; T.Munzner, Stanford; W. Sweldens, Bell Labs; Jeff Vitter, Duke University; Kiem-Phong Vo, AT&T Labs.

  We are witnessing a proliferation of applications that need to process large amounts of information. I/O between main and external memory is becoming an increasingly significant bottleneck. How, where and in what circumstances to involve the user are central questions and visualization may provide some of the answers. This workshop aimed at fostering the cooperation among theoreticians and practitioners in the fields of External Memory Algorithms and/or Visualization. Papers were presented on the following topics as they related to the workshop theme: Computational Models and Complexity; Memory Allocation and System Integration; Indexing Methods, Data Compression and Succinct Data Structures; External Memory Algorithms in Graph Theory, Computational Geometry, Combinatorics, Data Bases, Linear Algebra and Statistics; Visualization Metaphors and Multiresolution Frameworks; User Interfaces for External Memory Algorithms; Integrated Systems: Examples and Applications.

- **DIMACS Workshop on Combinatorial Clustering and Multi-Domain Protein Structure Analysis**
  - Dates: June 26-27, 1998
  - Location: Rutgers University
  - Attendance: 52
  - Organizers: Sylvia Spengler, Manfred Zorn, Inna Dubchak, Fred Roberts, Ilya Muchnik, Casimir Kulikowski

  The combinatorics, graph theory and algorithms branches of discrete mathematics are powerful techniques for use in analysis of diverse sets of data. Combinatorics introduces clustering models, forms a basis for the interpretation of results, and brings new ideas to clustering methods. The area of combinatorial clustering provides new opportunities for systematic (routine) and comprehensive study of very large databases containing highly complex nonregular elements. However, researchers developing methods for combinatorial clustering often do not collaborate with specialists in the creation, management and analysis of large, complex databases. The goal of this workshop was to foster and facilitate such a collaboration by focusing on a particular application. The specific problem addressed was in the area of protein structure and analysis of multi-domain proteins (as an example for other type of interactions among large bio-molecules). By hosting this workshop, DIMACS' aim was not only to stimulate research at the interface between these disciplines, but also to stimulate development of ideas and methods in each field.

Papers presented at the workshop can be loosely grouped into three areas: academic research in clustering methods (combining discrete math, computer science, operations research, and statistics); academic research in applications (protein multidomain dissection and function identification, secondary structure prediction, optimal threading, etc.); and research related with real application problems (targeted expression of gene families, receptor-ligand interactions, biochemical maps of protein surfaces, etc.). The papers on methods covered the latest dm/tcs approaches to clustering (visual clustering, clustering with model-based neural networks, quadratic Euclidean clustering) and the applied papers related to challenging issues in bioinformatics (threading by combinatorial optimization methods, protein databases through hidden Markov models, distant evolutionary relationships from protein sequence and structure).

The general perspective of the field became more clear as a result of the workshop. It appears that there is not much experience in systematic C2P-supported handling of massive data sets in bioinformatics. The large data sets currently maintained are maintained primarily through clustering tools developed 20 to 30 years ago. It is necessary to bring the power of modern combinatorial clustering to all classification, search, and database design problems in bioinformatics. Bioinformatics presents a variety of challenges to computer science, in particular on geometrical classification of 3D-protein structures, pre/post-processing stages of data sets, and analyses such as recognizing which segment of an amino acid sequence is a separate domain.

- **9th Annual Conference on Combinatorial Pattern Matching**
  - Dates: July 20-22, 1998
  - Location: Rutgers University
  - Attendance: 60
  - Organizers: Martin Farach-Colton, Rutgers University; Udi Manber, University of Arizona

  This conference emphasized combinatorial pattern matching and its applications. Specific topics covered included string algorithms, pattern recognition, applications in molecular biology, text searching, information retrieval, symbolic computing, natural language processing and data compression.

  Proceedings were published in the Springer-Verlag series Lecture Notes in Computer Science and distributed at the meeting.

  CPM is a small, specialty conference which alternates between Western Europe and North America. The North American conference had had declining attendance over the last few years, and this year reversed the trend. It is reasonable to believe that this was because of the strong synergy between the topic of CPM and various Massive Data and Computational Biology activities at DIMACS.

- **The Fourth International Seminar on Average-Case Analysis of Algorithms**
  - Dates: July 20-24, 1998
  - Location: Princeton University
  - Attendance: 47
  - Organizers: Philippe Flajolet, INRIA; Rainer Kemp, Johann Wolfgang, Goethe-Universitat; Hosam Mahmoud, George Washington University; Helmut Prodinger, Vienna University of Technology; Robert Sedgewick, Princeton University; Wojciech Szpankowski, Purdue University

The International Seminar on Average-Case Analysis of Algorithms is a biannual event. In 1998 DIMACS hosted the fourth of these seminars. The theme of the workshop was average-case analysis of algorithms, whereby one assesses the quality of an algorithm according to its underlying stochastic behavior. It is sometimes rather difficult to characterize the behavior of an algorithm under the various instances of its input. In such cases, experience shows that while the deterministic task may be formidable, by assuming a reasonable probability measure on the input, one can say something meaningful about the average behavior of the algorithm. More recently, research has gone into the more ambitious avenue of finding the complete probability distribution of characteristics of algorithms. Two major schools of thought approach the subject from two different angles. Given a problem, the "analytic school" sets out to represent the problem in terms of generating functions obtained from the structure of the problem. This approach usually leads to non-standard functional equations and recent developments indicate that a miscellany of integral transforms are best suited for the task. The "probabilistic school" seeks formulations in terms of sums of stochastic components to which laws of large numbers may be sought. This approach is sometimes simpler, but often gives less information than the analytic approach.

The algorithmic problems considered at the workshop can be broadly categorized as follows: Basic search and data structures (hashing tables, search trees, neighbor search, balanced trees); sorting and complex search problems (partial match queries, adaptive search); strings and textual data (exact and approximate string matching algorithms); graph algorithms and combinatorial optimization (spanning trees, graph partitioning, packing); systems algorithms (caching) and models of computation (circuits); seminumerical algorithms (greatest common divisors, polynomial factorization); information theory (entropy, faulty components detection, games).

The highlights of the meeting can be summarized in the following points. Efficient algorithms to cope with massive data sets should provide pragmatic guarantees of fast running time in practice. Traditional worst-case analysis should be a tool in every computer scientist's and practitioner's repertoire, but our area is providing and developing the most useful tools. It is better to have a practical algorithm with good expected performance and "low" variance than an impractical algorithm with worst-case guarantees. Constant factors and lower order terms *do* matter. A precise mathematical analysis may prove essential to improve an algorithm. The analysis points out the weak points of an algorithm and suggests ways to improve them, because the analysis requires deep understanding of the internal mechanics of the algorithm. The analysis can also predict what is to be quantitatively expected from the potential improvements. The community is developing more refined and realistic random data models. This is a key step towards the design of efficient algorithms (in pragmatic terms): one designs an algorithm to work well on typical instances; one does not spend one's efforts to cope with unlikely ill-behaved instances.

- **DIMACS Workshop on Codes and Trees: Algorithmic and Information Theoretic Approaches**
  - Dates: October 5-7, 1998
  - Location: Rutgers University
  - Attendance: 75
  - Organizers: Mordecai Golin, Hong Kong University of Science and Technology; Julia Abrahams, DIMACS, Rutgers University

This workshop had the aim of making connections between distinct research communities interested in code tree problems of which Huffman coding is the prototype. We were successful in bringing together groups representing the information theory, computer science algorithms, and formal languages communities, thereby promoting research linkages among people and perspectives which are typically disjoint.

Although Huffman coding is a very classical problem, interest in the area continues, prompted by its central role in data compression. In recent years, as the need to store and retrieve massive amounts of data increases, compression has become a key technology. In addition, many other problems in diagnostic testing, database search, and circuit design are isomorphic to the Huffman optimization problem and its variants. Furthermore, new developments in optimization theory provide techniques for the resolution of long-standing open questions in the design of optimal coding algorithms.

Three tutorials presented at the workshop were designed to acquaint representatives of the various distinct communities with issues and approaches they may not have been familiar with, from the discussion of Huffman problem variants typical of the information theoretic community as discussed by Julia Abrahams, to the discrete optimization techniques centered around Monge properties typical of contemporary approaches to optimization algorithms among the computer science community by Wolf Bein, to a discussion of automata theoretic problems and solutions by Veronique Bruyere.

Among the other speakers, Wojtek Szpankowski discussed the asymptotics of a tree structured approach to the problem of electing a leader. Kingo Kobayashi presented a method for representing the set of all k-ary trees through a hyper-tree encoding technique. Rudolf Fleischer discussed self organizing data structures, an area of on line algorithms, and its application in data compression. Mike Fredman described a problem in multiple parallel processors searching through a tree together with a reduction to a Huffman-like problem from which he was able to obtain a good but not optimal solution. Mordecai Golin presented a computationally efficient algorithm for the optimal solution of a constrained coding problem of significance in synchronization. Bob Sedgewick discussed quicksort, a different type of algorithmic problem. The program continued with Ruy Milidiu on maximum length constraints, Hiro Morita on using code trees for random number generation, Ian Munro and David Benoit on the succinct representation of trees, Karthik Visweswariah on parse trees, Erik Demaine on fast intersection for web search engine queries, Roberto De Prisco on the connection between redundancy and data expansion for in place Huffman encoding, Akiko Kato on infinite source alphabets, and Ion Mandoiu on the problem of optimally encoding source symbols in the presence of some prespecified codewords. Additional speakers included Andrej Brodnik on fast decoding techniques, Teresa Przytycka on heuristics for good trees within a constrained tree structure, Eduardo Laber on dynamic Huffman coding, Paul Tucker on applications in searching, Mihai Sipitca on applications in image compression, Amiya Bhattacharya on applications in wireless network location management, Helmut Jurgensen on language theoretic approaches to error correction, Yuriy Reznik on random search trees, Mamoru Hoshi on homophonic coding, Gil Shamir on context tree methods, and Artur Pessoa on efficient encoding techniques.

Events Planned for the Rest of 1998-99

- **Workshop on Algorithm Engineering and Experimentation (ALENEX99)**
  - Dates: January 15-16, 1999
  - Location: Omni Hotel, Baltimore, MD
  - Organizers: Catherine C. McGeoch, Amherst College; Mike Goodrich, Johns Hopkins University

- **Sixth Implementation Challenge Workshop on Near Neighbor Searches (a special session of ALENEX99)**
  - Dates: January 15-16, 1999
  - Location: Omni Hotel, Baltimore, MD
  - Organizers: Michael Goldwasser, Princeton University

Selected Research Results

Logical Analysis of Data and Clustering Peter Hammer (PM,RU) and Christian Lenart (V) investigated the relationship between logical analysis of data (LAD) and clustering. The objective of LAD is to produce a list of logical patterns (based on a given training set) which provide structural information about datasets; in particular, patterns are used to distinguish between so-called positive and negative observations (e.g., presence or absence of a certain disease). Based on experiments with various sets of data (taken from the Repository of Machine Learning Databases and Domain Theories maintained by the University of California at Irvine), they discovered that this relationship has two aspects. First of all, clustering can improve the performance of the LAD procedures developed by Hammer and his collaborators. The main idea is to apply a clustering procedure to the raw data and then to use the centers of the clusters to generate the patterns as logical descriptors of the clusters; the second step is based on a Boolean optimization procedure. Hammer and Lenart observed that this two-step procedure improves the results of the best algorithms for the diabetes dataset which appeared to be the most difficult for LAD. On the other hand, by a slight modification of the LAD procedures, one can obtain patterns corresponding to clusters whose overlap does not exceed a certain prescribed limit; hence LAD can be used for clustering purposes.

Pattern Generation is Hard Lijie Shi (GS,RU) examined the problem of creating a good pattern for a given set of data, where a pattern is defined as a collection of inputs that makes a given Boolean function take the value 1. This problem was proved to be NP-hard by Boros, Ibaraki, and Makino. Shi proved that this problem does not have any polynomial approximation algorithm with ratio $N^{1-\epsilon}$ unless P=NP, where $N =$ the number of positive data points, and further proved that if the pattern must satisfy one of the positive data points, the maximum pattern is still hard to approximate unless P=NP. This result answered an open problem of Peter Hammer. From the approximation viewpoint, the greedy algorithm for this problem is optimal.

Finding Optimal Boolean Classifiers John Franco (V) worked on the problem of finding optimal Boolean Classifiers. From a given data set consisting of a number of 0-1 vectors, each with a value of 0 or 1, one can develop a Boolean expression F which evaluates to true iff an assignment of values to variables of F matches a vector of value 1. Well studied procedures for creating such Boolean classifiers are identified with a research area known as the logical analysis of data. These procedures can output many classifiers for a given data set. The question is, which is the best classifier to use from the perspective of trying to minimize the error in classification among future data points? Franco proposed a statistical approach to solving this problem based on cross moments and covariances of existing data. The idea was to develop a reasonable parametrized model of data generation, use the data to set the parameters, then use one of a number of methods to generate an estimate, at each point, which minimizes the probability of error in some sense at that point. The need for a model seemed to be critical since it is easily possible to get poor results by estimating cross moments directly from the data if the underlying distribution of data points is not stationary.

Hence he experimented with data sets in a quest for reasonable models.

Optimization Algorithms for Separable Functions with Tree-Like Conjugacy of Variables and their Application to the Analysis of Massive Sets of Seismic Data In work of Vadim Mottl (V) and Ilya Muchnik (V), a massive data set, in particular, that of seismic data, was considered as a set of experimentally acquired values of a number of variables each of which is associated with the respective node of an undirected conjugacy graph that presets the fixed structure of the data set. The class of data analysis problems under consideration can be described by the assumption that the ultimate aim of processing can be represented as a transformation of the original data array into a secondary array of the same structure but with node variables of, generally speaking, different nature, i.e. different ranges. Such a generalized problem was interpreted as the formal problem of optimization (minimization or maximization) of a real-valued objective function of all the node variables. The objective function wass assumed to consist of additive constituents of one or two arguments, respectively, node and edge functions. The former carry the data-dependent information on the sought-for values of the secondary variables, whereas the latter are meant to express the a priori model constraints. For the case when the graph of the pairwise conjugacy of the data set elements has the form of a tree, an effective global optimization procedure was proposed which is based on a recurrent decomposition of the initial optimization problem over all the node variables into a succession of partial problems each of which consists of the optimization of a function of only one variable like the Bellman function of classical dynamic programming. Two kinds of numerical realization of the basic optimization procedure were considered on the basis of parametric representations of Bellman functions, respectively, for discretely defined and quadratic node and edge functions. The proposed technique of optimization of separable functions supported by trees underlies the algorithmic solution of a group of seismic data analysis problems which includes the problems of segmentation of explorative seismic sections through layered underground rock mass, texture analysis of plane and three-dimensional explorative seismic data arrays, and detection of seismic signals.

A Functional Approach to External Graph Algorithms James Abello (PM, ATTL), Adam Buchsbaum (PM, ATTL) and Jeffrey R. Westbrook (PM, ATTL) developed a divide-and-conquer approach for designing external graph algorithms, i.e., algorithms on graphs that are too large to fit in main memory. Their approach builds a succession of graph transformations that reduce to sorting, selection and a recursive bucketing technique. They applied their techniques to devise external algorithms for computing connected components, minimum spanning trees (BMSTs), and maximal matchings in undirected graphs and multi-graphs. These algorithms are purely functional. That is, each algorithm is specified as a sequence of functions applied to input data and producing output data, with the property that information, once written, remains unchanged. The function is then said to have no "side effects." A functional approach has several benefits. External memory algorithms may run for hours or days in practice. The lack of side effects on the external data allows standard checkpointing techniques to be applied, increasing the reliability of any real application. A functional approach is also amenable to general purpose programming language tranformations that can reduce running time. Abello, Buchsbaum, and Westbrook also considered a semi-external model for graph problems, in which the vertices but not the edges fit in memory. This is not uncommon in practice, and when vertices can be kept in memory, significantly more efficient algorithms are possible. They designed new algorithms for external grouping and sorting with duplicates and applied them to produce better I/O bounds for the semi-external case of connected components.

External Memory Algorithms on Massive Data Sets Data sets in large applications are often too massive to fit completely inside the computer's internal memory. The resulting input/output communication (or I/O) between fast internal memory and slower external memory (such as disks) can be a major performance bottleneck and raises fascinating questions about the efficiency of algorithms on massive data sets. Jeffrey S. Vitter (V) conducted research on ways to solve massive data problems efficiently in external memory. He worked in particular on ways to do fast range searches (given a query rectangle, report all the points it contains) and on ways to do approximate queries in On-Line Analytic Processing (OLAP) in databases. He worked on tree-based search methods that are efficient for various special cases of range queries, such as when two or three sides of the rectangle are specified and the other sides extend to infinity. The OLAP work deals with a new approach based on wavelet approximations of the data cube. The issue in practical use is to deal with the massive size of the data cube. Programming tools and environments are available for simplifying the programming task. Vitter has been doing experiments on some newly developed algorithms for spatial databases incorporating these paradigms, implemented using TPIE (Transparent Parallel I/O programming Environment), and the results are very promising.

Amortized Analysis for Pairing Heaps The pairing heap was introduced in the mid 1980's as a self-adjusting and computationally more efficient alternative to Fibonacci heaps. Highly efficient in practice and well regarded in this respect, it is included in the GNU C++ library. Strikingly simple in design, the pairing heap data structure nonetheless seems difficult to analyze. It has been previously conjectured that the pairing heap provides an implementation of decrease key operations with constant amortized running time, the raison d'etre of Fibonacci heaps, and several published experimental studies have supported this conjecture. Efficient implementation of decrease key operations plays an important role in the design of many highly efficient network optimization algorithms, such as shortest path and minimum spanning tree algorithms. Michael Fredman (PM, RU) obtained a significant new result which demonstrates that the pairing heap requires more than constant amortized time to perform decrease key operations, contrary to conjecture. Nevertheless, although Fibonacci heaps are asymptotically superior, pairing heaps continue to be of great practical importance. Aside from the practical usefulness of the pairing heap, the question of its efficiency has theoretical interest deriving from the connection between pairing heaps and splaying. The mechanism underlying the pairing heap can be viewed as an application of the self-adjusting splay heuristic of Sleator and Tarjan. The splay heuristic has been demonstrated to have surprising power in the context of search trees. A recent and notable example concerns the work of Cole on the dynamic finger conjecture for splay trees. As an issue, the competitiveness of the pairing heap thus constitutes an interesting and significant challenge for splaying, testing the limits of this remarkable and powerful heuristic. More precisely, splaying is a general purpose technique having the surprising feature that in all of the domains in which it can be appropriately applied, it seems to be competitive with the alternative specialized techniques previously developed for that particular domain. For the domain concerning priority queues which efficiently implement decrease key operations (that for which Fibonacci heaps were developed) Fredman's new result shows that splaying is asymptotically dominated by the specialized method, and this is the first time a result of this kind has been established. The techniques developed to analyze pairing heaps may find application in the analysis of splay trees and other problems involving amortized analysis of algorithms. Fredman's approach was based on the following concept. An attractive aspect of splaying is that it selects the operands for the atomic operations that get performed (in this instance the operations are comparisons between quantities) in a highly uniform manner. The intuition behind the new result is to consider the information gain that these comparisons can potentially generate and then show that as a consequence of the uniformity intrinsic to the splaying technique, the resulting loss of flexibility, in choosing which comparisons to perform as the computation proceeds, inevitably creates inefficiencies, at least within an appropriately constructed probabilistic setting. However, in order to make this framework tractible, it has been necessary to devise a combinatorial parody of efficiency, which highly distorts the usual notion, but nonetheless suffices for this purpose.

Compressing and Indexing Massive Text Archives Yossi Matias (PM, BL), S. Muthukrishnan (PM, BL), Suleyman Cenk Sahinalp (V), and J.Ziv (V, BL) worked on compressing and indexing massive text archives. This research resulted in a linear time algorithm for the HYZ compression method - which is in fact the first linear time algorithm for any context based compresion method. The algorithm was proven to outperform conventional compression schemes when the memory in use is limited, both in theory and practice. This work also provided insight into solving the document listing problem: given a set of documents to be preprocessed, index them to respond to list queries efficiently. The list query inquires the list of documents in which a given pattern occurs. The method is especially interesting for information retrieval applications in which many documents have to be preprocessed.

Bucket Sorting In the general area of distributions of sorting algorithms, one looks at the exact or limiting probability distributions of cost parameters of sorting algorithms (algorithms to order data either completely or partially in the case we are interested in finding order statistics). Exact distributions are sometimes too complex to deal with, in which case one seeks approximations by simpler functions that may be relevant only for very massive data sets. Knowing the distribution puts in good perspective the average-case analysis and tells us how well concentrated the algorithm's behavior is or how likely it is to deviate substantially from its expectation. Bucket sorting algorithms are sorting algorithms based on data classification according to "throwing data in buckets." Philippe Flajolet (V), Philippe Jacquet, Hosam Mahmoud (V), and Mireille Regnier used analytic probability as a general approach to bucket sorting. Functional equations for the cost parameters were set up and asymptotically solved by analytic techniques such as Poissonization and de-poissonization and saddle point methods. This work was definitive because it dealt with all the possible flavors of bucket sorting as well as their adaptations for finding order statistics. For each case Flajolet, et al. were successful in finding limit distributions of cost.

Stochastic Analysis of Shell Sort Hosam Mahmoud (V) and Robert Smythe conducted a stochastic analysis of shell sort. Shell sort is an algorithm that generalizes the method of sorting by insertion. It is essentially several stages of insertion sorting. The probabilistic analysis of shell sort stood as a formidable tantalizing challenge. Most research on shell sort had previously been in the direction of making good choices for the parameters of the stages to obtain good worst-case behavior. Mahmoud and Smythe took the research in a different direction. They analyzed the stochastic structure of the algorithm and identified connections of the algorithm to order statistics and ultimately to standard stochastic processes such as the Brownian bridge process.

Visualization of Large Spatial Data Sets Octavian Procopiuc (V) worked on visualization of large spatial data sets with telephone data in mind. Telephone data can be thought of as a directed graph, with telephone numbers being vertices and calls being edges. A method of visualizing this graph is by constructing a height field based on the weights of the edges, as follows: the vertices are arranged in a certain order on the x axis and on the y-axis of the 3-dimensional Euclidean plane, and each edge from a vertex on the x-axis to a vertex on the y-axis is represented by a spike of height given by the weight of the edge. This height field is obviously very large and sparse. The goal of the research was to design and implement an I/O efficient hierarchical data structure for such multidimensional data, which would increase the performance of a visualization program. The main features that can benefit immensely from such an indexing data structure are zooming in and out and panning. The main issues that Procopiuc had to deal with in designing such a data structure were: preserving the inherent hierarchy existent in the telephone numbers, preserving spatial locality, building the structure incrementally from the initial passes over the data in order to give the user immediate feedback. The data structure that resulted from extensive analysis of the problem and its peculiarities is based on the cross tree of Italiano and Grossi, modified to handle the hierarchy in vertices (telephone numbers) and simplified by observing that deletes don't need to be handled.

The Suffix Tree of a String The suffix tree of a string is a fundamental data structure for string processing. Recent focus on massive data sets has sparked interest in overcoming the memory bottleneck of known algorithms for building suffix trees. Paolo Ferragina (V), Martin Farach-Colton (PM, RU) and S. Muthukrishnan (PM, BL) showed how to choreograph almost all disk accesses to be via the sort-scan paradigm, thus achieving the first I/O-optimal algorithm for building a suffix tree on disk devices.

3-D Object Recognition One approach to 3-D object recognition is to represent an object by a set of 2-D views. To minimize the number of views required to encode an object, one seeks a 2-D representation that is highly viewpoint invariant. Sven Dickinson (PM, RU) recently proposed two new view-based object recognition approaches which offer high viewpoint invariance. The first approach was based on a hierarchical wavelet decomposition of an image into salient regions, giving rise to a directed acyclic graph, called the *saliency map graph*. Object recognition was formulated as matching an image graph to a database of model object graphs. He developed two coarse-to-fine matching algorithms based on a fast approximation to finding the maximum cardinality, minimum weight matching in a bipartite graph spanning an image graph and a model graph. One algorithm compares only the topological structure in a pair of graphs, while the other algorithm includes both topological and geometric structure. Initial results yielded a recognition accuracy of better than 95%. Dickinson's second approach to view-based object modeling was based on a skeleton-like decomposition of an object's silhouette according to a taxonomy of qualitative part classes. This description can also be represented as a graph which can, in turn, be mapped in linear time to a unique rooted tree. Object recognition, in this case, was formulated as matching an image shock tree to a database of model shock trees. In this work, he introduced a novel compact encoding of a tree's topology based on an eigenvalue characterization of the tree's structure. This topological signature not only provided a fast index into a database of model shock trees, but formed the basis of an efficient algorithm which computes, in the presence of occlusion and noise, the largest isomorphic subtree of two trees. Initial results showed that the matching and indexing algorithms support both generic recognition (focusing primarily on tree topology) and exemplar recognition (focusing on both topology as well as node geometry).

The List Update Problem Bela Csaba (GS, RU) and Sachin Lodha (GS, RU) investigated the list update problem. The problem can be briefly stated as follows: There are n different items in a doubly linked list. Every item has its own unique key. Access requests are coming for these items, one at a time, in the form of their key values. The request is satisfied by finding an item of the corresponding key value in the doubly linked list. If the position of the item requested is i in the list, then the cost of this operation is i units. While satisfying such a request, the requested item can be moved to any of the first i positions at no cost. The goal is to satisfy all requests at minimum possible cost. In the off line setting, the request sentence is known in advance. On the other hand, in the on line setting, a request must be satisfied without knowledge of any future request. Let ρ denote a finite length request coming from $\{1, 2, ..., n\}$. Let $C_{OPT}(\rho)$ denote the cost of the optimal satisfaction of ρ. Let A be a deterministic on line algorithm and let $C_A(\rho)$ denote the cost of A while satisfying ρ. If there exists a constant K such that $\forall\rho \;\; C_A(\rho) \leq cC_{OPT}(\rho) + K$ then A is said to be $c-$competitive. It is natural to ask whether knowing the structure of the request sequence could be of any help or not, i.e., beating the barrier of 2 on the competitive ratio of on line algorithms. This assumption may not be impractical, since in practice it may be reasonable to expect some kind of pattern in the incoming requests. Csaba and Lodha pursued this approach and considered two diffferent kinds of restrictions on the requests. For these they gave two different optimal on line strategies.