D I M A C S

Center for Discrete Mathematics and Theoretical Computer Science

DIMACS SPECIAL YEAR ON MASSIVE DATA SETS: 1999 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 continued through 1999. The workshop activity has wound down, but there is still a great deal of interest among DIMACS members, visitors, postdocs, and graduate students in the topics of the special year. Massive data sets remain a major area of interest at DiMACS and many massive data sets problems fall at the interface of this special year and the Special Year on Next Generation Networks Technologies and Applications that we are launching early in 2000. Interest in this area has also led us to begin serious planning for a follow-up Special Focus on Data Analysis and Data Mining starting in Fall 2001.

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
Institute for Defense Analyses/Center for Communications Research (IDA/CCR)
NEC Research Institute
Princeton University
Rutgers University
Telcordia Technologies

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, credit card transactions, telephone calling data, 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, February 1999 - January 2000

- **Distributed Data and Structures**
  - Dates: May 10 - 11, 1999
  - Location: Princeton University
  - Attendance: 23
  - Organizers: Yuri Breitbart, AT&T Labs, Sajal Das, University of North Texas, Nicola Santoro, Carleton University, Peter Widmayer, ETH Zurich

The *1999 Workshop on Distributed Data and Structures, (WDAS'99)* continued the series successfully started with WDAS'98, which was held in Orlando as a satellite workshop of the *12th International Parallel Processing Symposium (IPPS'98).* The aims and the scope of WDAS'99 were characterized as follows: Currently, the research on distributed data structures is carried out in many fields, from parallel systems to distributed computing, from AI to GIS. Their systematic design and analysis has just started: in the database literature, dynamic file structures for distributed object management have attracted some attention, and in the algorithms literature, data structures have been studied from a complexity oriented point of view. However, this research is mostly "hidden", relegated to the side of each field, obscured by the weight of the application domain, especially in databases but also in the algorithms community. As databases are growing steadily, applications become more and more demanding, and distributed computer systems are becoming rather easily available, the problem of how to efficiently maintain large datasets gains importance. An important aspect of this problem is the design, implementation, and operation of a data structure in a distributed system. At the same time, in the constantly expanding net-centric universe, an increasing amount of data is available, distributed among sites. The structuring of the data for accessing, manipulation and processing is a crucial task which can ultimately affect the performance, integrity and usefulness of the entire system. These two developments bring the research on distributed data and structures to the forefront. This workshop was intended to bring together application oriented developers and theoretical researchers concerned with the maintenance of distributed data and the organization of the interaction between the computing nodes. Topics of interest included, but were not limited to: design, implementation and operation of distributed data structures in general; measures of efficiency for distributed data structures; complexity analysis (lower and upper bounds) for distributed data structures; particular data maintenance problems (distributed token, mutual exclusion, counter); connection between distributed data maintenance and mobile data maintenance; connection between the interaction structure of the nodes and the data maintenance; experience with distributed data structures on current systems. The program of the workshop included 18 presentations, either invited or contributed in response to the call for papers. As expected, a large majority of the talks addressed issues related to the design, analysis, implementation, application and testing of distributed data structures. Other topics covered in WDAS'99 were algorithms and data allocation problems for parallel disk systems, novel interconnection structures, and basic algorithmic primitives of distributed computing. The first day of the workshop was opened by Maurice Herlihy, who illustrated the design of a *directory*, a distributed data structure used to keep track of a mobile object. Some problems arising in practical scenarios were also at the core of Witold Litwin's talk. In the first of his two talks, Philip Gibbons introduced the concept of a *synopsis* as a small, cacheable "summary" of a large, remote data set. The first group of talks concerning distributed data structures was completed by Adriano di Pasquale, who discussed issues arising in the design and maintenance of logarithmic-cost, balanced, distributed search trees. External searching, sorting and data allocation were the subjects of the next three talks. First, Ekow J. Otoo discussed issues related to the implementation of *B+Tree indices* for storage systems made of multiple disks. Next, Sanguthevar Rajasekaran described the *Parallel Disk System (PDS)* model and gave an overview of a number of deterministic and randomized PDS algorithms for fundamental problems, such as sorting and selection. Finally, Maria Cristina Pinotti discussed an optimal disk allocation strategy for partial match queries on nonuniform cartesian-product files. In his talk, Nicola Santoro amply reported on the large body of past research on distributed sorting and selection, also highlighting several remaining open problems. Next, Paola Flocchini envisaged a distributed scenario where objects are dynamically added or removed, and there is the need for a consistent and efficient naming scheme so that the objects can cross-reference each other. The first day of WDAS'99 was closed by Nir Shavitt, who described two efficient and scalable implementations of priority queues for bounded-range keys, based on funnel-like structures. The second day was opened by Phil Gibbons' second talk. He presented "black-box" testing strategies to check the correctness of distributed data structures such as (priority) queues, stacks of counters. Linda Pagli introduced the *dense tree*, a recursive structure obtained as a natural extension of a tree. Data structures and their use in the realm of parallel computation was the focus of the next two talks. First, Geppino Pucci presented near optimal deterministic algorithms for the *branch-and-bound* problem on a distributed-memory, p-processor machine. Next, Andrea Pietracaprina explored the deterministic complexity of parallel *multisearch* on a distributed-memory, coarse-grained parallel machine. Three talks in WDAS'99 concentrated on the concept of *self-stabilization*, a feature that makes a distributed system tolerate transient faults by converging to the intended behaviour in finite time. In particular, Ajoy K. Datta concentrated on tree networks and derived optimal, self-stabilizing algorithms for the *Propagation of Information with Feedback (PIF)* problem. Next, Vincent Villain described a tool to achieve optimal, self-stabilizing algorithms in Dijkstra's *state model*. Finally, Franck Petit discussed the time and space optimality of self-stabilizing algorithms for distributed, depth-first token circulation. The proceedings of WDAS'99 will appear in the series "Proceedings in Informatics" of Carleton Scientific Press.

Selected Research Results

Engineering the Compression of Massive Tables: An Experimental Approach In a variety of problems, one has massive tables about which one can learn something by examining a small, fixed size of data. Adam Buchsbaum (PM, ATTL), Don Caldwell (ATTL), Ken Church (ATTL), Glenn Fowler (ATTL), and S. Muthukrishnan (PM, ATTL) showed how one can use such a training set to come up with a better way to organize the table, so that applying gzip (for example) to the newly reorganized table significantly outperforms just applying gzip to the table in the original order (by a factor of 2, in fact).

External Memory Graph Transversal Adam Buchsbaum (PM, ATTL), Michael Goldwasser (PD), Suresh Venkatasubramanian (PM, ATTL), and Jeff Westbrook (PM, ATTL) gave the first non-trivial external memory algorithm for *directed* breadth-first search and an improved external memory algorithm for directed depth-first search. They provided reductions between several common formulations of undirected bfs, showing them equivalent up to an external sort. Using those reductions, they gave the first optimal (constant sort) external memory algorithm for undirected bfs on trees. (This is trivial in core since a tree is a bfs tree, but externally, it's nontrivial to decide the bfs numbering if all one is given is a list of edges, even when told they form a tree.)

There was considerable cross-fertilization between our Massive Data Sets Special Year and our

Large Scale Discrete Optimization Special Year. The following two results are examples of such cross-fertilization.

Generating Transversals Generating transversals to a hypergraph (or the dual of a Boolean function) is a well-known problem, which became very important recently, due to its data-mining applications ("association rules" or more specifically "frequent sets" are transversals to a certain hypergraph defined by the input data set.) This problem is also interesting, since its true complexity (in the incremental sense) is not known. Due to an incrementally quasi-polynomial algorithm (Fredman and Khachiyan, 1996) it is very unlikely to be an NP-hard problem, but as of today we do not know any polynomial algorithm, either – it could even be the core to a separate new class, laying somewhere between P and NP-hard (unless of course these two coincide.) Endre Boros (PM, RU), Vladimir Gurvich (V), Leo Khachiyan (PM, RU), and Kazuhisa Makino (V) showed that several open versions of transversal generation are in fact equivalent with Boolean dualization, and hence treatable with an incrementally quasi-polynomial algorithm. Their results are based on several new inequalities, relating the number of maximal false points to the number of minimal true points of a Boolean function. For instance, we can prove that the number of maximal infeasible binary vectors to a monotone system of linear inequalities is polynomially (almost linearly) bounded in the number of minimal feasible vectors. Based on this result, they can generate all minimal feasible solutions to a monotone system of inequalities in incremental quasi-polynomial time.

Attribute Selection in Learning and Data-Mining Problems In the logical analysis of data, one of the first and most important steps is the selection of a small subset of binary attributes, to a given input data set, which will be used in later steps of the learning process. Such a reduction step, if done carefully, is useful to eliminate noise, and to reduce dimensions, making possible the application of more demanding computational methods. Recently similar considerations became popular in the learning community – so called "attribute efficient learning." Endre Boros (PM, RU) and Toshide Ibaraki (V) considered such an attribute selection problem for the case of binary data, and based on various measures of the "separating power" of the individual attributes and their small subsets. They showed that a wide variety of such models are in fact linearly constrained maximization problems with submodular objective functions, and hence efficient approximation algorithms with a guaranteed performance rate can be applied. They considered continuous relaxations of these problems, which turn out to be smooth convex minimizations over a convex set, and hence can also be solved efficiently.

The following two results come from cross-fertilization between our Special Years on Networks and Massive Data Sets.

Maximum Clique Problems in Very Large Graphs The proliferation of massive data sets brings with it a series of special computational challenges. Many of these data sets can be modeled as very large multidigraphs with a special set of edge attributes that represent special characteristics of the application at hand. Understanding the structure of the underlying digraph is essential for storage organization and information retrieval. Such problems arise for example from telephone billing records. One builds a multidigraph with vertices being telephone numbers and edges calls made from one number to another. James Abello (PM, ATTL) sudied the evolution of such a graph over one 20-day period and found that it grew to have 290 million vertices and 4 billion edges. In practice, this multidigraph is treated as if it were undirected and multiple edges are collapsed into a single edge. Even a computer with six gigabytes of main memory cannot hold the full call graphs arising in practice. Cliques in the call graph represent important structural information. Maximum cliques are important in such practical applications as project selection, classification theory, fault tolerance, coding theory, computer vision, economics, information retrieval, signal transmission theory, and molecular biology. Yet, identifying a maximum clique in a graph is computationally difficult even in a graph of modest size. Abello and collaborators Panos Pardalos (V), Mauricio Resende (PM, ATTL), and Sandra Sudarsky (GS, V) developed an approach to finding cliques in very large multidigraphs. They developed several graph decomposition schemes for breaking up the problem into manageable pieces and a semi-external greedy randomized adaptive search procedure for finding approximate solutions to the maximum clique problem and the maximum quasi-clique problem in very large sparse graphs. They experimented with their heuristic on real data, for instance a one-day call graph with 53,767,087 vertices and 170 million edges, finding large cliques of as many as 30 vertices.

Maintaining Hierarchical Graph Views In connection with a problem in visualizing large graphs, Adam Buchsbaum (PM, ATTL) and Jeff Westbrook (PM, ATTL) considered a hierarchy defined on the vertices, and how one would visualize aggregations of the graph by contracting vertices with respect to the hierarchy. For example, in the graph induced by telephone customers (vertices) and calls (edges), one can contract all the vertices in a single area code together. The problem then becomes one of maintaining this induced, contracted graph under expansion (area code into exchanges, for example) and contraction (the inverse) operations. They formalized the problem and provided data structures that allow navigation in optimal time (linear in the size of the change to the induced graph), although they use superlinear space.

A Multiple Prototype Model for Fuzzy Clustering In partitional fuzzy clustering, each cluster is characterized by two items, its centroid and its membership function, that are usually interconnected through distances between centroids and entities (as in fuzzy c-means). Susana Nascimento (GS, V), Boris Mirkin (V) and F. Moura-Pires proposed a different framework for partitional fuzzy clustering which suggests a model of how the data are generated from a cluster structure to be identified. In the model, they assume that the membership of each entity to a cluster expresses a part of the cluster prototype reflected in the entity. Due to many restrictions imposed, the model as is leads to removing of unneeded cluster prototypes and, thus, can serve as an index of the number of clusters present in data. A comparative experimental study of the method fitting the model, its relaxed version and the fuzzy c-means algorithm, suggested that their methods can be considered a model-based parallel to the fuzzy c-means approach. Moreover, their generic version can be viewed as a device for revealing "the natural cluster structure" hidden in data.

Massive Data Set Analysis in Seismic Explorations for Oil and Gas in the Crystalline Basement Interval Ilya Muchnik (V), Vadim Mottl (V) and Vladimir Levyant worked on an optimization-based approach to the analysis of massive ordered data sets. They introduced a new optimization method for computer-aided interpretation of seismic exploratory data from the so-called crystalline basement of the earth's mantle which underlies the relatively thin sedimentary cover which has been, up to now, essentially the only object of seismic explorations. The seismic exploratory data sets, seismic sections and cubes are a class of, respectively, two- and three-dimensional data arrays which are analyzed in the course of gas and oil reserves prospecting with the purpose of studying the structure of the underground rock mass. The seismic data sets consist of synchronous records of reflected seismic signals registered by a large number of geophones (seismic sensors) placed along a straight line or in the nodes of a rectangular lattice on the earth's surface. The vertical time axis forming the resulting two- or three-dimensional picture is identified with depth, so that the peculiarities of the reflected signal under the respective sensors carry information on the local properties of the rock mass at the respective points of the underground medium. The new method of seismic data analysis developed by Muchnik, Mottl, and Levyant is aimed at finding fractured zones of the basement rock mass capable of accumulating oil or gas. The essence of the method consists of numerically evaluating distinctions in the local spatial texture of the seismic picture that are caused by differences in physical properties of fractured and monolith rock.

Fully Consistent Extensions of Partially Defined Boolean Functions with Missing Bits The problem of finding patterns in data mining can be formalized as the problem of finding an extension of a partially defined Boolean function in which the input contains some missing bits. Endre Boros (PM, RU), Toshihide Ibaraki (V) and Kazuhisa Makino (V) considered four different definitions for an extension and showed that, for many general and reasonable families of function classes, three of these extensions are mathematically equivalent. They also demonstrated that such an equivalence does not hold for all classes.

Apparent Motion in Spatial Time-varying Data: A Variational Approach to Pointwise Tracking of Coherent Structures in Computational Fluid Dynamics A massive time-varying data set is an experimentally acquired or mathematically simulated function of spatial coordinates and time which is to be analyzed with the purpose of studying behavior of related distributed dynamical systems. Vadim Mottl (V), Alexander Blinov, Norman Zabusky and Ilya Muchnik (V) addressed problems of computational fluid dynamics that are concerned with the necessity to visualize and study overwhelmingly large time-varying data sets generated by numerical unsteady simulations. Of major interest is time tracking of persistent coherent structures of the fluid flow that arise, move in space for a time, fade, and are replaced by a new pattern building up from their remains. The spatial motion of persistent patterns in a time-varying data set is called apparent motion because it does not coincide, in the general case, with mechanical motion of any physical particles. With the purpose of tracking coherent structures, Mottl, et al. proposed to evaluate the apparent motion in terms of apparent velocity of space points. The apparent velocity is to be estimated in the form of a time-varying vector field as the result of processing the original time-varying data set or immediately in the course of numerically simulating the respective physical reality as the solution of an additional system of partial differential equations. In both versions, Mottl, et al. leaned upon the variational approach to the analysis of massive data sets and estimated the vector field of apparent velocity as a time-varying function that delivers the minimum value to a quadratic functional. In experiments with simulated two-dimensional time-varying spatial dynamics of the fluid's density and vorticity in shock wave interaction with an elliptical bubble, the estimated apparent velocity provided long-term pointwise tracking of persistent coherent structures.

A Geometric Model of Information Retrieval Systems This decade has seen a great deal of progress in the development of information retrieval systems such as Yahoo and Alta Vista. Unfortunately, we still lack a systematic understanding of the behavior of the systems and their relationship with documents. Myuung Ho Kim (V) developed a completely new approach towards the understanding of these information retrieval systems. Recently, it has been observed that retrieval systems in TREC 6 (TREC = Text Retrieval Conference) show some remarkable patterns in retrieving relevant documents. Based on the TREC 6 observations, Kim introduced a geometric linear model of information retrieval systems. He then applied the model to predict the number of relevant documents by the retrieval systems. The model is also scalable to a much larger data set. Although the model was developed based on the TREC 6 routing test data, Kim believes it can be readily applicable to other information retrieval systems.

Structural Properties and a Recognition Algorithm for Boolean Geodetic Functions A Boolean function is called *geodetic* if any pair of true points is connected by a shortest path consisting of true points (a path is a sequence of points of the Boolean hypercube B^n such that the Hamming distance between any two consecutive points of this sequence is 1). One of the main applications of the geodetic functions is in Logical Analysis of Data, since they provide models for the extensions of some data sets. Gabriela Alexe (GS, RU), Sorin Alexe (GS, RU), Peter Hammer (PM, RU), and Alexander Kogan (PM, RU) obtained structural characterizations of geodetic functions in terms of their set of true points and also in terms of their complete DNF. Based on these results, some recognition algorithms for the geodetic, co-geodetic and strongly geodetic functions were designed.

Data Access Hypergraphs and Applications in Database Modeling Gabriela Alexe (GS, RU) and Sorin Alexe (GS, RU) provided a hypergraph-based approach for the modeling of complex data structures and investigated some of its properties and operations. They defined the Data Access Hypergraphs, which are connected simple hypergraphs having the property that any two intersecting edges are comparable relative to set inclusion, and found several structural characterizations. They defined an abstract database model together with structural, transactional and mixed operations. By focusing on an object-oriented implementation model, and by defining data definition, data manipulation, and data query techniques, they found methods that can be efficiently used for the development of software applications.

Admission Control with Instant Notification When admission control is used, an online scheduler chooses whether or not to complete each individual job successfully by its deadline. An important consideration is at what point in time the scheduler determines if a job request will be satisfied, and thus at what point the scheduler is able to provide notification to the job owner as to the fate of the request. In the loosest model, often seen in real-time systems, such a decision can be deferred up until the job's deadline passes. In the strictest model, more suitable for customer-based applications, a scheduler would be required to give notification at the instant that a job request arrives. Michael Goldwasser (PD) and Boris Kerbikov (undergraduate, PU) studied the effect of the notification model on a specific problem, namely the non-preemptive scheduling of a single resource in order to maximize resource utilization. Although the immediate notification model is severely more restrictive, they were able to give alternate algorithms which provide such notification while matching most of the performance guarantees which are possible in the loosest model with no advance notification. In one case, they gave evidence that providing immediate notification may be more difficult, as they improved upon both the lower bound for the immediate notification model and the upper bound for the loosest notification model.

Concept Learning and Feature Selection Based on Square-error Clustering Based on a reinterpretation of the square-error criterion for classical clustering, Boris Mirkin (V) developed a "separate-and-conquer" version of K-Means clustering. The method determines a contribution weight for each variable of every cluster. The weight is used to produce conjunctive concepts that describe clusters and to reduce or transform the variable (feature) space.

Counting Special 0-1 Matrices, with Applications to Information Theory Mordecai Golin (V), Yuanping Zhang (GS, V) and Xuerong Yong (GS, V) examined 0-1 matrices which do not contain certain sub-matrices, for example do not contain any two consecutive horizontal or vertical ones. The problem they examined was how the number of such matrices grows with the size of the matrices. This problem has become of interest to the information theory community during the past few years because the matrices can model various physical constraints in 2-dimensional media and the growth of the number of such matrices can then describe the information carrying capacity of the media. Golin, et al. examined this quantity for various problems; in some they obtained better bounds than previously existed; in others, they derived new structural properties that will hopefully permit the derivation of better bounds in the future.