DIMACS Center
Rutgers University

Working Groups in Data Analysis and Mining

Final Report

November 2004

Ia. Participants in the program

Organizer:
Fred Roberts, DIMACS

Web masters for the Working Group web sites:
Gabriela Alexe, Rutgers University
Sorin Alexe, Rutgers University
Vashist Akshay, Rutgers University

Graduate Students doing research through the grant:
Khaled Elbassioni, Rutgers University
Ulas Akkucuk, Rutgers University
Igor Zverovich, Rutgers University

Working Group on Algorithms for Multidimensional Scaling, Meeting I
Dates: August 6 - 10, 2001

Organizers:
J. Douglas Carroll (chair). Rutgers University
Phipps Arabie, Rutgers University

Working Group on Algorithms for Multidimensional Scaling, Meeting II
Dates: June 11-12, 2003

Organizers:
J. Douglas Carroll (chairman), Rutgers University
Phipps Arabie, Rutgers University
Larry Hubert, University of Illinois
Michael Trosset, The College of William & Mary
Mike Brusco, Florida State University
Mel Janowitz, DIMACS

Working Group on Streaming Data Analysis and Mining, Meeting I
Dates: November 5 - 9, 2001

Organizer:
Adam Buchsbaum, AT&T Labs

Working Group on Streaming Data Analysis and Mining, Meeting II
Dates: March 24-26, 2003

Organizers:
Adam Buchsbaum, AT&T Labs
Rajeev Motwani, Stanford University

Working Group on Computer-Generated Conjectures from Graph Theoretic and Chemical Databases, Meeting I
Dates: November 12 -16, 2001

Organizers:

Patrick Fowler, University of Exeter
Pierre Hansen, GERAD - University of Montreal

Working Group on Computer-Generated Conjectures from Graph Theoretic and Chemical Databases, Meeting II
Dates: June 2 - 5, 2004

Organizers:

Patrick Fowler, University of Exeter
Pierre Hansen, GERAD - University of Montreal

Ib. Participating Organizations

Telcordia Technologies: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

AT&T Labs - Research: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning, research and Working Group leadership.

NEC Laboratories America: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Lucent Technologies, Bell Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Princeton University: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Avaya Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

HP Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

IBM Research: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Microsoft Research: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Centre de Recherches Mathematiques: Collaborative Research and Financial Support
Co-funded the second meeting of the Working Group on Computer-Generated Conjectures from Graph Theoretic and Chemical Databases.

GERAD: Group for Research in Decision Analysis, University of Montreal: Collaborative Research and Financial Support
Co-funded the second meeting of the Working Group on Computer-Generated Conjectures from Graph Theoretic and Chemical Databases.

1c. Other Collaborators

The project involved scientists from numerous institutions in numerous countries. The resulting collaborations also involved individuals from many institutions in many countries.

II. Project Activities

The project supported three interdisciplinary working groups of researchers addressing different aspects of problems involving the analysis of massive data sets. The groups were concerned with "streaming" data analysis and mining, multidimensional scaling and the generation of scientific conjectures by computers. Each working group met twice, and generated conjectures and results in its area of concern. These working group meetings were informal with plenty of time set aside for interactions among the participants.

Each working group consisted of researchers with expertise in the field and/or in one of the applications areas. We brought these working groups together for several days to a week for a first meeting. There were informal presentations and lots of time for discussion and interaction. At the end or beginning of this first get-together, we held a public workshop with more formal presentations. During the public workshop, others in the community interested in the field were given an opportunity to learn about the working group's efforts and the working group had the opportunity to learn about the efforts of others that they might not have known about. Because we wanted to be inclusive, we used this public meeting to identify individuals from the community who might wish to join the working group or have their students do so. The goals of this first get-together at DIMACS were to formulate problems, share ideas and approaches, and set an agenda for future interactions. These interactions took place via email and as such interactions take place normally in scientific collaborations. At the second meetings of each working group, the participants had an opportunity for intensive collaborations and also to make informal presentations to each other about their progress. The working groups have obtained exciting results, at least of a preliminary nature, and we have laid the groundwork for extensive future collaborations.

The working groups were interdisciplinary. Researchers who have addressed the problems we investigated include theoretical and applied computer scientists, statisticians, discrete and non-discrete mathematicians, chemists, astronomers, economists, psychologists, information theorists, management scientists, ecologists, molecular biologists, and others. Most of these fields were represented in our working groups. DIMACS has a long and successful history of getting researchers with different backgrounds and approaches together, stimulating new collaborations, helping to set the agenda for future research, and acting as a catalyst for major new developments at the interface among disciplines and we built on this tradition with these working groups.

Working Group on Algorithms for Multidimensional Scaling

The term “data mining” has come to describe any type of data analysis, with an emphasis on the need to discern new, useful information from the large quantities of data stored in today's databases. Though the term “data mining” is relatively new, the fundamental ideas behind it are not, and the potential approaches to it are grounded in the basic theoretical and algorithmic approaches to data analysis developed over the last several decades. What is new is the challenge of modifying these approaches and developing new ones that can scale to the very large size of the problems that are faced or that apply to new types of data and in new types of applications. Most algorithms currently used in data mining do not scale well when applied to very large data sets, often because they rely on random access to the data sets, which scales only while the data sets fit entirely in relatively small main memories. Multidimensional scaling (MDS) is a case in point. The methodology has its roots in work going back to the first half of the 20th century and its modern roots in work of Shepard, Kruskal, and others. Yet, the traditional algorithms of MDS do not always work efficiently with today's large data sets and the traditional models of MDS are not always the most appropriate for the complex data types that we are seeing today. There has been in recent years a major trend toward the development of totally new and different algorithmic procedures for MDS as well as the modification of traditional procedures; an emphasis on the development of new models of MDS and on the development of new algorithmic procedures for fitting these new models; and a stress on the application of MDS to new types of data and new fields. These new algorithms, new models and new applications are being developed from different points of view and by researchers with different backgrounds and interests. It was our plan to get the developers of new models of, new algorithmic approaches to, and new applications of MDS together to share ideas and explore new approaches. We also included experts in the applications areas and individuals who have not necessarily worked on MDS before but are expert in methodologies such as combinatorial optimization that are increasingly relevant to MDS.

Among the specific areas of emphasis for this working group were challenges to MDS research from:

- the field of graph layout/graph drawing;
- chemometrics;
- biomolecular conformation;
- telephone call graphs, scientific collaboration graphs and other large social networks;
- the application of methods of social networks to ecological data;
- the increasing stress on data visualization arising from such areas as graph layout, fraud and intrusion detection, and amino acid databases;
- the increasing importance for MDS of discrete models and hybrid models mixing spatial and discrete components;
- the application of methods from combinatorial optimization to fit MDS and MDS-related models to proximity data;
- the increasing emphasis on approximations and heuristic techniques that results from increasingly large data sets.
- the need to find a global optimum in “nonmetric MDS” when data sets are large and finding the best local optimum isn't feasible.

Sources of Data

MDS is widely used in psychology, the field in which it was originally developed, to determine the underlying perceptual structure of important classes of psychological stimuli in terms of important perceptual dimensions underlying these stimuli. For example, in the case of color vision, numerous MDS studies have confirmed that the color space is 3-dimensional, with one dimension corresponding to a red-green dimension, a second corresponding to a yellow-blue dimension, and a third corresponding to

lightness (or brightness) -- essentially the intensity, ranging from black to a very intense white. Similar studies have been done of perception of nations, of words describing personality traits, of kinship terms, of speech and acoustical stimuli, perception of human faces, and many other areas of psychology, often unearthing psychological dimensions underlying perception and other behavior that had not been anticipated in advance. We involved in our working group an expert on the application of MDS to new problems in social and clinical psychology.

In marketing, where MDS has also been widely applied, MDS of proximities is used to develop perceptual maps of products, which can be used to provide greater insight into the nature of that product class, and to devise predictive models of consumer choice. For example, an MDS map of cars will generally reveal, among others, a “sportiness” and a “luxuriousness” dimension, as well as more specific dimensions having to do with styling, fuel economy, overall size, etc. Data on preferences for cars by consumers can be used to fit either a vector or ideal point model or one can map either subject vectors or ideal points into the multidimensional perceptual space, in order to account for each individual consumer's preferences in terms of these perceptual dimensions. One application to marketing involves what is sometimes called “gap analysis”; i.e., looking for gaps or empty spaces in the product map which, if filled, would correspond to the most preferred product for a substantial subgroup of consumers. Such gaps might represent opportunities for a product manager to develop a new product or products that would command a significant market share, and thus result in significant profits to his or her company. As noted in the discussion of our streaming data analysis working group, massive data sets from supermarkets provide important challenges for data mining. We involved in our working group an expert on the use of MDS in marketing in the food industry.

MDS has been used for a long time in the study of social networks. Recently, there has been much interest in social networks called telephone call graphs, where the vertices are telephone numbers and the edges correspond to calls between them. Such call graphs also arise in the fraud and intrusion detection applications to be studied by the working group in streaming data analysis and mining. These graphs are extremely large. Buja, et al. use MDS to visualize a part of a call graph consisting of millions of vertices. Social networks also arise from scientific collaborations. Large scientific collaboration networks are visualized using MDS methodology, clearly identifying clusters. We included experts on these new, large social networks in our working group.

MDS is used regularly in biology, for example in ecology to compare communities or other biological assemblages, in biological taxonomy to discriminate between populations and/or species on the basis of morphometric or genetic data, and in evolutionary theory. J.D. Carroll made an early attempt to apply MDS methods to ecology. Ecological data of a multidimensional quantitative, semiquantitative, and qualitative nature of all kinds are described in the 1998 book Numerical Ecology by P. Legende and L. Legende. Recently, MDS methods have been used in conjunction with methods developed to study role assignments in social networks. This is a fascinating new idea and challenge for MDS and an exciting application of social-scientific methods in the biological sciences. We brought together in our working group those who have worked in the theory of social roles and those who are working on applying this theory to ecology, as well as experts in MDS and other classification methodologies useful in ecology.

MDS-like models arise in the field of biomolecular conformation in biochemistry. This work, concerned with reconstructing the geometry of molecules from neighborhood/distance information, goes back to the 1970's, but has found modern stimulus through the growing field of computational chemistry. While there are only 20 amino acids, the understanding of their interrelations, similarities, and differences is very complex and amino acid databases include hundreds of quantitative properties of such amino acids. To summarize properties of amino acids compactly, a readily visualizable mapping of these properties would be very helpful and could benefit from MDS-like methods. We brought together biologists, experts in classification/clustering working on biomolecular and protein databases, and MDS experts interested in

biochemistry in our working group. DIMACS is a natural place for such interdisciplinary work; our special foci on Mathematical Support for Molecular Biology and Computational Molecular Biology have brought together mathematicians, statisticians, computer scientists, molecular biologists, and computational chemists with great success since 1994.

Broader models of MDS developed by psychometricians, mathematicians, and statisticians have been applied in recent years to chemical data by a group of researchers called TRICAP (TRIlinear models in Chemistry and Psychology). A recent special issue of the journal *Chemometrics* (Vol. 14, #3, 2000) dealt with the methodology for and applications of multilinear models and the challenges for development of new, more sophisticated methods for handling 3-way and multi-way generalizations of factor and/or components analysis. Applications included data about polynuclear aromatic hydrocarbons, pesticides, and fluorescence. Our working group included individuals who are interested in the relation between MDS and chemometrics.

The field of graph layout/graph drawing has become a very important field from the point of view of information visualization. There is now an annual graph drawing conference and the drawing of large graphs arises in many contexts, including communication networks, electrical networks, wiring diagrams, etc. Graphs with thousands of vertices arise frequently in such applications. DIMACS has sponsored a variety of programs on graph layout, including summer tutorial programs on network visualization, and is well connected to the research community in this field. We involved people with interests in graph layout and MDS in the working group.

Among the specific areas of emphasis for this working group were challenges to MDS research from the field of graph layout/graph drawing; chemometrics; biomolecular conformation; telephone call graphs, scientific collaboration graphs and other large social networks; the application of methods of social networks to ecological data; the increasing stress on data visualization arising from such areas as graph layout, fraud and intrusion detection, and amino acid databases; the increasing importance for MDS of discrete models and hybrid models mixing spatial and discrete components; the application of methods from combinatorial optimization to fit MDS and MDS-related models to proximity data; the increasing emphasis on approximations and heuristic techniques that results from increasingly large data sets. There is a need to find a global optimum in “nonmetric MDS” when data sets are large and finding the best local optimum isn't feasible.

The working group web site contains a summary of the meetings, open questions, links to various types of software, and descriptions of the meetings.

This working group met as follows:

Algorithms for Multidimensional Scaling Meeting I
- Dates: August 6 - 10, 2001
- Location: DIMACS Center, CoRE Building, Rutgers University
- Organizers: J. Douglas Carroll and Phipps Arabie, Rutgers University
- Attendance: 30

MDS is widely used in the social and behavioral sciences. Its goal roughly is to take a multivariate data set and represent it in a low dimensional Euclidean space so as to minimize any distortion of the data. Often this is a representation in 2 dimensions. At its first meeting, the working group explored nonlinear and nonmetric versions of MDS, fitting of various non-Euclidean representations in both the two- and three- way cases, and the need to develop techniques that can be applied to massive data sets. This last problem, of dealing with massive data sets, is difficult because it will require the development of entirely new techniques, since most of the existing ones are extremely computationally intense and so tend to limit

the size of data arrays quite severely. One promising approach involves the random deletion of a substantial portion of the data. Preliminary results indicated that as much as 60% could be deleted without a serious effect on the output. Other approaches involve using heuristic approaches to get close to the solution and then trying to refine the output of the heuristic. This is work done by Willem Heiser and his colleagues from Leiden University. Since one well-known approach to fitting two-way Euclidean MDS models involves a singular value decomposition (SVD) of a derived matrix of scalar products, and since methods already exist for implementing the SVD on very large matrices, one approach, taken by the (unfortunately recently deceased) Mark Rorvig and David Dubin in some collaborative work with Douglas Carroll involved applying methods for SVD of massive data sets to solving this particular version of MDS for the case of extremely large matrices of proximities. This would involve proximity data on a very large number of stimuli or other objects. Various approaches were explored for extending such approaches to other, more complex MDS models and methods.

Algorithms for Multidimensional Scaling Meeting II

Dates: June 11-12, 2003
Location: Doubletree Hotel, Tallahassee, FL
Organizers: J. Douglas Carroll and Phipps Arabie, Rutgers University
Attendance: 21

The working group meeting had two themes. One was the development of computer efficient algorithms for multidimensional scaling so as to enable their use for massive data sets. The second was to investigate new and novel applications.

Two talks were given by Larry Hubert in which he illustrated the capabilities and background for a MATLAB toolbox he is developing. The purpose of the toolbox is to fit trees to proximity data by means of an efficient iterative projection strategy using the L_2 norm. The combination of being computationally efficient and user friendly became a topic of lively discussion among the participants. The talks by Hubert represented collaborative research with Phipps Arabie and Jacqueline Meulman.

Willem Heiser gave a talk on a formulation of correspondence analysis (CA) involving representing each row and column profile of the two-way MxN contingency table defining the basic data in barycentric coordinates, with each of the MN cells of the table providing a reference point for this barycentric system. He argued that this allows simultaneous definition of Euclidean distances among row points, among column points, and, most importantly, between row and column points, in contrast to the more conventional formulations of CA now in the literature.

Frank Busing, a graduate student working with Heiser, discussed a program called PREFSCAL, a specialization of the Heiser and Busing PROXSCAL program for both metric and nonmetric two- and three-way MDS via optimizing a natural generalization of the two measures of STRESS possible in Kruskal's KYST for two-way MDS. PREFSCAL allows analysis of an IxJxK array of preference (and other) data via the unfolding or "ideal point" model, in which each of the K IxJ matrices (for K "occasions" or other data sources, wherein each of the I subjects makes preference judgements for each of J stimuli, with preferences interpreted, as proposed by Coombs, as monotonically related to distances between each stimulus and each subject's "ideal stimulus point". These assumptions lead to each of these K matrices containing what Coombs called "off-diagonal conditional proximity data", which require very special variants of three-way metric or nonmetric MDS for analysis. Busing demonstrated that PREFSCAL accomplishes this special variety of MDS, leading to a procedure called "three-way unfolding".

Ulas Akkucuk, a graduate student working with Doug Carroll, in a paper coauthored with Carroll, discussed a method for nonlinear mapping of data whose topological dimensionality is lower than that of the linear vector space in which the data points are embedded; e.g., points on the surface of a sphere, which is a locally two-dimensional surface embedded in a three dimensional linear space, leading to a problem closely analogous to the cartographer's problem of producing flat two-dimensional maps of the earth's surface. Another highly nonlinear structure they discussed was a highly symmetric torus (or donut) embedded in four dimensional space, whose local topological dimensionality is also two.

Michael Trosett and Robert Lewis presented talks related to molecular conformation. The idea is to take actual inner-atomic measurements of molecules, and to use MDS to construct a feasible model of the molecule.

Suzanne Winsberg presented work that involves data mining applications of MDS. With massive data sets one may wish to first aggregate the data to make it a more manageable size. But then distances between objects may more naturally be viewed as being interval valued or even a distribution function.

By using techniques from distance geometry, Zhijun Wu presented a linear time algorithm for certain MDS problems, and discussed techniques for handling build-up errors, inconsistent distances, and distance bounds.

David Dubin reported on uses of MDS in the analysis of bug reports for software development.

Allistair Morrison is a graduate student at the University of Dublin. He presented some of his graduate work on fast MDS methods using hybrid non-linear algorithms. The work he presented was based on a paper that has appeared in the Information Visualisation Journal. The paper has been accepted for IEEE InfoVis 2003, and has recently been selected as one of six papers from that conference to be extended for publication in Information Visualization journal. Links for these are http://infovis.org/infovis2003/ and http://www.palgrave-journals.com/ivs/

Michael Brusco described his research on the application of MDS to a collection of proximity relations. The issues come up in many different areas. One that comes to mind is evolutionary biology, and another is cluster analysis. Does one cluster first and then summarize, or does one first summarize and then cluster? Just replace the word "cluster" with the phrase "apply MDS" and there is the research problem. Brusco investigates a third approach related to the
Tucker-Messick "Points of View" approach -- the first method, proposed in the 1960's for what is called "three-way" or individual-difference MDS, clustering the proximity matrices first and then applying various methods of MDS to the resulting clusters of subjects.

James Corter talked about fitting a mixture of directed and undirected tree structures to asymmetric proximities on the premise that one often loses important structure if one first converts to a symmetric proximity matrix before analyzing the data.

Willem Heiser, the editor of the *Journal of Classification* (JoC) was present, and invited Working Group members to submit papers presented at the meeting, or related work, to JoC, promising a rapid and "friendly" review process for such papers. A special issue of the *Journal of Classification* will be dedicated to the work of this working group and include the talks presented in the meetings.

Working Group on Streaming Data Analysis

Many critical applications require immediate (within seconds) decision making based on current information from a stream of data: e.g. intrusion detection and fault monitoring. Data must be analyzed as

it arrives, not off line after being stored in a central database. Processing and integrating the massive amounts of data generated by a number of continuously operating, heterogeneous sources poses many practical and theoretical challenges. At some point data sets become so large as to preclude most computations that require more than one scan of the data as they stream by. Analysis of data streams also engenders new problems in data visualization, and in the design of automatic response systems. The researchers in this working group addressed these issues.

The working group web site contains a summary of the meetings, open questions, links to various types of software, and descriptions of the meetings.

This working group met as follows:

Streaming Data Analysis and Mining Meeting I
Dates: November 5 - 9, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Adam Buchsbaum, AT&T Labs
Attendance: 43

The Problems of Streaming Data Analysis and Mining
The use of the computer in scientific research and as an essential ingredient in commercial systems has led to the proliferation of massive amounts of data. Researchers in a myriad of fields face daunting computational problems in organizing and extracting useful information from these massive data sets. Because of the sheer quantity of the data arising in various applications or because of their urgency, it becomes infeasible to store the data in a central database for future access and, therefore, it becomes necessary to make computations involving the data, and decisions about the data (like what to keep), during an initial scan as the data "stream" by. This working group was concerned with data mining in a "streaming" environment. It led to ideas that were later developed in much more detail in a DIMACS project on "Monitoring Message Streams" funded by the intelligence community.

Examples of applications requiring immediate decision making based on current information are intrusion detection and fault monitoring. Data must be analyzed as it arrives, not off-line after being stored in a central database, because the problems involved are so urgent from a time-to-react point of view. Some other applications require such quick reactions for theoretical (as well as practical) reasons because of the issues involved in processing and integrating the massive amounts of data generated by a myriad of continuously operating sources. For example, external memory algorithms are motivated by the fact that classical algorithms do not scale when data sets do not fit in main memory. At some point, data sets become so large as to preclude most computations that require more than one scan of the data, as they stream by.

Transactional and time-series applications exemplify current streaming data analysis systems.

Transactional applications exploit data recording individual events correlating two or more discrete entities. Examples are phone calls between people (data also studied by the multidimensional scaling working group) and purchases over a credit-card network. One common problem is to maintain behavioral profiles of individual entities (customers, for example). Goals include flagging aberrant transactions, i.e., those not indicated by the models and thus potentially being fraudulent; and detecting paradigm shifts in prevailing trends. In these applications as well as others, analysis of data streams also engenders difficult new problems in data visualization. For example, how is time-critical information best displayed? Can automatic response systems be created to deal with common cases?

Time-series applications exploit sequences of unitary observations taken over time. Examples are reports from sensors monitoring network equipment; inventory levels of parts in a warehouse; positions of objects in successive celestial surveys; and records of prices of commodities, stocks, and bonds. Analyzing and mining time-series data presents many new challenges. What are similar time-series data? How can they be clustered, e.g., to isolate seminal events that cause many simultaneous or near-simultaneous disruptions among the observed elements? How can we find interesting trends?
We exploited recent attention to streaming models of data analysis by the theoretical community as well as recent successes in real-time or near-real-time analysis by practitioners. We brought together an interdisciplinary group of researchers to share their ideas and experiences, in the hopes of initiating new approaches to and motivating others to attack core problems in streaming data analysis.

Existing Systems and Prior Work
We built on past DIMACS activities and present DIMACS strengths in various areas related to streaming data analysis and mining. We made use of the DIMACS partnership with AT&T Labs to draw upon the expertise of a variety of researchers at that organization and of the systems that they have developed and we investigated whether these systems might apply to other applications or more general approaches to streaming data analysis. Analysis of data streams is used at AT&T to detect many types of fraudulent behavior; stolen calling cards and impending bill defaults are two examples. Similarly, the JAM system developed at Columbia detects fraudulent activity in financial systems. Gecko, developed by Hume, Daniels, and MacLellan, is a system that audits the largest billing process in AT&T. Using data feeds from passive taps throughout the system, Gecko tracks individual billing records from the network (originating phone call) through settlement (account collection). The main purpose is to detect abnormal processing of records which, if uncorrected, can lead to significant revenue loss. Many open problems remain in such work. For example, how can many heterogeneous data feeds be integrated in one system? How are relevant queries best expressed, giving the user sufficient power while implicitly restraining him/her from incurring unwanted computational overhead? Can domain-specific programming languages be used to provide uniform interfaces to disparate streaming data mining applications? We investigated these.

Critical to the proper functioning of today's massive computer and communications networks is real-time network monitoring. Network monitoring infrastructure creates a sequence of network events and alarms, the correlation of which can facilitate the location of root problems. This is an example of a time-series application. Analysis of data streams is used to detect network faults when they occur, in order to direct timely corrective action. Security trace audits of IP logs are used to detect and react to network attacks, e.g., denial of service attacks. Such an event might trigger some automatic monitoring system, which then prompts intense analysis of recent logs. Among the open problems in this area are the following. How are the results of such monitoring systems best reported and visualized? To what extent can they incur fast and safe automated responses?

The amount of data being collected through scanners at supermarkets and other retail outlets is awesome and marketing research faces the task of making use of these gigantic data sets. Of great interest in marketing is research on “market basket” models and in particular on what items tend to be bought concommitantly. This area of research is moving from off-line settings to more on-line scenarios. For example, performed hourly or even more frequently, such analysis can be used for “just-in-time” provisioning of markets. On-line approaches to such market basket models are discussed in the papers by Ullman and Ganti, Gehrke, and Ramakrishnan.

The DIMACS “Special Year on Massive Data Sets” (1997-1999) featured a workshop on Astrophysics and Algorithms motivated by the huge data sets arising from 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. That special year

led to the beginnings of collaborations between computer scientists and astronomers, dealing with the "paradigm shift" in astronomy toward a situation where many researchers spend their time data mining a "digital sky" compiled from a vast array of multi-wavelength sky surveys. Current large scale cosmological simulations generate data of order 100 GB/per simulation. With the advent of higher bandwidth and faster computers, distributed data sets in the petabyte range are being collected. The problem of obtaining information quickly from such databases requires new and improved mathematical methods. Parallel computation and scaling issues are important areas of research. Techniques such as decision trees, vector-space methods, Bayesian and neural nets, and data compression have been utilized.

Financial markets provide time-series data, in particular, the prices of commodities, stocks, and bonds, as functions of time. Spotting and exploiting trends in such data is of great interest, e.g., to trading firms willing to risk capital and to companies that want to hedge against fluctuating international currencies. While DIMACS has not been heavily involved in financial modeling, we have organized a workshop on the topic and some of our members are actively involved in the field. Some of the general issues involving streaming data analysis for time series data, such as issues of similarity, clustering, and trend-spotting, present serious challenges in modeling of this kind of data and provided another motivating application for our working group.

Some of the Issues
We gathered at DIMACS researchers and developers who work on streaming data mining systems, for the purpose of exchanging ideas, problems, methods, and conjectures, and initiating joint research activities. We included theoreticians and practitioners from multiple disciplines (e.g., computer science, statistics, astrophysics). Some of the more specific issues addressed by the group have already been mentioned above. Among the broader questions addressed were the following: What is the scope of streaming data analysis? What are the relevant problems in data collection, transmission, processing, and visualization? How do current theoretical models for analyzing massive data streams correlate to methods used in practice? What are the core problems faced by practitioners, and how can theoretical insights be used to attack them? Our primary goals were to build bridges between various communities working on problems in streaming data analysis and mining; to initiate collaborative research; to develop a list of core problems that will motivate future work by the communities; and to produce an infrastructure for sharing results as we go forward.

Streaming Data Analysis and Mining Meeting II
Dates: March 24-26, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Adam Buchsbaum, AT&T Labs and Rajeev Motwani, Stanford University
Attendance: 35

The best way to summarize this meeting is to describe the presentations and discussions by topic.

DATA STREAMS FILTERS:

Saar Cohen and Yossi Mattias presented an extension to multi-sets of Bloom Filters called Spectral Bloom filters (SBF'S). SBF's support queries on the multiplicities of individual keys with a guaranteed small error probability. An efficient data structure to build and maintain SBF's overstreaming data was introduced. SBF's represent a high quality synopsis of a data stream that allow ad-hoc queries for individual items enabling a range of new applications.

SYNOPSIS DATA STRUCTURES:

An efficient back traversal of a unidirectional list using small memory and with an almost negligible slowdown in forward steps is possible by using a novel pebbling technique over a virtual binary tree that can be traversed only in a pre-order fashion. Yossi Matias and Ely Porat showed how this list traversal synopsis extends to general directed graphs and how it can be applied to memory efficient hash-chaining implementations.

AGGREGATE QUERIES AND FREQUENCY STATISTICS OVER DATA STREAMS:

Internet routers and Gateways benefit from accurately maintained frequency statistics. A class of frequency statistics algorithms use counters as their primary mechanism of operation. Prosenjit Bose, Evangelos Kranakis, Pat Morin, and Yihui Tang (of Carleton University) addressed the issue of accuracy in packet counting algorithms and show that a very large class of randomized algorithms can not be significantly more accurate than their deterministic counterparts.

Time-decaying aggregates are used in applications where the significance of data items decreases over time. Edith Cohen and Martin Strauss (of AT&T Labs) offered a formalization of the problem of maintaining time-decaying aggregates and statistics of a data stream and provided storage-efficient algorithms for important families of decay functions.

Johannes Gehrke (of Cornell University) described the use of probabilistic "skecthes" to obtain approximate answers to aggregate queries over a data stream. Sketch sharing techniques allow the improvement of the overall space utilization among multiple queries. (This is part of joint work with Al Demers, Alin Dobra, and Mirek Riedewald of Cornell University and Minos Garofalakis and Rajeev Rastogi from Lucent Bell Labs.)

MONITORING AND DIAGNOSIS:

Irina Rish, Mark Brodie, Sheng Ma, Genady Grabarnik and Natalia Odintsova approached monitoring and diagnosis in a distributed system as a probabilistic inference problem. They used the graphical framework of Bayesian networks, where probe results correspond to observed nodes (evidence), while the problems to be diagnosed are represented by hidden nodes. In order to represent temporal dependencies they considered a "factored" HMM model where both hidden and observed states are multivariate. They derive some theoretical conditions on the minimum number of probes required for an asymptotic error-free diagnosis and develop efficient search techniques for probe set selection that can greatly reduce the probe set size while maintaining its diagnostic capability.

Xin Guo and Bonnie Ray proposed a dynamic sampling strategy to efficiently determine if a collection of metrics, representing a system status, have exceeded a set of pre-determined threshold values. The technique's objective is to minimize the probability of undetected threshold crossings. Numerical results were presented regarding the performance under storage and cost constraints.

Brian Babcock, Shivnath Babu, Mayur Datar and Rajeev Motwani showed how the choice of an operator scheduling policy can have significant impact on peak system memory usage and present several scheduling strategies. One of their strategies has near-optimal instantaneous memory usage at every time instant. The authors have conducted a thorough experimental evaluation comparing competing scheduling strategies and validating their analytical conclusions.

SLIDING WINDOWS:

Phillip B. Gibbons (Intel Research) and Srikanta Tirthapura (Iowa State) used a family ofsynopsis data structres called "waves" to obtain algorithms for estimating aggregate functions over a "sliding window" of the k most recent data items in one or more data streams.

A stochastic stream traffic model assumes that the frequencies of items approximately conform to a multinomial distribution in every instance of the sliding window. Two algorithms for identifying frequent items in this model were introduced by David DeHaan, Erik D. Demaine, Lukasz Golab, Alejandro L'opez-Ortiz, and J. Ian Munro. These algorithms use a generalization of previous work of D. Zhu and D. Shasha on the statistical monitoring of thousands of data streams in real time.

The diameter problem of a set of 2d points in the streaming and the sliding-window model was discussed by J. Feigenbaum, S. Kannan, and J. Zhang. Certain parametrized versions of this problem help to view several proposed sliding-window approximations as extensions of the more traditional dynamic versions of these problems.

GRAPHICS HARDWARE AIDED ALGORITHMS

Suresh Venkatasubramanian (AT&T Labs) surveyed the streaming computing capabilities of the graphics pipeline and presented examples of problems that can currently be solved by its careful use.

Stream caching is a mechanism to support multi-record computations within a stream processing architecture. Building on experiences with graphics hardware, Nat Duca, Jonathan Cohen and Peter Kirchner, suggested that stream caching greatly expands the capabilities of stream processing. The extent to which this is the case is an open question that the authors posed to the stream processing community as a whole.

STREAMING, PSEUDORANDOM GENERATION AND MARKOV CHAINS

Sudipo Guha presented a black box proof of correctness for small-space and poly-log time sketch algorithms even when the sketch is reused many times. This result is interesting since simulation by pseudo-random number generators is non-trivial.

Inferring mixtures of Markov chains from observing a stream of their interleaved outputs was one of the problems considered by Tugkan Batu, Sudipto Guha and Sampath Kannan. They studied the case where the mixing process chooses one of the Markov chains independently according to a fixed set of probabilities at each point, and only that chain makes a transition and outputs its new state. For this case, if the individual Markov chains have disjoint state sets the authors showed that a polynomially-long stream of observations is sufficient to infer arbitrarily good approximations to the correct chains. If the state sets are not disjoint only the special case of two Markov chains is well understood. In the case where the output function of the states is not a one to one mapping as above the authors posed the following question: "Under which circumstances is the observed behavior of such a chain itself a Markov chain?". The problems considered in this work are motivated by applications such as gene finding and intrusion detection.

COMPUTATIONAL MODELS

The current formal streaming models are overly pessimistic in modeling the capabilities of today's computing platforms. Extending the classical streaming model by providing a sort box it can be shown that undirected connectivity, minimum spanning trees and suffix array construction can be solved efficiently. If a machine now is allowed to read two streams in parallel then the model becomes computationally more powerful. This gives rise to the notion of "Streaming Networks". These are directed

acyclic graphs where each node represents a machine with a small local memory, and edges are streams transmitted between nodes. Matthias Ruhl, Gagan Aggarwal, Mayur Datar, and Sridhar Rajagopalan, showed that the power of Streaming Networks depends quite directly on the manner the nodes can access their input strings. By allowing them to freely interleave accesses to their input streams the network's computational power increases dramatically. They formalized these different models and present results that relate Streaming Networks to more classical models like the I/O disk model used in the analysis of external memory algorithms. These models are motivated by recent networking applications for distributed graphics rendering that use streams to exchange data in a network.

CLUSTERING

Thanks to the work of Moses Charikar, Liadan O'Callaghan, and Rina Panigrahy now we have a randomized algorithm that works with high probability for the streaming k-median problem. It uses O(k poly log n) space and produces an O(1)-approximation. Given an upper bound D on the ratio of the maximum distance to the minimum distance, a streaming algorithm that makes O (log D) passes and uses space O((k**2)(log D/eps**2)) to produce a solution which is a (1 +eps)-approximation using O(k(log D)/eps) centers. This work also presents O(1)-approximations for stream clustering problems where it is allowed to exclude at most delta fraction of the points (or "outliers" from the clustering).The algorithms use O(k log n) space and slightly increase the outlier fraction (a bicriterion guarantee). It is worth mentioning that even in the offline setting, it is not known how to obtain a solution with a constant factor approximation while maintaining the same fraction of outliers as the optimal solution.

HASHING

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni offered a novel version of a locality-sensitive hashing algorithm. It is simple and quite easy to implement and unlike the previous algorithm it works directly on the points in d dimensional space without embeddings. The query time bound is free of large polylogarithmic factors and it works for any Lp norm for 0 < p <= 3D 2.

PATTERN MATCHING IN XML DOCUMENT TREES

Finding all occurrences of a tree pattern in an XML document tree can be done optimally if one is allowed to perform offline sorting on the nodes and to keep redundant nodes. This proves the optimality of the so-called Holistic Join Algorithm and is the joint work of Byron Choi and Malika Mahoui (University of Pennsylvania).

Working Group on Computer Generated Conjectures from Graph Theoretic and Chemical Databases

The primary goals of the working group were to foster an interchange of ideas among those working on different approaches to the use of computers to generate scientific conjectures, mainly in graph theory and in chemistry, and to initiate new collaborative research in this area. There was an exchange of views both about methods and about specific conjectures.

The working group web site contains a summary of the meetings, open questions, links to various types of software, and descriptions of the meetings.

This working group met as follows:

Computer-Generated Conjectures from Graph Theoretic and Chemical Databases Meeting I
Dates: November 12 -16, 2001

Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Patrick Fowler, University of Exeter; Pierre Hansen, GERAD - University of Montreal
Attendance: 43

The process of scientific discovery is a very complex one. Computers can aid human beings in the process and, as has been a goal of researchers in artificial intelligence, in an automatic way. In the mathematical sciences, discovery can be thought to have three components: development of conjectures, formation of new concepts, and proving of theorems or disproving of conjectures. There has been a large amount of research done on automatic theorem proving. Here, we concentrate on the use of the computer as a tool in generating conjectures, whether in an automated way or as a tool used interactively by a person.

Over the years, a variety of programs have been written to produce mathematical conjectures, either automatically or interactively. Some of these programs have led to very interesting new conjectures and new theorems. An area of particularly widespread activity and interest has been graph theory and related areas of chemistry. Researchers from a variety of groups around the world have been working in this area and are engaged in the tasks of using computational power to develop new conjectures, eliminate uninteresting ones, publicize the remaining conjectures, and aid in their solution by working scientists, with or without the aid of the computer. Many of these efforts start with large databases, for example of graph invariants and of relations among them, or of chemical structures. We brought together a group of researchers working on computer generated conjectures from graph-theoretic and chemical databases, including developers of some of the most important programs being used, to share their ideas and systems and initiate joint research efforts. We also gave a larger group of researchers an opportunity to try these systems out.

The primary goals of the working group were to foster an interchange of ideas among those working on different approaches to the use of computers to generate scientific conjectures, mainly in graph theory and in chemistry, and to initiate new collaborative research in this area. We encouraged exchange of views both about methods and about specific conjectures. Among the questions addressed were the following:

* Can we reinforce existing systems that are knowledge-based and put some knowledge-base into other systems?

* Some of the systems consist of protocols for computing graph invariants. How can these be expanded to include broader concepts of graph invariant and combine various constraints?

* Some of the systems in use depend upon a huge number of heuristics and others upon only a few? Are there general principles as to the usefulness of adding more heuristics?

* Are there metaheuristics such as variable neighborhood search that can be usefully employed by a variety of systems?

* What would be involved in broadening the usefulness of existing systems, applying them for example to biological databases?

* Can we define interesting lists of conjectures in terms of novelty, simplicity, generality?

* Can we specialize systems according to the types of conjectures we wish to find, e.g., 'tightest' ones (those with best possible upper or lower bounds)?

* Can we build 'self-improving' systems that use results of their search to improve the search process?

* Can any of these systems be applied to cluster analysis, which tries to form subsets of a set that share some common pattern?

* Can any of these systems be applied to, or benefit from, consensus theory, which looks at several objects and tries to form one or more objects that are 'close' to the original ones?

Software for Automatic Generation of Conjectures

The use of computers in scientific discovery, and in particular in the development of mathematical conjectures, is not new. A paper by Larsen in the DIMACS volume resulting from this working group surveys automatic conjecture-generating software, tracing back to the work of Wang in the 1950's. Graffiti, a program developed by Siemion Fajtlowicz, was introduced in 1986. For a given set of invariants, conjectures are automatically generated and tested on a database of examples. The system discards many of the conjectures automatically, for instance those implied by others. The most interesting conjectures are proposed to the community of researchers and counterexamples that are produced are added to the database. There are now over 900 conjectures developed through Graffiti, many in graph theory and related areas of chemistry, but others in geometry, number theory, and other fields. See the URL http://www.math.uh.edu/~clarson/#fajt. This website provides information about the current status of these conjectures -- open, refuted, or proved -- and on developments related to these conjectures, such as new conjectures or theorems. Some of the world's most distinguished graph theorists have worked on the conjectures generated by Graffiti. They include Noga Alon, Bela Bollobas, Fan Chung, Paul Erdos, Jerry Griggs, Daniel Kleitman, Laszlo Lovasz, Paul Seymour, and Joel Spencer. Among the more exciting products of Graffiti are conjectures in chemistry on the structure of fullerenes (as represented by graphs). The papers by Patrick Fowler and colleagues, one motivated by Graffiti, show how some very simple ideas of symmetry, geometry, and graph theory can be used to count, construct, and classify fullerenes and make predictions about the types and structures of their chemical derivatives. As an aside, we comment on the chemistry involved. The main point of much of the conjecture-generating software we are summarizing is to use theory to give models and physical pictures that can be used to understand complicated chemical systems. The all-carbon fullerenes give an example of a relatively new class of molecules for which these conjecture-generating systems have already been very useful. Among the questions of interest about these molecules are: How many fullerenes are there? Which have greatest overall stability? What are their likely patterns of reactivity? Graph theory can be used to answer some of these questions since fullerenes can be viewed as planar cubic graphs having only faces of size 5 or 6. Collaborations between chemists and graph theorists in the working group followed on collaborations stimulated by the DIMACS workshop on Discrete Mathematical Chemistry and through workshops on computational chemistry in the Special Year on Mathematical Support for Molecular Biology.

The system INGRID was developed by Robert Brigham, Ronald Dutton, and Felix Gomez in the late 1980's. INGRID was used to generate patterns, thus suggesting conjectures to its users, with an emphasis on parameters of graph theory. INGRID automatically bounded invariants of graphs from values for some given invariants stored in a database and by manipulation of a database of formulas between graph invariants. Whether or not INGRID can be said to have "automatically generated conjectures" or even "generated conjectures" is a point of dispute to which we return below.

The AutoGraphiX (AGX) system was developed more recently by Pierre Hansen and co-workers at GERAD, Universite de Montreal. The AGX system is designed to find (heuristically) extremal graphs for some invariant. To that effect, Caporossi and Hansen have made use of a "variable neighborhood search" metaheuristic within AGX, developed by viewing the conjecture generation task as a combinatorial optimization problem on an infinite family of graphs. By making use of parametrization, the system produces a set of presumably extremal graphs in an automated way. Conjectures can then be found either interactively or automatically. In the latter case, the process exploits a numerical approach based upon

principle components analysis, a geometric approach using the notion of convex hull in a space of chosen invariants, or an algebraic approach based on graph recognition and known relations among families of graphs. In applications in graph theory, AGX has been used, for example, to obtain results about the largest eigenvalue for color-constrained trees. In applications related to chemistry, the system AGX has been used to obtain bounds for the connectivity of chemical trees and for the "energy" of a graph, a problem for which much more complicated bounds had previously been obtained. It has also found tight bounds on extremal chemical graphs for the much-studied Randic index.

An earlier system we should mention is called Graph. This system, due to Dragos Cvetkovic and colleagues, is aimed at enhancing the capacity of the graph theorist to explore graphs and families of related graphs through computation of invariants and visualization. In a review of Graph in the DIMACS volume resulting from this working group, 92 papers using it are mentioned. Brief mention should also be made of the system HR (for Hardy-Ramanujan) that was developed by Simon Colton and others to form concepts, make conjectures, and use a theorem prover and model generator to prove or disprove those conjectures. HR has so far produced conjectures in group theory, number theory, and graph theory.

Other Relevant Software

We mention a number of other systems that do not generate conjectures automatically, but have been widely used as tools in scientific discovery/conjecture generation and also in education in discrete mathematics.

The LINK software system was developed at DIMACS starting in 1991. (See the URL http://zhivago.elon.edu/~berryj/LINK.html.) LINK resulted from the DIMACS workshop on Software for Discrete Mathematics, during which we brought together researchers working on software from different points of view, using different platforms, emphasizing different fields of discrete mathematics. LINK is a software system designed to be a general-purpose environment for experimentation with discrete mathematics. The environment is friendly enough so that non-technical users will be able to visualize problems easily, yet powerful enough for computer scientists and mathematicians to express complicated computations. The system is designed to support educational pursuits, to be used as a research tool, and to be useful in solving real-life problems. LINK grew out of the experiences of the authors of Combinatorica, due to Steven Skiena, NETPAD, due to Nate Dean and others at Bellcore, SetPlayer, due to Mark Goldberg and his students, and GraphLab, due to Greg Shannon, et al. Skiena, Dean, Goldberg, and Shannon came together through DIMACS to develop LINK. Also, Jonathan Berry, a DIMACS postdoc, played an important role in developing LINK and currently maintains the LINK website. LINK has been used as a tool to develop and analyze graph-theoretical conjectures, as well as a tool to explain graph-theoretical concepts and a tool for practical applications of graph theory and discrete mathematics. For example, Brenda Latka has used the system to generate conjectures about infinite antichains of tournaments (complete directed graphs). Of some interest are some of the non-mathematical uses of LINK. Berry and Dean used LINK to analyze correlations between purchases using supermarket data. An interesting offshoot of LINK has been its adoption by the Internal Revenue Service in fraud detection. This resulted from the system's ability to detect common patterns.

Briefly, here are some other systems of interest. VEGA, of which Tomasz Pisanski is a major developer, is a system for "doing" discrete mathematics that is based on Mathematica and springs from the Combinatorica system developed by Steve Skiena and the nauty system developed by Brendan McKay (see the URL http://cs.anu.edu.au:80/people/bdm/nauty). VEGA allows researchers, teachers, and students to quickly test ideas on small and mid-size examples. (See the URL http://vega.ijp.si/Htmldoc/vinfo.htm.)

Plantri, developed by Brendan McKay and Gunnar Brinkmann, is a program for generating planar triangulations and planar cubic graphs. It includes a fullerene generator fullgen, written by Brinkmann.

Fullerenes, as we have noted, can be viewed as planar cubic graphs having only faces of size 5 or 6. See http://cs.anu.edu.au/people/bdm/plantri for general information about plantri and fullgen.

LEDA (Library of Efficient Data Types and Algorithms) was started in 1988. Stefan Naeher is a primary developer. LEDA provides a collection of data types and algorithms for "combinatorial computing". (See http://www2.informatik.uni-halle.de/~naeher/Manual/Preface.html.) The Graph Theorist, GT, is an older system developed by Susan Epstein. It is a learning system that uses algorithmic class descriptions to discover and prove relations among mathematical concepts. CABRI, developed by Jean Marie Laborde, has recently been aimed at the use of the computer to learn about and explore concepts of geometry. However, it has also been widely used for exploring graphs and families of related graphs.

Some Controversial Issues

The literature is filled with illustrations of the fact that it is extremely difficult to define what constitutes "intelligence" and exactly what would constitute "automatic generation" of new scientific ideas. Indeed, there is even controversy in what it means for a computer to "generate" new scientific ideas, let alone whether the generation is "automatic." Siemion Fajtlowicz, in an early version of his paper "Postscript to fully automated fragments of graph theory," argues: "When the act of 'making a conjecture' is attributed to a machine, the author of the program should be expected to clearly explain how the program, as opposed to its users, reached these conclusions." In an early version of this paper and elsewhere, he explains how this was done with Graffiti. In a paper in the DIMACS volume resulting from this working group, Pierre Hansen argues that several steps in the process of finding conjectures with Graffiti (i.e., mainly, but not only, finding and adding counter-examples) are not automated. He therefore views such conjectures as obtained interactively, with Graffiti, not by Graffiti. Siemion Fajtlowicz, in early versions of the paper "Toward fully automated fragments of graph theory," disagrees, viewing such steps as happening between "rounds" of the use of Graffiti. He notes that a system can, as does a mathematician, formulate several successive conjectures on the same topic, based on increasing knowledge. Fajtlowicz therefore wishes to "attribute to Graffiti all of its conjectures." In their paper in the DIMACS volume, Hansen and his collaborators observe, using the definition of conjecture in Bouvier and George's *Dictionnaire des Mathematiques*, that conjectures in graph theory may take many different forms. As a consequence, they view two of the six functions of the system INGRID, mentioned above, as providing conjectures automatically. Again, Simeon Fajtlowicz disagrees. In an early version of "Postscript to fully automated fragments of graph theory," referring to AGX, he says: "It is self-evident that attributing to a program conjecture-making abilities automatically excludes from this category programs written for the purpose of assisting humans in making conjectures, or for the purpose of verification or confirmation of existing conjectures." It is reasonable to infer that he feels the same way about INGRID. The authors of INGRID have never claimed that the program generates conjectures without the intervention of human beings. However, whether or not it is reasonable to refer to such a program as a "conjecture-making program" remains a bone of contention.

Computer-Generated Conjectures from Graph Theoretic and Chemical Databases Meeting II

Dates: June 2 - 5, 2004
Location: Centre de Recherches Mathematiques (CRM), Universite de Montreal, Quebec, Canada
Organizers: Patrick Fowler, University of Exeter; Pierre Hansen, GERAD-University of Montreal
Attendance: 53

This meeting had several aims:
(i) To survey main results obtained in computer-aided or automated discovery in various fields of mathematics such as number theory, geometry, graph theory, algebra, etc., as well as in various sciences such as chemistry, physics, bioinformatics, economics, ecology, etc.;

(ii) To present and discuss main tools of computer-aided or automated discovery;
(iii) To illustrate the working of software for discovery through demonstrations and discussions;
(iv) To stimulate the initiation of collaborative research between teams using different techniques and/or working in different fields.

The meeting was truly an international event. The presentations were mostly on discovery in mathematics (number theory, geometry, game theory, graph theory) but also in chemistry, biochemistry, physics, and economics.

A large variety of techniques were described and there were demonstrations of several systems, including Mikhail Klin's system COCO, Ermelinda Delavina's system Graffiti.pc, Simon Colton's system HR, Gilles Caporossi's system AGX2, and the first public demonstrations of Stevanovic's NewGRAPH and of Melot's GraPHedron.

Jonathan Borwein, Dalhousie University, opened the meeting with a talk entitled "Experimentation in mathematics: computational paths to discovery." Following along the same lines was the talk entitled "Experimental mathematics: discovering new formulas and theorems" given by David H. Bailey, Lawrence Berkeley National Laboratory. Shang-Ching Chou, Wichita State University, Kansas, talked about machine proofs and discovery in geometries, giving a general introduction to this topic. Several talks dealt with particular methods for computer-aided discovery. Simon Colton, Imperial College, London, spoke about the HR project, its hits and misses. Ermelinda Delavina, University of Houston, Texas, spoke about "The Dalmation heuristic," that arises in Graffiti.

Still other talks dealt with some concrete applications of computer-aided discovery. Zigzags and central circuits for 3- and 4-value plane graphs was the topic of a talk by Mathieu Dutour, École Nationale Supérieure, Paris and Einstein Institute, Jerusalem. Pat Langley, Institute for the Study of Learning and Expertise, Palo Alto, California, spoke about the computational induction of explanatory process models. Stephen Muggleton, Imperial College, London, made a presentation called "The robot scientist."

Additional speakers and their presentations were:

Mark Goldberg, Rensselaer Polytechnic Institute, "Experimental Asymptotics: how much experimentation is enough"

Gunnar Brinkmann, Gent University, "Generating benzenoids and fusenes with perfect matchings"

Mikhail Klin, University of Delaware, "Regular subgroups of collineation groups of proper finite loops: From a computer experiment to an infinite series of examples"

Hadrien Mélot, University of Mons, "Facet defining inequalities among graph invariants: the system GraPHedron"

Jack. E. Graver, Syracuse University, "The independence numbers of fullerenes and their duals"

Dragan Stevanovic, University of Nis, "Using NewGRAPH in research and teaching"

Mustapha Aouchiche, École Polytechnique, "Conjectures about average distance in graphs"

Charles Audet, GERAD and École Polytechnique, "Vincze's wife's octagon is suboptimal"

David Avis, GERAD and McGill University, "All meals for a dollar, Nash Equilibria and other vertex enumeration problems"

Simon Plouffe, Montréal, "A search for a mathematical expression of mass ratio using a large database"

Patrick W. Fowler, Exeter University, "Non-bonding orbitals: much ado about nothing"

Reinhard Laue, University of Bayreuth, "Challenges for group actions from t-design construction problems"

Wendy Myrvold, University of Victoria, "Generating small combinatorial objects"

Nair Abreu, Federal University of Rio, "Bounds on the algebraic connectivity of a graph"

Pierre Hansen, GERAD and HEC Montréal, "Exploring graph theory with AutoGraphiX"

Yoshua Bengio, University of Montréal, "Learning the density structure of high-dimensional data"

Shengrui Wang, University of Sherbrooke, "Cluster analysis on graph data"

Robert Cowen, Queens College, Flushing, NY, "Computer-assisted investigations for the paper 'Odd Neighborhood Transversals for Grid Graphs'

Sandra Kingan, Penn State University, "Excluded minor results in matroids"

Vladimir Brankov, University of Nis, "NewGRAPH architecture"

Claudia Justus, Bielefeld University, "Numbers of faces in disordered patches"

Gilles Caporossi, GERAD and HEC Montréal, "Automated proof of simple conjectures in graph theory"

III. Project Findings

Working Group on Algorithms for Multidimensional Scaling

This working group met in August 2001 and June 2003. The main accomplishment of this group was the development and enhancement of cross-disciplinary research efforts. Here are the highlights of these endeavors.

Mathematical Programming and MDS
Larry Hubert (Psychology, University of Illinois), Phipps Arabie and Douglas Carroll (Graduate School of Management, Rutgers) together with Michael Brusco (School of Business, Florida State University) began to explore various mathematical programming techniques to fit MDS models, including various possible collaborative efforts. Partly as an outgrowth of this working group, Phipps Arabie, Larry Hubert, (University of Illinois, Champaign) and J. Meulman have been working on a monograph, *The Structural Representation of Proximities with MATLAB*. The 230-page monograph is now under review with the Society of Industrial and Applied Mathematics, generally considered this country's foremost non-profit publisher in mathematics and includes applications in the behavioral sciences). They are including extensive software with the monograph.

Approaches to MDS of Massive Data Sets
David Dubin (Library Science, University of Illinois) Douglas Carroll and Michael Trossett (Math., William and Mary) are all exploring various approaches to MDS of massive data sets including possible extensions of some already established research in this area.

Non-linear Mapping of Data Whose Topological Dimensionality is Lower than that of the Space into which Data Points are Embedded
Ulas Akkucuk and Doug Carroll have been developing a method for nonlinear mapping of data whose topological dimensionality is lower than that of the linear vector space in which the data points are embedded, e.g., points on the surface of a sphere, which is a locally two-dimensional surface embedded in a three dimensional linear space, leading to a problem closely analogous to the cartographer's problem of producing flat two-dimensional maps of the earth's surface. Another highly nonlinear structure they discussed was a highly symmetric torus (or donut) embedded in four dimensional space, whose local topological dimensionality is also two.

A technique previously proposed by Carroll called Parametric Mapping of Nonlinear Data Structures (or PARAMAP) had worked quite well on "flattening" out these structures, with regularly spaced points and no added error into appropriate flat two-dimensional "maps", but had found that this method broke down completely when the spacing of points was irregular or when significant error was added. It appeared at the time (the late 1960's) that the failure of PARAMAP with irregularly spaced or "noisy" data was due to the existence of a particularly severe local optimum problem. In those days it was possible to run the gradient based optimization procedure used in PARAMAP from only a relatively small number of different starting points, and it appeared that the problem was an extremely severe local optimum (minimum) problem. Motivated in part by the development of other nonlinear mapping methods, such as Tenenbaum et al's ISOMAP procedure, which solves a less complex version of this problem (but definitely does not work for completely closed surfaces, such as the points on the complete sphere or torus), Akkucuk and Carroll began their re-exploration of PARAMAP, dealing primarily with perturbed or "noisy" points on the sphere, by simply running the gradient-based method for optimizing the measure of fit (called "kappa", a measure of "continuity" or "smoothness" of the empirical function mapping the points in the flat space onto the sphere, in this case) defining the loss function in PARAMAP from many more different random starting points and for many more iterations than was possible with the primitive equipment of the late 1960's. This simple extension generated many more local minima, among which it was much more probable that one was very close to the solution providing the global minimum of kappa. The solution from these initial runs with the lowest value of kappa was then run to full convergence, and it was found that this conformed very well to the map one would expect for the perturbed set of points on the circle, as gauged by an independent measure of preservation of local structure. Steps are now being taken to extend this approach to handle a larger number of points, to synthesize the best aspects of PARAMAP with the best aspects of ISOMAP, and to improve this approach in numerous other ways.

A working paper entitled "Nonlinear mapping: Approaches based on optimizing an index of continuity and applying classical metric MDS to revised distances" has been written by Akkucuk and Carroll, and is available on the website for the MDS-2 working group. In the discussion in the working group meeting after presentation of the paper by Akkucuk, there were a number of suggestions made for improving the optimization procedure being used, with the goal of increasing the likelihood of obtaining a global rather than a merely local minimum of kappa. Some further work is now underway to implement many of these suggested changes, with the intention of submitting the resulting paper to the *Journal of Classification.*

Akkucuk and Carroll have been working with data that involves imaging of points on the surface of the cerebral surface of the brain, which in fact is essentially (at least to a first approximation) a highly convoluted two-dimensional manifold embedded in the three dimensional structure of the brain. Other methods of "brain flattening" have been developed but found to be less than fully satisfactory. It is hoped

that use of a hybrid PARAMAP/ISOMAP approach designed to deal with a large number of points will produce flat maps much more useful in studies of the structure of the brain than these existing techniques. This work will be pursued as a follow-up to discussions in the working group.

MDS and Molecular Conformation
Trosett and Lewis have begun to collaborate on MDS and molecular conformation. The idea is to take actual inner-atomic measurements of molecules, and to use MDS to construct a feasible model of the molecule. Plans are being prepared to organize a small working group meeting at DIMACS that will bring together chemists who are doing the measurements and mathematicians, computer scientists who are developing the computer algorithms. Lewis has begun to use the second-order sensitivity work he discussed to implement some specialized nonlinear optimization algorithms for the problem of reconstructing molecular structure from bound constrained nuclear magnetic resonance (NMR) data. Trossett and Lewis are also beginning to look more closely at Zhijun Wu's approach to the same sorts of NMR MDS problems. Likewise, in her talk and in discussions with Trossett and Lewis, Suzanne Winsberg suggested that some of her work may be useful in getting them pretty good starting points for optimization, and they are investigating this direction, too. They owe their exposure to Zhijun's and Suzanne's work entirely to the working group.

Comparision of Metric INDSCAL with Nonmetric PROXSCAL
Willem Heiser reports that the working group has "provided very important new impetus" for his research. He started a new collaboration with Doug Carroll, to compare his metric INDSCAL method with the nonmetric PROXSCAL method that was developed by Frank Busing and himself. He and Carroll are planning to write a joint paper on the project, using a very interesting historical dataset from Frank Klingberg that Heiser presented at the second working group meeting.

Properties of SINDCLUS
Chaturvedi and Carroll have proposed the SINDCLUS method for fitting the INDCLUS model. It is based on splitting the two appearances of the cluster matrix in the least squares fit function and relying on convergence to a solution where both cluster matrices coincide. Kiers has proposed an alternative method that preserves equality of the cluster matrices throughout. Jos ten Berge and Henk Kiers have shown that the latter method is generally to be preferred. However, because the method has a serious local minimum problem, alternative approaches should be contemplated.

Industrial Connections
There was also the start of or enhancement of collaborations among academic participants and industrial scientists such as Anil Chaturvedi of Kraft Foods and Andreas Buja of AT&T Laboratories.

Working Group on Streaming Data Analysis

This working group met in November 2001 and March 2003. Apart from the valuable brainstorming sessions held with the meetings, new and continuing research relationships were begun. Here are some samples.

A Windowed Data Stream Model for Detecting Patterns in Streaming Data
In many areas of application, the sheer volume of data requires us to make a quick decision about it as it 'streams' by. This is particularly true in applications of fraud detection in telecommunications, in surveillance in homeland security, and in financial transactions, as well as in astrophysics, environmental modeling, etc. Methods are needed to detect patterns in streaming data. S. Muthukrishnan (Rutgers) and Mayur Datar (Stanford) developed a unique 'windowed data stream model' that is based on observing a 'window' of the last N observations. With storage space less than the window size N, they studied the

problem of estimating the 'rarity' of items in the window, a crucial problem in telephone credit card fraud detection and anomaly detection in counter-terrorism surveillance. They also studied the problem of estimating the 'similarity' between two data stream windows, an important problem in sharing data between datasets obtained by different agencies/companies, a vital problem in the homeland security area. Dr. Muthukrishnan visited the NJ Office of Insurance Fraud in the Division of Criminal Justice to discuss analogous problems they are having. Muthukrishnan and Datar found novel, simple algorithms for estimating rarity and similarity in this windowed situation.

Constructing the First Known Data Stream Algorithm for Estimating Dominance of Multiple Signals
Graham Cormode (University of Warwick) and S. Muthukrishnan (Rutgers) considered streams of multiple signals (i,a_i,j) where the i's correspond to the domain, the j's index the different signals and i,a_i,j gives the value of the j'th signal at point i. They addressed the problem of determining the dominance norms over the multiple signals, in particular the max-dominance norm defined as $\Sigma_i \max_j a_i,j$. Besides finding many applications, such as in estimating the `worst case influence' of multiple processes, for example in IP traffic analysis or electrical grid monitoring, this norm is a natural measure: it generalizes the notion of union of data streams and may be alternately thought of as estimating the L1 norm of the upper envelope of multiple signals. Cormode and Muthukrishnan constructed the first known data stream algorithm for estimating max-dominance of multiple signals. The algorithm is simple and implementable; its analysis relies on using properties of stable random distributions with small parameter alpha, which may be a technique of independent interest. They also showed that other dominance norms -- min-dominance $\Sigma_i min_j a_i,j$, count-dominance ($|\{i: a_i > b_i|$) or relative-dominance ($\Sigma_i a_i/max1,b_i$) -- are all impossible to estimate accurately with sublinear space.

Space Lower Bounds for Distance Approximation in the Data Stream Model
Michael Saks (Rutgers) and Xiaodong Sun (Rutgers) investigated the problem of approximating the distance between two d-dimensional vectors **x** and **y** in the classical data stream model. In this model, the $2d$ coordinates are presented as a "stream" of data in some arbitrary order, where each data item includes the index and value of some coordinate and a bit that identifies the vector (**x** or **y**) to which it belongs. The goal is to minimize the amount of memory needed to approximate the distance. For the case of L^p-distance with $p \in [1,2]$, there were known good approximation algorithms that run in polylogarithmic space in d (here one assumes that each coordinate is an integer with $O(\log d)$ bits). Saks and Sun proved that they do not exist for $p>2$. In particular, they proved a nearly optimal approximation-space tradeoff of approximating L^∞ distance of two vectors. They showed that any randomized algorithm that approximates L^∞ distance of two length d vectors within factor of d^δ requires $\Omega(d^{1-4\delta})$ space. As a consequence they showed that for $p>2/(1-4\delta)$, any randomized algorithm that approximates L^p distance of two length d vectors within a factor d^δ requires $\Omega(d^{[1-\frac{2p-4}{\delta}]})$ space. The data stream model has been recognized recently as an important model because it abstracts the problem of analyzing vast amounts of data "streaming" into a collection point. As for any model of computation, it is of interest to draw a dividing line between things that are efficiently computable in the model and things that are not. Previously other researchers had shown that approximating the distance of two vectors in this model could be done efficiently for standard Euclidean distance and L^1 (Manhattan) distance. These algorithmic techniques did not help for the case of L^∞ (max) distance. The work of Saks and Sun demonstrated that this latter problem is inherently unsolvable efficiently, provided a striking contrast with the other two cases. The lower bound follows from a lower bound on the two-party one-round communication complexity of this problem. This lower bound is proved using a combination of information theory and Fourier analysis.

k-*Median Clustering over "Sliding Windows"*
Brian Babcock, Mayur Datar, and Laidan O'Callaghan are developing techniques for k-median clustering over 'sliding windows'. Thus, a stream of data points is arriving, and the most recent N points are

considered ‘relevant’, but memory is not large enough to store *N* points. Their algorithm maintains a running constant-approximate *k*-median clustering of the relevant points in polylog space.

Monte Carlo Algorithm for Clustering
Adam Meyerson, Laidan O'Callaghan, and Serge Plotkin have a Monte Carlo algorithm that produces a clustering that with high-probability is approximately optimal under the k-median measure of clustering. They have produced lower bounds on the running time required and on the sizes of the samples that must be clustered.

Operator Scheduling for Memory Minimization in Data Stream Systems
In many applications involving continuous data streams, data arrival is bursty and data rate fluctuates over time. Systems that seek to give rapid or real-time query responses in such an environment must be prepared to deal gracefully with bursts in data arrival without compromising system performance. Brian Babcock, Shivnath Babu, Mayur Datar and Rajeev Motwani found one strategy for processing bursty streams --- adaptive, load-aware scheduling of query operators to minimize resource consumption during times of peak load. They showed that the choice of an operator scheduling strategy can have significant impact on the run-time system memory usage. They present Chain scheduling, an operator scheduling strategy for data stream systems that is near-optimal in minimizing run-time memory usage for single-stream queries involving selections, projections, and foreign-key joins with stored relations. Chain scheduling also performs well for queries with sliding-window joins over multiple streams, and multiple queries of the above types. A thorough experimental evaluation demonstrated the potential benefits of Chain scheduling, compared it with competing scheduling strategies, and validated their analytical conclusions.

Maintaining Time-Decaying Stream Aggregates
Edith Cohen and Martin Strauss formalized the problem of maintaining *time decaying* aggregates and statistics of a data stream: the relative contribution of each data item to the aggregate is scaled down by a factor that depends on, and is non-decreasing with elapsed time. Time-decaying aggregates are used in applications where the significance of data items decreases over time. They developed storage-efficient algorithms, and established upper and lower bounds. Surprisingly, even though maintaining decayed aggregates have become a widely-used tool, their work seems to be the first both to explore it formally and to provide storage-efficient algorithms for important families of decay functions, including polynomial decay.

On the Optimality of Holistic Algorithms for Twig Queries
Byron Choi, Malika Mahoui, and Derick Wood's work on Streaming XML documents has many emerging applications. They showed that the restrictions imposed by data streaming are too restrictive for processing twig queries - the core operation for XML query processing. The previous proposed algorithm, **TwigStack**, is an optimal algorithm for processing twig queries with only descendent edges over streams of nodes. The cause of the suboptimality of the **TwigStack** algorithm is the structural recursions appearing in XML documents. They showed that without relaxing the data streaming model, it is not possible to develop an optimal holistic algorithm for twig queries. Also the computation of the twig queries is not memory bounded. This motivated them to study two variations of the data streaming model (1) offline sorting is allowed and the algorithm is allowed to select the correct nodes to be streamed and (2) multiple scans on the data streams are allowed. They computed the lower bounds of the two variations.

Better Streaming Algorithms for Clustering Problems
Moses Charikar, Rina Panigrahy and Liadan O'Callaghan studied clustering problems in the streaming model, where the goal is to cluster a set of points by making one pass (or a few passes) over the data using a small amount of storage space. Their main result is a randomized algorithm for the *k*--Median

problem which produces a constant factor approximation in one pass using storage space $O(k\ poly\ \log n)$. This is a significant improvement of the previous best algorithm which yielded a $2^{O(1/\epsilon)}$ approximation using $O(n^{\epsilon})$ space. They also found a streaming algorithm for the *k*--Median problem with an arbitrary distance function. They also studied algorithms for clustering problems with outliers in the streaming model. They gave bicriterion guarantees, producing constant factor approximations by increasing the allowed fraction of outliers slightly.

Sharing Sketches among Concurrent Queries
Alin Dobra, Minos Garofalakis, Johannes Gehrke, and Rajeev Rastogi studied the problem of using data-stream sketches to process multiple data stream queries concurrently. Their work shows that intelligent sharing of sketches among concurrent queries can result in substantial improvements in the utilization of the available space and the quality of the resulting error guarantees. They prove that optimal sketch sharing typically gives rise to NP-hard questions, and they propose novel heuristic algorithms for finding good sketch-sharing configurations in practice.

Gathering Internet Statistics using Relatively Little Memory
Prosenjit Bose, Evangelos Kranakis, Pat Morin, and Yihui Tang considered the problem of approximating the frequency of frequently occurring elements in a stream of length n using only a memory of size m<<n. This models the process of gathering statistics on Internet packet streaming using a memory that is small relative to the number of classes, e.g. IP address, of packets. They found an upper bound on the error in the frequency of occurrence of a data item in a stream of length n as estimated by the FREQUENT algorithm of Demain et al. They also computed a lower bound on the accuracy of any deterministic packet counting algorithm, which implies the FREQUENT algorithm is nearly optimal. Finally, they showed that randomized algorithms can not be significantly more accurate.

Data Streaming Algorithms for Efficient and Accurate Estimation of Flow Distribution
Knowing the distribution of the sizes of traffic flows passing through a network link helps a network operator to characterize network resource usage, infer traffic demands, detect traffic anomalies, and accommodate new traffic demands through better traffic engineering. Previous work on estimating the flow size distribution has been focused on making inferences from sampled network traffic. Its accuracy is limited by the (typically) low sampling rate required to make the sampling operation affordable. Abhisehk Kumar, M. Sung, Jim Xu, and J. Wang found a novel data streaming algorithm to provide much more accurate estimates of flow distribution, using a "lossy data structure" which consists of an array of counters fitted well into SRAM. For each incoming packet, their algorithm only needs to increment one underlying counter, making the algorithm fast enough even for 40 Gbps (OC-768) links. The data structure is lossy in the sense that sizes of multiple flows may collide into the same counter. Their algorithm uses Bayesian statistical methods such as Expectation Maximization to infer the most likely flow size distribution that results in the observed counter values after collision. Evaluations of this algorithm on large Internet traces obtained from several sources (including a tier-1 ISP) demonstrate that it has very high measurement accuracy (within 2%). Their algorithm not only dramatically improves the accuracy of flow distribution measurement, but also contributes to the field of data streaming by formalizing an existing methodology and applying it to the context of estimating the flow-distribution.

Space-Code Bloom Filter for Efficient Per-Flow Traffic Measurement
Per-flow traffic measurement is critical for usage accounting, traffic engineering, and anomaly detection. Previous methodologies are either based on random sampling (e.g., Cisco's NetFlow), which is inaccurate, or only account for the "elephants". Kumar, A., Xu, J., Wang, J., Spatschek, O., Li, L. introduced a novel technique for measuring per-flow traffic approximately, for all flows regardless of their sizes, at very high-speed (say, OC768). The core of this technique is a novel synopsis data structure called Space Code Bloom Filter (SCBF). A SCBF is an approximate representation of a *multiset*; each element in this multiset is a traffic flow and its multiplicity is the number of packets in the flow. The

multiplicity of an element in the multiset represented by SCBF can be estimated through either of two mechanisms -- Maximum Likelihood Estimation (MLE) or Mean Value Estimation (MVE). Through parameter tuning, SCBF allows for graceful tradeoff between measurement accuracy and computational and storage complexity. SCBF also contributes to the foundation of data streaming by introducing a new paradigm called blind streaming. They evaluated the performance of SCBF through mathematical analysis and through experiments on packet traces gathered from a tier-1 ISP backbone. Their results demonstrate that SCBF achieves reasonable measurement accuracy with very low storage and computational complexity.

Working Group on Computer Generated Conjectures from Graph Theoretic and Chemical Databases

This working group met in November 2001and June 2004. Here is a selection of research results.

Connecting Discrete Mathematics and Chemistry
As a result of these meetings, Fowler, Stevanovic, Hansen and Capporossi initiated collaborations on graph eigenvalues and connection to chemistry. Fowler and Brinkmann are collaborating on the spiral construction of general cubic polyhedra on the systematic mathematically complete cataloguing of fullerene-to-fullerene transformations. Some of this work is still in progress, but several papers appeared as a direct result of the discussions held and contacts made at the workshop.

Proving a Computer-generated Conjecture about the Separator of Fullerenes
It has long been a goal of artificial intelligence to design machines that can think like human beings. A small but quite remarkable research area related to this is concerned with designing machines that can aid in and eventually replace the process of scientific discovery by human beings. During the working group meeting in November 2001, a demonstration of the computer program 'Graffiti' led to the computer generation of several conjectures about the chemical structures known as fullerenes. Fullerenes are molecular graphs of carbon isomers. The separator of the fullerene is the difference between the two largest eigenvalues of the graph and is very useful in understanding the structure of these important compounds. 'Graffiti' generated a conjecture that the separator of fullerenes was always at most $1 - 3/n$, where n is the number of vertices of the graph, and then the conjecture that the separator is always at most one. The latter, stronger conjecture ended up being true and was proved during the working group meeting by two of the participants, Gilles Caporossi from Montreal and Dragan Stevanovic from Belgrade. The proof led to insights that gave rise to the result that the dodecahedron has the largest separator of all fullerenes. It was fascinating to see the computer generate new scientific conjectures that captured the attention of leading researchers and led to new scientific knowledge.

Bounding the Irregularity of a Connected Graph: Proving Conjectures Generated by Computer
The program AutoGraphiX was one of the conjecture-generating programs studied by the working group. In order to understand the ability of this program to find conjectures in an automated way, help to find further conjectures in an assisted way and help to prove conjectures (or to conjecture proof strategies) the problem of the irregularity of a graph, defined by Albertson as $\Sigma_{(i,j)\in E} ||d_i - d_j||$ where d_j is the degree of vertex j was studied by Pierre Hansen and Hadrien Melot. An upper bound as a function of order n and size m for the irregularity of a connected graph was found, and it is best possible in the strong sense, i.e., attained by a graph for all n and m compatible with the existence of a connected graph. Hanson and Melot's paper "Variable Neighborhood Search for Extremal Graphs 9. Bounding the Irregularity of a Graph" will appear in the DIMACS volume resulting from the working group.

Limits on Conjecture Making in Graph Theory
The working group enabled the participants to get thoroughly acquainted with several systems: GRAPH, Graph Theorist, Graffiti, AutoGraphiX, etc. and to compare them. This led Pierre Hansen to the realization that these systems had some similar and many different functions and that a bright future

might be expected due to cross-fertilization. Reflections along those lines led him to write a paper on "How far Should, Is and Could Conjecture Making in Graph Theory be Automated." This paper will appear in the DIMACS volume resulting from the working group.

What Makes a Conjecture Interesting?
A recurrent theme at the working group meetings, and which is quite general, is 'What makes a conjecture interesting?' This appears to be hard to answer, and a preliminary question would be 'What Forms Have Interesting Conjectures in Graph Theory Taken?'. Beginning with the observation that famous (and less famous) theorems in Graph Theory were first conjectures (if only in the minds of those who found them), Mustapha Aouchiche, Gilles Caporossi, Dragan Stavanovic, and Pierre Hansen investigated this question. It appears that a large variety of forms have been considered, that no single system considers all of them and that many forms are untouched. So automated computer-assisted conjecture-making in graph theory, while already quite successful, is still pretty much at its beginning. A paper on this topic by the four authors mentioned will appear in the DIMACS volume resulting from the working group.

Minimum Energy Carbon Nanotubes
Many different sizes and shapes of nanotubes have been found experimentally and many others proposed because of their expected energetic, structural, and electronic properties. Even with the most advanced algorithms, sophisticated data structures, and parallel computing, directly determining the energetic stability of nanotubes is a quantum mechanical problem that is still beyond our reach. Nate Dean attacked the problem of determining minimum energy configurations of single-walled carbon nanotubes through the use of a nonlinear programming model. The model includes a potential energy function that is minimized subject to constraints on the angular resolution and bond lengths. His first implementation of this approach seems to consistently produce stable configurations. Because of its importance to physicists and material scientists he will continue this study.

An Approach to Molecular Electronics
Molecular electronics is an emerging field that seeks to build faster, cheaper, denser computers from nanoscale devices. The nanocell is a molecular electronics design wherein a random, self-assembled array of molecules and metallic nanoparticles is addressed by a relatively small number of input/output pins. The challenge then is to program, or train the nanocell post-fabrication. As the nanocell can be modeled as a graph (the vertices are nanoparticles and the edges are molecules), discrete algorithms and graph theoretical approaches are useful in solving many problems arising in nanocell training. Dean and his collaborators have addressed, solved and published their findings regarding several of these problems.

Windowed Correlation Problem.
Given many time series and a window size w, which pairs of w-sized subsequences from the same or different time series are highly correlated? Dennis Shasha's Courant Institute time series analysis group has developed techniques that use a mixture of Fourier, Wavelet, and random projection embedding techniques to achieve near on-line performance. The data streaming work at NYU is summarized in the book *High Performance Discovery in Time Series*, Springer Verlag 2004.

Multi-scale Burst Detection.
Given a sequence of non-negative values coming in over time and a set of pairs ($w1$, $t1$), ... (wk, tk) where wi represents a window size and ti represents a threshold, the multi-scale burst detection problem is to declare that there is a burst if for some subsequence of length wj for some j ($1 <= j <= k$) the sum of the values > tj. If the likelihood of a burst is low enough, Dennis Shasha's shifted binary tree algorithm can determine whether and where there is a burst in time linear in the size of the sequence. Linear means "independently of k".

No Projective Planes of Order Six
Reinhard Laue and Wendy Myrvold had been doing some work together on some problems regarding permutations and projective planes. They wrote some computer programs at the first working group meeting. Also, Myrvold was able to better describe the problem to Laue and he was able to teach her some group theory. They currently have a very elegant proof that there are no projective planes of order 6. It uses group theoretic concepts as opposed to the standard approaches, which argue by exhaustion that Latin squares of order 6 have no mates. They hope to extend the technique.

The (m,k)-patch boundary code problem
A simple (non-overlapping) region of the hexagonal tessellation of the plane is uniquely determined by its boundary. This seems also to be true for 'regions' that curve around and have a simple overlap. However, X. F. Guo, P. Hansen and M. L. Zheng constructed a pair of nonisomorphic (self-overlapping) regions of the hexagonal tessellation which have the same boundary. These regions overlapped themselves several times. Jack Graver proved that any region not uniquely determined by its boundary must cover some point three or more times. He is extending this work to the continuous case.

Encoding Fullerenes, Geodesic Domes, and Nanotube Caps
Jack Graver extended Coxeter's classification of the highly symmetric geodesic domes (and, by duality, the highly symmetric fullerenes) to a classification scheme for all geodesic domes and fullerenes. Each geodesic dome is characterized by its signature: a plane graph on twelve vertices with labeled angles and edges. In the case of the Coxeter geodesic domes, the plane graph is the icosahedron, all angles are labeled one, and all edges are labeled by the same pair of integers (p,q). Edges with these "Coxeter coordinates" correspond to straight line segments joining two vertices of Λ, the regular triangular tessellation of the plane, and the faces of the icosahedron are filled in with equilateral triangles from Λ whose sides have coordinates (p,q). Graver described the construction of the signature for any geodesic dome and how each geodesic dome may be reconstructed from its signature: the angle and edge labels around each face of the signature identify that face with a polygonal region of Λ and, when the faces are filled by the corresponding regions, the geodesic dome is reconstituted. The signature of a fullerene is the signature of its dual. For each fullerene, the separation of its pentagons, the numbers of its vertices, faces, and edges, and its symmetry structure are easily computed directly from its signature. Also, one can identify nanotubes by their signatures. Graver is preparing a complete catalog of nanotube caps.

Creating a Database of Graph Drawings
Reinhard Laue and Axel Kohnert are experimenting with a database of graphs, which they will continue to discuss with the working group members. A preliminary version that demonstrates some of the principles can be found under http://btm2xg.mat.uni-bayreuth.de/GRAPHDB/. The main point is to use an XML-version of the data for exchange and to access the database with popular browsers via the internet. Mathematically, they compute a canonical form to search for a graph. Their interest lies in finding stored drawings showing symmetries for a graph that stems from theoretical considerations.

IV. Project Training/Development

The section on Contributions to Education and Human Resources describes the training of Graduate and Undergraduate students through their involvement in the various phases of the grant.

V. Outreach Activities

The P.I. gave talks on the research in the project to high school teachers at the DIMACS Connect Institute (DCI) in July 2002 and July 2003. Project participant James Abello did the same in July 2002.

Wendy Myrvold, who was assigned by the Association for Women in Mathematics to mentor junior faculty member Sandra Kingan, met Sandra for the first time at the Working Group on Computer-Generated Conjectures. Sandra later returned to DIMACS to participate in the DCI Program.

Ermelinda DeLaVina, who works at a predominantly minority institution, University of Houston - Downtown, came to DIMACS for the first time for the Conjectures Working Group meeting. She and the P.I. talked about finding ways to involve more minorities in DIMACS programs and she then became involved in the DIMACS Reconnect program. That program is aimed at reconnecting teachers to research in 2- and 4-year colleges with heavy teaching loads and at "teaching" others to organize such programs. Ermelinda was an observer at one of our summer Reconnect programs.

Earlier observers from minority institutions have gone on to plan similar programs at Spelman College and Morgan State University and we hope that Ermelinda will as well.

VI. Papers/Books/Internet

Book(s) of other one-time publication(s):

P. Arabie, L. J. Hubert and J. Meulman, *The Structural Representation of Proximities with MATLAB*, under review with the Society of Industrial and Applied Mathematics.

S. Fajtlowicz, P. Fowler, P. Hansen, M. Janowitz, and F. Roberts (eds), *Graphs and Discovery*: *Proceedings of the DIMACS Workshop in Computer Generated Conjectures from Graph Theoretic and Chemical Databases*, American Mathematical Society, in preparation.

D. Shasha and Y. Zhu, *High Performance Discovery in Time Serie: Techniques and Case Studies,* Monographs in Computer Science, Springer Verlag 2004.

A special issue of the *Journal of Classification* will be dedicated to the work of the Working Group on Algorithms for Multidimensional Scaling.

Journal and Proceedings Articles from all Three Working Groups

U. Akkucuk and J. D Carroll, "PARAMAP vs. ISOMAP: A comparison of two nonlinear mapping algorithms," submitted to *Journal of Classification*.

B. Babcock, S. Babu, M. Datar and R. Motwani, "Chain: Operator scheduling for memory minimization in data stream systems," *Proc. of the ACM International Conference on Management of Data (SIGMOD)*, (2003), 253-264.

P. Bose, E. Kranakis, P. Morin and Y. Tang, "Bounds for frequency estimation of packet streams," in *Proceedings of the 10th International Colloquium on Structureal Information and Communication Complexity (SIROCCO 2003)*, (2003), 33-42.

G. Brinkmann, G. Caporossi, and P. Hansen, "A survey and new results on computer enumeration of polyhex and Fusene hydrocarbons," *Journal of Chemical Information and Computer Sciences*, **43** (2003), 842-851.

G. Brinkmann, G. Caporossi, and P. Hansen, "A constructive enumeration of Fusenes and Benzenoids," *Journal of Algorithms*, **45** (2002), 155-166.

G. Brinkmann, P.Hansen, and D. Stevanovic, “On the independence of fullerenes”, in preparation.

G. Brinkmann, P.W. Fowler and C. Justus, “Catalogue of isomerisation transformations of fullerene polyhedra,” *J. Chem. Inf. Comp. Sci.*, **43** (2003), 917-927.

M. Charikar, L. O’Callaghan and R. Panigrahy, “Better streaming algorithms for clustering problems,” *Proceedings of the Thirty-Fifth ACM Symposium in Theory of Computing*, ACM Press, New York, NY, USA, (2003), 30-39.

B. Choi, M. Mahoui and D. Wood, “On the Optimality of the Holistic Algorithms on Twig Queries,” *Proceedings of the 14th International Conference on Database and Expert Systems Applications (DEXA)*, Springer, LNCS, 2003, 28-37.

E. Cohen and M. Strauss, “Maintaining time-decaying stream aggregates,” *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2003)*, 2003, 223-233.

J. E. Corter, “Representing asymmetric proximities using directional trees,” in preparation.

D. Cvetkoviæ, P.W. Fowler, P. Rowlinson and D. Stevanovic, “Constructing fullerene graphs from their eigenvalues and angles,” *Linear Algebra and its Applications*, **356** (2002) 37-56.

N. Dean, “Mathematical programming model of bond length and angular resolution for minimum energy carbon nanotubes,” *Proceedings of the 2001 1st IEEE Conference on Nanotechnology*, (2001), 513-515.

P.W. Fowler, “Complexity, spanning trees and relative energies in fullerene isomers,” *MATCH Communications in Mathematics and in Computer Chemistry*, **48** (2003) 87-96.

P.W. Fowler, P. Hansen and D. Stevanoviæ, “A note on the smallest eigenvalue of fullerenes,” *MATCH Communications in Mathematics and in Computer Chemistry*, **48** (2003), 37-48.

P.W. Fowler, “Hückel spectra of Möbius pi systems,” *Phys. Chem. Chem. Phys.*, **4** (2002), 2878-2883.

P.W. Fowler, “Resistance distances in fullerene graphs,” *Croat. Chem. Acta*, **75** (2002), 401-408.

J. Graver, “The (m,k)-patch boundary code problem,” *MATCH-Communications in Mathematics and in Computer Chemistry*, **48** (2003), 189-196.

J. Graver, “Encoding fullerenes and geodesic domes,” *SIAM J. Discrete Math.*, **17** (2004), 596-614.

J. Graver, “A complete catalog of nanotube caps,” in preparation.

J. Graver, “An extension of the (m,k)-patch boundary code problem to the continuous case,” in preparation.

P. L. Hammer and I. E. Zverovich, “Splitoids,” *New York Graph Theory Notes*, to appear.

P. L. Hammer and I. E. Zverovich, “Construction of maximal stable sets with k-extensions,” *Combinat., Probab. and Comput.*, **13** (2004) 1-8.

S. M. Husband, C. P. Husband, N. Dean and J. M. Tour, "Mathematical details for the nanocell approach to molecular electronics," submitted to *Discrete Applied Mathematics*.

L. Hubert, P. Arabie, and J. Meulman, "Modelling dissimilarity: Generalizing untramatic and additive tree representations," *British Journal of Mathematical and Statistical Psychology*, **54** (2001), 103-123.

L.J. Hubert, P. Arabie, and J. Meulman, "Linear unidimensional scaling in the L2-norm: Basic optimization methods using MATLAB," *Journal of Classification*, **19** (2002), 303-328.

R. J. Kingan, S. R. Kingan, and Wendy Myrvold, "On matroid generation," *Proceedings of the Thirty-Fourth Southeastern International Conference on Combinatorics, Graph Theory, and Computing*, Congressus Numerantium, **164** (2003), 95-109.

A. Kumar, M. Sung, J. Xu, and J. Wang, "Data streaming algorithms for efficient and accurate estimation of flow distribution," *Proc. of ACM Sigmetrics*, 2004.

A. Kumar, J. Xu, J. Wang, O. Spatschek, and L. Li, "Space-code bloom filter for efficient per-flow traffic measurement," *Proc. of IEEE Infocom*, 2004.

M. Saks and X. Sun, "Space lower bounds for distance approximation in the data stream model," *Proceedings 34th Annual ACM Symposium on Theory of Computing (STOC)*, 2002, 360-369.

J.M.F. ten Berge, "Partial uniqueness in CANDECOMP/PARAFAC," *J. Chemometrics*, **18** (2004), 12-16.

J.M.F. ten Berge, "Simplicity and typical rank of three-way arrays, with applications to TUCKER-3 analysis with simple cores," *J. Chemometrics*, **18** (2004), 17-21.

J.M.F. ten Berge, N.D. Sidiropoulos, and R. Rocci, "Typical rank and INDSCAL dimensionality for symmetric three-way arrays of order Ix2x2 or Ix3x3," *Linear Algebra and Applications*, **388** (2004), 363-377.

J. M.F. ten Berge and H. A.L. Kiers, "A comparison of two methods for fitting the INDCLUS model," *Journal of Classification*, submitted.

I. E. Zverovich and I. I. Zverovich, "A characterization of superbipartite graphs," *New York Graph Theory Notes*, (2004), to appear.

I. E. Zverovich, "A finiteness theorem for primal extensions," *Discrete Math.*, to appear.

I. E. Zverovich, "A solution to a problem of Jacobson, Kezdy and Lehel," *Graphs Combin.*, **20** (2004).

I. E. Zverovich and V. E. Zverovich, "Basic perfect graphs and their extensions," *Discrete Math.*, to appear.

I. E. Zverovich, "Characterizations of closed classes of Boolean functions in terms of forbidden subfunctions and Post classes," *Discrete Appl. Math.*, to appear.

I. E. Zverovich and I. I. Zverovich, "Closure of $K_1 + 2K_2$-free graphs," *Austral. J. Combin.*, to appear.

I. E. Zverovich, “Generalized matrogenic graphs,” *Ann. Combin.*, to appear.

I. E. Zverovich and I. I. Zverovich, “Negative results on the stability problem within co-hereditary classes,” *J. Combin. Math. Combin. Comput.*, to appear.

I. E. Zverovich and O. I. Zverovich, “Polar SAT and related graphs,” *J. Discrete Algorithms*, to appear.

I. E. Zverovich and I. I. Zverovich, “Ramseian partitions and weighted stability in graphs,” *Bulletin Inst. Combin. Appl.*, to appear.

I. E. Zverovich and I. I. Zverovich, “Ratio of generalized stability and domination parameters,” *Austral. J. Combin.*, to appear.

I. E. Zverovich, “A characterization of domination reducible graphs,” *Graphs Combin.*, **20** (2004), 281-289.

I. E. Zverovich, “Weighted well-covered graphs and complexity questions,” *Moscow Math. J.*, **4** (2004), 521-526.

I. E. Zverovich and O. I. Zverovich, “Stability number in subclasses of P_5-free graphs,” *Appl. Math. J. Chinese Univ. Ser. B*, **19** (2004), 125-132.

I. E. Zverovich and I. I. Zverovich, “Forbidden induced subgraph characterization of cograph contractions,” *J. Graph Theory*, **46** (2004), 217-226.

I. E. Zverovich and O. I. Zverovich, “Dominant-matching graphs,” *Discuss. Math. Graph Theory*, **24** (2004), 485-490.

I. E. Zverovich and O. I. Zverovich, “Bi-induced subgraphs and stability number,” *Yugoslav J. Oper. Res.*, **14** (2004), 27-32.

I. E. Zverovich, “Stabex method for extension of α-polynomial hereditary classes,” *Europ. J. Oper. Res.*, **155** (2004), 792-795.

I. E. Zverovich and Yu. L. Orlovich, “Implementation of the reducing copath method for specific homogeneous sets,” *Vestsi Nats. Akad. Navuk Belarus Ser. Fiz.-Mat. Navuk*, **2** (2004), 48-55. (in Russian)

I. E. Zverovich, “Minimum degree algorithms for stability number,” in *Special Issue on Graph Stability and Related Topics, Discrete Appl. Math.*, **132** (2003), 211-216.

I. E. Zverovich and V. E. Zverovich, “Bipartition of graphs into subgraphs with prescribed hereditary properties,” *Graph Theory Notes of New York*, **XLIV** (2003), 22-29.

I. E. Zverovich and V. E. Zverovich, “Locally well-dominated and locally independent well-dominated graphs,” *Graphs Combin.*, **19** (2003), 279-288.

I. E. Zverovich, “Extension of hereditary classes with substitutions,” *Discrete Appl. Math.*, **128** (2003), 487-509.

I. E. Zverovich, “Perfect connected-dominant graphs,” *Discuss. Math. Graph Theory*, **23** (2003), 159-162.

I. E. Zverovich, "The domination number of (K_p, P_5)-free graphs," *Australas. J. Combin.*, **27** (2003), 95-100.

Talks

Gabriela Alexe, Sorin Alexe, Peter L. Hammer, and Bela Vizvari, "Pattern-Based Feature Selection in Genomics and Proteomics," Classification Society of North America Annual Meeting (CSNA 2003), Tallahassee, Florida, June 12-15, 2003.

Gabriela Alexe, Department of Statistical Genetics, Rockefeller University, April 22, 2003.

Gabriela Alexe, School of Natural Sciences, Institute for Advanced Study, Princeton, May 23, 2003.

Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, "Chain : Operator Scheduling for Memory Minimization in Data Stream Systems**,"** 2003 ACM SIGMOD International Conference on Management of Data, San Diego, California, USA, June 9-12, 2003.

Prosenjit Bose, Evangelos Kranakis, Pat Morin and Yihui Tang, "Bounds for Frequency Estimation of Packet Streams," 10th Colloquium on Structural Information and Communication Complexity (SIROCCO 2003), Umeå University, Sweden, June 18-20, 2003.

Moses Charikar, Liadan O'Callaghan and Rina Panigrahy, "Better streaming algorithms for clustering problems," Thirty-Fifth ACM Symposium in Theory of Computing, San Diego, California, June 9, 2003.

Byron Choi, Malika Mahoui, Derick Wood, "On the Optimality of the Holistic Algorithms on Twig Queries," 14th International Conference on Database and Expert Systems Applications (DEXA), Czech Technical University**,** Prague, Czech Republic, September 1-5, 2003.

Edith Cohen and Martin Strauss, "Maintaining time-decaying stream aggregates," *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2003)*, San Diego, California, June 9-12, 2003.

Nate Dean, "Mathematical programming model of bond length and angular resolution for minimum energy carbon nanotubes," *1st IEEE Conference on Nanotechnology*, Outrigger Wailea Resort, Maui, Hawaii, Oct. 28-30, 2001.

P. W. Fowler, "Fullerenes: a Favourable Case for Interaction between Mathematics and Chemistry," Com^{2}MaC Mini-Workshop on Two-face embeddings of graphs and applications, University of Pohang, South Korea, January, 2004.

Sandra Kingan, "Catalog of finite linear spaces," Special Session on Matroids at the 2002 AMS Sectional Meeting at Portland State University. June 20 - 22, 2002.

Sandra Kingan, "Computational Matroid Theory," Women of Applied Mathematics Research and Leadership Workshop, University of Maryland, MD. October 8 - 10, 2003.

Sandra Kingan, "Matroids from an experimental perspective," Conference on Matroid Structure Theory in honor of W. T. Tutte at Ohio State University. July 1 - 5, 2002.

Sandra R. Kingan, “On matroid generation,” Thirty-Fourth Southeastern International Conference on Combinatorics, Graph Theory, and Computing, Florida Atlantic University, Boca Raton, Florida, March 3-7, 2003

Abhishek Kumar, Min-Ho Sung, Jun (Jim) Xu (Georgia Institute of Technology), Jia Wang (AT&T Labs - Research), “Data streaming algorithms for efficient and accurate estimation of flow distribution,” Joint International Conference on Measurement and Modeling of Computer Systems ACM Sigmetrics, Columbia University, New York, June 12-16, 2004.

Abhishek Kumar (Georgia Institute of Technology), Jun Xu (Georgia Tech), Jia Wang, Oliver Spatscheck (AT&T Labs - Research), Li Li (Bell Labs, Lucent Technologies), “Space-code bloom filter for efficient per-flow traffic measurement," IEEE Infocom, Hong Kong, March 7-11, 2004.

Michael Saks and Xiaodong Sun, “Space lower bounds for distance approximation in the data stream model,” 34th Annual ACM Symposium on Theory of Computing (STOC), Montréal, Québec, Canada, May 20, 2002.

VII. Other Products

http://dimacs.rutgers.edu/SpecialYears/2001_Data/

This is the main web page for the Special Focus on Data Analysis and Mining.

Working Group Websites

http://dimacs.rutgers.edu/SpecialYears/2001_Data/Algorithms/AlgorithmsMS.htm

This is the website for the Working Group on Algorithms for Multi-Dimensional Scaling with links to their goals, meeting schedule, people, relevant literature references, software, scientific results of the working group, and other useful links.

http://dimacs.rutgers.edu/SpecialYears/2001_Data/Streaming_Web/Streaming_Home.htm

This is the website for the Working Group on Streaming Data Analysis with an introduction to this group and links to committee members, talks given at group meetings, papers, and other useful links.

http://dimacs.rutgers.edu/SpecialYears/2001_Data/Conjectures/index.html

This is the website for the Working Group on Computer-Generated Conjectures from Graph Theoretic and Chemical Databases with links to their goals, participants, meeting programs, references, selected accomplishments, and other useful links.

VIII. Contributions within Discipline

This is an inherently interdisciplinary project. Still, if the ‘discipline’ is defined to be ‘data analysis,’ there have been important developments in each of the working groups. In all of them, the interconnections among people working in data analysis but from different points of view were particularly striking. In the case of multidimensional scaling, we matched up experts in cluster analysis with researchers in combinatorial optimization, to give one example. In streaming data analysis, we matched up those working on frequency statistics with those working on probabilistic “sketches.” In the case of computer-

generated conjectures, the most important thing was to connect up developers of different methods to compare and contrast their software, with lively discussions as to the pros and cons, advantages and limitations of each method.

The meetings of the working group on Computer Generated Conjectures appear to have had a long term impact on the field. Due to the large number of collaborations resulting, there are already plans to have a third meeting on this topic at the University of Ghent, with very likely more to follow.

IX. Contributions -- other Disciplines

This is an inherently interdisciplinary project. Each of the working groups was multi-disciplinary. Some of the connections between computer science/data mining and other disciplines that have been forthcoming from this project are the following:

1). In the MDS working group, participants represented areas such as chemometrics, marketing, social networks, ecology, biomolecular databases, social and clinical psychology. Of particular note is the connection to industry, with participation from Kraft Foods, AT&T Labs, etc. Several new applications of MDS have been jump-started or significantly enhanced; image visualization, structure of molecules, MRI imaging, distance geometry, datamining.

2). In the Streaming Data Analysis and Mining working group, applications areas discussed have included financial modeling, astrophysics, and homeland security.

3). In the Computer-generated Conjectures working group, there was a strong mix of computer scientists, mathematicians, and chemists. A large number of the problems worked on dealt with chemical structures and one of the most exciting results obtained was proof of a computer-generated conjecture about the separator of fullerenes.

This project spawned a related research project on "Monitoring Message Streams" that is an outgrowth of the Streaming Data Analysis And Mining working group's work. The project, in close collaboration with the intelligence community, deals with finding "new events" from large datasets of text messages. A second and related project on Author Identification has also recently ensued.

X. Contributions -- Human Resource Development

The project has had impact on both graduate and undergraduate students. The P.I. gave talks on the research in the project to undergraduates in the DIMACS REU program in July 2002 and July 2003. James Abello served as a mentor in the REU program and also gave a lecture to the REU students. Several of the undergraduate students in our REU program worked on data mining projects closely related to and stimulated by this project. The students, affiliations, project subject, and mentors were:

Thomas Schneider, Rose-Hulman Institute of Technology, Text Categorization Using Bayesian Networks, David Madigan, Department of Statistics.

Sabyasachi Guharay, Princeton University, Text Categorization of South Asian Names, David Madigan, Department of Statistics.

Vishal Gupta, Yale University, Visibility Graphs of Simple Polygons, James Abello, DIMACS.

Adam Kirsch, Brown University Dynamic Matrix Approximations for Data-Streaming Applications, Graham Cormode, DIMACS postdoc and S. Muthu Muthukrishnan, DIMACS and Computer Science Department.

Diana Michalek**,** University of California at Berkeley, Authorship Identification, Paul B. Kantor, School of Communication, Information, and Library Studies

Several graduate students were involved in the project and ended up doing research that was closely related. Rutgers students heavily involved in the project were:

Khaled Elbassioni, Computer Science Department, worked on "Learning monotone binary functions in products of lattices and its applications in data mining" and gave a talk on the topic at the seventh International Symposium on artificial Intelligence and Mathematics, January 2-4, 2002, Fort Lauderdale, Florida.

Igor Zverovich, RUTCOR, worked on computer-generated conjectures concerning hereditary properties of graphs and produced a number of papers.

Sorin Alexe, RUTCOR, worked on feature selection in data analysis.

Gabriela Alexe, RUTCOR, worked on a consensus-type algorithm for spanned pattern generation. Gabriela also worked on cancer data.

Akshay Vashist, Computer Science, worked on structural analysis of repeat elements.

Ulas Akkucuk, Graduate School of Management, worked on nonlinear mapping of data whose topological dimensionality is lower than that of the linear vector space in which the data points are embedded.

Other graduate students were heavily involved in the working groups. Particularly important contributions were made by Craig Larson, University of Houston, Mayar Datar, Stanford, Alistair Morrison, University of Dublin, and Frank Busing, University of Leiden.

These working groups contributed to the career development of the participants by giving them the opportunity to share ideas with researchers across disciplines. This often had significant impact in their teaching also. We let the participants report the impact in their own words.

We received the following from MDS participants:

From Jos ten Berge, University of Groningen:
"The Dimacs workshop of August 2001 was very stimulating for my work on three-way algebra." It resulted in three publications. See Papers/Books/Internet. "In addition, Doug Carroll drew my attention to numerical problems of SINDCLUS (proposed by Chaturvedi and Carroll), a method for extracting binary clusters from symmetric proximity matrices. A revised paper on properties of SINDCLUS has recently been submitted to the Journal of Classification."

From James E. Corter, Teachers College, Columbia University:
"The MDS working group meeting organized in Tallahassee in June 2003 was a major impetus for me to pull together some thoughts on the graphical representation of asymmetric proximity data, and to assemble them into a talk for the meeting. The talk seemed well received, and I received encouragement

from the the Editor of the Journal of Classification. I am currently preparing a manuscript, based on this talk, for submission to JOC or a like journal."

We received the following from the Streaming Data Analysis And Mining participants:

From Jim Xu, College of Computing, Georgia Institute of Technology:
"Our (me and my Ph.D. student, Abhisehk Kumar) first exposure to data streaming is to attend the NSF DIMACS workshop on data streaming in March 2003. Since then, we have worked on various data streaming problems in networking and security, and produced two papers that appeared in top conferences. Thanks for running these great programs!"

We received the following from the Computer Generated Conjectures participants:

From Dennis Shasha, Courant Institute
A new result specifically from the DIMACS meetings is a fledgling collaboration between Dennis Shasha's Courant Institute time series analysis group (fast correlation) and Amy Braverman of the Jet Propulsion Laboratory. "The MISR satellite gives data at multiple frequency bands, multiple angles, and at 1.1 km x 1.1 km resolution over the entire earth surface. The data constitutes a time series because every data point is hit every few days. We are applying our high performance correlation techniques to the data in collaboration with Dr. Amy Braverman at JPL."

From Wendy Myrvold, University of Victoria:
"One of the main benefits for me of the meeting was the opportunity to meet with Patrick Fowler. I had been working with him by e-mail but we had never had a chance to meet in person before. He gave me a much better understanding as to how the graph theoretic algorithms relate to chemical molecules. We came up with a new idea for an independent set algorithm for fullerenes. I currently have an Honours Project student working on coding that algorithm so its performance can be tested.

I especially enjoyed the talks describing how Graffiti had been used in an education setting. That sounded like a very fun and exciting way to teach graph theory.

I enjoyed discussing with Gunnar Brinkmann the techniques he has been using for generating objects. Some of the ideas we discussed are being used in my Masters student's thesis (Hongmei Wang, Generating Small Embeddings).

I had been assigned as a mentor to Sandra Kingan by the AWM (Association for Women in Mathematics). This conference was my first (and only) opportunity to meet her in person. She is interested in generating small matroids. We discussed that problem and made substantial progress (together with R. Laue) towards formulating a potential generation algorithm.

The combined effect of all the talks at the conference was to give me a vision as to how one might put together a very powerful research tool which uses the ideas that the participants had presented and combines them into one package. I frequently use the computer to investigate conjectures, but I generally end up writing code from scratch for each application (as well as using nauty) and I have not had the visual interfaces that the participants demonstrated.

So, the conference was extremely beneficial to me. The format permitted interaction with my colleagues. Also, I enjoyed every talk which was given at the conference."

Additionally Wendy reported

"At the Montreal meeting, I started some collaborations with Dr. Patrick Fowler. Since then I have had several students start new projects, which are joint research with Dr. Fowler.

These include 3 undergraduate Honours project supervisions.

Fall 2004 (in progress):
Erin Delisle: Independent sets in the 120-cell
Daniel Horspool: Vertex Spirals of Fullerenes

Anticipated for Spring 2005:
Warren Shenkenfelder: Edge Spirals of Fullerenes

I had a Ph.D. student, Aaron Williams, start under my supervision in Fall 2004. His research topic is Independent Sets in Fullerenes and this is also joint work with Dr. Fowler.

I have proposed teaching a new graduate class: Applications of Graph Theory and Graph Algorithms to Chemistry in the 2005-06 academic year which was inspired by the collaborations I have had with Dr. Fowler.

Dr. Fowler visited me after the Montreal conference to continue our work together August 6-18. He is planning another visit to continue this work December 4-19, 2004.

The Montreal meeting was critical to starting these collaborations."

From Sandra Kingan, Penn State, Harrisburg:
"I attended the first and second meeting of the working group on computer generated conjectures from graph theoretic and chemical databases. My work is in matroids, which is a generalization of several combinatorial objects such as graphs, matrices, designs, linear spaces etc. While computers have long been used in graph theory research, it is somewhat of a new direction in matroid theory research. I got many ideas from the formal talks and the informal discussions. I met several people with whom I hope to collaborate in future. These meetings and the interactions with graph researchers helped me a great deal in my research. I was also able to direct a Masters level student project on Visualization of Linear Spaces.

At the first meeting we had just a prototype version of the matroid software system Oid. Since then we released a completed version. It can be downloaded from the web at http://cs.hbg.psu.edu/~srk1/matroids/software.html. A paper on Oid and its software design called 'A software system for matroids' is scheduled to appear in the Proceedings of the first workshop.

I met Wendy Myrvold, who had been assigned by AWM as my mentor, at the first meeting. Since then we kept in touch and published a paper together: 'On matroid generation'. We are working on another paper."

From Jack Graver, Syracuse University:
"It is a pleasure to write a few words about the DIMACS meeting last November on Computer Generated Conjectures and its effect on my research. Having just moved into a new area of research (using graph theory to model carbon atoms), the interactions with chemists and computer scientists working in that area was extremely valuable. I am not a chemist nor computer scientist and my research is purely theoretical; but its foundation is the body of their results, many of which were computer generated. As a direct result of my interactions with this group, I have submitted four papers and have made significant progress on two others."

From Craig Larson, University of Houston (graduate student):
"The working group meetings were very helpful. I had never met a number of the participants before, many of whom I have since corresponded with by email regarding various questions, including Myrvold, Lisken and Brinkmann. I have a much better understanding now of what Hansen and Caporossi are doing with their program -- which was useful when I was penning my conference paper.

A paper I am just finishing up on a connection between graph-theoretic independence and the stability of fullerenes grew directly out of comments of Fajtlowicz and Fowler at the conference (I expect to be talking about this at NanoSpace 2003). Most everything I've been working on has some connection to the program at DIMACS and I expect collaborations with the participants in the near future."

From Dragan Stevanovic, University of Nis, Yugoslavia:
"New research collaborations (with Gilles Caporossi) and proofs for (a) couple of Graffiti's conjectures happened already during the conference and will appear in standard proceedings, as the paper 'On Separators of Fullerenes'.

During that meeting I also made fruitful contacts with Pierre Hansen, Patrick Fowler and Gunnar Brinkmann, and during this year we wrote a paper 'On the smallest eigenvalue of fullerenes' (with Pierre Hansen and Patrick Fowler, accepted for MATCH) and a preprint 'On the independence of fullerenes' (with Pierre Hansen and Gunnar Brinkmann, still in preparation).

Besides that, DIMACS meeting also served as a showroom of software for helping research in graph theory. That enabled me to take all their good and bad features into account when specifying a new version of Dragos Cvetkovic's system GRAPH (which is a project financed by Serbian Ministry for Science, Technology and Development)."

From Jonathan Berry, Lafayette:
"The workshop inspired me to formally write my thoughts up on the future of software systems for discrete mathematics. This paper was submitted to the DIMACS volume based on the working group meeting."

From Gabriela Alexe, Rutgers University (Graduate Student):
"My research activity is focused on Data Mining/Bioinformatics area, where I concentrate on the problem of reducing the size of genomic or proteomic datasets by selecting an optimal subset of attributes which collectively are capable of describing the biological phenomenon with high accuracy. The main directions of my research are the following:

i. Detection of Relevant Sets of Attributes. We have shown in preliminary studies that the irredundant sets of attributes which offer the most accurate analyses of a dataset, do not necessarily consist of attributes having the highest correlations with the outcome. There are no currently known methods to predict the relative accuracies of models built on various irredundant subsets of attributes, and to systematically search for the optimal sets. I am currently working in developing such evaluation methods, and in incorporating them into an optimal support set detection method.

ii. Synthetic Risk Factors in Epidemiology. Space transformation/reduction methods for developing compact representations of observations in terms of new, `synthetic' variables are known in statistics (e.g., principal component analysis), machine learning (e.g., support vector machines), the logical analysis of data (logical patterns), etc. It was noticed in preliminary studies that the collection of logical patterns which can be extracted from a dataset can provide a new, improved way of representing data; in this representation, each observation is represented as a 0-1 vector, which specifies the particular patterns

covering the given observation. In particular, this representation seems advantageous in clustering studies, and has the potential of leading to the detection of new classes of observations. I am studying the possibility of using space transformations for the representation of observations in a dataset for (i) the compression of datasets, and (ii) an enhanced clustering of observations, with a view to the discovery of new classes.

I have presented some of the recent results obtained in collaboration with DIMACS members and with medical doctors from the Cancer Institute of New Jersey and the National Institutes of Health, to the Annual Meeting of the Classification Society of North America (organized in connection with the Second Workshop of the DIMACS Working Group in Multidimensional Scaling, June 11-15, 2003, Florida, US), as well as in scientific seminars organized by the Department of Statistical Genetics, Rockefeller University (April 22, 2003), and by the School of Natural Sciences, Institute for Advanced Study, Princeton (May 23, 2003). I intend to defend my thesis in Fall 2003 and to continue my research in finding collective biomarkers for breast and prostate cancer as a postdoctoral fellow in the Program of Theoretical Biology at the Institute for Advanced Study, Princeton. I am very grateful to DIMACS for the partial support provided for my research; the support was very beneficial, especially for establishing collaboration relations with DIMACS members and with experts in biology from other institutions."

From Robert Cowen, Dept. of Math, Queens College, CUNY:
I attended the Dimacs/Gerad conference in Montreal in June (2004) and found it very stimulating! After hearing some of the talks and talking to several of the participants, I became interested again in research I had done about 12 years ago on automated theorem proving with a student of mine (published in the Notre Dame J.Formal Logic). I have since been working on this again, with two new students, one undergraduate and one graduate student. I am using data provided by the TPTP project at the U. of Miami and in fact Geoff Sutliffe has been very kind and has written a translation program so I can use the data with my Mathematca programs. Also I have applied for a PSC/CUNY grant to continue working on this project. I am certain that I would not have gone back to this work if it weren't for this conference.

In addition, I found presenting my paper on using Mathematica to do research in Graph Theory stimulating and talking to people about using the Mathematica programming language to motivate students to do research was also very worthwhile and was pleased that people seemed interested in these endeavors.

Finally I induced Dragan Stevanovic to develop his graph theory exploration program for the Mac OS and have recently received a copy. We will use this with our graph theory students in our Advanced Math Lab. It should be quite helpful.

All in all a wonderful conference!"

From Gunnar Brinkmann, Applied Mathematics and Computer Science, Ghent University:
"For me it definitely had a big impact. Right next month the followup conference will be and maybe after that the impact will be even bigger.

But already the Dimacs conference resulted in some cooperation with e.g. Ermelinda delaVina and Siemion Fajtlowicz, which lead to a grant application which is currently processed here in Belgium and also to courses for students and a workshop for high school teachers with Ermelindas program that I would not have given otherwise. I also met Dragan Cvetkovicz there with whom (and Pierre Hansen, whom I knew before) I now have a pending paper that will hopefully also be worked on a bit next month in Montreal...).

… in short I can say that it was a very important conference for me."

XI. Contributions to Resources for Research and Education

A unique research skill that participants went away with was the use of computers to generate conjectures. Especially useful is a website to related software that is included in the web pages of the Computer-generated Conjectures Working Group. See:

http://dimacs.rutgers.edu/SpecialYears/2001_Data/Conjectures/GC_Discovery/Resources.htm

Also worth mentioning is the introduction given to people working on applications in chemometrics, marketing, psychology, etc. of new applications for MDS.

Astrophysicists and financial modelers were introduced to the latest algorithmic methods for data analysis.

Online Software

http://dimacs.rutgers.edu/SpecialYears/2001_Data/Algorithms/NewSoftware.htm

This website contains links to many software programs, including systems for multivariate data visualization /multidimensional scaling and graph layout in any dimension, Clustering software (source + executables) and documentation, program to fit extended trees to proximity data, among others.

http://dimacs.rutgers.edu/SpecialYears/2001_Data/Conjectures/GC_Discovery/

Graph is an expert system for the classification and extension of knowledge in the field of graph theory.

Other Resources

http://dimacs.rutgers.edu/SpecialYears/2001_Data/Conjectures/GC_Discovery/

This web page contains links to pages describing open problems in graph theory and discrete mathematics.

XII. Contributions Beyond Science and Engineering

Our interest in data analysis and mining in this project spawned two related research projects on "Monitoring Message Streams" and "Author Identification" that are an outgrowth of the Steaming Data Analysis and Mining working group's work. The projects, worked out in close collaboration with the intelligence community, deal with finding "new events" from large datasets of text messages and determining which of a given set of authors created a document. These large interdisciplinary, inter-institutional projects are funded by the intelligence community through NSF. They have applications to practical problems of homeland security.