Working Groups in Data Analysis and Mining

Fred S. Roberts, PI
The Center for Discrete Mathematics and Theoretical Computer Science (DIMACS)
Rutgers University

Contact Information:

Fred S. Roberts, Director of DIMACS
DIMACS Center
Rutgers University
96 Frelinghuysen Road
Piscataway, NJ 08854-8018
Email: froberts@dimacs.rutgers.edu

WWW Page:

http://dimacs.rutgers.edu/About/Reports/DataAnalysis.html

Project Award Information:

Award Number IIS-0100921
Duration: 9/15/01 - 8/31/03
Title: Working Groups in Data Analysis and Mining

Keywords:

Massive data sets, data mining, multidimensional scaling, computer conjectures, streaming data, fullerenes.

Project Summary:

The project is supporting three working groups of researchers who are addressing different aspects of problems involving the analysis of massive data sets. The groups are concerned with "streaming" data analysis and mining, multidimensional scaling and the generation of scientific conjectures by computers. Each working group has met once, and has generated conjectures and results in its area of concern. The idea is that there will be a second meeting in roughly a years time at which results will be presented and further ideas explored. These working group meetings are informal with plenty of time set aside for interactions among the participants. In addition to this, the DIMACS program in general, and these working groups in particular, tend to attract high quality research visitors who produce important results related to the meetings.

Publications and Products:

A. Computer Generated Conjectures. The first meeting of this group was a very lively meeting in which several open questions were solved -- an impressive accomplishment! Researchers presented a number of different approaches to the generation of computer conjectures, and were able to compare their results and interact with each other. Indeed, researchers from Montreal, Houston, and Bielefeld, Germany demonstrated independently designed software. They are now cooperating with each other and using the best features of each of their software projects. The working group web site can be found at

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

It contains a summary of the meeting, open questions, links to various types of software, and descriptions of the meeting.

A good part of the meeting was devoted to the study of fullerenes. These are planar 3-regular graphs having either pentagons or hexagons for their faces. From the chemical point of view they are molecular graphs of carbon isomers.

Here are some examples of the exciting discoveries that were made in and shortly after this meeting.

At the end of Pierre Hansen's talk on Computers in Graph Theory a member of the audience asked about the maximum number of K4's in a graph with a given number of K3's. Gilles Caporossi used the AutoGraphiX (AGX) system to find extremal graphs for this criterion on the spot. The values found turned out to be equal to the best known ones and larger examples than previously known were found, so the values were conjectured to be optimal.

In his talks at the workshop preceding the working group meeting, which we called "Graph Theory Day," and at the working group meeting itself, Siemion Fajtlowicz mentioned the conjecture 895 of Graffiti: The separator of a fullerene is at most 1. Dragan Stevanovic proved this was true during the workshop, using Cauchy's Interlacing Theorem and the interactive component of System Graph; while he was presenting his proof, Gilles Caporossi simplified it using the interactive component of AGX. Both of them joined forces to write the paper [1], in which they prove that the dodecahedron has the largest separator among all fullerenes. The same proof technique was used later by Patrick Fowler, Pierre Hansen and Dragan Stevanovic in [2] to show, among other results, that buckminsterfullerene has the maximum smallest eigenvalue in the class of IPR-fullerenes.

Reference [3] gives an upper bound of an eigenvalue of NEPS as an eigenvalue of its component, and gives a new sufficient condition for the almost cospectrality of components of NEPS of connected bipartite graphs. NEPS is an acronym for "non-complete, extended p-sum", a graph-theoretic construction that is explicitly defined in the report.

In [4], integral graphs are characterized in the family of complete split graphs, and in [5] antipodal graphs of diameter 3 are characterized and it is shown that almost every graph is an induced subgraph of such a graph. A proper metric space is called antipodal if for every x in X there is a y such that X is the closed interval [x,y].

[1] Stevanovic,D., Caporossi,G. "On the separator of fullerenes." Les Cahiers du GERAD G-2001-55, and DIMACS TR 2001-48.
[2] Fowler,P.W., Hansen,P., Stevanovic,D. "A note on the smallest eigenvalue of fullerenes." Les Cahiers du GERAD G-2002 (to appear).
[3] Stevanovic, D., On the components of NEPS of connected bipartite graphs, DIMACS TR 2001-45
[4] Hansen, P., Melot, H., Stevanovic, D., Integral complete split graphs, DIMACS TR 2001-47.
[5] Stevanovic, D., Antipodal graphs of small diameter, DIMACS TR 2001-46.

B. 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.

Apart from the valuable brainstorming sessions held within the first meeting, new and continuing research relationships were begun. An indication of this can be found on the web site:

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

Here are two samples of research in progress.

1) 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.

2) 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. Lower bounds are produced on the running time required and on the sizes of the samples that must be clustered.

C. Multidimensional Scaling (MDS): 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 are being explored for extending such approaches to other, more complex MDS models and methods. The web site associated with this working group is at:

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

The main accomplishment of the first meeting of this group was the development and enhancement of cross-disciplinary research efforts. Here are the highlights of these endeavors.

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) are all exploring various mathematical programming techniques to fit MDS models, including various possible collaborative efforts.

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

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.