DIMACS Center
Rutgers University

Three Special Focus Programs

Annual Report

October 2004

Ia. Participants from the program

Participants:

Computational Information Theory and Coding Special Focus Organizers:

Robert Calderbank, Chair, AT&T Labs
Chris Rose, Rutgers University
Amin Shokrollahi, Digital Fountain
Emina Soljanin, Bell Labs
Sergio Verdu, Princeton University

Workshop Organizers:

Alexei Ashikhmin, Bell Labs
Alexander Barg, Bell Labs
Adam Buchsbaum, AT&T Labs - Research
Iwan Duursma, University of Illinois
Lance Fortnow, NEC Research
Jerry Foschini, Bell Labs
Michael Fredman, Rutgers University
Chris Fuchs, Bell Labs
Vivek Goyal, Digital Fountain
Piyush Gupta, Bell Labs
Mark Hansen, Bell Labs
Janos Komlos, Rutgers University
Jelena Kovacevic, Bell Labs
Gerhard Kramer, Bell Labs
Joel Lebowitz, Rutgers University
S. Muthukrishnan, AT&T Labs Research and Rutgers University
Suleyman Cenk Sahinalp, Case Western University
Amin Shokrollahi, Digital Fountain
Peter Shor, AT&T Labs - Research
Paul Siegel, Univ. of California
Emina Soljanin, Bell Labs
Dan Spielman, MIT
Jim Storer, Brandeis University
Ruediger Urbanke, EPFL
Sergio Verdu, Princeton University
Paul Vitanyi, University of Amsterdam
Jeff Vitter, Duke University
Adriaan van Wijngaarden, Bell Labs
Bane Vasic, University of Arizona
David P. Williamson, IBM Almaden
Bin Yu, University of California, Berkeley

Computational Information Theory and Coding Special Focus Visitors:

Christina Fragouli, University of Athens, 2/1/03-6/13/03 and 10/12/03-10/19/03
Gregory Kabatianski, Institute for Information Transmission Problems, 5/20/02-6/4/02
Viacheslav Prelov, Russian Academy of Sciences, 10/1/01-1/31/02
Eran Sharon, Tel Aviv University, 4/5/04-4/15/04
Gilles Zemor, E'Cole National Supérieure des Télécommunications, 8/18/01-8/31/01 and 8/18/02-8/31/02

Computational Information Theory and Coding Special Focus Postdoc:

Xuerong Yong, Hong Kong University of Science and Technology, 2002 - 2003

Computational Geometry and Applications Special Focus Organizers:

Steven Fortune, Chair, Bell Labs
Bernard Chazelle, Princeton
Bill Steiger, Rutgers University

Computational Geometry and Applications Special Focus Advisory Committee:

Pankaj Agarwal, Duke
Ken Clarkson, Bell Labs
David Dobkin, Princeton
Chris Hoffmann, Purdue
Joe Mitchell, SUNY Stony Brook
Joe O'Rourke, Smith
Janos Pach, CUNY
Ricky Pollack, NYU
Jarek Rossignac, Georgia Tech
Jack Snoeyink, UNC

Workshop Organizers:

Pankaj Agarwal, Duke University
Nina Amenta, University of California – Davis
Fausto Bernardini, IBM - T. J. Watson Research Center
Herve Bronnimann, Polytechnic University
Danny Chen, University of Notre Dame
Scott Drysdale, Dartmouth College
Debasish (Deba) Dutta, University of Michigan
Steve Fortune, Bell Labs
Leonidas Guibas, Stanford University
Ravi Janardan, University of Minnesota
Jean-Claude Latombe, Stanford University
Shelly Leibowitz, Wheaton College
Regina Liu, Rutgers University
Joe Mitchell, SUNY Stony Brook
Janos Pach, New York University, Courant Institute of Mathematical Sciences
Robert Serfling, University of Texas at Dallas
Diane Souvaine, Tufts University
Michiel Smid, Carleton University
Yehuda Vardi, Rutgers

Computational Geometry and Applications Special Focus Visitors:

Pankaj Agarwal, Duke University, 5/18/03-5/25/03
Sergio Cabello Justo, Universiteit Utrecht, 11/10/02-11/23/02
Siu-Wing Cheng, Hong Kong University of Science and Technology, 11/14/02-11/28/02
Bhaskar Das Gupta, University of Illinois at Chicago, 8/10/02-8/17/02 and 11/10/02-11/17/02
Ioannis Emiris, University of Athens, 11/30/02-12/7/02
Raffaele Esposito, Universita di L'Aquila, 2/25/01-3/25/01
Marina Gavrilova, University of Calgary, 12/2/02-12/8/02
Subir Kumar Ghosh, Tata Institute of Fundamental Research, 9/16/02-10/6/02
Ravi Janardan, University of Minnesota, 4/2/03-4/4/03 and 10/7/03-10/9/03
Vladlen Koltun, University of California at Berkeley, 9/29/02-10/5/02
Pino Persiano, Universitá di Salerno, 7/15/01-8/31/01

Val Pinciu, Southern Connecticut State University, 8/29/04-9/25/04
Sylvain Pion, Max Planck Institut fur Informatik, 12/2/02-12/6/02
Bruce Reed, McGill University, 4/3/03-5/3/03
Guenter Rote, Institut fur Informatik, 9/16-10/2/02
Micha Sharir, Tel Aviv University, 5/17/03-5/24/03
Alexander Soifer, University of Colorado at Colorado Springs, 1/1/03-8/31/04
Ileana Streinu, Smith College, 9/30/02-10/4/02
Roberto Tamassia, Brown University, 12/5/02-12/6/02
Monique Teillaud, INRIA - Sophia Antipolis, 12/2/02-12/6/02
Emo Welzl, ETH Zurich, 5/18/03-5/25/03
Ming Zhang, University of Texas, 3/30/03-4/12/03

Next Generation Networks Technologies and Applications Special Focus Organizers:

B. R. Badrinath, Rutgers University
Sandeep Bhatt, Akamai Technologies
Joan Feigenbaum, co-chair, Yale University
Muthukrishnan, co-chair, Rutgers University and AT&T Labs

Workshop Organizers:

Micah Adler, University of Massachusetts
Rajeev Agrawal, Motorola
Nina Amenta
Venkat Anantharam, University of California, Berkeley
Matthew Andrews, Bell Labs
B.R. Badrinath, Rutgers University
Michael Berry, University of Tennessee
Endre Boros, Rutgers University
Tom Buckman, MITRE Corporation
Adam Buchsbaum, AT&T Labs – Research
John Byers, Boston University
Ran Canetti, IBM Watson Research Center
John Doyle, California Institute of Technology
Funda Ergun, Case Western Reserve University
Deborah Estrin, UCLA
Joan Feigenbaum, Yale University
Rudolf Fleischer, HKUST
M. Franklin, UC Berkeley
Warren Greiff, Mitre Corporation
Piyush Gupta, Lucent Technologies
D. Frank Hsu, Fordham University
David Hull, WhizBang
Paul Kantor, Rutgers University
P.R. Kumar, University of Illinois, Urbana
Liz Liddy, Syracuse University
Lyle McGeoch, Amherst College
Ueli Maurer, ETH Zurich
Rafail Ostrovsky, Telcordia Technologies
Debasis Mitra, Lucent Technologies
Michael Mitzenmacher, Harvard University
S. Muthukrishnan, Research and Rutgers University
Noam Nisan, Hebrew University
Andrew Odlyzko, AT&T
Rajmohan Rajaraman, Northeastern University
Danny Raz, Technion and Bell Labs

Fred Roberts, Rutgers University
S. Cenk Sahinalp, Case Western Reserve University
Yuval Shavitt, Tel-Aviv University and Bell Labs
Jack Snoeyink, UNC, Chapel Hill
Aravind Srinivasan, Bell Labs
Divesh Srivastava, AT&T Labs - Research
Matt Stallman, NC State University
Jim Storer, Brandeis University
David Tse, University of California, Berkeley
Vijay Vazirani, Georgia Tech
Jeff Vitter, Duke University
Dorothea Wagner, University of Konstanz
Ruth Williams, University of California, San Diego
Walter Willinger, AT&T
Roy Yates, Rutgers University
Bulent Yener, Bell Labs

Next Generation Networks Technologies and Applications Visitors:

James Abello, AT&T Labs Research, 6/15/02-8/31/02, 6/15/03-9/15/03 and 1/1/04-9/30/04
Ziv Bar-Yossef, University of California at Berkeley, 5/13-5/17/02
Graham Cormode, University of Warwick, 9/23/00-9/29/00, 3/28/01-4/12/01, 7/28/01-8/10/01, 6/17/02-6/28/02
Bhaskar Das Gupta, University of Illinois at Chicago, 8/10/02-8/17/02 and 11/10/02-11/17/02
Mayur Datar, Stanford University, 9/13/01-9/26/01
Erik Demaine, University of Waterloo, 1/10/01-1/19/01
Camil Demetrescu, University of Rome "La Sapienza", 12/1/00-2/1/01
Paolo Ferragina, Universitá di Pisa and the Max-Planck Institute fur Informatik, 1/10/01-1/15/01
Lixin Gao, Smith College, 11/12/01-11/18/01
Stuart Haber, Intertrust - Star Lab, 6/17/02-9/15/02
Mohamed Haouari, Ecole Polytechnique de Tunisie, 10/22/01-11/4/01
Frank Hsu, Fordham University, 2/28/01-5/31/02
Nicole Immorlica, MIT, 8/25/03-9/5/03
Olga Kapitskaia, Pôle Universitaire Leónard de Vinci, 3/8/01-3/22/01
Jonathan Katz, Columbia University, 3/11/02-5/3/02
Malwina Joanna Luczak, Wadham College, 8/20/01-9/14/01
Filippo Menczer, University of Iowa, 2/13/02-2/15/02
Alvaro Monge, California State University at Long Beach, 2/13/02-2/15/02
Vincent Mousseau, University of Paris-Dauphine, 2/1/01-8/16/01
Jayesh Pandey, Purdue University, 7/7/03-8/15/03
Gopal Pandurangan, Brown University, 6/17/02-8/1/02
Benny Pinkas, Intertrust - Star Labs, 5/29/02-8/23/02
Bartosz Przydatek, Carnegie Mellon University, 6/16/03-8/15/03
Suleyman Cenk Sahinalp, University of Warwick, 1/5/03-1/11/03
Dana Shapira, Brandeis University, 12/2/02-12/13/02
Nitin Thaper, MIT, 7/23/01-8/17/01
Andrea Vitaletti, University of Rome "La Sapienza", 11/27/00-12/20/00
Jian Zhang, Yale University, 12/2/02-12/13/02

Next Generation Networks Technologies and Applications Special Focus Postdocs:

Nicolas Schabanel, Ecole Normale Superieure de Lyon, 2000-2001
Mahesh Viswanathan, University of Pennsylvania, 2000-2002

Other Visitors (Cutting Across Special Foci or in Related Areas)

Boris Aronov, Polytechnic University of Brooklyn, 1/13/04-2/12/04
David Assaf, The Hebrew University, 9/1/02-10/5/02
Tullio Ceccherini-Silberstein, Universita degli Studi del Sannio di Benevento, 2/3/03-2/9/03
Han Hyuk Cho, Seoul National University, 10/10/03-10/15/03
Yves Crama, Ecole d'Administration des Affaires, Universite de Liege, 7/17/02-7/31/02
Tanka Nath Dhamala, Tribhuvan University, 6/4/03-6/30/03
Matthias Fitzi, University of California-Davis, 11/4/02-11/15/02
Stephan Foldes, Tampere University of Technology, 7/31/02-8/8/02 and 8/21/02-8/28/02
Fanica Gavril, CEMA, 9/1/03-6/17/04
Alexander Gnedin, Mathematical Institute, 3/1/03-3/31/03
Anupam Gupta, Carnegie Mellon University, 7/16-7/17/04
Anupam Gupta, Bell Labs, 10/1/02-12/15/02
Vladimir Gurvich, Russian Academy of Science, 8/1/02-8/31/02 and 8/1/03-8/31/03
Jin Jinji, Kyung Hee University, 10/10/03-10/15/03
Ilya Kapovich, University of Illinois at Urbana-Champaign, 2/22/02-2/24/02
Claire Kenyon, Universite de Paris-Sud, 3/5/02-3/16/02
Suh-Ryung Kim, Kyung Hee University, 10/10/03-10/15/03
Brenda Latka, Lafayette College, 9/1/01-8/31/02
Angus MacIntyre, University of Edinburgh, 3/17/03-4/15/03
Kazuhisa Makino, Osaka University, 9/22/04 - 10/14/04
Benjamin Martin, Hebrew University of Jerusalem, 2/11/02-2/18/02 and 2/16/03-2/22/03
Boris Mirkin, Birkbeck College, 10/24/03-10/31/03
Vahab Mirrokni, MIT, 8/10/03-8/22/03
Seffi Naor, Technion, 8/18/03-9/11/03
Rafail Ostrovsky, University of California - Los Angeles, 6/28/04 - 7/9/04
Dmitry Pasechnik, J.W. Goethe-Universitaet, 10/15/02-10/19/02
Dieter Rautenbach, Lehrstuhl II für Mathematik, 10/1/02-11/30/02
Tim Roughgarden, Cornell University, 7/28/02-8/3/02
Rahul Santhanam, University of Chicago, 6/29/03-7/25/03
Serap Savari, Bell Labs, 8/1/03-10/31/03
Warren Smith, NEC Research Institute, 7/1/02-9/30/02
Ronald Solomon, Ohio State University, 9/1/03-9/21/03
Vera Sos, Mathematical Institute of Hungary, 11/9/03-11/14/03
Venkatesh Srinivasan, Max Planck Institute, 3/15/03-4/15/03
Joel Tropp, University of Texas, 6/3/02-8/9/02
Alexis Tsoukias, Universite Paris Dauphine, 3/15/03-6/3/03
Rebecca Wright, AT&T Labs - Research, 3/18/02-8/15/02
Ke Yang, Carnegie Mellon University, 8/19/02-8/30/02
Jing Zou, Princeton University, 6/24/02-8/30/02
Ziv Bar-Yossef, University of California at Berkeley, 5/13-5/17/02
Meeyoung Cha, Advanced Networking Lab, 8/2/04-8/27/04
Guido Consonni, Universitá di Pavia, 1/13/02-2/9/02
Yves Crama, Ecole d'Administration des Affaires, Universite de Liege, 2/20/01-3/6/01
Zsolt Csizmadia, Eotvos Lorand University, 6/22/04-9/3/04
Nando de Freitas, University of British Columbia, 4/9/02-4/13/02
Vladimir Gurvich, Russian Academy of Science, 7/1/01-8/31/01
Panagiotis Ipeirotis, Columbia University, 2/1/02-2/28/02
Martin Anthony, London School of Economics, 3/18/02-3/31/02
Mattias Kreuseler, University of Rostock, 11/8/01-11/30/01
Vadim Mottl', Tula State University, 1/16/02-3/2/02
Alberto Roverato, Universitá di Economia Politica, 1/13/02-2/9/02
Paul Sant, Kings College, London, 8/6/03-8/20/03
Iryna Skrypnyk, University of Jyvaskyla, 3/1/03-4/30/03 and 7/3/04-5/3/05
Bela Vizvari, Eotvos Lorand University, 2/6/04-3/7/04 and 6/22/04-9/3/04
Wei Wang, University of North Carolina, 7/28/03-8/1/03

Mutsonari Yagiura, Kyoto University, 11/26/01-12/2/01

Monitoring Message Streams Participants

Endre Boros, Rutgers, RUTCOR
Wen-Hua Ju, Avaya Labs
Paul Kantor, Rutgers, SCILS
David Lewis, Consultant
Michael Littman, Rutgers, Computer Science
David Madigan, Rutgers, Statistics
Ilya Muchnik, DIMACS
Muthu Muthukrishnan, Rutgers, Computer Science
Rafail Ostrovsky, Telcordia
Fred Roberts, DIMACS and Rutgers Mathematics
Martin Strauss, AT&T Labs

Andrei Anghelescu, Rutgers, Computer Science, graduate student
Dmitriy Fradkin, Rutgers, Computer Science, graduate student
Aynur Dayanik, Rutgers, Computer Science, graduate student
Suhrid Balakrishnan, Rutgers, Computer Science, graduate student

Alex Genkin, programmer
Vladimir Menkov, programmer

Author Identification: Identifying Real-Life Authors in Massive Document Collections Participants

Paul Kantor, Rutgers, SCILS
David Lewis, Consultant
David Madigan, Rutgers, Statistics
Fred Roberts, DIMACS and Rutgers Mathematics
Alex Genkin, DIMACS
Li Ye, Rutgers, Statistics
Diana Michalak, UC Berkeley, REU student
Ross Sowell, University of the South, REU student

Knowledge Discovery: Mining Multilingual Resources Using Text Analytics

Salim Roukos, IBM T. J. Watson Research Center
Scott Fahlman, Language Technologies Institute and Computer Science, Carnegie Mellon University

Consultants

Alexander Barg, Bell Labs
Mel Janowitz, DIMACS
Kim Factor, Marquette University
David Lewis
Vladimir Menkov

Ib. Participating Organizations

Telcordia Technologies: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops. Subcontractor on Monitoring Message Streams project.

AT&T Labs - Research: Facilities; Personnel Exchanges

Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

NEC Laboratories America: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

Lucent Technologies, Bell Labs: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

Princeton University: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

Rutgers University: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

Avaya Labs: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

HP Labs: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

IBM Research: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops. Subcontractor on Mining Multilingual Resources Using Text Analytics project.

Microsoft Research: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

Office of Naval Research: Facilities; Personnel Exchanges
ONR partially funded one of the workshops.

The New Jersey Commission on Science and Technology
Funded all three foci.

The National Security Agency
NSA partially funded two of the foci, NGN and Comp. Geom.

1c. Other Collaborators

The project involved scientists from numerous institutions in numerous countries. There were hundreds of attendees at our workshops, coming from a variety of types of institutions and disciplines. The resulting collaborations also involved individuals from many institutions in many countries.

II. Project Activities

Computational Information Theory and Coding Special Focus Project Activities

Theme of the Program:

The main purpose of coding theory is the reliable transmission or storage of data. It is common to think about information theory as a science concerned with theoretical limits on rates at which reliable transmission or storage of data is possible. During the course of the 50-year life of information theory, these limits (known as channel capacities) have been computed for various important telecommunications channels based on their underlying physics, which was considered given. Computational Information Theory is concerned with techniques (such as channel coding) for achieving channel capacities. Computational information theory has achieved dramatic scientific breakthroughs in recent years, and codes that come close to theoretical limits have been discovered.

This special focus explored the interconnections among coding theory, theoretical computer science, information theory, and related areas of computer science and mathematics, and addressed some of the challenges for the next 50 years - in particular wireless communication, magnetic/optical storage, the role of signal processing in networks, network information theory, and the creation of quantum information theory.

Today, the development of coding and information theory is closely related to the explosion of information technology, with applications to the Internet and the next generation of networks technologies. The rapid development of a myriad of networked devices for computing and telecommunications presents challenging and exciting new issues for coding and information theory.

In 1948 Shannon developed fundamental limits on the efficiency of communication over noisy channels. The coding theorem asserts that there are block codes with code rates arbitrarily close to channel capacity and probabilities of error arbitrarily close to zero. Fifty years later codes for the Gaussian channel have been discovered that come close to these fundamental limits. Today, theoretical limits on rates at which reliable transmission or storage of data is possible are known for many telecommunications challenges. The same types of challenges face us in today's and tomorrow's new eras of networked, distributed computing and communications.

Coding theory has long used deep mathematical methods. We explored the use of such mathematical ideas from linear systems theory, automata theory, algebraic geometry, etc. in understanding the issues at the interface among coding theory, information theory, and theoretical computer science.

There are many connections between coding theory and theoretical computer science. We explored the connections between coding theory and: upper bounds on the number of random bits used by probabilistic algorithms and the communication complexity of cryptographic protocols; lower bounds on the complexity of approximating numerous combinatorial optimization functions; characterizations of traditional complexity classes such as NP and PSPACE; and the emerging theory of program testing.

Connections between information theory/coding and other parts of science, in particular physics, go back to the need to understand the relations between Shannon's perspective on entropy and the laws of thermodynamics. Today, attention has turned to quantum mechanics and the processing of intact quantum information states, and we will explored such connections.

Coding theory is essential in numerous applied areas such as deep space communication, the theory of wireless channels, and optical/magnetic recording. Such applied areas are posing new and challenging problems for coding theorists and were a central focus of this special focus.

The development of the ``next generation'' of networks technologies leads to many new challenges for coding and information theory. Thinking of the network as a channel should help us to understand these challenges.

This special focus began in summer 2001 and will be ending in 2005.

Workshops, Working Groups, Tutorials:

Course: A Crash Course on Coding Theory: Madhu Sudan, MIT
November 6 - 10, 2000
Location: IBM Almaden, San Jose, California
Organizer: David P. Williamson, IBM Almaden
Co-sponsored by IBM Almaden and DIMACS.

Attendance: 56

This course, a “pre-special focus” event, was a fast-paced introduction to the theory of error-correcting codes aimed at computer scientists. This theory, dating back to the works of Shannon and Hamming from the late 40's, overflows with theorems, techniques, and notions of interest to the computer scientist. The course focused on results of asymptotic or algorithmic significance. Topics included:

1. Introduction to the theory of error-correcting codes;
2. Construction and existence results for ECCs;
3. Limitations on the combinatorial performance of ECCs;
4. Decoding algorithms - Part 1. The decoding problem: Unambigous and list-decoding versions;
5. Decoding algorithms - Part 2. The linear time decoding problem: worst-case and random error settings;
6. Complexity results;
7. Applications in computer science.

Computational Information Theory and Coding "Kickoff"
October 26, 2001
Location: Arnold Auditorium, Bell Labs, Murray Hill, NJ
Organizers: Emina Soljanin and Adriaan van Wijngaarden, Bell Labs
Attendance: 120

DIMACS "kicked off" its Special Focus on Computational Information Theory and Coding with a half day conference at Bell Labs in Murray Hill, New Jersey. The program included three special talks:

- "Combinatorics, Quantum Computing and Cellular Phones" by Dr. Robert Calderbank, AT&T Research Laboratories. This talk explored the connection between quantum error correction and wireless systems that employ multiple antennas at the base station and the mobile terminal.

- "Quantum Error Correcting Codes and Quantum Cryptography" by Dr. Peter Shor, AT&T Research Laboratories. This talk introduced the fundamentals of quantum error correction and showed its relationship to quantum key distribution.

- "Frames Everywhere" by Dr. Jelena Kovacevic, Bell Laboratories. Frames started as a mathematical theory by Duffin and Schaeffer, who provided an abstract framework for the idea of time-frequency atomic decomposition by Gabor. The theory then laid largely dormant until 1986 with the publication of work by Daubechies, Grossman and Meyer. Since then, frames have evolved into a state-of-the-art signal processing tool. Frames, or redundant representations, have been used in different areas under different guises. Perfect reconstruction oversampled filter banks are equivalent to frames in the space of square-summable sequences, described by Cvetkovic and Vetterli. Frames show resilience to additive noise as well as numerical stability of reconstruction. They have also demonstrated resilience to quantization. Several works exploit the resilience of frame expansions to coefficient losses as well as the greater freedom to capture significant signal characteristics. Frames have been used to design unitary space-time constellations for multiple-antenna wireless systems. Finally, although a well-known result by a Russian mathematician M.A. Naimark -- Naimark's Theorem -- has been widely used in frame theory in the past few years, only recently have researchers recast certain quantum measurement results in terms of frames.

Workshop on Codes and Complexity
December 4 - 7, 2001
Location: DIMACS Center, CoRE Building, Room 431, Rutgers University
Organizers: Amin Shokrollahi, Digital Fountain; Dan Spielman, MIT; Ruediger Urbanke, EPFL
Attendance: 90

Ever since Shannon's paper on information theory more than 50 years ago, the construction of codes that have efficient encoders and decoders with performance arbitrarily close to Shannon's bound has been the supreme goal of coding research. The last few years have witnessed tremendous progress towards achieving this goal. One of the most intriguing aspects of recent developments has been a cross fertilization of ideas in coding and information

theory, theoretical computer science, and physics. Especially the connection to theoretical computer science ultimately led to an asymptotic analysis of many classes of codes based on graphs; conversely, coding theory has turned out to be an indispensable tool in many exciting developments in theoretical computer science, such as the design of probabilistically checkable proofs, or pseudo random generators.

This workshop brought together researchers in coding and information theory, theoretical computer science, and physics in the hope of further stimulating cross-collaboration. The workshop was preceded by a tutorial on low-density parity-check codes intended to bring graduate students and other interested researchers with little or no previous background up to speed in this important area of research.

Working Group Meeting: Computational Complexity, Entropy, and Statistical Physics
December 12 - 13, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Michael Fredman, Janos Komlos and Joel Lebowitz, Rutgers University
Attendance: 33

See the write up below for the companion workshop on this topic.

Workshop: Computational Complexity, Entropy, and Statistical Physics
December 14, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Michael Fredman, Janos Komlos and Joel Lebowitz, Rutgers University
Attendance: 82

A working group is an interdisciplinary group of researchers that meets to discuss a research area, has informal presentations, and defines research issues for further discussion and collaboration. The working group met just prior to a larger more formal workshop.

The working group and workshop explored some underlying connections among biology, computational complexity, discrete mathematics, dynamical systems and statistical physics. All these disciplines use in one way or another something called entropy, a word first introduced by Rudolf Clausius in 1865 when he enunciated his famous two laws:

1. The energy of the universe stays constant
2. The entropy of the universe always increases.

But just what is entropy? It is frequently said that entropy is a measure of disorder, and while this needs many qualifications and clarifications it does represent something essential about it. By comparing the uses of entropy in these very different contexts we were able to gain new insights into some universal aspects common to all of them.

The public workshop featured a variety of connections between entropy and other fields, including computer science. Jennifer Chayes, Microsoft Research, talked about the role of entropy in statistical mechanics, the field where this concept originated. Yuval Peres, UC Berkeley, described how this concept has come to play a role in discrete mathematics. David Zuckerman, University of Texas, Austin, talked about computational complexity and entropy. Yasha Sinai, Princeton University, related entropy to dynamical systems. Though we usually think of biology as studying systems with increasing organization, John Hopfield, Princeton University, explained how entropy also plays an important role in that field.

Workshop: C-CR Quantum Planning
January 17 - 18, 2002
Location: Hampton Inn, Elmsford, New York
Organizers: Alexander Barg, Bell Labs; Chris Fuchs, Bell Labs; Lance Fortnow, NEC Research; Peter Shor, AT&T Labs - Research
Attendance: 23

Though funded by a separate NSF grant, this event was connected to this special focus. For most of the history of computer science, researchers have considered information mostly from a binary point of view: bits are either true or false. Quantum mechanics tells us that these bits may lie in some "superposition" of true and false. Considering quantum mechanics fundamentally changes the way we must consider computation, communication and information in ways that we are only beginning to understand. The workshop report recommended that the NSF Division of Computer-Communications Research (C-CR) develop a new initiative in "Theory of Quantum Computing and Communication" to understand these issues from two directions: how the theory communities can contribute to more effective techniques in quantum information processing and how the emerging availability of quantum information processing raises research questions central to the theory communities.

Computer science has played a critical role in the development of quantum information processing. Most notably, Shor's algorithm for efficient factoring on quantum computers has shown that quantum computation can give exponential speed-up on some natural problems. Computer scientists have also led the way to develop error-correcting protocols to handle decoherence problems in quantum computers.

During this workshop, a detailed exposition of several important questions about quantum information was developed that requires careful examination by computer science researchers. These questions are in four categories:

- Algorithms and Complexity: What is the power of quantum computation?
- Cryptography: How can quantum information lead to better cryptographic protocols?
- Information and Communication: How do entanglement and other properties of quantum bits allow us to improve communications and information storage?
- Implementations, Models and Applications: How can theoretical computer science help in the design and structure of quantum machines? What are the theoretical questions arising in the application of quantum mechanics to a wide variety of computer science applications?

See http://dimacs.rutgers.edu/Workshops/QuantumPlanning/ccr-quantum.pdf for the full report.

Working Group Meeting: Data Compression in Networks and Applications
First meeting March 18 - 20, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Adam Buchsbaum, AT&T Labs - Research; S. Muthukrishnan, AT&T Labs Research and Rutgers University; Suleyman Cenk Sahinalp, Case Western University; Jim Storer, Brandeis University; Jeff Vitter, Duke University
Attendance: 53

Recent advances in compression span a wide range of applications and these were the subject of this working group meeting. Homogeneous data sources invite specific compression methods, such as for Java bytecodes, XML data, WWW connectivity graphs, and tables of transaction data. WWW infrastructure also benefits from compression. Cache sharing proxies can exploit recent work on compressed Bloom filters. Search engines can extend the idea of sketches that work for text files to music data. Examples also abound in more heterogeneous domains, such as database compression, compression of biosequences and other biomedical data which are becoming of key importance in the context of telemedicine. Additionally, new general compression methods are always being developed, in particular those that allow indexing over compressed data or error resilience. The application of compression to new domains continues, e.g., the use of multicasting to reduce information communicated during parallel and distributed computing, and the installation of general compression methods deep in the network stack, allowing transparent, stream-level compression, independent of the application. Compression also inspires information theoretic tools for pattern discovery and classification, especially for biosequences.

Workshop on Source Coding and Harmonic Analysis
May 8 - 10, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Vivek Goyal, Digital Fountain and Jelena Kovacevic, Bell Labs
Attendance: 52

Lossy compression is ubiquitous in the digital age, as evidenced by the photographs, songs, and video clips that are routinely exchanged on the Internet. Though the technologies that underlie this compression are quite complicated, well-educated generalists know that compression and digital communication are based on information theory. But this tells only part of the story.

The practice of compression has never been driven only by information theory. At least at an elementary level, information theory suggests techniques that require complete statistical knowledge of a class of data and even with this knowledge are computationally infeasible. Compression has benefited, for example, from efficient signal representations from the field of harmonic analysis, in particular advances in nonlinear approximation; computational techniques from signal processing; advanced techniques in statistical modeling and statistical inference; and insights and innovations from hands-on engineering.

This workshop continued and accelerated the exchange of ideas between the various groups that contribute to compression. This included participation by researchers in the following overlapping fields:

- source coding theory (information theorists);
- compression practice (signal processors);
- statistical modeling and inference (statisticians);
- harmonic analysis (applied mathematicians).

Workshop on Signal Processing for Wireless Transmission
October 7 - 9, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Sergio Verdu, Princeton University and Jerry Foschini, Bell Labs
Attendance: 137

In contrast to the voiceband telephone channel, the wireless channel suffers from interference from other users and from fading due to destructive addition due to multipath propagation. This workshop explored the ultimate limits that information theory puts on spectral efficiency, as well as the best means of striving toward that efficiency. Multiuser detection and "Dirty Paper Coding" are among the key signal processing countermeasures that promise substantial improvements over existing systems. This workshop investigated such approaches from both a link and network level perspective. The enhancements from multiple antennas were also explored, including means of space-time coding.

Workshop on Network Information Theory
March 17 - 19, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Piyush Gupta, Gerhard Kramer and Adriaan van Wijngaarden, Bell Labs
Attendance: 132

The DIMACS Workshop on Network Information Theory focused on the area of efficient and reliable communication in multi-terminal settings. This field has recently attracted renewed attention because of key developments that have spawned a rich set of challenging research problems. Applications such as wireless cellular and LAN data services, ad hoc networks and sensor networks should benefit from these developments. The aim of the workshop was to achieve a better understanding of the underlying information theoretic problems and their solutions.

The workshop topics included, but were not limited to, the following areas:

- Large communication networks: analysis, design, asymptotics
- Multi-terminal capacity/coding: relays, multi-access, random access
- Multi-terminal source coding: distributed sources, multiple descriptions
- Network coding: efficiency and reliability

Workshop on Complexity and Inference
June 2 - 5, 2003

Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Mark Hansen, Bell Labs; Paul Vitanyi, University of Amsterdam; Bin Yu, University of California, Berkeley
Attendance: 73

The notion of algorithmic complexity was suggested independently by Kolmogorov, Chaitin, and Solomonoff in the 1960's. Both Kolmogorov and Chaitin introduced the concept as a way to formalize notions of entropy and randomness, building on results from theoretical computer science dealing with partial recursive functions. Independently, Solomonoff defined algorithmic complexity in the pursuit of universal priors for statistical inference. In recent years, Rissanen expanded the applicability of these ideas, employing well-established concepts from information theory to frame his principle of Minimum Description Length (MDL) for statistical inference and model selection.

Each of these lines of research has developed methods for describing data (through coding and compression, or by analogy with some formal computing device); and each of these lines has employed some concept of an efficient representation to guide statistical inference. This workshop explored both the foundational aspects of complexity-based inference as well as applications of these ideas to challenging modeling problems. Participants were drawn from the fields of statistics, information and coding theory, machine learning, and complexity theory. Application areas included biology, information technologies, physics and psychology. The following specific topics were covered by the workshop:

- Kolmogorov complexity and inference
- MDL (MML) theory and applications
- Lossy compression and complexity theory
- Complexity and Bayesian methods
- Individual sequence/on-line prediction and predictive complexity
- Compression methods for clustering
- Machine learning and computational complexity
- Complexity and cognitive science
- Applications

Workshop on Algebraic Coding Theory and Information Theory
December 15 - 18, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Alexander Barg, DIMACS/Rutgers University; Alexei Ashikhmin, Bell Labs; Iwan Duursma, University of Illinois
Attendance: 77

Discoveries made in coding theory in the 1990s brought forward a number of new topics that presently attract the attention of specialists. Present-day developments in coding theory are concentrated around low-complexity code families, algebraic and lattice decoding, code design for multiple access channels, interplay with the theory of random matrices, quantum error correction, quantization, nontraditional applications of coding theory such as alternatives to routing in networks, coding of correlated sources and problems in theoretical computer science. Expanding into these areas enriched coding theory research with new problems and ideas. On the other hand, a number of recent studies in information theory attempt at developing structured code families as an alternative to random coding arguments usually employed to establish performance limits.

This workshop established and strengthened links between algebraic coding and information/communications theory. The workshop featured theoretical contributions in some of the named areas, both new results and tutorial presentations. It had a substantial educational component, exposing coding theorists to a new range of problems and presenting constructive methods to the information theory community. It is hoped that as a result of the workshop, experts in coding theory will be able to identify problems in information theory that can be addressed with coding theory methods, and information theorists will become more familiar with ideas used for code construction.

Working Group Meeting: Theoretical Advances in Information Recording

March 22 - 24, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Emina Soljanin, Bell Labs; Paul Siegel, Univ. of California, Bane Vasic, University of Arizona; Adriaan J. van Wijngaarden, Bell Labs
Attendance: 22

See write the up for the companion workshop below.

Workshop on Theoretical Advances in Information Recording
March 25 - 26, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Emina Soljanin, Bell Labs; Paul Siegel, Univ. of California, Bane Vasic, University of Arizona; Adriaan J. van Wijngaarden, Bell Laboratories
Attendance: 25

The last twenty years have witnessed a tremendous increase in recording densities and data rates of man-made recording systems, as well as a steady progression in a variety of methods for information storage and retrieval. The continuing strong demands for more storage capacity and faster access will further drive the development of even more sophisticated storage systems and new materials. Further advances in physics, chemistry and biology may provide new possibilities to further push the limits. Information theory will be playing an important role in identifying principles unique to both natural and artificial storage systems and in building a model and framework for storage in these new materials. These principles are as old as the time "when life divided the labour between two separate activities: chemical work and information storage, metabolism and replication" (as noted by Matt Ridley in "Genome"), and an effort to better understand and mathematically formulate these mechanisms will offer new directions in information and coding theory. It will also motivate a review of existing storage system models and coding and detection methodologies, which, in turn, can provide improvements to new and existing systems as well.

This working group and workshop brought together experts on information storage from a range of different fields. In order to facilitate a synthesis of ideas, the workshop was organized around half-day sessions, each consisting of one or two short presentations followed by active discussion. The six primary topics were:

- recent advances in coding for magnetic storage
- two-dimensional and higher dimensional storage systems
- recent advances in coding for optical recording systems
- sofic systems and constrained coding
- error control coding and iterative decoding for storage systems
- information storage in DNA and combinatorics in computational biology.

Princeton-Rutgers Seminar Series in Communications and Information Theory 2002 - 2003

Tuesday, October 1, 2002
Speaker: **P. R. Kumar**, University of Illinois
Title: Ad hoc wireless networks: Analysis, protocols, architecture, and towards convergence
Location: DIMACS Center, Rutgers University, Busch Campus, Piscataway, NJ

Thursday, October 3, 2002
Speaker: **Jacob Ziv**, Professor of Electrical Engineering, Technion--Israel Institute of Technology: President, Israeli Academy of Sciences and Humanities
Title: Classification with Finite Memory Revisited
Location: Princeton University, Friend 101

Thursday, October 3, 2002
Speaker: **Shlomo Shamai (Shitz)**, Technion
Title: On Information Theoretic Aspects of Multi-Cell Wireless Systems
Location: Princeton University, Friend 004

Thursday, November 7, 2002
Speaker: **Michael Luby**, Digital Fountain
Title: LT Codes
Location: Princeton University, Friend 101

Thursday, November 14, 2002
Speaker: **Upamanyu Madhow**, University of California, Santa Barbara
Title: Information-Theoretic Prescriptions for Outdoor Wireless Communication
Location: DIMACS Center, Rutgers University, Busch Campus, Piscataway, NJ

Thursday, November 21, 2002
Speaker: **Tsachy Weissman**, Hewlett-Packard Labs
Title: Universal Discrete Denoising: Known Channel
Location: Princeton University, Friend 101

Monday, March 3, 2003
Speaker: **Michael Honig**, Northwestern University
Title: Asymptotic Methods in Wireless Communications
Location: DIMACS Center, Rutgers University, Busch Campus, Piscataway, NJ

Thursday, March 6, 2003
Speaker: **Babak Hassibi**, California Institute of Technology
Title: Some Asymptotic Results in Wireless Networks
Location: Princeton University, Friend 006

Thursday, March 13, 2003
Speaker: **Robert J. McEliece**, Professor of Electrical Engineering, California Institute of Technology
Title: Belief Propagation on Partially Ordered Sets
Location: Princeton University, Friend 006

Friday, March 14, 2003
Speaker: **Venkat Anantharam**, University of California, Berkeley
Title: Message Passing Algorithms for Marginalization
Location: CoRE Bldg, Rutgers University, Busch Campus, Piscataway, NJ

Monday, April 14, 2003
Speaker: **David N. C. Tse**, University of California, Berkeley
Title: Diversity and Multiplexing: A Tradeoff in Wireless Systems
Location: DIMACS Center, Rutgers University, Busch Campus, Piscataway, NJ

Thursday, April 24, 2003
Speaker: **Andrea Goldsmith**, Stanford University
Title: Capacity Limits of Wireless Channels with Multiple Antennas: Challenges, Insights, and New Mathematical Methods
Time: 4:30-5:30pm
Location: DIMACS Center, Rutgers University, Busch Campus, Piscataway, NJ

Thursday, May 1, 2003
Speaker: **Thomas Richardson**, Flarion Technologies
Title: Principles and Design of Iterative Coding Systems
Time: 4:30-5:30pm
Location: Princeton University, Friend 006

Princeton-Rutgers Seminar Series in Communications and Information Theory 2003 - 2004

Thursday, November 13, 2003
Speaker: **Vaidyanthan Ramaswami**, AT & T Labs
Title: Providing Dial Tone in the Presence of Circuit Congestion due to Long Holding Time Internet Dial-Up Calls - Assuring Emergency Services Access
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, November 20, 2003
Speaker: **Paul Henry**, AT & T Labs
Title: 4G Cellular: Now That's Mobile Data
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, December 4, 2003
Speaker: **Larry Peterson**, Princeton University
Title: A Blueprint for Introducing Disruptive Technology into the Internet
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, December 11, 2003
Speaker: **Nirwan Ansari**, NJIT
Title: IP Traceback by Deterministic Packet Marking
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, February 5, 2004
Speaker: **Sekhar Tatikonda**, Yale University
Title: Markov Channels, Sufficient Statistics, and Capacity
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, February 12, 2004
Speaker: **P. R. Kumar**, University of Illinois
Title: Wireless Networks: From Information Transfer to Sensing and Control
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, February 19, 2004
Speaker: **Sandy Fraser**, Fraser Research
Title: A Packet Switch to Serve One Million Households
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, February 26, 2004
Speaker: **John Tsitsiklis**, MIT
Title: A Game Theoretic View of Efficiency Loss in Network Resource Allocation
Time: 4:30 - 5:30pm
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, March 4, 2004
Speaker: **Lang Tong**, Cornell University
Title: On Cross-Layer Design of Wireless Networks
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, April 15, 2004
Speaker: **Steve Sposato**, SBC Network Systems Engineering
Title: Network Design and Optimization for Large Scale Network Deployments
Location: Princeton University, E-Quad B205, Princeton, NJ

Thursday, April 22, 2004
Speaker: **Ness Shroff**, Purdue University
Title: Simplification of Network Dynamics in Large Systems
Location: Princeton University, E-Quad B205, Princeton, NJ

Thurday, May 27, 2004
Speaker: **Venkat Anantharam**, U. C., Berkeley
Title: The Role of Common Randomness in Communications and Control
Location: Princeton University, E-Quad B205, Princeton, NJ

Computational Geometry and Applications Special Focus Project Activities

Theme of the Program:

Computational applications that manipulate geometric models are ubiquitous in science and technology. Efficient algorithms are essential to allow increases in the size, complexity, and completeness of the models. Computational geometry provides efficient algorithms when the models can be described as large collections of relatively simple objects such as points and polygons. Computational geometry has its roots in algorithm analysis and draws heavily from both classical Euclidean geometry and more modern discrete and combinatorial geometry; it includes a strong theoretical foundation of algorithms, data structures, analysis techniques, and conceptual models, as well as an emerging software library.

A few examples illustrate the success of computational geometry methods. First, the solution of partial differential equations is fundamental in many areas, e.g., stress analysis, fluid flow, electromagnetic modeling. A mesh is invariably required to solve the partial differential equation numerically. Delaunay triangulations form the basis of essentially all three-dimensional simplicial mesh generators. Computational geometry has provided a solid theoretical and practical understanding of Delaunay triangulations, including asymptotically efficient algorithms, robust implementations, and connections with other geometric structures. Second, prediction of radio propagation is fundamental to the design of wireless communication networks (a topic emphasized in the special focus on Next Generation Networks Technologies and Applications). Classical geometric optics and the uniform theory of diffraction provide a sophisticated physical model of radio propagation based on ray-tracing. A naive simulation of the model on a large problem instance might require a day or a week of computing time; computational geometry techniques can provide interactive simulations, allowing a system designer to experiment with variations in network parameters. Third, the prospective technology of `3d photocopying' attempts to characterize an object by sampling points on its surface. Computational geometry has provided algorithms that reconstruct a surface model from the sample points, minimizing the number of required sample points and the deviation from the actual object as well as the required computing time.

This special focus on computational geometry and applications followed by 13 years the first DIMACS special year on discrete and computational geometry. During the 13 years, computational geometry evolved into a mature field with rigorous mathematical foundations. At the same time, there was a continuing demand for geometric algorithms from affiliated disciplines (graphics, geographic information systems, robotics, VLSI design, computer-aided design and manufacturing, among others). A strong desire developed within the computational geometry community to improve connections with these and other disciplines. It was timely to have another special focus on computational geometry to exploit these trends, and DIMACS, with its local strength and international connections in the field, was a natural place to have this.

The special focus deepened the connections between computational geometry and other disciplines. For example, computational geometry and graphics have traditionally been closely tied. However, with changes in technology, e.g., fast cheap PCs and global but relatively slow networks, it was appropriate to reexamine the algorithmic connections between the two fields. The compression, transmission, and adaptive rendering of meshed models involve combinatorial and topological ideas very much of the flavor of computational geometry; similarly, object-space-based rendering algorithms may become more attractive as the size of models increases. As another example, the solid modeling community has developed in parallel to the computational geometry community. Both communities are concerned with geometric algorithms and have had to deal with many of the same issues, e.g., complexity of representation data structures. The solid modeling community has traditionally been more closely tied to industry and more concerned with curved surfaces; the computational geometry community has been more mathematical and more concerned with linear objects. It is clear that there is considerable potential interaction between the fields, for example on such fundamental conceptual issues as how to describe geometric shapes and

how to deal with geometric tolerances. Finally, there are emerging connections with other fields, such as molecular biology (for example through geometric searching).

The special focus also extended the connections between computational geometry and mathematical geometry. The connection with discrete and combinatorial geometry is already well-established, but is sufficiently fundamental to develop further; an intriguing new connection is the one made with computational real algebraic geometry. In addition, the special focus addressed issues of importance within computational geometry, such as the challenging engineering required to implement geometric algorithms.

This special focus began in Fall 2002 and will end in smmer 2005.

Workshops, Working Groups, Tutorials:

Special Program: Reconnect Conference 2002: Voronoi Diagrams - Properties, Algorithms and Applications
Dates: August 11 - 17, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Scott Drysdale, Dartmouth College; Shelly Leibowitz, Wheaton College
Attendance: 27

This conference, funded by a different NSF grant and acting as a "pre-special focus" activity, exposed faculty teaching undergraduates to the mathematical sciences research enterprise by introducing them to a current research topic relevant to the classroom through a series of lectures by a leading expert and involving them in writing materials useful in the classroom. Participants had the possibility of following up by preparing these materials for publication in the DIMACS Educational Modules Series.

This conference was also aimed at reconnecting faculty to the mathematical sciences research enterprise by involving them in DIMACS and through opportunities to follow up after the conference by getting connected to DIMACS researchers and other DIMACS programs throughout the year.

Workshop on Geometric Graph Theory
September 30 - October 4, 2002
Location: Rutgers University
Organizer: Janos Pach, New York University, Courant Institute of Mathematical Sciences
Attendance: 56

The rapid development of computational geometry in the past two decades presented a powerful new source of inspiration for combinatorial geometry, the theory of geometric arrangements. Many curious extremal problems in recreational mathematics turned out to be crucially important for the analysis of a wide range of geometric algorithms. Perhaps the best example of this is the so-called k-set problem of Erdos, Lovasz, Simmons, and Straus, which is more than thirty years old. Given n points in general position in the plane, what is the maximum number of k-tuples that can be separated from the remaining n-k points by a straight line? After almost twenty years of stagnation and ten years of slow progress, Dey and Toth achieved two breakthroughs by substantially improving the best known upper and lower bounds, respectively.

The k-set problem belongs to a newly emerging discipline, geometric graph theory. A geometric graph is a graph drawn in the plane such that its vertices are points in general position and its edges are straight-line segments. The first result on geometric graphs was proved more than seventy years ago by Hopf and Pannwitz: if a geometric graph has no two disjoint edges, then its number of edges cannot exceed its number of vertices. According to Conway's famous Thrackle Conjecture, the same assertion remains true for graphs drawn by (not necessarily straight) continuous arcs with the property that any two of them have at most one point in common. Many recent results in this area are relevant to proximity questions, bounding the number of incidences between points and curves, designing various graph drawing algorithms, etc.

This workshop explored the consequences of the new results, and discussed their extensions and some related questions. Many of the researchers were brought together who contributed to the development of this area.

Workshop on Computational Geometry
November 14 - 15, 2002
Location: Rutgers University
Organizer: Joe Mitchell, SUNY Stony Brook
Attendance: 73

This was the twelfth in a series of annual fall workshops on Computational Geometry. This workshop series, founded initially under the sponsorship of the Mathematical Sciences Institute (MSI) at Stony Brook (with funding from the U. S. Army Research Office), continued during 1996-1999 under the sponsorship of the Center for Geometric Computing, a collaborative center of Brown, Duke, and Johns Hopkins Universities, also funded by the U.S. Army Research Office. In 2000, the workshop returned to the campus of the University at Stony Brook. In 2001, it was held at Polytechnic University in Brooklyn. In 2002, as part of the DIMACS Special Focus on Computational Geometry and Applications, the workshop was hosted and sponsored by DIMACS.

This workshop brought together students and researchers from academia and industry, to stimulate collaboration on problems of common interest arising in geometric computations. Topics covered include, but were not limited to:

- Logorithmic methods in geometry
- Geometric data structures
- Implementation issues
- Robustness
- Computer graphics
- Solid modeling
- Geographic information systems
- Applications to computational biology and chemistry
- Computational metrology
- Graph drawing
- Experimental studies
- Computer vision
- Robotics
- Computer-aided design
- Mesh generation
- Manufacturing applications of geometry
- I/O-scalable geometric algorithms
- Animation of geometric algorithms.

Following the tradition of the previous workshops on Computational Geometry, the format of the workshop was informal, extending over 2 days, with several breaks scheduled for discussions. There was also an Open Problem Session in order to promote a free exchange of questions and research challenges.

Workshop on Algorithmic Issues in Modeling Motion
November 18 - 20, 2002
Location: Rutgers University
Organizers: Pankaj K. Agarwal, Duke University and Leonidas Guibas, Stanford University
Attendance: 55

Motion, like shape, is one of the fundamental modalities to be modeled in order to represent and manipulate the physical world in a computer. As such, motion representations and the algorithms that operate on them are central to all computational disciplines dealing with physical objects: computer graphics, computer vision, robotics, etc. Modeling motion is also crucial for other disciplines dealing with temporally varying data, including mobile networks, temporal data bases, etc. Motion algorithms require computational resources, and frequently sensing and communication resources as well, in order to accomplish their task. Despite the prominent position that motion plays in so many computer disciplines, little had been done previously to provide a clean conceptual framework for representing motion, describing algorithms on moving objects, and analyzing their behavior and performance.

This workshop brought together people from the different research communities interested in algorithmic issues related to moving objects. The workshop addressed core algorithmic issues as well as aspects of modeling and analyzing motion. The issues in representing, processing, reasoning, analyzing, searching, and visualizing moving objects were debated and discussed; the key research issues were identified, and relationships that can be used to strengthen and foster collaboration across the different areas were established.

Workshop on Implementation of Geometric Algorithms
December 4 - 6, 2002
Location: Rutgers University
Organizers: Herve Bronnimann, Polytechnic University and Steve Fortune, Bell Labs
Attendance: 42

It is notoriously difficult to implement geometric algorithms. This difficulty arises in part from the conceptual complexity of geometric algorithms, the proliferation of special cases, the dependence of combinatorial decisions on numerical computation, and frequent theoretical focus on worst-case asymptotic behavior.

This workshop addressed research issues related to the implementation of geometric algorithms. Typical, but not exclusive topics included:

- Numerical issues
- Noisy data and data repair
- Geometric data structures
- Massive geometric data sets
- Algorithm library design
- Algorithm engineering
- Experimental studies.

Numerical issues have long been an important concern in the implementation of geometric algorithms. In the last decade the issue has become a central research topic in computational geometry, and a reasonably successful approach based on the use of exact (extended-precision) arithmetic has been developed. However, many significant problems remain---high-level rounding, extension to curved objects, performance---and the practical impact of the research is not yet clear.

Geometric data sets based on physical measurements are inherently noisy. If such geometric data also has combinatorial structure, the geometric and combinatorial information may be inconsistent. To be useful, geometric algorithms must be able to repair such data, that is, in some fashion eliminate inconsistencies. Unfortunately, there is little relevant theory, and current data repair is heuristic at best.

Geometric data structures are known that can represent complex structures in any dimension. However massive data sets in two dimensions, or even modest data sets in high dimension, can require enormous amounts of memory. A challenging research topic is to design algorithms and data structures that are cognizant of the memory hierarchy---cache, main memory, disk---and to provide appropriate implementations.

These are just some of the problems faced by general purpose geometric algorithms libraries. Considerable effort has been expended developing the geometric algorithm library CGAL, which is now reasonably mature. CGAL,just together with the LEDA algorithm library, provides an unparalleled resource for users of geometric algorithms. Further development of algorithm libraries requires attention to many issues---those mentioned above, but also functionality, interface, performance, and support for specific application areas.

This workshop brought together both researchers and practitioners. The practitioners benefited from discussions of the state of the art in research, and the researchers benefited by being exposed to implementation issues of practical importance.

Workshop on Medical Applications in Computational Geometry
April 2 - 4, 2003

Location: Rutgers University
Organizers: Danny Chen, University of Notre Dame and Jean-Claude Latombe, Stanford University
Attendance: 66

Computer technology plays an increasingly important role in modern medicine and life sciences. Many medical problems are of a strong geometric nature and may benefit from computational geometry techniques. The DIMACS workshop on Computational Geometry and Medical Applications provided a forum for researchers working in computational geometry, medicine, and other related areas to get together and exchange ideas, and to promote cross-fertilization and collaborations among these areas. The theme of the workshop was the exploration of the applicability of computational geometry to medical problems and the new challenges posed by the current medical research and practice to geometric computing. Examples of topics included surgical simulation and planning, geometric representation and modeling of medical objects and human-body tissue structures, geometric problems in medical imaging, computational anatomy, registration and matching of medical objects, etc.

Workshop on Surface Reconstruction
April 30 - May 2, 2003
Location: Rutgers University
Organizers: Nina Amenta, University of California - Davis; Fausto Bernardini, IBM - T. J. Watson Research Center
Attendance: 55

Surface reconstruction is the problem of producing a representation of a two-dimensional surface in 3D, give a set of sample points lying on or near the surface. There are a number of interesting systems which collect sample sets and construct surfaces, using laser range scanners, structured light and other techniques. Results in computational geometry have focused on algorithms for the surface reconstruction problem and finding ways to guarantee topologically and geometrically correct outputs, given good enough input samples. This workshop encouraged interaction between systems builders and the computational geometry community.

The workshop surveyed the state of the theory and practice of surface reconstruction, and looked at related problems arising in systems that reconstruct objects, such as the alignment of multiple laser range scans, integrating photometric data, reconstructing dynamic scenes and real-time object acquisition.

Workshop on Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications
May 14 - 16, 2003
Location: Rutgers University
Organizers: Regina Liu, Rutgers University; Yehuda Vardi, Rutgers; Robert Serfling, University of Texas at Dallas; Diane Souvaine, Tufts University
Attendance: 92

Multivariate statistical methodology plays a role of ever increasing importance in real life applications, which typically entail a host of interrelated variables. Simple extensions of univariate statistics to the multivariate setting do not properly capture the higher-dimensional features of multivariate data, nor do they yield geometric solutions because of the absence of a natural order for multidimensional Euclidean space. A more promising approach is the one based on “data depth”, which can provide a center-outward ordering of points in Euclidean space of any dimension. Extensive developments in recent years have generated many attractive depth-based tools for multivariate data analysis, with a wide range of applications. The diversity in approaches, emphases, and concepts, however, makes it necessary to seek unified views and perspectives that would guide the further development of the depth-based approach.

The concept of data depth provides new perspectives to probabilistic as well as computational geometries. In particular, the development of implementable computing algorithms for depth-based statistics has brought about many new challenges in computational geometry. This workshop created a unique environment for multidisciplinary collaboration among computer scientists, theoretical and applied statisticians, and data analysts. It brought together active researchers in these fields to discuss significant open issues, establish perspective on applications, and set directions for further research.

Workshop on Geometric Optimization
May 19 - 21, 2003
Location: Rutgers University
Organizers: Joe Mitchell, SUNY Stony Brook; Pankaj Agarwal, Duke University
Attendance: 62

Combinatorial optimization typically deals with problems of maximizing or minimizing a function of one or more variables subject to a large number of constraints. In many applications, the underlying optimization problem involves a constant number of variables and a large number of constraints that are induced by a given collection of geometric objects; these problems are referred to as geometric-optimization problems. Typical examples include facility location, low-dimensional clustering, network-design, optimal path-planning, shape-matching, proximity, and statistical-measure problems. In such cases one expects that faster and simpler algorithms can be developed by exploiting the geometric nature of the problem. Much work has been done on geometric-optimization problems during the last twenty-five years. Many elegant and sophisticated techniques have been proposed and successfully applied to a wide range of geometric-optimization problems. Several randomization and approximation techniques have been proposed. In parallel with the effort in the geometric algorithms community, the mathematical programming and combinatorial optimization communities have made numerous fundamental advances in optimization, both in computation and in theory, during the last quarter century. Interior-point methods, polyhedral combinatorics, and semidefinite programming have been developed as powerful mathematical and computational tools for optimization, and some of them have been used for geometric problems.

This workshop brought together people from different research communities interested in geometric-optimization problems. Various techniques developed for geometric optimization and their applications were discussed, key research issues that need to be addressed were identified, and relationships that can be used to strengthen and foster collaboration across the different areas were established.

Workshop on Computer-Aided Design and Manufacturing
October 7 - 9, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Debasish (Deba) Dutta, University of Michigan; Ravi Janardan, University of Minnesota; Michiel Smid, Carleton University
Attendance: 54

Computer-Aided Design and Manufacturing (CAD/CAM) is concerned with all aspects of the process of designing, prototyping, manufacturing, inspecting, and maintaining complex geometric objects under computer control. As such, there is a natural synergy between this field and Computational Geometry (CG), which involves the design, analysis, implementation, and testing of efficient algorithms and data representation techniques for geometric entities such as points, polygons, polyhedra, curves, and surfaces. On the one hand, CG can bring about significant performance improvements in CAD/CAM, while, on the other hand, CAD/CAM can be a rich source of interesting new problems that provide new impetus to research in CG. Indeed, such two-way interaction has already been witnessed in areas such as numerically-controlled machining, casting and injection molding, rapid prototyping and layered manufacturing, metrology, and mechanism/linkage design, to name just a few.

This workshop further promoted this interaction by bringing together researchers from both sides of the aisle to assess the current state of work at the interface of the two fields, to identify research needs, and to establish directions for collaborative future work.

Topics that were addressed included: geometric aspects of manufacturing processes (from traditional machining to layered manufacturing to nanoscale manufacturing), process planning and control, rapid prototyping technologies, computational metrology and tolerancing, geometric problems in mechanism design, geometric constraint systems, geometric modeling related to manufacturing, computer vision and robotics related to manufacturing, and geometric issues in standards development.

Next Generation Networks Technologies Special Focus Project Activities

Theme of Program:

Begun in January 2000, the Special Focus on Next Generation Networks Technologies and Applications continued many of the same themes introduced in the 1996-2000 Special Focus on Networks.

As the computer age has reached the end of the millennium, there has been a major shift, namely, to interconnect computation and global communications as well as share information. This convergence of computing, telecommunications and information is poised to take off even further with novel technology (e.g., cable, wireless and other high speed/bandwidth networking options) and phenomenally successful applications (e.g., web). We are seeing the proliferation of mobile computation units, networking of unprecedented scale, bandwidth demand growing at an exponential rate, and petabytes of distributed shared data. The emerging technological reality offers opportunities that will dramatically transform society.

While the underlying technologies continue to advance at a rapid pace, the fundamental principles governing the design, deployment and use of next-generation networks of unprecedented scale, heterogeneity, and complexity are not entirely clear. How are systems involving the emerging technologies (such as cable or wireless) best designed? This question pervades all levels of networking: at the network level (providing a personal network space, designing QoS for heterogeneous networks such as wired-cable and wired-wireless); at the application level (providing a thin-client view of the web for mobile units, secure electronic commerce, etc.); at the system level (providing interoperability, a homogeneous view of the system in the presence of mobility, etc.); at the protocol level (how to modify TCP for wired-wireless networks, how to multicast in wireless networks, etc.); at the link level (scheduling coding and power level for packet transmissions). New network services raise new challenges: Privacy can be threatened; small faults can cause cascading network failures with severe economic consequences; the sheer number of requests for information can slow the networks and information servers to a crawl; protocols created for previous generation networks may not be appropriate for the scale and complexity of newer technologies; achieving effective interoperability among systems remains extremely difficult; managing distributed systems with mobile software is poorly understood; multiple networks and software systems are being interconnected without a full understanding of the ramifications. This special focus took advantage of new opportunities to solve underlying problems by addressing the theoretical foundations on which the emerging environment of commingled communication, information, and computation is based.

The special focus concentrated on the following:

- Algorithms to achieve cost- and service-effective use of the resources in the emerging environment in virtually all areas, including communications (mobile or stationary, network design, protocols, routing), large scale planning and scheduling, and economic decision making. This involves expertise from fields such as optimization, approximation, probability and queuing theory, control theory and game theory.

- Models for massive data analysis of multimedia information (e.g., data warehousing and data mining) and traffic and information management in global intelligent networks. Because existing algorithms for data analysis are unlikely to scale, new methods are needed to organize, store, and retrieve data, discover patterns, and use them for planning and decision making. This involves expertise from diverse areas including data structures, databases, data compression and coding, statistics, learning, geometry and visualization.

- Fundamental theoretical work to address problems of reliability, security and privacy. The design of systems that maximize the potential for new technologies while protecting security and privacy requires expertise in fields such as verification, distributed computing, probability, coding theory, number theory, cryptography and security.

This special focus fostered closer working relationships between the algorithms and networking technology communities. With the help of such relationships, researchers can pursue the development of the fundamental theory, models, and scalable algorithms that cut through a variety of application domains and manage the present technology and design the future technology, from the physical layer to the application layer. Foundational work is required to solve existing problems and to lay the groundwork for the next generation networks technologies and applications, and this special focus was aimed at pinpointing the opportunities for theorists to make a contribution and at stimulating the research needed to take advantage of these opportunities.

Workshops, Working Groups, Tutorials:

Working Group Meeting: Policy-driven Decision Making and Dynamic Interoperability
December 8, 2000
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Tom Buckman, MITRE Corporation; Joan Feigenbaum, Yale University; Fred Roberts, Rutgers University
Co-sponsored by DIMACS, the Office of Naval Research and the National Science Foundation
Attendance: 30

Modern military operations emphasize a need to be able to rapidly assemble multiservice, multination coalitions. The modern, high speed, high tech information based economy has put a premium on the ability of companies to partner rapidly in joint ventures. Nations and corporations must quickly interface operational processes and federate supporting communication and information systems, in order to meet mission objectives. This calls for a new approach to achieving interoperability, one that emphasizes the ability of operational units and their supporting information systems to adapt to unanticipated changes in operational requirements that occur during the execution of a mission or that occur in the context of new joint ventures.

In order to respond to unanticipated changes in requirements, it will be necessary to rapidly translate policy decisions generated by mission commanders, senior executives, system operators, security authorities and others into coherent sets of commands or protocols that can be used directly by supporting information systems to reconfigure themselves to satisfy these policy requirements.

This initial working group meeting explored previous work done by theoretical computer scientists in the area of policy driven decision-making and its potential application to solving problems associated with enabling dynamic interoperability. Previous work done in this area related to scenarios in which mutually distrustful (or partially distrustful) parties wish to perform a security-critical joint action was studied.

The working group meeting concluded with a roundtable discussion on ideas. The discussion focused on identifying areas for further research that could serve as focal points for a follow-on workshop and working group meeting.

3rd Workshop on Algorithm Engineering and Experiments (ALENEX 2001)
January 5 - 6, 2001
Location: Wyndham City Center Hotel, Washington, DC
Organizers: Nina Amenta; Adam Buchsbaum, Co-Chair, AT&T Labs - Research; Rudolf Fleischer, HKUST; Lyle McGeoch, Amherst College; S. Muthukrishnan, AT&T Labs; Jack Snoeyink, Co-Chair, UNC, Chapel Hill; Matt Stallman, NC State University; Dorothea Wagner, University of Konstanz
Attendance: 125

The aim of the annual ALENEX workshops is to provide a forum for the presentation of original research in the implementation and experimental evaluation of algorithms and data structures.

Papers presented original research in all aspects of algorithm engineering and experimentation including:

- Implementation, experimental testing, and fine-tuning of discrete algorithms.
- Development of software repositories and platforms that allow use of, and experimentation with, efficient discrete algorithms.
- Methodological issues including standards in the context of empirical research on algorithms and data structures.
- Methodological issues regarding the process of converting user requirements into efficient algorithmic solutions and implementations.

In 2002 a special issue of the ACM Journal of Experimental Algorithmics was devoted to selected papers from this workshop.

Mini-Workshop: Quality of Service Issues in the Internet
February 8 - 9, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Funda Ergun, Co-Chair, Case Western Reserve University; Michael Mitzenmacher, Harvard University; S. Cenk Sahinalp, Case Western Reserve University; Aravind Srinivasan, Bell Labs; Bulent Yener, Co-Chair, Bell Labs
Attendance: 58

This workshop identified the fundamental problems in providing QoS in the IP networks, both from theoretical and application oriented points of view.

The explosive growth of the Internet is accompanied by two paradigms. First, the IP-based networking is expected to be the dominant infrastructure. Second, unconventional applications such as IP telephony, Virtual Private Networking, Real-time transactions, IP HDTV, are expected to be carried on the IP-based networks. These two paradigms are translated to the following challenge: how to provide Quality of Service (QoS) in the IP-based Internet. Fundamentally the IP-based networking is designed for delivering data traffic with best-effort (i.e., no guarantee) service, thus it is not capable of providing end-to-end QoS. The challenge is to devise new protocols and algorithms for providing QoS while maintaining the basic principles of IP-based networking.

This workshop included the following topics:

1. Overview: QoS on IP networks, both wired and wireless.
    - Emerging high-speed technologies of QoS
    - IP over ATM, IP over SONET, etc.
    - What is QoS? Requirements, resources, allocation/protection of the resources.
2. Current state of the field.
    - Diffserv
    - Intserv
    - MPLS
3. Quality of Service in Wireless IP Networks.
4. QoS Routing, provisioning, pricing.
5. Packet Filtering/Classification.
6. QoS Resource Reservations/Signaling (RSVP, LDP, ...).
7. Packet Scheduling to improve performance, as well as to satisfy QoS guarantees.
8. Related issues such as congestion control, multicast, network optimization, etc.

Workshop: Resource Management and Scheduling in Next Generation Networks
March 26 - 27, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Matthew Andrews, Bell Labs; M. Franklin, UC Berkeley; S. Muthukrishnan, AT&T Labs; Rajmohan Rajaraman, Northeastern University; Roy Yates, Rutgers University
Attendance: 51

The next generation of communication networks will have a number of features that give rise to new optimization problems. In particular, it is becoming clear that future networks will be characterized by increased mobililty and heterogeneity of users and increased distribution of content. Although these trends have obvious benefits, they also raise fundamental questions about the best way to manage network resources.

For example, the traditional packet scheduling literature assumes that servers have a fixed processing rate that must be shared between clients. However, in a mobile environment the rate at which we can transmit to clients is fundamentally affected by their position. New scheduling algorithms are needed to deal with this spatial aspect. The advent of distributed storage networks gives rise to a number of new problems. For instance, we must make sure that all new content is distributed to the storage sites in a timely fashion and we must ensure the consistency of the content. In addition, whenever a client wishes to retrieve some content, the network must work out which is the best storage site for that client.

This workshop focused on optimization problems that arise in new network settings. Specific topics included:

- Physical Layer Issues: channel allocation; rate and power control; contention resolution and multiple access schemes.
- Network and Transport Layer Issues: location management (role of kinetic geometric algorithms and dynamic data structures), distributed directory services.
- Application Layer Issues: data dissemination by broadcast servers, management of replicated data in wide-area heterogeneous networks,(caching, replication, and consistency), addressing the limited capability of mobile stations and adapting web content.

Workshop: Multicasting: Architecture, Algorithms, and Applications
May 2 - 4, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Micah Adler, University of Massachusetts; D. Frank Hsu, Fordham University
Attendance: 40

Multicasting plays an important role in the design, development, operation, and application of next generation networks that rely on the efficient delivery of packets to multiple destinations across a multifaceted and multi-domain network. Due to the advent of broadband, wireless and web-based system design technologies, it has become possible and feasible to design and construct large scale, heterogeneous and complex wireline and wireless communication networks that can support multimedia conferencing, streaming media distribution, distributed data sharing, distance learning, "push" oriented application, and QoS for wired-cable and wired-wireless applications. However, these technology advancements and applications and the convergence of computing, telecommunications and information processing also open-up many challenging problems and issues for both the theory community and practitioners.

This workshop brought together many of the best researchers and practitioners to present and discuss the recent evolution in the subject area, to interact on emerging issues of common interest, and to set directions and possible standards for future research of and implementation on multicasting networks and their infrastructures.

This workshop covered the following topics:

- IP Multicasting; concept, motivation, standardization, ``Host Group'' model, scoping and forwarding, IGMP for end stations.
- Multicast routing protocols and algorithms; RPM, source-based trees (DVM RP, PIM- dense mode, MOSPF), shared trees (Core-Based Trees(CBT), PIM-sparse mode), Internet multicast routing (Mbone vs. M-GBP or BGMP/GUM?), fast algorithms to compute the multicast tree.
- Multicast transport protocols, RMP and RAMP, interoperability frameworks, expanding-ring searches.
- Multicast Congestion Control; layering schemes such as RLM, multicast at the router and/or switch level.
- Topology inference and network monitoring using multicasting.
- Multicast in wireless systems, mobile computing, ad hoc networks.
- Case Studies and other issues; implementation cases, multicast in the enterprise, fairness in multicasting, pricing of multicasting.

Working Group: Data Analysis and Digital Libraries
May 17, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Paul Kantor and Endre Boros, Rutgers University; David Hull, WhizBang; Michael Berry, University of Tennessee; Warren Greiff, Mitre Corporation; Liz Liddy, Syracuse University
Attendance: 29

This working group meeting brought together researchers whose work bears on the question: "Since the content of a Digital Library is machine readable, to what extent can the computers that maintain the library also extract new and useful information from that content?"

The meeting aimed at better communication among members of the diverse communities that all have something important to offer about the problem. A key challenge in the course of the meeting was to develop a common language, enabling participants to bring their own knowledge to bear on the problems that interest others.

Among the application areas explored were:

- clustering of documents and collections
- collection selection
- topic detection
- adaptive filtering
- automatic thesaurus construction.

The tools discussed included:

- vector models and linear classifiers
- Boolean models and rule based models
- probabilistic models of relevance
- inference networks for information retrieval
- human rule finding capabilities
- bigram and n-gram based models for text
- NLP based models for organizing and mining libraries
- co-citation and other bibliographic linkages
- web-based and link-based methods of organization.

This meeting focused on researchers who are addressing one or more of these problems: image indexing, sound indexing, text indexing, clustering of (image, sound, text); collection selection; information retrieval; pattern finding in text collections. We also focused on new research on the problem of evaluating schemes for either retrieval or clustering and classification, that can bring some much needed rigor to this problem.

We also discussed the issue of "evaluating" systems. Two practically oriented "gold standards" have emerged. One is the TREC setting, a DARPA/NIST/ARDA sponsored annual workshop in which dozens of systems address the same fixed set of tasks, with performance on those tasks being assessed impartially at NIST. The other is the commercial web setting, where schemes compete to attract the eyeballs of surfers, and the methods are often tailored to hold the user's attention, while providing acceptable levels of organization and retrieval. The effectiveness (in, say, the TREC sense) of the commercial schemes is not well understood. Perhaps neither set of measures will ultimately prove most effective in defining the new technology of digital libraries.

Workshop: Pervasive Networking
May 21, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: B.R. Badrinath, Rutgers University; Deborah Estrin, UCLA
Attendance: 55

Progress in technology will make it possible to network and monitor every physical object that surrounds us. This large scale sensor network will become a reality in the near future due to the availability of cheap, low-power sensors with wireless access. This ubiquitous network will result in computing environments in places such as automobiles, homes, roads, infrastructures, and buildings. These pervasive computing environments will contain millions of components that are distributed throughout physical space allowing users to query the physical surroundings and get continuous information about the physical space at a level of detail that is not yet possible.

These future computing environments must be robust, deployable and maintainable. They must be self-configuring, self-monitoring and scalable. They must monitor performance, schedule resources, predict and recover and/or work around failures, and exhibit predictable behaviors. They must be long-lived and therefore low-power.

To realize the above vision, there are several research challenges in the area of low-cost networking, disposable networks, application level routing, information management and robust access protocols and design.

This workshop focused on research challenges and problems in this new paradigm of a networked physical world. Specifically, topics in the following areas were covered:

- important unsolved problems (e.g., paging channels, network management)
- new directions
- low-maintenance networks (disposable networks)
- inter-disciplinary approaches
- promising models/modeling tools
- application drivers
- networking approaches.

Workshop: Mathematical Opportunities in Large Scale Network Dynamics
August 6 - 7, 2001
Location: Institute for Mathematics and its Applications (IMA), University of Minnesota
Organizers: John Doyle, California Institute of Technology; Andrew Odlyzko, AT&T; Ruth Williams, University of California, San Diego; Walter Willinger, AT&T
Attendance: 73

The Internet is an example of a large-scale, massively-distributed, and highly-interacting network of devices (i.e., computers) characterized by explosive growth, extreme heterogeneity any which way one looks, and unpredictable or even chaotic dynamic behavior. Measuring, modeling, simulating, and especially analyzing such networks pose completely new and immensely challenging mathematical problems, where scale, complexity, robustness, adaptivity, and dynamics play key roles and need to be faced head-on. Solving these problems can be expected to have profound implications for the efficient design and effective engineering, control, and management of future communication networks such as the next-generation Internet and wireless networks, or sensor networks (i.e., networks of massively-distributed, dynamic, and physically-embedded devices).

This workshop was a follow-up meeting to an April 28-29, 2000 BMS-NRC Workshop on "The Interface between the Mathematical Sciences and Three Areas of Computer Science: Network Traffic Modeling, Computer Vision,and Data Mining." It introduced and attracted mathematicians to the field and pointed out some of the most significant mathematical challenges associated with measuring, simulating, modeling, and analyzing large-scale, heterogeneous, and complex communication networks. In addition to a small number of tutorial-type and cutting-edge research talks by experts in the field, there was a panel on the evolution of communication networks -- with panelists representing the various (not necessarily consistent) perspectives of backbone providers, Internet service providers, and content distribution network providers -- followed by a discussion about how the mathematical sciences can provide a framework for formulating and (hopefully) solving key issues related to network evolution. Also there was a presentation by NSF and other federal agency program directors and a discussion to acquaint the participants with funding opportunities in Network Dynamics and to help them apply for grants in this emerging field.

This workshop brought together mathematicians/statisticians interested in (or involved in) networking research and networking experts, to assess mathematical challenges in the field and funding opportunities for interdisciplinary research teams. The participants came away with a clearer idea of the mathematical issues and funding opportunities in network modeling, and they made useful new contacts with researchers in other disciplines and in industry who are working on aspects of the same problems.

"Hot Topics" Workshop: Wireless Networks
August 8 - 10, 2001
Location: Institute for Mathematics and its Applications (IMA), University of Minnesota

Organizers: P.R. Kumar, University of Illinois, Urbana; Rajeev Agrawal, Motorola; Venkat Anantharam, University of California, Berkeley; Piyush Gupta, Lucent Technologies; Debasis Mitra, Lucent Technologies; David Tse, University of California, Berkeley
Attendance: 106

Over the past three decades we have been witnessing the confluence of communications and computation. This revolution has been enabled by the wiring together of computers. The next phase of this revolution appears to be the effort to network mobile users and embedded devices, and more generally to allow the formation of spontaneous ad-hoc networks. Over the next few years it is anticipated that the dominant access networks will be wireless. It is this topic that was the subject of this workshop.

There are several issues, covering theoretical mathematical and computational aspects, that arise in the design and analysis of wireless networks. Many layers of the protocol stack need to be properly designed. These involve coding, modulation, multiple access, power control, routing, congestion control, and transport, among others.

In designing these protocols, several fundamental mathematical issues arise. Many questions surround how one may best utilize the wireless medium, which is fundamentally a shared medium. Others concern how messages should be routed between nodes whose location may be changing, and what the dynamics of such adaptive distributed routing algorithms are. Yet others deal with how users should regulate their own traffic streams so as to serve the common good. Still more relate to what kind of service quality can be provided by wireless networks to their users. Answering these questions involves the use of fields such as probability theory, discrete mathematics, information theory, dynamical systems, etc.

This workshop brought together leading experts from both academia and industry. We hoped also to have engaged the mathematics community, as this is a fertile area for the future.

Workshop: Computational Issues in Game Theory and Mechanism Design
October 31 - November 2, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Vijay Vazirani, Georgia Tech; Noam Nisan, Hebrew University
Attendance: 133

The research agenda of computer science is undergoing significant changes due to the influence of the Internet. Together with the emergence of a host of new computational issues in mathematical economics, as well as electronic commerce, a new research agenda appears to be emerging. This area of research is collectively labeled under various titles, such as “Foundations of Electronic Commerce”, “Computational Economics”, or “Economic Mechanisms in Computation” and deals with various issues involving the interplay between computation, game-theory and economics.

This workshop not only summarized progress in this area and defined future directions for it, but also helped the interested but uninitiated, of which there seem many, understand the language, the basic principles and the major issues.

The workshop was structured around survey talks by leading experts on some key topics in this area. In addition to these survey talks, we had short research presentations as well as panels and informal networking-time. We invited many of the currently active researchers in related fields. About half of the participants came from the computer science community and half from game theory, economics and operations research.

Workshop: Internet and WWW Measurement, Mapping and Modeling
February 13 - 15, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: John Byers, Boston University; Danny Raz, Technion and Bell Labs; Yuval Shavitt, Tel-Aviv University and Bell Labs
Attendance: 118

The Internet is capturing a central role in the social and economic fabrics of the global structure. While it is growing at a remarkable rate, there is currently no means by which users or network planners can track this growth. Mapping the network, namely, taking a snapshot of its current status, can help applications to better utilize the network. Analyzing maps taken over long periods of time can help in understanding how the Internet evolves. Understanding the Internet structure and evolution can help in designing and constructing better applications, and in the deployment of new network level services. This workshop examined the Internet structure and the structure of its most widely-used application, the WWW, and examined tools, methods, and instrumentations designed to map and understand the Internet structure. In particular, we focused on the following issues:

- Internet and WWW structure modeling: empirical studies, mathematical models, topology generators.
- Tools for mapping and measuring the Internet and the WWW: discovery techniques, measurement techniques, measurement infrastructure, visualization.
- Effect of mapping and measurement on application performance: application-level routing, network-adaptive applications, group communication, virtual topology construction.

Working Group Meeting: Data Compression in Networks and Applications
March 18 - 20, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Adam Buchsbaum, AT&T Labs - Research; S. Muthukrishnan, AT&T Labs - Research and Rutgers University; Suleyman Cenk Sahinalp, Case Western University; Jim Storer, Brandeis University; Jeff Vitter, Duke University
Attendance: 53

Recent advances in compression span a wide range of applications. Homogeneous data sources invite specific compression methods, such as for Java bytecodes, XML data, WWW connectivity graphs, and tables of transaction data. WWW infrastructure also benefits from compression. Cache sharing proxies can exploit recent work on compressed Bloom filters. Search engines can extend the idea of sketches that work for text files to music data. Examples also abound in more heterogeneous domains, such as database compression, compression of biosequences and other biomedical data which are becoming of key importance in the context of telemedicine. Additionally, new general compression methods are always being developed, in particular those that allow indexing over compressed data or error resilience. The application of compression to new domains continues, e.g., the use of multicasting to reduce information communicated during parallel and distributed computing, and the installation of general compression methods deep in the network stack, allowing transparent, stream-level compression, independent of the application. Compression also inspires information theoretic tools for pattern discovery and classification, especially for biosequences.

This working group explored the role of data compression in all layers of data networks, from the physical layer to the application and services layers, and addressed foundational and applied issues, as in (but not limited to) the examples above.

Workshop: Cryptographic Protocols in Complex Environments
May 15 - 17, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Ran Canetti, IBM Watson Research Center; Ueli Maurer, ETH Zurich; Rafail Ostrovsky, Telcordia Technologies
Attendance: 75

Designing secure protocols for complex and unpredictable multi-user environments like the Internet is a non-trivial task. Challenges range from developing appropriate notions of security, to the design and analysis of efficient multi-party cryptographic protocols that satisfy these notions, to the secure implementation of these protocols. Recent years have seen considerable advancements in research towards design of secure multi-party protocols. These include dealing with the security implications of protocol composition, coming up with protocols with improved efficiency, and communication complexity and designing protocols for a variety of tasks such as electronic voting, auctions, private information retrieval, e-commerce, secure unicasting and multicasting.

This workshop brought together experts on topics related to the design of cryptographic multi-party protocols to present their on-going work, survey previous work and discuss future research directions -- with strong emphasis on the design, analysis, and implementation of protocols that maintain their security in complex adversarial environments and in particular in the Internet.

Workshop on Management and Processing of Data Streams
(In conjunction with ACM SIGMOD/PODS and FCRC 2003)
June 8, 2003
Location: San Diego, California
Organizers: S. Muthukrishnan, Rutgers University; Divesh Srivastava, AT&T Labs - Research
Attendance: 72

This workshop was part of the Federated Computing Research Conference (FCRC03). Measuring and monitoring complex, dynamic phenomena -- atmospheric conditions, stellar patterns, movement of financial markets, traffic evolution in telephone and internet communication infrastructures, usage of world wide web, email and newsgroups -- produces stream data, i.e., data that arrives as a series of "observations", often very rapidly. These observations generally have intricate relationships between them, are often grouped in multiple ways into logical events, and are naturally collected via distributed platforms.

Databases and applications are increasingly required to manipulate stream data. This presents many challenges which are just beginning to be addressed.

- Fundamental algorithmic models and results for data streams are being studied in the theoretical computer science community.
- Data stream management systems are being designed and prototyped in the database community.
- IP traffic analysis is a powerful application domain for data streams, and is being addressed in the networking community.
- Programming language and software support needed for managing data streams is being determined in the systems community.
- Custom and off-the-shelf hardware support is being developed in the networking and graphics communities.

FCRC'03 provided a valuable opportunity to engage this large and diverse community of interest in drawing up the specification of data stream management systems, understanding their power and limitations, as well as motivating the applications that will drive the development of data stream systems. This workshop provided a forum to discuss the progress on all aspects of managing stream data. The audience included researchers, practitioners and graduate students interested in data stream management from the database (SIGMOD/PODS) community, as well as from other member FCRC conferences.

DIMACS Tutorial on Applied Cryptography and Network Security
August 4 - 7, 2003
DIMACS Center, CoRE Building, Rutgers University
Organizer: Rebecca Wright, Stevens Institute of Technology
Attendance: 46

This tutorial was a crash course on cryptography and its applications to secure networking and electronic commerce. It was designed to provide background knowledge to researchers and graduate students who wish to participate in the Special Focus on Communication Security and Information Privacy, or just to get an introduction to some of the fundamental issues in this field. The tutorial consisted of lectures and hands-on activities. Topics included:

- cryptographic primitives and protocols: symmetric key cryptography, public key cryptography, authentication, and key exchange protocols.
- key management and access control: public key infrastructures and trust management.
- network security: snooping, spoofing, distributed denial of service attacks, SSL, SSH, IPsec.
- electronic commerce: electronic payments protocols, auctions.

In each area, we introduced the most important and relevant topics. Lectures included basic mathematics, tools, and techniques, information about the current state of the art, and suggestions for further reading for more advanced topics.

Other Project Activities

Monitoring Message Streams

This project, supported by the KDD program through ITIC funding, aims to improve on existing methods for monitoring huge streams of textualized communications to automatically identify clusters of messages relating to significant "events." This project is concentrating on retrospective or "supervised" event identification where the system begins with a set of some labeled documents and automatically associates a relevant new message with one or more of these pre-existing events. The premise remains that significant performance improvements in automatic processing of messages will result from innovations involving more than one component of the process.

The focus is on the following five components of automatic message processing: (1) **compression** of text to meet storage and processing limitations; (2) **representation** of text into a form amenable to computation and statistical analysis; (3) **a matching scheme** for computing similarity between documents in terms of the representation chosen; (4) **a learning method** building on a set of judged examples to determine the key characteristics of a document cluster or "event"; and (5) **a fusion scheme** that combines methods that are "sufficiently different" to yield improved detection and clustering of documents. The goal is to identify promising combinations of methods through careful exploration of a variety of tools. The work so far has included running some 5,000 complete experiments, and we have developed an efficient language for intercommunication and exchange of data. This will facilitate further systematic exploration of the space of possibilities, and combination of methods originating with different paradigms.

Seminar listings:

Text Categorization: A Review, David Madigan, July 11, 2002.
The Rutgers MMS Project: A Parallel Assault on Monitoring Message Streams, Paul Kantor, Professor of Library and Information Science, Rutgers University, July 18, 2002.
Estimation and Retrieval of Documents Matching Specified Patterns, Muthu Muthukrishnan, August 1, 2002.
Computer Vision Methods for Analyzing Humans, Dimitris Metaxas, August 15, 2002.
Content-Sensitive Fingerprinting and Dimension Reduction, Rafail Ostrovsky, August 29, 2002.
Text Streaming , Muthu Muthukrishnan, Rutgers CS, October 27, 2003
Rutgers Warning and Indication System Engineering Laboratory-The WISE Lab, Rick Mammone, CAIP, December 1, 2003
Sequential Decision Making Algorithms in Filtering, Michael Littman, Rutgers CS, December 18, 2003.
Dynamic thresholding for network monitoring, Diane Lambert, Bell Labs, March 5, 2004
Who Was the Author? Using Statistics to Analyze Literary Style, David Holmes, College of New Jersey, April 9, 2004
Large Margin Generative Models, Tony Jebara, Columbia University, May 7, 2004
Network algorithms for homeland security, or How to find a small hidden group in a large communication network, Mark Goldberg, RPI, September 27, 2004.

Author Identification: Identifying Real-Life Authors in Massive Document Collections

This project, supported by the KDD program through ITIC funding, aims to develop techniques for identifying authors in large collections of textual artifacts (e-mails, communiques, transcribed speech, etc.). Author identification presents a significantly greater challenge than conventional text classification. The number of classes (i.e., possible authors) is large and imprecisely determined, training data are scarce or absent, and authors may be attempting deception.

The project focus is on probabilistic models, some old and some new, that combine the disparate features necessary for author identification with task-specific knowledge in a variety of forms. The project focus is on highly scalable,

sparseness-inducing models than can handle hundreds of thousands or even millions of features. The KDD-supported "Monitoring Message Streams" (MMS) project had considerable success with such models.

Test corpora with known authorship and differing textual properties are used to tune and validate our algorithms. By starting with known authors and "hiding" them in a variety of ways, a variety of operational authorship problems can be simulated.

The plan of research is to establish a testbed and experimental framework that build upon our experience in the MMS project. The project will deliver research quality software in Java or C++, technical reports or papers, and definitive experimental results on specific corpora.

Knowledge Discovery: Mining Multilingual Resources Using Text Analytics

This project aims to improve the detection of new and significant events or changes within a topic of interest from a multilingual stream of messages by developing a novel, **automatic,** probabilistic system to extract entities, events and relations and use them to populate an ontology (Knowledge Representation) that can be used to identify significant new information. The KR is a unifying framework to represent acquired knowledge independent of the source including sources from different languages; the system will encode both reinforcing and contradicting information by conditioning probabilities on source id and source reliability. The extracted facts will be used to build a **probabilistic** KR (PKR) that represents the system's view (or knowledge) of the world using probabilistic relational models. This PKR has many uses but we will evaluate our progress in this project by comparing our results to the current state-of-the-art in discovering significant new relations/attribute values in newly received data by building a First Story Detection system, as described below.

By understanding the incoming text beyond simple bag of words models as is commonly done in today's text search and classification systems, the project expects to improve the accuracy of discovering new information in a specific subtopic or answering a user's question about some entities or events. Current systems have a high false alarm rate with a low detection rate: incoming messages appear to be new though they do not provide significant new information.

By analyzing documents in multiple languages that describe the same event, called **comparable documents**, the project has a rich set of facts that may reinforce each other or may be contradictory. The richness of a comparable corpus will require the introduction of a probabilistic model to the KR to handle the information fusion. The PKR will be able to incorporate conditioning the probability of a fact on its source, and other related facts of the same entity or related entities. The PKR will learn its parameters from the comparable corpus.

III. Project Findings

In this section, we use the abbreviation PM for "permanent member" of DIMACS. Where individuals have shifted affiliations, we usually use the convention of listing their affiliation at the time of the result described.

Special Focus on Computational Information Theory and Coding Selected Research Results

Efficient Codes for Digital Fingerprinting. Alexander Barg (Bell Labs and DIMACS) and Gregory Kabatiansky (Visitor, IPPI RAN, Moscow) studied efficient constructions of codes for digital fingerprinting. Relying on the ideas of their earlier work (DIMACS report 2001-52) they gave a first construction of codes with the identifiable parent property in which identification of a member of the coalition can be accomplished in time logarithmic in the total number of users of the system. Previously known constructions relied basically upon exhaustive search and thus had exponentially slower identification procedures.

Input-Output of Mutual Information of Multidimensional Channels. Slava Prelov (Visitor, Russian Academy of Sciences) and Sergio Verdu (PM, Princeton) obtained a formula for the second-order expansion of the input-output of mutual information of multidimensional channels as the signal-to-noise ratio goes to zero. While the additive noise is assumed to be Gaussian, very general classes of input and channel distributions were considered. As special

cases, these channel models include fading channels, channels with random parameters and channels with almost Gaussian noise. When the channel is unknown at the receiver, the second term in the asymptotic expansion depends not only on the covariance matrix of the input signal but also on the fourth mixed moments of its components. The second-order asymptotics of mutual information finds application in the analysis of the bandwidth-power tradeoff achieved by specific (not necessarily optimum) input signaling in the wideband regime.

Lowering the Bias in Quantum Coin Flipping Protocols. Postdoc Ashwin Nayak and Peter Shor (PM, AT&T Labs) studied bit-commitment-based quantum coin flipping protocols, proving a new bound of 1/16 for the bias of any such protocol. In the same work they analyzed a sequence of multi-round protocols that tries to overcome the drawbacks of previously-proposed protocols in order to lower the bias. By exhibiting an intricate cheating strategy they showed a lower bound of 1/4 on the bias and conjectured that this is the optimal bias for such protocols.

Performance of Expander Codes on a Binary Symmetric Channel. Alexander Barg (Bell Labs and DIMACS) and Gilles Zemor (Visitor, ENST-Paris) studied performance of expander codes on a binary symmetric channel. This research was initiated during a visit of Zemor to DIMACS in 2001 when he, together with Barg, showed that these codes reach channel capacity under iterative decoding, producing the first example of capacity-attaining linear-time decodable codes. This work was extended during a DIMACS visit of Z**J**mor in August 2002 when they introduced another family of expander codes tailored for transmission at rates in the immediate neighborhood of capacity. For this region they estimated the decrease rate (the error exponent) of error probability of decoding for randomized ensembles codes. The resulting estimate gives a substantial improvement of previous results for expander codes and some other explicit code families.

A Fast Algorithm for Generating Pairs of Codes with Prescribed Pairwise Hamming Distances and Coincidences. Vince Grolmusz (Visitor, Eotvos University, Hungary) developed a fast algorithm for generating pairs of q-ary codes with prescribed pairwise Hamming-distances and coincidences (for a letter s 0$\{0,1, \ldots, q\text{-}1\}$, the number of s-coincidences between codewords a and b is the number of letters s in the same positions both in a and b). The method is a generalization of a method for constructing set-systems with prescribed intersection sizes, where only the case q=2 and s=1 was examined. He also generated codes with prescribed k-wise coincidences and Hamming-distances.

Interaction in quantum communication and the complexity of set disjointness. One of the most intriguing facts about communication using quantum states is that these states cannot be used to transmit more classical bits than the number of qubits used, yet in some scenarios there are ways of conveying information with exponentially fewer qubits than possible classically. Moreover, these methods have a very simple structure---they involve only little interaction, few message exchanges between the communicating parties. Ashwin Nayak (DIMACS postdoc), Amnon Ta-Shma and David Zuckerman looked more closely at the ways in which information encoded in quantum states may be manipulated, and considered the question of whether every classical protocol may be transformed to a ``simpler'' quantum protocol---one that has similar efficiency. By a simpler protocol, we mean a protocol that uses fewer message exchanges. Nayak, Ta-Shma and Zuckerman showed that for any constant k, there is a problem such that its k+1 message classical communication complexity is exponentially smaller than its k message quantum communication complexity, thus answering the above question in the negative. This in particular proves a round hierarchy theorem for quantum communication complexity, and implies, via a simple reduction, an $\Omega(N^{1/k})$ lower bound for k message protocols for Set Disjointness (for constant k).

Digital fingerprinting codes. A. Barg (Bell Labs), G. R. Blakley, and G. Kabatiansky considered a general fingerprinting problem of digital data under which coalitions of users can alter or erase some bits in their copies in order to create an illegal copy. Each user is assigned a fingerprint which is a word in a fingerprinting code of size M (the total number of users) and length n. They found binary fingerprinting codes secure against size-t coalitions which enable the distributor (decoder) to recover at least one of the users from the coalition with probability of error $exp(\text{-}Omega(n))$ for $M=exp(Omega(n))$. This is an improvement over the best known schemes that provide the error probability no better than $exp(\text{-}Omega(n^{1/2}))$ and for this probability support at most $exp(O(n^{1/2}))$ users. The construction complexity of codes is polynomial in n. They also found versions of these constructions that afford poly(n)=poly log(M) identification algorithms of a member of the coalition, improving over the best previously known complexity of $Omega(M)$. For the case t=2 they constructed codes of exponential size with even stronger performance, namely, the distributor can either recover both users from the coalition with probability 1-$exp(Omega(n))$, or identifies one traitor with probability 1.

Special Focus on Computational Geometry and Applications Selected Research Results

Reconstructing Sets from Interpoint Distances. Paul Lemke (Troy U.), Steven Skiena (SUNY, Stony Brook) and Warren D. Smith (Visitor, DIMACS) studied the problem of reconstructing sets from interpoint distances. The general question addressed by their work was: Which point sets realize a given distance multiset? Interesting cases include the ``turnpike problem" where the points lie on a line, the ``beltway problem" where the points lie on a loop, and their multidimensional versions. These problems have applications in genetics and crystallography. The research of Lemke, Skiena and Smith gives improved combinatorial bounds for the turnpike and beltway problems. The research further presents a pseudo-polynomial time algorithm as well as a practical $O(2^n n \log n)$-time algorithm that find all solutions to the turnpike problem, and shows that certain other variants of the problem are NP-hard.

Polynomiography
Bahman Kalantari (PM, Computer Science, Rutgers) defines polynomiography as the art and science of visualization in approximation of zeros of complex polynomials, via fractal and non-fractal images created using the convergence properties of iteration functions. As polynomials are a fundamental class of functions in vetually every branch of science and mathematics, Kalantari expects polynomiography to find a wide range of scientific applications. From the scientific point of view it provides not only a tool for viewing polynomials, present in virtually every branch of science, but also a tool to discover new theorems. In his presentation "Can Polynomiography be useful in Computational Geometry?" at the DIMACS Workshop on Computation Geometry, November 15, 2002, Kalantari introduced polynomiography to the computational geometry community and subsequently the computer graphics community. His work eventually appeared on the cover of the Journal of Computer Graphics, August, 2004 and the cover of ISAMA-BRIDGES 2003 Conference Proceedings.

Dynamic computation of depth contours
Statisticians have recently developed the notion of *data depth* for non-parametric multivariate data analysis. This new concept provides center-outward orderings of points in Euclidean space of any dimension and leads to a new non-parametric multivariate statistical analysis in which no distributional assumptions are needed. A data depth measures how deep (or central) a given point x in R^d is relative to F, a probability distribution in R^d (assuming $\{X_1,.., X_n\}$ is a random sample from F) or relative to a given data cloud. Some examples of data depth are *Half-space Depth*, *Simplicial Depth*, the *Convex Hull Peeling Depth* and *Regression depth* (which is the depth of a hyperplane relative to a set of points). All of these depths are affine invariant: each depth value remains the same after the data are transformed by any affine transformation. Different notions of data depth capture different statistical characteristics of the underlying distribution. Depth contours, constructed by enclosing all points of depth *d* or higher, are especially powerful for visualizing and quantifying data. Simple (in many cases 2D) graphs can be used to visualize these parameters for the data set. The potential is enormous for analysis of massive data sets in such areas as quality control and aviation safety analysis, clinical data mining, biological imaging analysis, and statistical process control. M. Burr, E. Rafalin. D. L. Souvaine (Computer Science Department, Tufts University, Medford, MA) identified two competing methods for computation of depth contours and analyzed their computational features. They created the first algorithm for dynamically computing the two methods of half-space depth contours.

Special Focus on Next Generation Networks Technologies and Applications Selected Research Results

Signature Computation on Insecure Devices such as Mobile Phones. Visitors Yevgenij Dodis (NYU), Jonathan Katz (Visitor, U-Marryland), Shouhuai Xu (Visitor UC-Irvine) and Moti Yung (Columbia) studied signature computation on insecure devices such as mobile phones, operating in an environment where the private (signing) key is likely to be exposed. Strong key-insulated signature schemes are one way to mitigate the damage done when this occurs. In the key-insulated model, the secret key stored on an insecure device is refreshed at discrete time periods via interaction with a physically-secure device which stores a ``master key". All signing is still done by the insecure device, and the public key remains fixed throughout the lifetime of the protocol. In a strong (t, N)-key-insulated scheme, an adversary who compromises the insecure device and obtains secret keys for up to t periods is unable to forge signatures for any of the remaining N-t periods. Furthermore, the physically-secure device (or an adversary who compromises only this device) is unable to forge signatures for *any* time period. The work of Dodis, Katz and Xu presents constructions of strong key-insulated signature schemes based on a variety of assumptions. It gives an

improved construction of a strong (t, N)-signature scheme whose security may be based on the discrete logarithm assumption in the random oracle model. This construction offers faster signing and verification than the generic construction, at the expense of $O(t)$ key update time and key length. These researchers also construct strong (N-1,N)-key-insulated schemes based on any ``trapdoor signature scheme''; their resulting construction in fact serves as an identity-based signature scheme as well. This leads to very efficient solutions based on, e.g., the RSA assumption in the random oracle model.

Random Graph Models of Sensor Networks. S. Muthu Muthukrishnan (PM, Rutgers) and Gopal Pandurangan (Visitor, Purdue) studied random graph problems motivated by sensor networks. Sensor networks, a technology likely to have a significant impact on our lives in the future, are modeled by random graphs $G(n,r)$ where the n vertices are represented by random points on the unit square. A pair of vertices at most r apart is connected by an edge. Muthukrishnan and Pandurangan studied three properties of such networks: coverage (how many nodes are needed to cover the entire area), high quality paths (those with a small stretch, i.e., the ratio of the shortest path between the nodes in the graph to the distance in the plane), and disjoint paths (the number of vertex-disjoint paths between a given pair of nodes). As a result of this work, bounds were established on r as a function of n for each of these properties. The bounds show that with each sensor (node) transmitting at power slightly greater than needed to ensure connectivity of the underlying random network, (a) the network will nearly cover the entire area in the plane, (b) shortest paths between any two nodes are high-quality paths, (c) there are many vertex-disjoint paths between any two nodes, all of high quality.

Mix Networks. Markus Jacobsson (PM, Bell Labs and, later, RSA Technologies) proposed and improved techniques for so-called mix networks. These are constructions useful for Internet voting; privacy preserving browsing; and controlled anonymity for payment transactions. The new results allow for more efficient and more versatile mix networks.

Collaboration between Low-Power Mobile Devices. Markus Jacobsson (Bell Labs and, later, RSA Technologies) also worked in the area of low-power mobile devices, such as military sensors, cellular phones, and radio-frequency identification (RFID) tags. The work has involved means for promoting collaboration between selfish devices (who want to maximize their own benefits, but may not care about the benefit of their peers). It has also involved work on providing more efficient computational methods for traversing one-way hash chains, a problem that often comes up in work relating to low-power mobile devices, due to its application for authentication.

The Economics of Multicasting. A supplier of multicast information services will often be faced with the following problem: Broadcasting to the whole customer base (including non-paying customers) is cheaper than multicasting only to the paying customers. However, broadcasting discourages potential customers from paying. The result is an economic game in which the supplier tries to maximize profit in the face of rational, but not omniscient, behavior by customers. Yuval Shavitt (PM, Bell Labs), Peter Winkler (DIMACS member, Bell Labs), and Avishai Wool (DIMACS member, Bell Labs) built a model for such environments in which the supplier's basic strategy is to broadcast every service for which the fraction of subscribed customers exceeds some threshold. They modeled the customers' behavior, taking into account the possibility of customers receiving services for free and found, under some mild assumptions on the supplier's cost structure, the optimal setting of the supplier's broadcast threshold. In all the examples they studied, their model predicts that the supplier's profits will be maximized if the supplier's broadcast threshold is set below 100%: The loss in revenue due to customers subscribing to fewer services is offset by the cost savings made possible by broadcasting the most popular services to all customers. The model should be of value to a supplier in devising a multicast/broadcast strategy and so should be the conclusion that broadcasting when subscriptions are sufficiently high is likely to be the approach of choice in maximizing profits.

Reducing the Servers' Computation in Private Information Retrieval. A Private Information Retrieval (PIR) protocol allows a user to retrieve a data item of his choice from a remote database while the database learns nothing about which item is retrieved. A trivial solution to this problem is to have the entire database contents sent to the user. However, the communication complexity of this solution is linear in the size of the data, which may be prohibitively large. Previous work on this problem has focused on minimizing the communication complexity of PIR, possibly by replicating the data in several non-communicating servers. However, in all previous solutions the servers' computation for each retrieval is at least linear in the size of the entire database, even if the user requires just one bit. Amos Beimel, Tal Malkin, and Yuval Ishai (DIMACS postdoc) initiated the study of the computational complexity of PIR. They showed that off-line preprocessing can help to significantly reduce the servers' work in replying to

each of an unlimited number of on-line queries. They also proved lower bounds on the work of the servers as a function of the number of extra bits that they are allowed to store at the end of the preprocessing stage.

Interdomain Routing in the Internet. Interdomain routing in the Internet is coordinated by the Border Gateway Protocol (BGP). BGP allows each autonomous system (AS) to apply diverse local policies for selecting routes and propagating reachability information to others. This flexibility is crucial in the decentralized and commercial environment of today's Internet. However, BGP permits AS'es to have conflicting policies that can lead to route divergence, which can cause an AS to oscillate between two or more routes to a particular destination. Lixin Gao (DIMACS Visitor) and Jennifer Rexford (DIMACS Member, AT&T Labs) proposed a set of guidelines for an AS to follow in setting its routing policies, without requiring coordination with other AS'es. Their approach exploits the Internet's hierarchical structure and the commercial relationships between AS'es to impose a partial order on the set of routes to each destination. The guidelines conform to conventional traffic-engineering practices of ISP's, and provide each AS with significant flexibility in selecting its local policies. Furthermore, the guidelines ensure route convergence even under changes in the topology and routing policies. They developed a formal model of BGP that includes AS'es with multiple BGP speakers, both interior BGP (iBGP) and exterior BGP (eBGP), and additional BGP attributes. Drawing on this model, they proved that following their proposed policy guidelines guarantee route convergence. Their approach has significant practical value since it preserves the ability of each AS to apply complex local policies without divulging its BGP configurations to others.

Approximate Nearest Neighbors and Sequence Comparison with Block Operations. Let D be a database of n sequences. We would like to preprocess D so that given any on-line query sequence Q, we can quickly find a sequence S in D for which $d(S,Q) \leq d(S,T)$ for any other sequence T in D. Here, $d(S,Q)$ denotes the distance between sequences S and Q, defined to be the minimum number of edit operations needed to transform one to another. These operations correspond to the notion of similarity between sequences that we wish to capture in a given application. Natural edit operations include character edits (inserts, replacements, deletes, etc.), block edits (moves, copies, deletes, reversals), and block numerical transformations (scaling by an additive or a multiplicative constant). S. Muthukrishnan (PM, AT&T Labs) and Suleyman Cenk Sahinalp (DIMACS Visitor) found the first known efficient algorithm for ``approximate" nearest neighbor search for sequences with preprocessing time and space polynomial in the size of D and query time near-linear in the size of Q. They also found an algorithm for exactly computing the distance between two sequences when edit operations of the type character replacements and block reversals are allowed. The time and space requirements of the algorithm are near linear; previously known approaches take at least quadratic time.

An Algebraic Approach to Internet Mapping. Distance estimation is important to many Internet applications, most notably for a WWW client that needs to select a server among several potential candidates. Current approaches to distance (i.e., time delay) estimation in the Internet are based on placing tracer stations in key locations and conducting measurements between them. The tracers construct an approximated map of the Internet after processing the information obtained from these measurements. Yuval Shavitt (PM, Bell Labs), Xiadong Sun (DIMACS graduate student, Rutgers), Avishai Wool (PM, Bell Labs), and Bulent Yener (PM, Bell Labs) developed a novel algorithm, based on algebraic tools, that computes additional distances that are not explicitly measured. As such, the algorithm extracts more information from the same amount of measurement data. The algorithm has several practical impacts. First, it can reduce the number of tracers and measurements without sacrificing information. Second, it is able to compute distance estimates between locations where tracers cannot be placed. This is especially important when unidirectional measurements are conducted, since such measurements require specialized equipment which cannot be placed everywhere. To evaluate the algorithm, it was tested both on randomly generated topologies and on real Internet measurements. The results show that the algorithm computes up to 50-200% additional distances beyond the basic tracer-to-tracer measurements.

A Scalable Distributed QoS Multicast Routing Protocol. Many Internet multicast applications such as teleconferencing and remote diagnosis have Quality-of-Service (QoS) requirements. Such requirements can be additive (end-to-end delay), multiplicative (loss rate) or with a bottleneck nature (bandwidth). For these applications, QoS multicast routing protocols are important in enabling new receivers to join a multicast group. However, current routing protocols are either too restrictive in their search for a feasible path between a new receiver and the multicast tree, or burden the network with excessive overhead. Shigang Chen and Yuval Shavitt (PM, Bell Labs) developed S-QMRP, a new Stateless QoS Multicast Routing Protocol that supports all three QoS requirement types. S-QMRP is scalable because it has very small communication overhead and requires no state

outside the multicast tree; yet, it retains a high success probability. S-QMRP achieves the favorable tradeoff between routing performance and overhead by carefully selecting the network subgraph in which it conducts the search for a path that can support the QoS requirement, and by auto-tuning the selection according to the current network conditions. S-QMRP does not require any global network state to be maintained and can operate on top of any unicast routing protocol. Extensive simulation has shown that S-QMRP performs better than previously suggested protocols.

Online Scheduling with Release Times and Set-ups. Srikrishnan Divakaran (Rutgers Graduate Student) and Michael Saks (PM, Rutgers) studied the problem of online scheduling of n independent jobs with release times and sequence independent set-ups on a single machine system with the objective of minimizing the maximum flow time. They found an $O(1)$-competitive online algorithm and proved the *NP*-completeness of the offline problem.

Approximation Algorithms for Offline scheduling with Set-ups. Srikrishnan Divakaran (Graduate Student, Rutgers) and Michael Saks (PM, Rutgers) found new approximation results for the offline problem of scheduling n independent jobs with sequence independent set-ups and common release times on a single machine system for the following optimality criteria: total weighted completion time and maximum lateness. For the total weighted completion time criterion, they found a *2*-approximation algorithm, the first known polynomial time constant approximation algorithm for this problem. For this criterion, when the number of job types is arbitrary, the computational complexity is open and prior to their results there were no known polynomial time constant approximation algorithms even when all jobs have unit weight and all job types have unit set-up times. For the maximum lateness criterion, they found an algorithm that obtains an optimal solution in a relaxed framework where their algorithm is given strictly more resources (i.e. a machine with higher processing speed) than the optimal offline algorithm to which it is compared. For this criterion, when the number of job types is arbitrary, this problem is known to be *NP*-hard even when there are two jobs in each type or when all job types have unit set-up times, there are at most three jobs in each type and there are at most three distinct due dates. Also, for this criterion, when the number of job types and the due dates are arbitrary, there are no constant approximation algorithms unless $P=NP$.

QuickSAND: Quick Summary and Analysis of Network Data. Monitoring and analyzing traffic data generated from large ISP networks imposes challenges both at the data gathering phase as well as the data analysis itself. Still, both tasks are crucial for responding to day to day challenges of engineering large networks with thousands of customers. Anna C. Gilbert (PM, AT&T Labs), Yannis Kotidis (visitor), S. Muthukrishnan (PM, AT&T Labs), and Martin J. Strauss (DIMACS member, AT&T Labs) built on the premise that approximation is a necessary evil of handling massive datasets such as network data. They proposed building compact summaries of the traffic data called sketches at distributed network elements and centers. These sketches are able to respond well to queries that seek features that stand out of the data. They call such features ``heavy hitters.'' They showed how to use sketches to answer aggregate and trend-related queries and identify heavy hitters. This may be used for exploratory data analysis of network operations interest. They supported their proposal by experimentally studying AT&T WorldNet data and performing a feasibility study on the Cisco NetFlow data collected at several routers.

On Performance Prediction of Cellular Telephone Networks. Linchun Gao (Graduate Student, Rutgers) and Andras Prekopa (PM, Rutgers) studied large cellular mobile networks in which the arrival rates of calls for different cells are different. The lower and upper bound of how much the given number of channels would satisfy the arriving calls were determined.

Networked Cryptographic Devices Resilient to Capture. Philip MacKenzie (DIMACS member, Bell Labs) and Michael K. Reiter (PM, Bell Labs) found a simple technique by which a device that performs private key operations (signatures or decryptions) in networked applications, and whose local private key is activated with a password or PIN, can be immunized to offline dictionary attacks in case the device is captured. Their techniques do not assume tamper resistance of the device, but rather exploit the networked nature of the device, in that the device's private key operations are performed using a simple interaction with a remote server. This server, however, is untrusted---its compromise does not reduce the security of the device's private key unless the device is also captured---and need not have a prior relationship with the device. They extended this approach with support for ``key disabling,'' by which the rightful owner of a stolen device can disable the device's private key even if the attacker already knows the user's password.

Compositional Reasoning. Traditional methods for verifying software systems, like model checking, systematically step through the global states of the system while checking various properties at each stage. Such state exploration methods, however, run into the well-known problem of state-space explosion, which severely limits the size of systems that can be successfully analyzed. One reason for the increase in the size of the state space is because the number of global states of a systems grows as the product of the number of states of the individual concurrently executing sub-components. So a system consisting of *n* parties has size that is exponential in the size of the individual systems. Thus, in order to verify such systems compositionally or hierarchically in which properties of sub-components (that are smaller in size) are verified in isolation and these ``local'' properties are composed to derive the property that the global system consisting of these components satisfies. Composing these properties require circular reasoning principles in which properties of other components need to be assumed in proving the properties of each individual component; the inherent circularity makes deriving sound rules a difficult and subtle problem. A number of such circular assume-guarantee rules have been proposed for different concurrency models and different forms of property specifications. Postdoctoral Fellow Mahesh Viswanathan formulated a framework that unifies and extends these results. He defined an assume-guarantee semantics for properties expressible as least or greatest fixed points, and a circular compositional rule that is sound with respect to this semantics. The utility of this rule is demonstrated by applying to some specific contexts where he found alternate proofs of the strongest reasoning principles known in those contexts, and furthermore derived new rules for more expressive classes of properties than had previously been considered.

Monitoring Message Streams

The following have been acomplished:

1. Achieved a 10-fold reduction in space required for management of large numbers of profiles with approximately 10% reduction in filtering effectiveness. This was done through speed-up of k-nearest-neighbor methods, taking advantage of a unique partnership between theoreticians and practitioners.

2. With the same k-NN methods, performed approximate matching at a speed 10 to 100 times faster with only 2-10% loss in effectiveness (macro-F1).

3. Achieved a state of the art performance on effectiveness with Bayesian feature selection and modeling using Bayesian Logistic regression with Laplace Priors that at the same time uses 200 to 2000 times less space than competing methods.

4. Developed and prototyped a novel online Bayesian algorithm for Bayesian Binary regression.

5. Developed a new combinatorial analogue of Principal Components Analysis. In experiments with 27,000 documents and 47,000 words, we obtained a reduction in dimensionality by a factor of 30 (i.e. 1,500 from 47,000) compared to the original space, with performance changes of 1-2% in macro-average F1. These changes were consistent across 1-NN and SVM classifiers, as well as 16 different data normalization schemes. (We know of no implementations of PCA that can handle data of this size.)

6. Conducted more than 5,000 specific experiments to explore the effectiveness of several methods and many parameterizations. These experiments have demonstrated that the performance of any method is determined more by the specific parameter settings than by the overall implementation choices and the "philosophy" of the method.

7. In a "streaming data context," developed rapid methods for a number of monitoring applications such as finding changing trends, outliers and deviants, unique items, rare events, heavy hitters, etc.

8. Established a parameterized design space for the important class of Rocchio Adaptive Filters making possible the systematic investigation of design choices and results such as those cited in (6).

9. Established several families of rigorous, decision-theoretic models to explore the interplay of key aspects of message filtering, including optimal selection of messages for human reading.

10. Developed and delivered software for a variety of filtering models including Adaptive Filtering: Rocchio and Centroid methods; Batch Filtering: Bayesian Feature Selection and Filtering, k-NN methods, Combinatorial variants of Principal Component Analysis and Support Vector Machines; also Homotopic Linkage of Different Rocchio Methods; and Fusion of Results from Multiple Systems.

k-NN:

We are investigating applying more sophisticated learning algorithms such as "reverse *k*NN learning" to the neighbors once they are found. This is called "local learning" and has been used in robotics and other areas, but has not been much explored in text classification. Extensive experiments have been conducted, and we are now also working specifically on greatly speeding up the kNN, so that a very large range of experiments becomes possible.

Bayesian methods:

We have developed an online algorithm that updates itself as new data arrive, implemented it for small scale problems and compared it with a published alternative. We will investigate extension of this work to very high dimensional applications.

All our work so far focuses on binary classification – each document either is or is not in each category. The reality is that in most of our target applications, documents can simultaneously belong to many stochastically dependent categories. We have designed three algorithms that extend our binary model to the multilabel case. We will implement and experiment with one or more of these algorithms.

Our target applications feature a paucity of labeled training data. We will explore highly informative Bayesian priors drawing on information in category descriptions. These have the potential to provide dramatic improvements on predictive accuracy with tiny training datasets.

We have completed and publicly released our Bayesian BBR software.
We have extended the Binary regression algorithms and software to the polychotomous case (i.e., multiclass classification)

We are applying our methods that extend our binary model to the multilabel case to the TREC Genomic Track challenges. Initial results show that BBR works as well as state-of-the-art SVM methods, and we have not yet exploited prior knowledge in these models.

Historic Data Analysis:

Previously we designed "sketches" that were more compact than all the others known thus far, and used them to analyze streams of data for heavy hitters, quantiles etc. Now, we have designed methods that rely on these sketches to summarize stream data as it goes by, so we can do historic analysis of the data from the summaries. For example, we can determine the words whose frequency of occurrence changed the most from one month to another, one week to another, etc.

It is a great challenge to use only small summaries of historic data to do historic, posteriori and root cause analysis to see if a current emerging phenomenon occurred any time in the past. We will develop for historic and posterior analysis the architecture that has been proposed in our earlier work. This involves implementing known methods for summarizing data streams, as well as developing novel methods that are suitable for historic analyses of interest.

We have used sketches to analyze streams of data so we can do historic analysis from the summaries. We can determine the words whose frequency of occurrence changed the most from one month to another, one week to another, etc.

Regarding the use of sketches to retrospectively identify individuals who have written a small number of messages on a topic that has later proved to be important, we are formulating rigorously the questions of how large the data structures must be to achieve acceptable levels of performance. We expect to have specific numerical results.

Fusion:

Fusion will be more effective when it is applied earlier in the chain of components to combine methods of representation, matching, compression, and learning, and we will explore approaches to accomplish this.

We are working on low-level methods of fusion. As an example, fusion of text based methods (bag of words) and word length measures has produced highly accurate classification of the disputed Federalist papers, using no more than 3 features in a fused set.

Decision-theoretic Adaptive Filtering

A key technical challenge is that if the filter fails to present any documents, then no feedback will be forthcoming and the filter will not improve over time. Our models for this process show considerable promise and we will extend both in various ways.

We will develop extensions in which several documents are batched before deciding which ones to present, making use of “Bandit Theorems.”

We will also marry our online sparse Bayesian model with the decision-theoretic framework.

We have conducted numerical experiments and theoretical investigations on the problem of "exploration". Preliminary results, using data developed at Carnegie Mellon, using the Lemur system, suggest that we can model the rate of convergence of key parameters towards their best values. These models are essential in optimizing the trade-off between exploration and exploitation.

We have studied a family of exploration methods. They sometimes helped and sometimes hurt. It was a net improvement, but a small one. We were learning a classifier and using uncertainty in the learned classifier to drive exploration. We can prove that under certain assumptions this leads to an algorithm with formal bounds on performance.

We have been refocusing on studying an idea that was more successful in TREC, namely adaptively setting the score threshold for a classifier. At TREC, several ad hoc schemes have done well and we'd like to provide a scheme with formal guarantees that does at least as well as the existing approaches.

Author Identification: Identifying Real-Life Authors in Massive Document Collections

Algorithms for Sparse Bayesian Multinomial regression have been developed and implemented. This BBR software, developed as part of the MMS project, deals with binary classification. Standard author attribution problems typically involve more than two authors so we extended the BBR algorithms and software to deal with multiclass (i.e., 1-of-K or polychotomous) classification. Our sparseness-inducing approach gives increasing improvements in classifier efficiency as the number of classes grows.

Initial experiments with RCV-1 corpus using function word and part-of-speech tag features have been completed. These include 1-versus-all experiments, 1-versus-1 experiments, and multiclass experiments.

Initial general-purpose Perl code for extraction of topic-free features from general corpora has been developed.

An extensive corpus of newsgroup postings, including 1.2 million messages by 82,000 authors from 63 usenet newsgroups has been amassed.

Though the initial results from the experiments for the “document pair” problem are disappointing, the initial results from the experiments for the “odd-man-out” problem are more promising.

Project presentations were made at the Symposium on Intelligence and Security Informatics (ISI-2004), the DIMACS Workshop on the Mathematics of Web Search in Italy, and the Chicago Chapter of the American Statistical Association, and KDD 2004 in Seattle.

Data from this project has been shared with the Smyth KDD project.

Knowledge Discovery: Mining Multilingual Resources Using Text Analytics

The following have been completed:

Complete acquisition of a total of 100 million pages of Arabic, Mandarin, and English text,
Finalize domain design as a base domain model of entities and relations using NETL2 KR,
Build English extractors for base domain model,
Build Chinese language extractor for the same base domain model,
Build automatic knowledge acquisition for selected set of relations domain from selected web pages,
Explore and define a probabilistic model for a selected subset of domain; estimate an initial probabilistic model to demonstrate feasibility of probabilistic approach for information fusion.

The following milestones are next:

Adapt entity/relation detectors for foreign language.
Automatic population of NETL2 ontology using two streams.
Estimate probabilistic relational model using extracted facts for small base ontology.

We anticipate that the resulting base domain model can be used in applications such as improving First Story Detection systems.

IV. Project Training/Development

Small research projects by graduate students were part of the project, aimed at involving them in all of the topics of the Special Focus. In addition, funds were set aside at each workshop for support of non-local students and non-local students were also invited to be in residence as visitors for periods of a week to several months. Here we list the research of the local students. Research of non-local students is included in description of research results in the "Findings" section.

Graduate Students Supported and Their Research Topics

Computational Information Theory and Coding Special Focus

Eva Curry, Mathematics, Rutgers University, Summer 2002: Characterization of Low-pass in *n* dimensions.

Navin Goyal, Computer Science, Rutgers University, Summer 2002: Parent Identifying Codes.

Computational Geometry and Applications Special Focus

Xiaomin Chen, Computer Science, Rutgers University, Summer 2004: Packing a set of circles into the plane so that the minimum spanning tree connecting them is minimized.

Xiaomin Chen, Computer Science, Rutgers University, Winter 03/04: Generalize some classical results in geometry.

Xiaoling Hou, RUTCOR, Rutgers University, Summer 2004: Two-stage data selection for the problem involving very large data sets.

Xiaolei Huang, Computer Science, Rutgers University, Winter 02/03: (1) Computational geometry algorithms that reconstruct a surface model from sample points, minimizing the number of required sample

points and the deviation from the actual object as well as the required computing time, (2) Triangulation, (3) Aligning Point Sets.

Aaron Lauve, Mathematics, Rutgers University, Summer 2004: Combining the theory of automata and quasideterminants to determine explicit characterizations of eigenvalues and eigenvectors for quaternionic matrices.

Chan Su Lee, CAIP, Rutgers University, Summer 2002: The Physically Accurate Haptic Rendering of Elastic Objects for a Haptic Glove.

Tongyin Liu, RUTCOR, Rutgers University, Summer 2004: Convex hull of *P* level efficient points in a stochastic programming problem; Summer 2002: 1) Efficient Representations of Graphs and 2) Drawing of Planar Graphs on Higher Surfaces; and Summer 2001: Research Thrust: 1) GraphTheory & Algorithms and 2) Models for communication network consisting of heterogeneous groupware; and Winter 00/01: Computational geometry and applications.

Ying Liu, RUTCOR, Rutgers University, Summer 2003: Planar case of the maximum box and related problems; Winter 02/03: Bounds on the size of the binary space partitioning of sets of axis-parallel rectangles and their applications; Summer 2002: Bounds on the size of the binary space partitioning of sets of axis-parallel rectangles and their applications; Winter 01/02: Computing the maximum monochromatic rectangle; and Summer 2001: Pattern selections in data analysis.

Mikhail Nediak, RUTCOR, Rutgers University, Winter 01/02: Deep convexity cut generation for 0-1 MIP.

Michael Neiman, Mathematics, Rutgers University, Summer 2004: Phase transitions in discrete structures.

Bin Tian, Computer Science, Rutgers University, Summer 2004: Complexity of certain classes of planar circuits.

Vince Vatter, Mathematics, Rutgers University, Summer 2004: Restricted permutations.

Lei Wang, Computer Science, Rutgers University, Winter 03/04: Determine the upper bound on the number of *k*-sets

Igor Zverovich, RUTCOR, Rutgers University, Summer 2004: Generalized threshold graphs and Winter 03/04: Computer-generated conjectures from graph theoretic and chemical databases.

Next Generation Networks Technologies and Applications Special Focus

Gabriela Alexe, RUTCOR, Rutgers University, Winter 01/02: A consensus-type algorithm for spanned pattern generation.

Sorin Alexe, RUTCOR, Rutgers University, Winter 01/02: Feature selection in data analysis.

Xuhui Ao, Computer Science, Rutgers University, Summer 2002: A Scalable Mechanism for Interorganization Collaboration Access Control.

Samir Goel, Computer Science, Rutgers University, Summer 2003: Sensors on wheels - towards a zero-infrastructure solution for intelligent transportation systems.

Xiaoling Hou, RUTCOR, Rutgers University, Winter 02/03: Data selection for the problem involving very large data sets and Summer 2002: Data selection for the problem involving very large data sets.

Tongyin Liu, RUTCOR, Rutgers University, Winter 03/04: *P*-level efficient points of design of stochastic networks and Winter 02/03: Efficient and reliable network design.

Marcelo Mydlarz, Computer Science, Rutgers University, Summer 2003: Traveling salesman problem; Winter 02/03: Maximum special 3d-matching; Summer 2002: Baseball matrices; and Winter 01/02: A baseball problem.

Swati Sharma, CAIP, Rutgers University, Summer 2002: How to determine the most frequent items in a data stream and how to find the most rare items in a data stream.

Anthony Ian Wirth, Computer Science, Princeton University, Summer 2002: Towards a constant factor approximation algorithm for the asymmetric *k*-center problem.

Igor Zverovich, RUTCOR, Rutgers University, Summer 2002: Computer-generated conjecture on Ramseian partitions of graphs.

Monitoring Message Streams

Andrei Anghelescu, Rutgers, Computer Science, Monitoring Message Streams

Dmitriy Fradkin, Rutgers, Computer Science, Monitoring Message Streams

Aynur Dayanik, Rutgers, Computer Science, Monitoring Message Streams

Suhrid Balakrishnan, Rutgers, Computer Science, Monitoring Message Streams

Miscellaneous Research Topics

Diogo Andrade, RUTCOR, Rutgers University Winter 03/04: Establish some necessary and some sufficient conditions for the CIS-property to hold and Winter 02/03: Verify for the circulants of two recent conjectures of graph theory: The Hayward-Reed (2001) conjecture on even-hole-free graphs and the Hoang-McDiarmid (2001) conjecture on odd-hole-free graphs.

Tiberius Bonates, RUTCOR, Rutgers University, Summer 2004: LAD models based on maximum patterns. Winter 02/03: Solving 0/1 integer linear programs by a boolean equations approach.

Jeffrey Burdges, Mathematics, Rutgers University, Summer 2003: Simple groups of finite Morley rank.

Corina Calinescu, Mathematics, Rutgers University, Winter 03/04: Combinatorial identities and vertex operator algebra.

Jie Chen, Computer Science, Princeton University, Summer 2003: Objective functions and data characteristics in clustering algorithms.

Laura Ciobanu, Mathematics, Rutgers University, Summer 2003: The complexity of the endomorphism problem for free groups. Summer 2004: On the generic complexity of the endomorphism problem for free groups.

Khaled Elbassioni, Computer Science, Rutgers University, Winter 01/02: Travel support to 7th Intl. Symposium on Artificial Intelligence and Mathematics; "Learning monotone binary functions in products of lattices and its applications in data mining".

Cem Iyigun, RUTCOR, Rutgers University, Summer 2003: Data analysis and classification with singular or ill-conditional covariance matrices.

Pandurang Kamat, Computer Science, Rutgers University (Special Focus on Communication Security and Information Privacy), Summer 2003: Development of an agent-based distributed security infrastructure in sensor networks.

Marcin Kaminski, RUTCOR, Rutgers University, Winter 02/03: Calculus of Boolean function and graph theory.

Klay Kruczak, Mathematics, Rutgers University, Summer 2003: Positional game theory.

Aaron Lauve, Mathematics, Rutgers University, Summer 2003: Research for eigenvalues and eigenvectors of quaternionic matrices using the theory of quasideterminants.

Miguel Mosteiro, Computer Science, Rutgers University, Summer 2004: To perform a competitive analysis of the dirty page aggressive write back model both theoretically and empirically. Summer 2003: Cache Oblivious B-trees are not worse than their cache aware counterpart under prefetching and TLB techniques.

David Nacin, Mathematics, Rutgers University, Summer 2004: Non-commutative combinatorics.

Sasa Radamirovic, Mathematics, Rutgers University, Summer 2004: Birch and Swinnerton-Dyer conjecture over function fields.

Eric Sundberg, Mathematics, Rutgers University, Summer 2003: Positional game theory.

Nicholas Weininger, Mathematics, Rutgers University, Summer 2003: Investigation of correlation and monotonicity properties of Ising models.

Igor Zverovich, RUTCOR, Rutgers University, Summer 2003: Computer generated conjectures related to perfect graphs.

Undergraduate Students and Their Research Projects

Author Identification: Identifying Real-Life Authors in Massive Document Collections Participants

Diana Michalak, UC Berkeley, REU student, Authorship Identification
Ross Sowell, University of the South, REU student, Authorship Identification

V. Outreach Activities

Special Focus visitors, graduate students, and senior faculty were available to interact with 2- and 4-year college faculty in the DIMACS "Reconnect" program, and with high school teachers in the DIMACS Connect Institute and the DIMACS Bio-Math Connect Institute.

VI. Papers/Books/Internet

Books and One-Time Publications

Special Focus on Computational Information Theory and Coding

Multiantenna Channels: Capacity, Coding and Signal Processing, Editors: Gerard J. Foschini and Sergio Verdú, American Mathematical Society, DIMACS, Volume 62, 2003.

Advances in Network Information Theory, Editors: Piyush Gupta, Gerhard Kramer, and Adriaan J. van Wijnggaarden, American Mathematical Society, DIMACS, Volume 66, 2004.

Algebraic Coding Theory and Information Theory, Editors: Alexei Ashikhmin, Alexander Barg, and Iwan Duursma, American Mathematical Society, DIMACS, to appear.

Theoretical Advances in Information Recording, Editors: Adriaan van Wijngaarden, Emina Soljanin, Paul Siegel, and Bane Vasic, American Mathematical Society, DIMACS, in preparation.

Special Focus on Computational Geometry and Applications

Geometric and Algorithmic Aspects of Computer-Aided Design and Manufacturing, Editors: Ravi Janardan, Michiel Smid, and Debasish Dutta, American Mathematical Society, DIMACS, to appear.

Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, Editors: Regina Liu, Robert Serfling, Diane Souvaine and Yehuda Vardi, American Mathematical Society, DIMACS, in preparation.

Special Focus on Next Generation Networks Technologies and Applications

In 2002 a special issue of the ACM Journal of Experimental Algorithmics was devoted to selected papers from the workshop **Workshop on Algorithm Engineering and Experiments (ALENEX 2001)**.

Security Analysis of Protocols, Editors: Ran Canetti and John Mitchell, American Mathematical Society, DIMACS, in preparation.

Other Projects

Graphs and Discovery, Editors: Siemion Fajtlowicz, Patrick Fowler, Pierre Hansen, Mel Janowitz, and Pierre Hansen, American Mathematical Society, DIMACS, to appear.

Computer Generated Conjectures, Editors: Patrick Fowler and Pierre Hansen, American Mathematical Society, DIMACS, in preparation.

Data Mining in Epidemiology, Editors: James Abello and Graham Cormode, American Mathematical Society, DIMACS, in preparation.

Journal Articles

Special Focus on Computational Information Theory and Coding

A. Barg and G. Zémor, "Error exponents of expander codes under linear-time decoding," *SIAM Journal of Discrete Mathematics*, **17**, no. 2 (2004), 426-445.

A. Barg and G. Zémor, "Distance properties of expander codes," preprint at http://www.arxiv.org/abs/cs.IT/0409010.

A. Barg and G. Zemor, "Multilevel expander codes," in preparation.

Rudi Cilibrasi and Paul Vitanyi, "Clustering by compression," submitted to *IEEE Trans Inform Th*.

Rudi Cilibrasi, Paul Vitanyi, and Ronald de Wolf, "Algorithmic Clustering of Music," *Computer Music Journal*, to appear.

Rudi Cilibrasi, Paul Vitanyi, and Ronald de Wolf, "Algorithmic Clustering of Music," *Proc. IEEE Conf on Web Delivery of Music, 2004 ,* to appear.

C. Fragouli and E. Soljanin, "Information Flow Decomposition for Network Coding," submitted to *IEEE Trans. Inform. Theory,* June 2004.

G. Kramer and S. A. Savari, "Cut sets and information flow in networks of two-way channels," 2004 IEEE Int. Symp. Inform. Theory, (Chicago, IL, USA),p. 33, June 27-July 2, 2004.

G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," IEEE Trans. Inform. Theory, submitted Feb. 2004.

J. N. Laneman and G. Kramer, “Window decoding for the multiaccess channel with generalized feedback,” 2004 IEEE Int. Symp. Inform. Theory, (Chicago, IL, USA), p. 281, June 27-July 2, 2004.

Ming Li, Xin Chen, Xin Li, Bin Ma and Paul Vitanyi, “The similarity metric,” *Proc. 14th ACM-SIAM Symp. Discrete Algorithms*, 2003.

Ming Li, Xin Chen, Xin Li, Bin Ma and Paul Vitanyi, “The similarity metric,” *IEEE Trans Inform. Th.,* to appear.

V. V. Prelov and S. Verd, “Second-order Asymptotics of Mutual Information," *IEEE Trans. Information Theory*, vol. 50, no. 8, 1567-1580, Aug. 2004.

Special Focus on Computational Geometry and Applications

P. K. Agarwal, M. Sharir, and E. Welzl, “Algorithms for center and Tverberg points,” *Proc. 20th Annual Sympos. Comput. Geom.*, 2004.

M. Burr, E. Rafalin, D. Souvaine, “Simplicial Depth: An Improved Definition, Analysis, and Efficiency for the Discrete Case," DIMACS Technical Report, 2003-28.

M. Burr, E. Rafalin. D. L. Souvaine, “Dynamic computation of half-space depth contours,” in preparation.

Jason H. Cantarella, Erik D. Demaine, Hayley N. Iben, and James F. O'Brien, “An Energy-Driven Approach to Linkage Unfolding,” *The Proceedings of the 2004 Symposium on Computational Geometry*, Brooklyn, New York, , pages 134-143, 2004.

Artur Czumaj and Hairong Zhao, “Fault-Tolerant Geometric Spanners ,” *Proceedings of the nineteenth annual symposium on Computational geometry*, 1-10, 2003.

Artur Czumaj and Hairong Zhao, “Fault-Tolerant Geometric Spanners ,” *Discrete and Computational Geometry*, Vol. 32, pages 207-230, 2004.

Sariel Har-Peled and Yusu Wang, “Shape fitting with outliers,” *SIAM Journal on Computing*, Volume 33, Number 2, 269-285, 2004.

Sariel Har-Peled and Yusu Wang, “Shape fitting with outliers,” *Proceedings of the nineteenth annual symposium on Computational geometry*, 29-38, 2003.

J. Hugg, M. Ouyang, “Analyzing High-Dimensional Clusters with the L1 Depth Metric,” in preparation.

John Iacono, “A 3-D visualization of Kirkpatrick's planar point location algorithm,” In *Symposium on Computational Geometry (SoCG)*, page 377, 2003.

John Iacono and S. Langerman, “Proximate planar point location,” *Symposium on Computational Geometry (SoCG)*, pages 220-226, 2003.

S. Luan, C. Wang, D.Z. Chen, X. S. Hu, S. A. Naqvi, C. X. Yu, and C. L. Lee, “A New MLC Segmentation Algorithm/Software for Step-and-Shoot IMRT Delivery," *Medical Physics*, Vol. 31, No. 4, April 2004, 695--707.

E. Rafalin and D. Souvaine, “Data Depth Contours - a Computational Geometry Perspective,” *Tufts University Technical report TR-2004-1*.

E. Rafalin, D. Souvaine, “Computational Geometry, Data Depth and Robust Statistics, " *http://www.cs.tufts.edu/r/geometry/data_depth/geom_Interface.ps*

E. Rafalin, M. Burr, D. Souvaine, “Box depth – a computational and combinatorial analysis”, in preparation.

E. Rafalin, M. Burr, D. Souvaine, “Dominance depth – a computational and combinatorial analysis”, in preparation

H. Yu, P. K. Agarwal, R. Poreddy, and K. Varadarajan, “Practical methods for shape fitting and kinetic data structures using core sets,” *Proc. 20th Annual Sympos. Comput. Geom.*, 2004.

Special Focus on Next Generation Networks Technologies and Applications

Aaron Archer, Joan Feigenbaum, Arvind Krishnamurthy, Rahul Sami, and Scott Shenker, "Approximation and Collusion in Multicast Cost Sharing," *Games and Economic Behavior* **47** (2004), pp. 36-71. (Abstract appeared in *Proc. of ACM EC'01*.)

Graham Cormode, S. Muthukrishnan, "The String Edit Distance Matching Problem with Moves," *Proceedings of the Symposium on Discrete Algorithms (SODA)*, 2002, 667-676.

Graham Cormode and S. Muthukrishnan, "The String Edit Distance Matching Problem with Moves," DIMACS Tech Report 2001-26.

Joan Feigenbaum, Rahul Sami and Scott Shenker, "Mechanism Design for Policy Routing," *Proceedings of the 23rd Symposium on Principles of Distributed Computing*, ACM Press, New York, 2004, pp. 11-20.

Joan Feigenbaum, Christos Papadimitriou, Rahul Sami, and Scott Shenker, "A BGP-based Mechanism for Lowest-Cost Routing," to appear in *Distributed Computing (special issue of selected papers from Proc. of ACM PODC'02*).

Joan Feigenbaum, Lance Fortnow, David Pennock, and Rahul Sami, "Computation in a Distributed Information Market," to appear in *Theoretical Computer Science*. (Preliminary version appeared in *Proc. of ACM EC'03*.)

Joan Feigenbaum, Arvind Krishnamurthy, Rahul Sami, and Scott Shenker, "Hardness Results for Multicast Cost Sharing," *Theoretical Computer Science* **304** (2003), pp. 215-236. (Extended Abstract appeared in *Proc. of FST-TCS'02*.)

Joan Feigenbaum and Scott Shenker "Distributed Algorithmic Mechanism Design: Recent Results and Future Directions," *Proceedings of the 6th International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, ACM Press, New York, 2002, pp. 1-13.

Joan Feigenbaum, Christos Papadimitriou and Scott Shenker, "Sharing the Cost of Multicast Transmissions," *Journal of Computer and System Sciences* **63** (2001), pp. 21-41. (Preliminary version appeared in *Proc. of ACM STOC'00*.)

P. Ferragina, N. Koudas, S. Muthukrishnan and D. Srivastava, "Two-dimensional Substring Indexing," ACM Symposium on Principles of Database Systems PODS '01, Santa Barbara (USA), 2001.

P. Ferragina, N. Koudas, S. Muthukrishnan, D. Srivastava. "Two-dimensional substring indexing." Journal of Computer and System Sciences: Special Issue on ACM PODS '01, 66(4):763--774, 2003.

A. Barg, G. Cohen, S. Encheva, G. Kabatiansky and G. Zémor, "A Hypergraph Approach to the Identifying Parent Property: The Case of Multiple Parents," SIAM J. Discrete Math, 2001, vol. 14, 423-431.

Ayman Khalfalah, Sachin Lodha and Endre Szemerdi, "Tight Bound for the Density of Sequence of Integers the Sum of No Two of which is a Perfect Square," Discrete Mathematics, 256(1-2):243--255, 2002.

"Tight Bound for the Density of Sequences of Integers the Sum of No Two of which is a Perfect Square," Ayman Khalfalah, Sachin Lodha and Endre Szemeredi. DIMACS Tech Report 2000-39

F. Menczer, N. Street, N. Vishwakarma, A. Monge, M. Jakobsson, "IntelliShopper: A Proactive, Personal, Private Shopping Assistant," Proc. 1st ACM Int. Joint Conf. on Autonomous Agents and MultiAgent Systems (AAMAS 2002), Bologna, July 2002, pp. 1001-1008 .

Adam L. Buchsbaum, Jack Snoeyink (Eds.): Proceedings of Algorithm Engineering and Experimentation, Third International Workshop, ALENEX 2001, Washington, DC, USA, January 5-6, 2001, Revised Papers. Lecture Notes in Computer Science 2153 Springer 2001, ISBN 3-540-42560-8

Monitoring Message Streams

Monitoring Message Streams Journal/Conference Articles:

Andrei Anghelescu and Ilya Muchnik, “Combinatorial PCA and SVM Methods for Feature Selection in Learning Classifications (Applications to Text Categorization),” *Proceedings of the International Conference on Integration of Knowledge Intensive Multi-Agent Systems*, 2003, pg. 491-496.

Andrei V. Anghelescu, Endre Boros, Dmitriy Fradkin, David D. Lewis, Vladimir Menkov, David Neu, Ng, and Paul Kantor, "Prospective Data Fusion for Batch Filtering," submitted to *Journal of the Intelligence Community Research and Development.*

Andrei Anghelescu and Ilya Muchnik, "Optimization of SVM in a Space of Two Parameters: Weak Margin and Intercept. Applications in Text Classifier Design," submitted to *Journal of the Intelligence Community Research and Development.*

Endre Boros and David Neu, "Feature Selection and Batch Filtering," in preparation.

G. Cormode, F. Korn, S. Muthukrishnan and D. Srivastava. Finding Hierarchical Heavy Hitters in Data Streams. VLDB 2003.

G. Cormode, F. Korn, S. Muthukrishnan, D. Srivastava. Diamond in the Rough: Finding Hierarchical Heavy Hitters in Multi-Dimensional Data. SIGMOD 2004.

G. Cormode and S. Muthukrishnan. What is Hot and What is Not: Tracking Most Frequent Items Dynamically. PODS 2003.

G. Cormode and S. Muthukrishnan. What is New: Finding Significant Differences in Network Data Streams. INFOCOM 2004.

G. Cormode and S. Muthukrishnan. An Improved Data Stream Summary: The Count-Min Sketch and its Applications. LATIN 2004.

G. Cormode, S. Muthukrishnan, M. Preetham, C. Tacioni. “Managing blog texts using sketches,” in preparation.

F. Ergun, S. Muthukrishnan and S. Sahinalp. Comparing Sequences with Segment Rearrangements. FST & TCS 2003.

S. Eyheramendy, David D. Lewis, and David Madigan. On the Naive Bayes Model for Text Classification. In Proceedings of The Ninth International Workshop on Artificial Intelligence and Statistics, 2003, C.M. Bishop and B.J. Frey (Editors), 332-339.

D. Fradkin and David Madigan. Experiments with Random Projections for Machine Learning. In Proceedings of KDD-03, The Ninth International Conference on Knowledge Discovery and Data Mining, to appear.

Alexander Genkin, David D. Lewis, Susana Eyheramendy, Wen-Hua Ju, David Madigan, "Sparse Bayesian Classifiers for Text = Categorization," submitted to *Journal of the Intelligence Community Research and Development.*

A. Genkin, David D. Lewis and David Madigan, Large-scale Bayesian Logistic Regression for Text Categorization. *Journal of Machine Learning Research*, submitted.

A. Genkin, D. Lewis, D. Madigan. "Large Scale Bayesian Sparse Polychotomous Regression," in preparation.

A. Genkin, D. Lewis, D. Madigan. "Online Sparse Bayesian Classification," in preparation.

T. Johnson, G. Cormode, F. Korn, S. Muthukrishnan, O. Spatscheck, D. Srivastava. Holistic UDAFs at streaming speeds. SIGMOD 2004

Paul Kantor and Fred Roberts, "Monitoring Message Streams: Algorithmic Methods for Automatic Processing of Messages," submitted to *Journal of the Intelligence Community Research and Development.*

David D. Lewis and Vladimir Menkov, "A Closer Look at Nearest Neighbor Text Classification," submitted to *Journal of the Intelligence Community Research and Development.*

David D. Lewis, V. Menkov, D. Madigan, A. Genkin. "Leveraging Prior Knowledge: Informative Prior Distributions for Text Classification," in preparation.

David D. Lewis and Vladimir Menkov. "Inverted List Heuristics in Nearest Neighborhood Text Classification," in preparation.

S. Muthukrishnan and M. Strauss. Maintenance of Multidimensional Histograms. FST & TCS 2003.

Monitoring Message Streams Reports:

1. Andrei V. Anghelescu and Ilya B. Muchnik. Optimization of SVM in a Space of Two Parameters: Weak Margin and Intercept. May 2003.

2. Andrei Anghelescu, Endre Boros, Dmitriy Fradkin, David D. Lewis, Vladimir Menkin, David J. Neu, Kwong Bor Ng., and Paul Kantor. Prospective Data Fusion for Batch Filtering. May 2003.

3. Endre Boros and David J. Neu. Feature Selection and Batch Filtering. May 2003.

4. Susana Eyheramendy, Alexander Genkin, Wen-Hua Ju, David D. Lewis, and David Madigan. Sparse Bayesian Classifiers for Text Categorization. May 2003.

5. David D. Lewis and Vladimir Menkov. A Closer Look at Nearest Neighbor Text Classification. May 2003.

6. Andrei Anghelescu, Dmitiry Fradkin, David D. Lewis, Vladimir Menkov, Kwong Bor Bg, and Paul Kantor. Prospective Data Fusion for Batch Filtering. February 2003.

7. David D. Lewis and Vladimir Menkov. A Closer Look at Nearest Neighbor Classification. January 2003.

8. Paul Kantor and David Madigan. Towards a Formal Framework for Adaptive Filtering. December 2002.

9. David D. Lewis, S. Muthukrishnan, R. Ostrovsky, M. Strauss. Towards Fast, Effective Nearest Neighbor Text Classification. November 2002.

10. Martin Strauss. Distances from Sketches. November 2002.

11. Alexander Genkin, David D. Lewis, and David Madigan. Sparse Bayesian Classifiers for Text Categorization - Batch Learning. November 2002.

12. Wen-Hua Ju and David Madigan. Sparse Bayesian Classifiers for Text Categorization - Online Learning. November 2002.

13. Andrei Anghelescu and Ilya Muchnik. Combinatorial Clustering for Textual Data Representation in Machine Learning Models. November 2002.

14. Andrei Anghelescu and Ilya Muchnik. An Experimental Evaluation of a Combinatorial Clustering Model for Textual Data Representation. November 2002.

15. S. Muthukrishnan. A Simple, Coarse Vector Nearest Neighbors Algorithm. November 2002.

16. S. Muthukrishnan. Some Statistical Analysis Algorithms on Data Streams, November 2002.

17. Endre Boros. Notes on Term Selection. November 2002.

18. Andrei Anghelescu, David D. Lewis, Vladimir Menkov, and Paul Kantor. Training a Rocchio classifier for adaptive filtering. TREC 2002.

19. Endre Boros and David J. Neu. Feature Selection and Batch Filtering. TREC 2002.

20. Bob Grossman, Paul Kantor, and Muthu Muthukrishnan. CCR-DIMACS Workshop. June 2002.

21. Dmitriy Fradkin and Paul Kantor, February 2004 A Design Space Approach to Analysis of IR/AF Systems.

22. Dmitriy Fradkin, Paul Kantor, and Kwong Bor Ng, February 2004. Methods for Learning Classifier Combinations: No Clear Winner.

23. Dmitriy Fradkin and Ilya Muchnik, A Study of k-Means Clustering for Improving Classification Accuracy of Multi-Class SVM. DIMACS Technical Report #2004-02.

24. Sundara Venkataraman, Dimitris Metaxes, Dmitriy Fradkin, Casimir Kulikowski, Ilya Muchnik, Distinguishing Mislabeled Data from Correctly Labeled Data in Classifier Design. Submitted to International Conference on Tools with AI (ICTAI), 2004.

25. Dmitriy Fradkin, Paul Kantor. A Design Space Approach to Analysis of Information Retrieval Adaptive Filtering Systems, in preparation.

26. Dmitry Fradkin and Michael Littman. Exploration in Adaptive Text Filtering, in preparation.

Talks

Special Focus on Computational Information Theory and Coding

A. Barg and G. Zémor, “Concatenated Codes: Serial and Parallel,”International Symposium on Information Theory, Yokohama, Japan, June, 2003.

A. Barg and G. Zémor, “Distance Properties of Expander Codes,” International Symposium on Information Theory, Chicago IL June 28, 2004.

E. Soljanin, "Network Coding: From Graph Theory to Algebraic Geometry," Columbia University, Feb. 2004.

Special Focus on Next Generation Networks Technologies and Applications:

Sorin Alexe, "Accelerated Algorithm for Pattern Detection," Third International Conference on Discrete Mathematics and Theoretical Computer Science, DMTCS'01, Constanta, Romania, July 2-6, 2001. (invited) (refereed)

Camil Demetrescu, "Distance Sensitivity Oracles", 2nd Workshop of the Research Project "Algorithms for Large Data Sets: Science and Engineering" of the Italian Ministry of University and Scientific Research, February 24, 2001.

Funda Ergun, S. Cenk Sahinalp, Jon Sharp, and Rakesh K. Sinha, "Biased Skip Lists for Highly Skewed Access Patterns," 3rd Workshop on Algorithm Engineering and Experiments, ALENEX 01, January 5-6, 2001, Washington, DC. (invited) (refereed)

Joan Feigenbaum and Scott Shenker, "Incentives and Internet Computation," PODC 2003 Tutorial, Sunday 13 July 2003, Yale University

Ashish Goel, "Delay-sensitive Unicast and Multicast Routing," DIMACS Mini-Workshop on Quality of Service Issues in the Internet, February 8 - 9, 2001.

Gregory Kabatianksky, "On Digital Fingerprinting Codes," International Symposium on Information Theory, ISIT'01 (Washington, June 2001). (invited)

Kirk Pruhs, "Scheduling Broadcasts in Wireless Networks," DIMACS Workshop on Resource Management and Scheduling in Next Generation Networks, March 26 - 27, 2001. (invited) (refereed)

P. Ferragina, "Two-dimensional Substring Indexing," ACM Symposium on Principles of Database Systems PODS '01, Santa Barbara (USA), 2001.

Software:

Computational Geometry and Applications

The Geometry Factory provides flexible, reusable and cross-platform geometric software components and expertise. It gives development teams a head-start on building applications that solve business problems, increasing productivity and the ability to deliver products on time. By offering field-proven C++ components, along with expert services it saves customers valuable development time. Application domains include industrial users in areas such as digital maps, medical imaging, satellite image processing, computer graphics, structural geology, etc.

Monitoring Message Streams

Classic method: Rocchio

Classic method: Centroid

*k*NN with IFH (inverted file heuristic)

Sparse Bayesian (Bayesian with Laplace priors)

cPCA

Homotopic Linking of Widely Varying Rocchio Methods

aiSVM

Fusion

Public release of the sparse Bayesian software "Bayesian Binary Regression (BBR)" at the address http://stat.rutgers.edu/~madigan/BBR. Windows and Linux versions are available. The software has been downloaded over 1500 times.

Revised and extended version of kNN code.

Substantially extended version of BBR, including use of prior knowledge.

Count-Min (CM) Sketch: This is the basic sketch on which historic data analysis is based, with many applications.

cgt.h and cgt.c: Using CM sketch to finding frequent items/heavy hitters.

Websites

Websites for Computational Information Theory and Coding
http://dimacs.rutgers.edu/SpecialYears/2001_COD/

Websites for Computational Geometry and Applications
http://dimacs.rutgers.edu/SpecialYears/2002_CompGeom/

Websites for Next Generation Networks and Applications
http://dimacs.rutgers.edu/SpecialYears/2000_NGN/

Websites for Monitoring Message Streams
http://www.stat.rutgers.edu/~madigan/mms/

Websites for Author Identification: Identifying Real-Life Authors in Massive Document Collections
http://www.stat.rutgers.edu/~madigan/AUTHORID/

VIII. Contributions within Discipline

This project has been inherently interdisciplinary. The three special foci part of the project involves both fundamental theoretical issues and major outreach components to other communities and other parts of computer science and mathematics. This diversity of programs, feeding off of each other, led to many of the synergies that make DIMACS successful. One of our topics, Next Generation Networks Technologies and Applications, is very tied to technology and practice. A second, Computational Information Theory and Coding, explores the interface among coding theory, information theory, and parts of computer science and mathematics. A third, Computational Geometry and Applications, is closer to core theory, but has several strong applied components. This part of the project has led to important new directions of research in computer science, in particular theoretical computer science, and related areas of discrete mathematics. It also laid the groundwork for substantial new directions in theoretical computer science, for example through its emphasis on the network structure of the internet and on game theory and mechanism design in the Next Generation Networks special focus, its emphasis on new applications of computational geometry such as to medical applications and the connections between computational geometers and statisticians in areas such as data depth, and connections between network concepts and information theory concepts in the Computational Information Theory and Coding special focus. The special projects, Monitoring Message Streams, Author Identification, and Mining Multilingual Resources Using Text Analytics have led to important new results in machine learning and text classification.

Here are a few selected examples of the collobarations that have been fostered among participants of the three special foci.

As part of the Computational Geometry and Applications special focus, Herve Bronnimann organized the DIMACS Workshop on Implementation of Geometric Algorithms, December 2002, in collaboration with Steve Fortune. The main goal of the workshop was to put together people interested in the topic to discuss technical issues and foster collaborations. Following the workshop, there was a renewed collaboration between Sylvain Pion and Herve Bronnimann, under the title Genepi, which is currently funded by INRIA under the associated team program. Notably, also, another INRIA collaboration team called CALAMATA was set up earlier and continues being funded between Geometrica / Sylvain Pion, Galaad / Monique Teillaud, and Ioannis Emiris, all participants to the workshop. The action for the commercialization of CGAL, presented at the workshop by Andreas Fabri, has continued and led to the creation of GeometryFactory.

The workshop Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, held under the auspices of the Computational Geometry and Applications special focus, fostered a collaboration between Robert Serfling of the Department of Mathematical Sciences at the University of Texas at Dallas and Ovidiu Daescu of the Department of Computer Science at the same institution. These faculty members had previously not had an association, but they both attended the workshop and afterwards began meeting regularly to explore mutual interests. These in-depth discussions over a 10-month period led to a joint research proposal submitted to NSF and funded as a two-year award. The project, entitled *Outlier Identification and Handling in Computational Geometry Problems*, involves outlier issues arising in shape fitting problems and dimension reduction. Treatment of outliers from general principles is a major problem receiving high interest in computer science. The workshop was instrumental in bringing about this interdisciplinary project involving a statistical scientist and a computer scientist.

Several other collaborative efforts were initiated among a number of statisticians and computer scientists who participated in the workshop or who were exposed to work related to the workshop. J. Romo (Mathematics Department, Universidad Carlos III de Madrid, Spain) proposed a new concept of depth for functional data at the DIMACS workshop. D.L. Souvaine, E. Rafalin, M. Burr (Computer Science Department, Tufts University) studied the concepts of functional depth from a computational and combinatorial perspective and proposed efficient algorithms. The collaboration began at the workshop and resulted in two publications, given in the list of journal articles. Their research is ongoing.

D.L. Souvaine, J. Hugg, E. Rafalin (Computer Science Department, Tufts University) and M. Ouyang (University of Medicine and Dentistry of New Jersey, Informatics Institute) began collaboration at the workshop on the application of data-depth based statistical methods to the study of gene expression data. This resulted in one publication, given in the list of journal articles. Their research is ongoing.

M. Burr (Computer Science Department, Tufts University) and C. Small (Department of Statistics and Actuarial Science, University of Waterloo, Canada) initiated a collaboration based on published work related to the workshop dealing with unimodality in Simplicial depth.

Emina Soljanin and one of the Computational Information Theory and Coding Special Focus visitors Christina Fragouli started working on (at that time very new) area known as Network coding. Subsequently, Emina won a 5 year NSF medium ITR grant for Network Coding (No. CCR-0325673) with G. Kramer and P. Gupta (Bell Labs), R. Koetter (Univ. Ill., Urbana-Champaign), M. Effros (Caltech), M. Medard (MIT) and D. Karger (MIT)) . The DIMACS visitor Christina Fragouli won an FNS (Swiss counterpart to NSF) award (No. 200021-103836/1) to do research on network coding for a year.

IX. Contributions -- other Disciplines

As noted in the section on Contribuitons Within Discipline, the three special foci part of this project was inherently interdisciplinary. Among the most important new directions of work stressed were connections of network analysis to the social sciences (e.g., through game theory) and to intellectual property considerations, connection of computational geometry to a wide variety of applications in engineering, design, and the biological sciences, and the connection of information theory and coding to statistical physics.

The projects on Monitoring Mesasage Steams, Author Identification, and Mining Multilingual Resources Using Text Analytics focused computer scientists on important problems of the intelligence community and created important partnerships among computer scientists, statisticians, and practitioners.

X. Contributions -- Human Resource Development

Many of the comments in the section on Contributions within Discipline illustrate the human resource development contributions of this project. A major contribution is the impact on the research programs and careers of the participants. This project fostered new collaborations both within disciplines and among disciplines, both within academe and between academe and industry.

Here we describe some of the effects on teaching and education of the project. The workshop Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, held under the auspices of the Computational Geometry and Applications special focus, influenced training of school teachers during the summer of 2003 at the Computer-Science department at Tufts University. This was part of the "Research experience for teachers" program where teams of science, math and technology teachers from a school or district participate in a university level research and, based on their experience, develop a curricular activity they implement the following school year. For details see the "CEEO – Center for engineering educational outreach" http://www.ceeo.tufts.edu.

Gerhard Kramer developed and taught a course on "Topics in Multi-terminal Information Theory." This topic was directly related to the subject matter of the Workshop on Network Information Theory that he co-organized as part of the Computational Information Theory and Coding Special Focus. Gerhard taught the course internally (for Bell Labs researchers, summer students, and visitors) and externally. He then developed a university course on this topic.

The three special foci supported thirty-eight graduate students and the Monitoring Message Stream Project supported 4 graduate students. This is described in detail in the section on Project Training/Development. Many other students attended special focus workshops and tutorials. Two undergraduates did projects on author identification as part of the DIMACS REU in 2004.

XI. Contributions to Resources for Research and Education

The Author Identification project has amassed an extensive corpus of newsgroup postings, including 1.2 million messages by 82,000 authors from 63 usenet newsgroups.

XII. Contributions Beyond Science and Engineering

Polynomiography and its applications in science and mathematics are described in the section on Project Findings. But Bahman Kalantari's work on polynomiography also has applications in the visual arts. Polynomiography is the art and science of visualization in the approximation of zeros of complex polynomials. Informally speaking polynomiography allows one to take colorful pictures of polynomials. These images can subsequently be recolored in many ways using one's own creativity and artistry. It has tremendous applications in visual arts, education, and science. From the artistic point of view polynomiography can be used to create quite a diverse set of images reminiscent of the intricate patterning of carpets and elegant fabrics; abstract expressionist and minimalist art; and even images that resemble cartoon characters. From the educational point of view polynomiography can be used to teach mathematical concepts, theorems, and algorithms, e.g. the algebra and geometry of complex numbers; the notions of convergence, and continuity; geometric constructs such as Voronoi regions; and modern notions such as fractals. A polynomiography exhibition was held at Lawrenceville School in New Jersey and reviewed by the Trenton Times, January 9, 2004. The education program of SIGGRAPH 2003 featured Kalantari and his work on polynomiography. Polynomiography artwork was featured in the Spring 2003 edition of New Jersey Savvy Living Magazine. An October 2002 article in the New Jersey Star-Ledger let to a visit by Kalantari to Randolph Middle School. A polynomiograph was even displayed on the cover of the 2001-2003 Rutgers-New Brunswick Course Catalog.